International Journal of Computer Vision (2019) 127:930–956
https://doi.org/10.1007/s11263-019-01151-x

Deep, Landmark-Free FAME: Face Alignment, Modeling, and Expression Estimation

Feng-Ju Chang[2] · Anh Tuan Tran[2] · Tal Hassner[3] · Iacopo Masi[1] · Ram Nevatia[2] · Gérard Medioni[2]

Received: 23 February 2018 / Accepted: 17 January 2019 / Published online: 13 February 2019

Abstract
We present a novel method for modeling 3D face shape, viewpoint, and expression from a single, unconstrained photo. Our method uses three deep convolutional neural networks to estimate each of these components separately. Importantly, unlike others, our method does not use facial landmark detection at test time; instead, it estimates these properties directly from image intensities. In fact, rather than using detectors, we show how accurate landmarks can be obtained as a by-product of our modeling process. We rigorously test our proposed method. To this end, we raise a number of concerns with existing practices used in evaluating face landmark detection methods. In response to these concerns, we propose novel paradigms for testing the effectiveness of rigid and non-rigid face alignment methods without relying on landmark detection benchmarks. We evaluate rigid face alignment by measuring its effects on face recognition accuracy on the challenging IJB-A and IJB-B benchmarks. Non-rigid, expression estimation is tested on the CK+ and EmotiW'17 benchmarks for emotion classification. We do, however, report the accuracy of our approach as a landmark detector for 3D landmarks on AFLW2000-3D and 2D landmarks on 300W and AFLW-PIFA. A surprising conclusion of these results is that better landmark detection accuracy does not necessarily translate to better face processing. Parts of this paper were previously published by Tran et al. (2017) and Chang et al. (2017, 2018).

Keywords 3D face modeling · Face alignment · Facial expression estimation · Facial landmark detection

1 Introduction

Successful methods for single-view 3D face shape modeling were proposed nearly two decades ago (Blanz et al. 2002; Blanz and Vetter 2003; Paysan et al. 2009; Romdhani and Vetter 2003). These methods, and the many that followed, often claimed high fidelity reconstructions and offered parameterizations for facial expressions besides the underlying 3D facial shape.

Despite their impressive results, they and others since have suffered from problems when processing images taken under unconstrained viewing conditions (Blanz et al. 2002; Blanz and Vetter 2003; Chu et al. 2014; Paysan et al. 2009; Romdhani and Vetter 2003; Tang et al. 2008; Yang et al. 2011). Many of these methods relied, to some extent, on facial landmark detection performed either prior to reconstruction or concurrently, as part of the reconstruction process. By using landmark detectors, these methods were sensitive to face pose and, aside from a few recent exceptions [e.g., 3DDFA (Zhu et al. 2016b)], could not operate well on faces viewed in extreme out-of-plane rotations (e.g., near-profile).

✉ Tal Hassner
talhassner@gmail.com

Feng-Ju Chang
fengjuch@usc.edu

Anh Tuan Tran
anhttran@usc.edu

Iacopo Masi
iacopo@isi.edu

Ram Nevatia
nevatia@usc.edu

Gérard Medioni
medioni@usc.edu

1 Information Sciences Institute (ISI), USC, Marina Del Rey, CA, USA

2 Institute for Robotics and Intelligent Systems, USC, Los Angeles, CA, USA

3 Open University of Israel, Ra'anana, Israel

Fig. 1 Results of our FAME approach. We propose deep networks which regress 3DMM shape, expression, and viewpoint parameters directly from image intensities. We show this approach to be highly robust to appearance variations, including out-of-plane head rotations (top row), scale changes (middle), and ages (bottom)

Fig. 2 The problem with manually labeled facial landmarks. Images and annotations from the AFW (Zhu and Ramanan 2012) (left two columns) and iBug (Sagonas et al. 2013) benchmarks. One of each pair shows manually annotated landmarks; the other, a high-error prediction of our FacePoseNet (Sect. 3.3) Which one is which? (Images one, three, and five are ground truth.). Rather than measuring the accuracy of predicting human annotated landmarks, we propose evaluating the effect of different face alignment methods on the bottom line performance of the face processing systems

Scale changes and occlusions were also problems: either because landmarks were too small to be accurately localized or were altogether invisible due to occlusions, detection and 3D reconstruction were not handled well. Finally, many methods applied iterative *analysis-by-synthesis* steps (Bas et al. 2016; Huber et al. 2016; Romdhani and Vetter 2005). This approach was not only computationally expensive, but also hard to distribute and run in parallel on dedicated hardware offered by graphical processing units (GPU).

We offer a novel, efficient, and accurate alternative to these methods by describing a deep learning-based approach for *face alignment, modeling, and expression estimation* (FAME). We show how deep networks can separately estimate each of the following components of a 3D morphable face model (3DMM) representation (Sect. 3): 3D face shape, six degrees of freedom (6DoF) viewpoint, and 3D face expression (Fig. 1). We see access to sufficient labeled training data as a key concern in such an approach and explain how we mitigate this problem and obtain massive, labeled data sets for the supervised training of our networks.

Contrary to others, our approach does not require facial landmark detection at test time and instead models faces directly from image intensities. Still, if facial landmarks are required, we show how they can be estimated as a by-product of our modeling, rather than as part of our modeling process (Sect. 4).

Before testing our method, we discuss the shortcomings of facial landmark detection benchmarks (Sect. 5). In particular, we claim that manual landmark annotations used as ground truth by such benchmarks can be arbitrary and inaccurate. Thus, better detection accuracy may actually reflect better estimation of uncertain human labels rather than, say, better face alignment (Fig. 2). A similar observation was recently made by others Dong et al. (2018b) and is reflected in the design of the PIFA protocol Jourabloo and Liu (2015).

To address these concerns, we propose a simple, alternative test paradigm which considers the bottom line performances of the systems employing these methods. Thus, because one of the most popular applications of facial landmark detectors is face alignment in face recognition pipelines (Chang et al. 2017; Masi et al. 2018b), we evaluate different methods by measuring their effect on face recognition accuracy. To this end, we use the challenging, unconstrained images of the IJB-A (Klare et al. 2015) and IJB-B (Whitelam et al. 2017) benchmarks (Sect. 6.1). Both sets contain images with viewing conditions which are typically far more challenging than those in landmark benchmarks.

To further compare accuracy of non-rigid face deformation estimations, we propose tests on facial emotion classification benchmarks (Sect. 6.2). For this purpose, we use both the controlled images in the Extended Cohn-Kanade (CK+) benchmark (Lucey et al. 2010) and the unconstrained images in the EmotiW-17 benchmark (Dhall et al. 2017). We use both benchmarks to test how different expression estimates, obtained by different methods, affect the quality of descriptors used for emotion classification.

Finally, we evaluate the facial landmarks obtained by our method and others on the popular 300W benchmark (Sagonas et al. 2013) and AFLW-PIFA dataset (Jourabloo and

Liu 2015) for 2D landmark detection and the AFLW2000-3D benchmark for 3D landmark detection (Zhu et al. 2016b) (Sect. 6.3). A surprising conclusion of our tests is that older facial landmark detectors often outperform newer, presumably state-of-the-art, methods when used in face processing pipelines. Our approach is not necessarily more accurate as a landmark detector than state-of-the-art detectors, though its accuracy is comparable to theirs. Our method is, however, far faster and more accurate than existing landmark detectors when evaluated for its bottom line performance on face recognition and emotion classification.

To promote reproduction of our results, our code and deep models are publicly available from: https://github.com/fengju514/Expression-Net.

2 Related Work

2.1 Facial Landmark Detection

Facial landmark detection is a general problem which has applications in numerous face-related systems. Landmark detectors are very often used to align face images by applying rigid (Eidinger et al. 2014; Everingham et al. 2006; Wolf et al. 2011) and non-rigid transformations (Hassner 2013; Jeni et al. 2015; Zhu et al. 2016b) transformations in 2D and 3D (Masi et al. 2014, 2016b, 2017, 2018a). Other facial landmark applications involve estimating 3D face shape, expression, emotion, and many others.

To address the problem of varying 3D poses, 3DDFA (Zhu et al. 2016b) learns the parameters of a 3DMM representation using a deep convolutional neural network (CNN). Unlike us, however, they describe an iterative, analysis-by-synthesis technique. Tweaked CNN (TCNN) (Wu et al. 2017) optimize multiple versions of the final network layers, each specialized to produce landmark estimates for different face viewpoints, determined in an unsupervised manner. DCLM (Zadeh et al. 2016), by comparison, introduced an ensemble of convolutional expert networks to capture complex landmark appearance variations.

Very recently, Bhagavatula et al. (2017) proposed their 3D-STN approach which shares some of our design goals. In fact, we use a modified version of their regression sampler component to optimize facial landmark localization. Their approach, however, learns a 3D, thin plate splines (TPS) warping matrix and a 11DoF camera projection matrix (11 DoF) to modify and fit a generic 3D face model to the input. Our approach is very different: we use a 3DMM representation, estimating facial shape, expression, and 6DoF pose directly from image intensities with three deep networks.

Kumar et al. (2017) proposed the KEPLER system, an iterative method based on three submodules: a rendering module that stacks previously predicted key-point locations in the current image, a CNN that predicts key-point location updates towards the ground-truth, and a final step which applies these increments to generate new landmark estimates. Inspired by cascaded regression, the method iterates over these three modules in order to get final key-points predictions along with a visibility confidence and 3D head pose estimation. Unlike KEPLER, our method offers the speed of a direct regression method and is self-supervised, without the need for manual landmark annotations for training.

Bulat and Tzimiropoulos (2017b) discuss the importance of data set size, proposing the largest annotated training set for landmark detection (LS3D-W), consisting of $\sim$ 230,000 samples. They additionally provide ablation studies of pose, landmark initialization, face resolution, and more such factors. Bulat and Tzimiropoulos (2017a) studied how to redesign the bottleneck layer of a CNN in order to obtain substantial improvements in localization accuracy while constraining the learned model to be lightweight, fast, and compact. While most methods are designed to be robust to typical appearance nuisance factors—pose and illumination—(Bulat and Tzimiropoulos 2017b) and Dong et al. (2018a) designed a method to make landmark detection systems robust to different image *styles*.

Finally, Kumar and Chellappa (2018) used a pose estimating network similar to our FacePoseNet (Chang et al. 2017) in their facial landmark detection system. Our detection accuracy on landmark benchmarks may be lower than theirs. Unlike them, however, our FAME approach provides a complete 3D face reconstruction, with landmarks only being a by-product.

2.2 3D Face Modeling

Estimating the 3D shape of a face appearing in a single image is a problem with a history now spanning over two decades. Some proposed example-based approaches (Hassner and Basri 2006; Vetter and Blanz 1998; Hassner 2013) where the 3D shape was estimated using the shapes of similar reference faces. These methods were typically very robust to viewing conditions, but were not designed to offer accurate 3D reconstructions. To estimate fine facial details, others used shape from shading (SfS) for face reconstruction (Kemelmacher-Shlizerman and Basri 2011; Li et al. 2014). Though SfS reconstructions were detailed, they were often limited to rather constrained viewing settings.

Possibly the most popular methods of estimating 3D facial shapes involved 3DMM. These statistical representations were introduced by Blanz and Vetter (1999) and then later improved by others (Blanz et al. 2004; Chu et al. 2014; Paysan et al. 2009; Romdhani and Vetter 2003; Tang et al. 2008; Yang et al. 2011). We provide a brief overview of these representations in Sect. 3.1. Whereas classical 3DMM fitting methods used an *analysis-by-synthesis* approach which involved com-

putationally expensive rendering cycles, we estimate 3DMM parameters directly from image intensities with deep networks.

Facial landmarks were also used to produce 3D face shape estimates (Jourabloo and Liu 2016; Zhu et al. 2016b). Because these methods often focused on landmark detection accuracy, they were typically very fast, but not necessarily accurate in the quality of their reconstructions.

Finally, deep learning methods were also recently proposed for 3D face shape estimation (Booth et al. 2017; Dou et al. 2017; Jackson et al. 2017; Sela et al. 2017; Richardson et al. 2016, 2017; Sengupta et al. 2018; Tewari et al. 2018; Tran et al. 2017, 2018). Our work extends the method proposed by Tran et al. (2017) by adding deep 6DoF pose and 29D facial expression estimation to their accurate 3D faces.

2.3 Face Alignment and Pose (Viewpoint) Estimation

The term *face alignment* is often used in papers presenting *facial landmark detection* methods (Asthana et al. 2014; Cao et al. 2014; Ren et al. 2014), implying that the two terms are used interchangeability. This reflects an interpretation of alignment as *forming correspondences between particular spatial locations in two face images*. A different interpretation of *alignment*, and the one used here, refers not only to establishing these correspondences but also to *warping* the two face images in order to bring them into alignment, thereby making them easier to compare and match. Such methods using 2D or 3D transformations are well known to have a profound impact on the accuracy of face recognition systems (Hassner et al. 2015, 2016; Huang et al. 2007).

We describe a deep network trained to estimate the 6DoF of 3D faces viewed in single images. Deep learning is increasingly used for similar purposes, though typically focusing on general object classes (Bansal et al. 2016; Poirson et al. 2016; Su et al. 2015). Some recently addressed faces in particular, though their methods are designed to estimate 2D landmarks along with 3D face shapes (Jourabloo and Liu 2016; Kumar et al. 2017; Kumar and Chellappa 2018; Zhu et al. 2016b). Unlike our proposed pose estimation, many of these regress poses by using iterative methods which involve computationally costly face rendering. We regress 6DoF directly from image intensities without such rendering steps.

In all these previous efforts, absence of training data was cited as a major obstacle for training effective deep models for alignment. In response, some turned to larger 3D object data sets (Xiang et al. 2014, 2016) or using synthetically generated examples (Richardson et al. 2017). We propose a far simpler alternative and show it to result in robust and accurate face alignment.

2.4 Facial Expression Estimation

We first emphasize the distinction between the related, yet different tasks of *emotion classification* and *expression regression*. The former seeks to classify images or videos into discrete sets of facial emotion classes (Dhall et al. 2015; Lucey et al. 2010) or action units (Fabian Benitez-Quiroz et al. 2016; Zafeiriou et al. 2016). In the past, this problem was addressed by considering the precise locations of facial landmarks. In recent years, however, a growing number of state-of-the-art methods have instead adopted deep networks, applying them directly to image intensities rather than estimating landmark positions as a proxy step (Kosti et al. 2017; Levi and Hassner 2015; Zhang et al. 2016).

The latter task, expression regression, concerns estimation of the non-rigid facial deformations produced by facial expressions. These expressions are often expressed as active appearance models (AAM) (Lucey et al. 2010) and Blendshape model coefficients (Richardson et al. 2017; Zhu et al. 2016b, 2015b). In this work we focus on estimating 3D expression coefficients, using the same 29D representation of the 3D facial deformation described by 3DDFA (Zhu et al. 2016b). 3DDFA, however, obtains these expression parameters using an iterative analysis-by-synthesis approach, whereas we estimate these parameters directly using a single forward pass through a dedicated deep network.

3 Our Proposed FAME Framework

We describe a deep, landmark-free 3D face modeling method. Our approach, illustrated in Fig. 3, is designed to provide an alternative means of obtaining the same goals previously attained with the help of facial landmark detectors. Namely, it allows for face alignment in 3D and 2D, expression modeling (deformations) in 3D, and 3D face shape estimation. Our method uses three deep networks which separately estimate subject-specific 3D face shape, viewpoint, and 3D expression deformations directly from the input image. As we later show, rather than using landmark detectors in this process, accurate facial landmarks can actually be obtained as a by-product of our FAME approach.

We emphasize that although we use landmark detection methods during training, our test time system is completely landmark free. We use landmark detectors as a cheap yet effective alternative to the manual labor required by others for labeling training images. The term "landmark-free" therefore refers to the absence of landmark detection at test time.

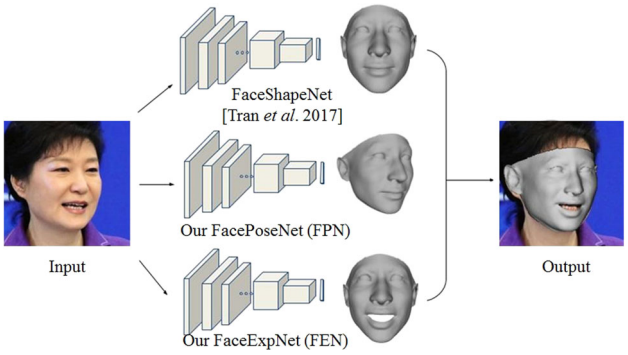


Fig. 3 Proposed framework for 3D face modeling. Given an input face photo, we process it using three separate deep networks. These networks independently estimate, from top to bottom: the 3D face shape (Tran et al. 2017), 6DoF viewpoint, and 29D expression coefficients (last two described here). The output is an accurate 3D face model, aligned with the input face

3.1 Preliminaries

We model a 3D face shape $\mathbf{F} \in \mathbb{R}^{3n}$ using the following standard, linear 3DMM representation (for now, ignoring parameters representing facial texture and 6DoF pose):

$$\mathbf{F} = \widehat{\mathbf{s}} + \sum_{i=1}^{s} \alpha_i \mathbf{S}_i + \sum_{j=1}^{m} \eta_j \mathbf{E}_j, \tag{1}$$

where $\widehat{\mathbf{s}} \in \mathbb{R}^{3n}$ represents the average 3D face shape. The first summation expresses shape variation as a linear combination of shape principal components $\mathbf{S} \in \mathbb{R}^{3n \times s}$ with the coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^{s}$. Here, $\mathbf{S}_i$ indicates the direction to deform the face, following the variation in the data learned by principle component analysis (PCA). This deformation is regulated by the corresponding scalar α_i. 3D expression deformation is represented as a linear combination of expression components $\mathbf{E} \in \mathbb{R}^{3n \times m}$ with the coefficient vector $\boldsymbol{\eta} \in \mathbb{R}^{m}$. Here, $3n$ represents the 3D coordinates for the n vertices of the 3D face shape. The numbers of components for shape, s, and expression, m, define the dimensionality of the 3DMM coefficients.

Our representation uses the Basel Face Model (BFM) 3DMM shape components (Paysan et al. 2009), where $s = 99$ and the expression components defined by 3DDFA (Zhu et al. 2016b), with $m = 29$. The vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$ control the intensity of deformations provided by the principal components. Given estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, it is thus possible to reconstruct the 3D shape of the face appearing in the input image by Eq. (1).

3.2 Modeling Subject-Specific 3D Shape

Given image $\mathbf{I}$, we estimate $\boldsymbol{\alpha}$ with the recent deep 3DMM approach of Tran et al. (2017), using their publicly available code and pre-trained model. They regress 3DMM coefficients, $\boldsymbol{\alpha}$, directly from image intensities using a ResNet architecture with 101 layers (He et al. 2016) which we refer to here as FaceShapeNet.

The decision to use this method to estimate a subject-specific 3D face shape is not matter of fact. This method was extensively tested, and the results reported in Tran et al. (2017) demonstrate it to be both invariant and discriminative under the harshest viewing conditions, previously considered only by face recognition systems. In particular, it has been shown to work well and produce invariant 3D shapes even on photos where the face is heavily occluded. In fact, it is this property which led to its use for occluded 3D face reconstruction by Tran et al. (2018).

We note that FaceShapeNet was designed to infer only the 3D face shape, without estimating its viewpoint or expression (Tran et al. 2017); i.e., the 3D face shape is produced in fixed 3D coordinates and is not modified to account for different viewpoints and expressions. Estimating these missing components—viewpoint and expression—is described next.

3.3 Modeling 6DoF Face Viewpoint

We propose to infer a global, 6DoF 3D face viewpoint directly from image intensities using a deep neural network. For a given face photo, our FacePoseNet (FPN) regresses 6DoF viewpoint parameters. We next describe FPN and the novel method used to produce sufficient training data, along with the pose labels required to train it.

Viewpoint representation We define the viewpoint, $\mathbf{h}$, as the 6DoF transformation components of the extrinsic camera matrix involved in projecting a 3D face head onto the face in the photo. These components are the three rotation angles, $\mathbf{r} = (r_x, r_y, r_z)^T$, represented as Euler angles (pitch, yaw, and roll), and translation vector, $\mathbf{t} = (t_x, t_y, t_z)^T$. Thus:

$$\mathbf{h} = (r_x, r_y, r_z, t_x, t_y, t_z)^T. \tag{2}$$

Given m 2D facial landmark coordinates on an input image, $\mathbf{p}_{m\times 2}$, and their corresponding, reference 3D coordinates, $\mathbf{P}_{m\times 3}$—selected on a fixed, generic 3D face model—we can obtain a 3D to 2D projection of the 3D landmarks onto the 2D image by solving the following equation for the standard pinhole model:

$$[\mathbf{p}, \mathbf{1}]^T = \mathbf{A}[\mathbf{R}, \mathbf{t}][\mathbf{P}, \mathbf{1}]^T \doteq \boldsymbol{\Pi}[\mathbf{P}, \mathbf{1}]^T, \tag{3}$$

where $\boldsymbol{\Pi}$ is defined as the camera projection matrix, with $\mathbf{A}$, the intrinsic camera matrix, $\mathbf{R}$ and $\mathbf{t}$ the 3D rotation matrix

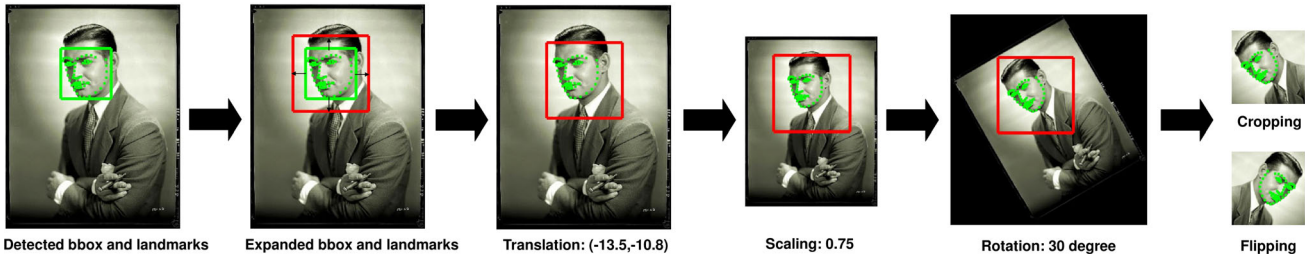


Fig. 4 Augmenting appearances of images from the VGG Face dataset (Parkhi et al. 2015). After detecting the face bounding box (bbox) and landmarks, we augment facial appearances by applying a number of simple planar transformations—including translation, scaling, rotation, and flipping. The same transformations are applied to the landmarks; this is followed by 6DoF viewpoint parameters being extracted from the transformed landmarks. Note that these images are often too challenging for existing landmark detectors to directly process (see Fig. 5)

and translation vector, and $\mathbf{1}$ a constant vector of 1. The intrinsic camera parameters in $\mathbf{A}$ are kept fixed with the focal length set to 2880 pixels and the principal point set to (112,112). We extract a rotation vector $\mathbf{r} = (r_x, r_y, r_z)^T$ from $\mathbf{R}$ using the Rodrigues rotation formula:

$$\mathbf{R} = \cos\theta \mathbf{I} + (1 - \cos\theta)\mathbf{r}\mathbf{r}^T + \sin\theta \begin{pmatrix} 0 & -r_z & r_y \\ r_z & 0 & -r_x \\ -r_y & r_x & 0 \end{pmatrix}, \quad (4)$$

where we define $\theta = ||\mathbf{r}||_2$. A similar process is used in many face processing pipelines to align faces (see, e.g., Masi et al. (2014, 2016b, 2017, 2018a) and many others).

Obtaining sufficient viewpoint training examples Training a deep network typically requires large quantities of labeled training data. Here, this implies large numbers of face photos, each associated with a ground truth 6DoF pose label. Ostensibly, existing data sets annotated for facial landmarks can be used to provide these viewpoint annotations by using their landmark annotations to estimate viewpoint for each image. We found, however, that the number of images in standard data sets is too small for this purpose. A key problem is therefore obtaining a large enough training set.

To produce our training labels, we turn to the facial landmark detection methods we ultimately seek to replace. Specifically, we propose to synthesize 6D, ground truth pose labels by running the existing facial landmark detector of Baltrušaitis et al. (2016) on a large image set. For this purpose, we use the 2.6 million images in the VGG Face dataset (Parkhi et al. 2015). The detected landmarks are then used to compute the 6DoF labels for the images in this set, as described above.

Viewpoint training set augmentation In practical use-cases, pose estimation is performed on face bounding boxes obtained from a face detector. Bounding boxes returned by such detectors can vary in how tightly they are positioned over the face: bounding boxes can either be tightly placed over the face region or loosely placed over the entire head. These variations translate to scale changes when the face is ultimately processed by the alignment method. In order to train our network to be robust to these variations, as well as potential variations in face location or rotation, we augment the training set used for training.

Table 1 Summary of augmentation transformation parameters used to produce FPN training data: $\mathscr{U}(a, b)$ samples from a uniform distribution ranging from a to b and $\mathscr{N}(\mu, \sigma^2)$ samples from a normal distribution with mean μ and variance σ^2

Transformation	Range
Horizontal translation	$\mathscr{U}(-0.1, 0.1) \times width$
Vertical translation	$\mathscr{U}(-0.1, 0.1) \times height$
Scaling	$\mathscr{U}(0.75, 1.25)$
Rotation (°)	$30 \times \mathscr{N}(0, 1)$

The values $width$ and $height$ are the face detection bounding box dimensions

Specifically, we apply a number of face augmentation techniques to the images in the VGG Face set, substantially enriching the appearance variations it provides. Figure 4 illustrates this augmentation process. Following face detection using the method of Yang and Nevatia (2016) and landmark detection (Baltrušaitis et al. 2016), we transform detected bounding boxes and their detected facial landmarks using a number of simple in-plane transformations. The parameters for these transformations are selected randomly from fixed distributions (provided in Table 1). The transformed faces are then used for training, along with their horizontally mirrored versions. The same transformations are applied to the landmarks used to produce the 6DoF training labels. Thus, our 6DoF training labels match the augmented faces.

Some example augmented faces are provided in Fig. 5. Importantly, the augmented face images used for training are often *too challenging for existing landmark detectors*, due to

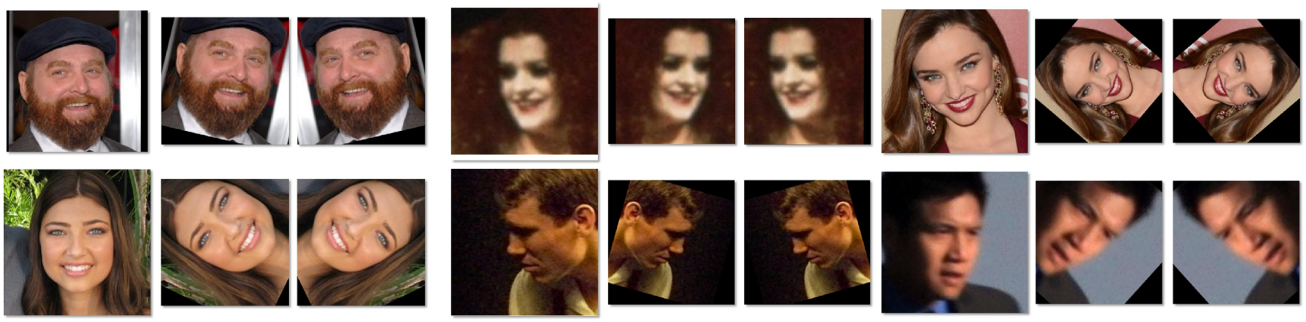


Fig. 5 Example augmented training images. Example images from the VGG Face data set (Parkhi et al. 2015) following data augmentation. Each triplet shows the original detected bounding box (left) and its augmented versions (mirrored across the horizontal axis). Both flipped versions were used for training FPN. Note that in some cases, detecting landmarks would be highly challenging on the augmented face, due to severe rotations and scalings not normally handled by existing methods. Our FPN is trained with the original landmark positions, transformed to the augmented image coordinate frame (Fig. 4)

extreme rotations or scaling. This, of course, *does not affect the accuracy of the ground truth labels* which were obtained from the original images. It does, however, force our CNN to learn to estimate poses even on such challenging images.

A simple FPN architecture We experimented with two network architectures for our FPN. The first, originally presented in Chang et al. (2017), is a simple AlexNet (Krizhevsky et al. 2012), initialized using weights provided by Masi et al. (2016a). We modify this network by defining a new output layer, FC8, which uses ℓ_2-loss to regress a 6D floating point output—representing the 6DoF viewpoint—rather than predict one-hot encoded, multi-class label. We initialize this new layer with parameters drawn from a Gaussian distribution with zero mean and standard deviation 0.01. All biases are initialized with zero. During training, batch size is set to 60, and the initial learning rate is set to 0.01. Learning rate is decreased by an order of magnitude every 5000 iterations until the validation accuracy for the fine-tuned network saturates.

Note that during training, each dimension of the head pose labels is normalized by the corresponding mean and standard deviation of the training set, compensating for the large value differences among dimensions. The same normalization parameters are used at test time.

Improved FPN architecture In addition to the shallow architecture described above, we tested a deeper network: a ResNet architecture with 101 layers (He et al. 2016) (ResNet101). This deeper network was trained on a larger training set—the 300W-LP dataset (Zhu et al. 2016b)—which offers richer appearance variations including, in particular, a larger proportion of profile views (Masi et al. 2018a). In this improved version, instead of using the pinhole model detailed in Eq. (3), the weak perspective projection is regressed for the 300W-LP dataset. All other settings were similar to the ones used for the AlexNet version.

2D and 3D face alignment with FPN Given a test image, we first apply the same face detector used in training (Yang and Nevatia 2016), then crop the face and scale it to the dimension of the network's input layer. The 6D network output is then converted to a projection matrix. Specifically, the projection matrix is produced by the (constant) camera matrix **A**, rotation matrix **R**, and the translation vector **t** in Eq. (3). With this projection matrix we can render new views of the face, aligning it across 3D views as was recently proposed by others (Masi et al. 2016b, 2017).[1]

3.4 Modeling Facial Expressions

Deformations of the 3D face shape due to changing facial expressions are estimated separately from the underlying subject-specific 3D shape and the viewpoint. As in the previous two sections, we again use a deep network for this purpose, our FaceExpNet (FEN), applying it directly to image intensities. As with the other two networks, availability of labeled training data is a primary concern. For our purposes, training labels are 29D real-valued vectors of expression coefficients (Sect. 3.1). These labels do not have natural interpretations that may easily be used by human operators to manually collect data. To obtain our required training data, we follow a process similar to the one used for training our FPN to estimate viewpoint in Sect. 3.3.

Obtaining sufficient expression training examples To our knowledge, there is no publicly available data set containing a sufficient number of face images, labeled with expression coefficients. Presumably, here again, one way of mitigating this problem is to use face landmark detection benchmarks. That is, taking the face images in existing landmark detection benchmarks and computing their expression

[1] FPN, bundled with rendering and alignment code, publicly available from: https://github.com/fengju514/Face-Pose-Net.

coefficients using their ground truth landmark annotations. As in Sect. 3.3, however, the number of examples we would gain would be far too small. The popular 300W data set, for example, offers 3026 images in its combined training and testing splits, and this is likely too few to train a deep CNN to regress 29D real valued output vectors.

To obtain training data, we again generate training labels by using existing methods to estimate labels for a large collection of face images. We begin by estimating 99D 3DMM coefficients for the 0.5 million face images in the CASIA WebFace collection (Yi et al. 2014) using the FaceShapeNet of Tran et al. (2017) (see also Sect. 3.2). For every CASIA image, we thus obtain the 3D face shape using the first two terms of Eq. (1).

We assume that all images belonging to the same subject should have the same, single 3D shape. We therefore apply the same shape coefficients pooling method of Tran et al. (2017) to average the 3DMM estimates for all images belonging to the same subject, thereby obtaining a single 3DMM estimate per subject. Viewpoint was then estimated using our FPN (Sect. 3.3). From this 6DoF pose estimate, we extract a projection matrix $\boldsymbol{\Pi}$ [Eq. (3)] using standard techniques (Hartley and Zisserman 2003).

Given a projection matrix $\boldsymbol{\Pi}$ that maps from the recovered 3D shape (determined by $\mathbf{F}'$ and $\boldsymbol{\eta}\mathbf{E}$) to the 2D points of an input image, we can solve the following optimization problem to get expression coefficients:

$$\begin{aligned} \boldsymbol{\eta}^{\star} &= \arg\min_{\boldsymbol{\eta}} ||\mathbf{p} - \boldsymbol{\Pi}\left(\mathbf{F}' + \boldsymbol{\eta}\mathbf{E}\right)||_2, \\ &\text{subject to} \quad |\eta_j| \leq 3\,\delta_{Ej}, \end{aligned} \tag{5}$$

where δ_{Ej} is the deviation of the j-th principal components of the 3DMM expression; $\mathbf{p}$ is a set of landmarks detected by standard facial landmark detection methods, here, CLNF (Baltrusaitis et al. 2013). The optimization itself is performed by standard Gauss-Newton optimization. This step produces a 29D expression coefficient estimate for every image in CASIA.

Training FEN to Predict Expression Coefficients The expression coefficients obtained by applying Eq. (5) are used as ground truth labels when training our FEN to regress 29D expression coefficients. In practice, we use a ResNet101 architecture (He et al. 2016) for this purpose. Our FEN is trained to regress a parametric function $f(\{\mathbf{W}, \mathbf{b}\}, \mathbf{I}) \mapsto \boldsymbol{\eta}$, where $\{\mathbf{W}, \mathbf{b}\}$ represent the parametric filters and weights of the CNN. Note that we did not experiment with smaller network structures; a more compact network architecture may work just as well here.

We note that our FEN is similar to the one used by Tran et al. (2017) (and Sect. 3.2). We use the same network architecture and training parameters. Unlike them, we used a standard ℓ_2 reconstruction loss between the FEN predictions and the expression coefficients used as ground truth. Standard stochastic gradient descent (SGD) was used for training with a mini-batch of size 144, momentum set to 0.9, and weight decay of 5e−4. FEN weights are updated with a learning rate set to 1e−3. When the validation loss saturated, we decreased learning rates by an order of magnitude until the validation loss stopped decreasing. Finally, in order to expedite the training, we removed the empirical mean from all the input faces.

Estimating expression coefficients with FEN Existing methods for expression estimation often take an analysis-by-synthesis approach to optimizing facial landmark locations. Contrary to them, we estimate expressions in a single forward pass of our FEN. To estimate an expression coefficients vector, $\boldsymbol{\eta}_t$, we evaluate $f(\{\mathbf{W}, \mathbf{b}\}, \mathbf{I}_t)$ for test image, $\mathbf{I}_t$. Similarly to 3D shape and viewpoint estimation, here too we assume that the face was detected using the method of Yang and Nevatia (2016). Here, however, we found better results were obtained with the face detection bounding box scaled by $\times 1.25$, which approximates the loose bounding boxes provided for CASIA images.

3.5 Discussion: Training Labels from Landmark Detections?

Training of both our FPN (Sect. 3.3) and our FEN (Sect. 3.4) followed a similar theme in which training labels were automatically synthesized by using facial landmark detectors. This approach raises a natural concern: Wouldn't our trained networks be only as good as the landmark detectors used to produce their labels?

First, recall that FPN was trained on images which sometimes underwent significant augmentations (Sect. 3.3). The same transformations were applied to the landmarks, producing training examples—images and viewpoint labels—which could often be too challenging for existing landmark detectors to process (Fig. 5). By training our FPN on these examples, we obtain a network that can handle such challenging viewing conditions and may therefore be more capable than the original landmark detector.

More generally, however, is the well-known robustness of deep networks to training label noise (here, errors in landmark detections and consequent viewpoint or expression estimates). This robustness is especially true in large training sets such as the 2.6 million images in the VGG Face dataset (Parkhi et al. 2015) and the 0.5 million face images in the CASIA WebFace collection (Yi et al. 2014) respectively used in training FPN and FEN. This phenomenon was demonstrated by Xie et al. (2016), who introduced label errors to *improve* training, and by Parkhi et al. (2015), who reported better face recognition accuracy with a network trained with noisy labels. A similar effect is the basis for

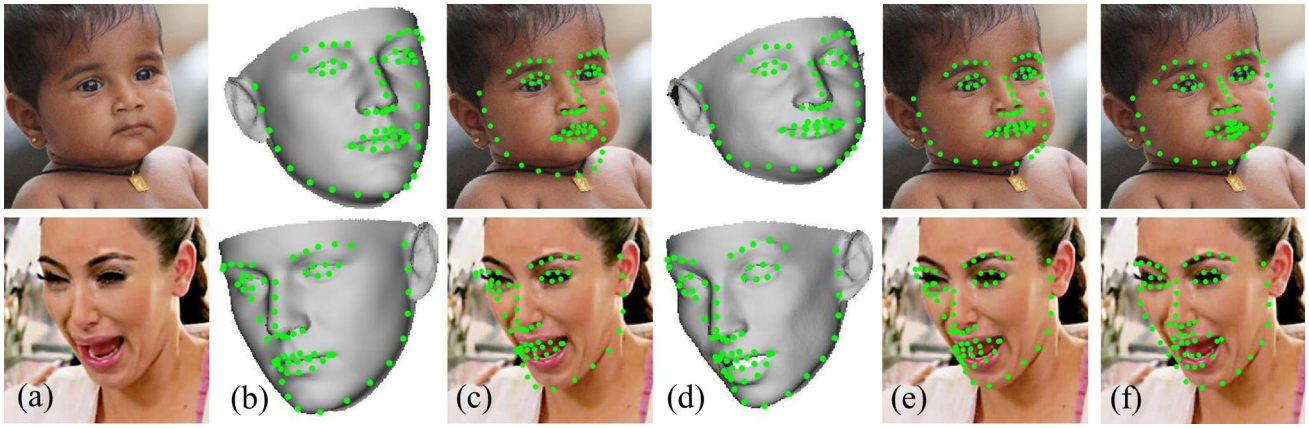


Fig. 6 From FAME to landmarks. Two examples of landmarks detected using our FAME framework. In each example: **a** Input image, **b** generic 3D face shape aligned using FPN (Sect. 3.3); reference 3D landmarks (including occluded ones) rendered in green, **c** reference 3D landmarks projected onto image (Sect. 4.1), **d** 3D shape estimated with FaceShapeNet (Tran et al. 2017), adjusted for pose with FPN, and expression with FEN (Sect. 3.4), **e** reference 3D landmarks on adjusted 3D shape, projected onto image (Sect. 4.1), finally, **f** 2D landmarks after refinement (Sect. 4.2). Note refined landmarks moved from self-occluded locations **e** to the contour landmarks **f** used by 2D landmark detection benchmarks (Color figure online)

semi-supervised methods using *pseudo-labels* (Dong et al. 2018c), though their approach is different from our own.

In our work, this robustness is reflected in our trained networks learning to generalize beyond any errors in their training labels, and so beyond the capabilities of the landmark detectors used to produce their labels. These improvements are demonstrated in Sect. 6.

4 From FAME to Landmark Detection

For a given face photo, the process described in Sect. 3 provides estimates for the 3D face shape, its viewpoint, and any deformation of its surface due to facial expressions. For many face processing applications, these estimates would suffice: For example, we later show that our FPN alone provides an effective means of aligning faces in 2D and 3D for face recognition, contributing to improved recognition accuracy. Whenever a complete 3D modeling is desired, all three networks can be used jointly.

Nevertheless, some applications may still require 2D facial landmark detection. In such cases, landmark coordinates can be obtained from our 3D face modeling as a by-product of our approach, rather being a step towards modeling.

4.1 Landmark Projections

We estimate 2D facial landmarks from a 3D face shape and a projection matrix, relating points on the 3D surface with coordinates in the input image. Our FPN (Sect. 3.3) provides an estimate for the projection matrix. 3D reference landmarks can be specified on the surface of, say, a generic 3D face shape (Hassner et al. 2015). 2D landmarks can then be estimated by projecting the reference 3D landmarks using the projection matrix of Sect. 3.3, Eq. (3) and Fig. 6b and c.

Projecting points from a generic shape would undoubtedly cause large errors, whenever the face in the image does not share the same face shape as the generic (e.g., is wider or slimmer than the generic shape) or in cases where the facial expression in the image affects landmark positions. These errors are illustrated in Fig. 6c. A more accurate estimate of landmark locations would then project 3D landmarks from a 3D face shape which matches the one in the image and obtained using FaceShapeNet (Sect. 3.2) and modified for expressions using FEN (Sect. 3.4).

Specifically, given the estimated shape coefficients $\boldsymbol{\alpha}$ (Sect. 3.2), expression coefficients $\boldsymbol{\eta}$ (Sect. 3.4), and the projection matrix derived from the 6DoF viewpoint (Sect. 3.3), we can reconstruct the 3D shape, $\mathbf{F}$, by Eq. (1). Because we use a standard 3DMM representation, all faces naturally have corresponding 3D vertices: 3D points on the face surface are consistently indexed and the same index always refers to the same facial feature (e.g., tip of the nose), no matter the particular face shape or expression.

A consequence of this correspondence is that we can select reference 3D facial landmarks of interest, $\mathbf{P}$, on an ideal 3D face shape, $\mathbf{F}$, once, at preprocessing. Given a novel face image and following our 3D modeling, we can project $\mathbf{P}$ to obtain the corresponding 2D landmark points by simply applying Eq. (3). Examples of such results are provided in Fig. 6e. We test the accuracy of landmarks predicted using FPN alone (and a generic face shape) as well as by additionally estimating expression and face shape in Sect. 6.3.

Importantly, unlike some landmark detection methods, our method of obtaining facial landmarks can easily be modified for different landmark numbers and locations without requiring re-training or redesign of the system. Instead, to obtain detections for different landmarks, we only need to select different reference 3D points $\mathbf{P}$ on $\mathbf{F}$.

4.2 Landmark Refinement

Our results in Sect. 6.3 show that projecting 3D landmarks from our estimated face model, already provides reasonable landmark accuracy. As we later discuss (Sect. 5), however, standard benchmarks for 2D facial landmark detection often measure the accuracy of landmark prediction along face contours—landmarks which, unlike our 3D reference points, change according to facial pose. If such points are desired, and to better optimize our detections to other variabilities of the manual annotations of these benchmarks, we further perform image-based, 2D landmark refinement.

To this end, we use a modified version of the *regression sampler* described by Bhagavatula et al. (2017). In our framework, we use this component to estimate 2D offsets for our projected landmarks (Sect. 4.1) and refer to it as our *offset regression network*. For details on the regression sampler, we refer to Bhagavatula et al. (2017). We found it necessary, however, to make the following changes to their design when using it in our framework.

Specifically, they used their *shared feature extractor network* to obtain local representations at landmark coordinates. We, instead, use our FPN (Sect. 3.3) for the same purpose. FPN was selected as it is trained to estimate pose, and so we expect it to capture viewpoint-related information. We found that the FPN conv2 layer is suitable for this purpose and use it as the input to our offset regression network.

Bhagavatula et al. (2017) used two convolutional layers in their regression sampler, with 3×3 and 1×1 convolutions. We found that our results improved if the input to our offset regression network was a 5×5 convolutional layer, which is then followed by their 3×3 and 1×1 layers. Finally, we appended another fully connected layer with ReLU activation to the end of our offset regression network. This layer added additional nonlinearities which were also empirically determined to improve performance.

Our offset regression network was trained separately for 2D and 3D landmark detection: for 2D training, we used AFW (Zhu and Ramanan 2012), the training splits of LFPW (Belhumeur et al. 2013), HELEN (Le et al. 2012), and AFLW-PIFA (Liu et al. 2017). We used 300W-LP (Zhu et al. 2016b) as a training set for 3D landmark refinement. In both cases, training labels for our offset regression network were obtained by subtracting the projected landmark locations (2D landmarks obtained by projecting 3D reference points; Sect. 4.1) from the ground truth landmark coordinates provided by each benchmark. The ℓ_1 loss is used here to prevent our model from being affected by few landmarks with large offsets from ground truth keypoints.

All layers in our offset regression network are initialized with parameters drawn from a Gaussian distribution with zero mean and standard deviation 0.01. All biases are initialized with zero. During training, batch size is set to 30 and the initial learning rate is set to 0.001. Learning rate is decreased by a factor of 0.8 per epoch until the validation accuracy converges. Examples of these refined landmarks are provided in Fig. 6f.

5 A Critique of Existing Test Paradigms

Our framework can be used in face processing systems as an alternative to existing facial landmark detectors. These methods are typically tested on benchmarks designed to measure landmark detection accuracy. Some such popular benchmarks include AFW (Zhu and Ramanan 2012), LFPW (Belhumeur et al. 2013), HELEN (Le et al. 2012), iBUG (Sagonas et al. 2013), and 300W (Sagonas et al. 2016). Before testing our method, we pause to consider how accuracy is evaluated on these benchmarks and raise a number of potential problems with these evaluation paradigms.

Detection accuracy measures Facial landmark detection accuracy is typically measured by considering the distances between estimated landmarks and ground truth (reference) landmarks, normalized by the reference inter-ocular distance of the face (Dantone et al. 2012):

$$e(\mathbf{L}, \hat{\mathbf{L}}) = \frac{1}{m\|\hat{\mathbf{p}}_l - \hat{\mathbf{p}}_r\|_2} \sum_{i=1}^{m} \|\mathbf{p_i} - \hat{\mathbf{p}}_\mathbf{i}\|_2, \tag{6}$$

Here, $\mathbf{L} = \{\mathbf{p_i}\}$ is the set of estimated m 2D facial landmark coordinates, $\hat{\mathbf{L}} = \{\hat{\mathbf{p}}_\mathbf{i}\}$ their ground truth locations, and $\hat{\mathbf{p}}_l$, $\hat{\mathbf{p}}_r$ the reference left and right eye outer corner positions. These errors are then translated to a number of standard quantities, including the mean error rate (MER), the percentage of landmarks detected under certain error thresholds (e.g., below 5% or 10% error rates) or the area under the accumulative error curve (AUC).

The ground truth compared against is manually specified, often by mechanical turk workers. As we detail next, these manual annotations can be misleading. Moreover, Eq. (6) itself can also be misleading: Normalizing detection errors by inter-ocular distances biases against images of faces appearing at non-frontal views. When faces are near-profile, perspective projection of the 3D face onto the image plane shrinks the distances between the eyes—thereby unnaturally inflating the errors computed for such images.

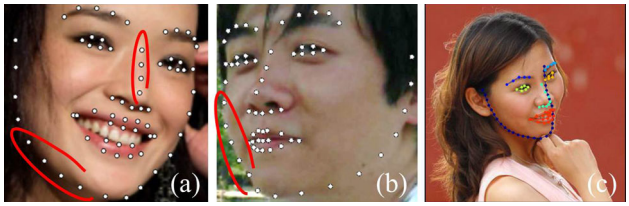


Fig. 7 Visualizing potential problems with facial landmark detection benchmarks. Three example photos along with their ground truth landmark annotations. **a** Annotations on the jawline and bridge of the nose do not correspond to well-defined facial features [from LFPW (Belhumeur et al. 2013)]. **b** Landmarks represent points on the face *contour* and so represent different facial locations in different views [AFW (Zhu and Ramanan 2012)]. **c** 3D landmark annotations represent occluded facial regions [3D Menpo (Zafeiriou et al. 2017)]

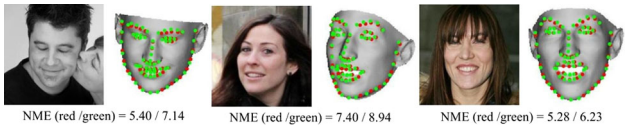


Fig. 8 Reference landmark selection versus detection accuracy. 68 reference points projected from a 3D face shape to the input image, along with landmark detection accuracy. Two sets of 3D reference points are considered: red points were manually annotated on the reference 3D face, green points were obtained by projecting 2D landmarks, detected by dlib (King 2009; Kazemi and Sullivan 2014) on a frontal face image, onto the 3D reference shape. Both sets of landmarks can be equally used for face alignment and processing, but red points produce far lower landmark prediction errors. Does this imply that the red points are better than the green? (Color figure online)

Ill-defined facial locations Recent benchmarks provide annotations for 49 or 68 facial landmarks. These annotations are presumed to offer more stability than the far fewer (typically five) landmarks originally used by older sets such as AFW (Zhu and Ramanan 2012) and LFPW (Belhumeur et al. 2013). By doing so, however, these extended annotations include many facial locations which do not correspond to well-defined facial features, such as points along the jawline or the bridge of the nose (illustrated in Fig. 7a).

It is well known that human annotators tend to vary greatly in the positions they choose for these ill-defined landmarks (Sagonas et al. 2016). This variance, however, is not reflected in the measures used to report accuracy [e.g., Eq. (6)]. Thus, estimating plausible positions for jawline landmarks may raise or lower the error depending on the ground truth annotations, despite any uncertainty of the ground truth.

Viewpoint-dependent facial locations 2D landmark detection benchmarks provide landmark annotations on face contours (Fig. 7b). These landmark locations correspond to different facial features in different views. Although there may be applications which require detection of facial contours, the use of these landmarks for face alignment (i.e., by matching them with corresponding points in reference views; Sect. 2.3) can introduce alignment errors.

Occluded points In response to this last problem, some have recently proposed labeling faces with 3D points, which correspond to the same facial features regardless of the viewpoint (Sagonas et al. 2016; Zafeiriou et al. 2017). The problem here is, of course, that because these points are occluded, it would be hard for human annotators to reliably localize these points, leading to increased uncertainty in their ground truth annotations (Fig. 7c).

Illustrative examples Figure 8 illustrates some of these problems, using landmarks estimated with our FAME approach and the landmark projection method of Sect. 4. In particular, it demonstrates the effect 3D reference landmark selections have on landmark estimation accuracies, measured on the 300W benchmark (Sagonas et al. 2016).

For each of the three example faces, we projected two different sets of 3D landmarks using the method described in Sect. 4. One was manually specified (red points), whereas the other (green) was obtained by using dlib (King 2009; Kazemi and Sullivan 2014) to detect 2D landmarks on a reference, frontal face photo (not shown) and projecting these landmarks onto the 3D surface.

Both sets of points are reasonable choices for reference facial landmarks and, of course, both sets of points can be used for the same purposes. For example, consistently using the same reference point selection (either red or green) would provide the landmark correspondences required for 2D or 3D face alignment between face photos (Sect. 2.3). These correspondences would be equally accurate regardless of which set of reference landmarks are used (so long as this selection is consistent).

We provide landmark detection accuracies for these two sets of landmarks on three images from the 300W benchmark, using its ground truth annotations to compute these errors. These errors are clearly very different despite the fact that there is no real practical difference between the two sets of points. These results support our concerns with landmark detection benchmarks. They suggest that errors reported on these benchmarks may not reflect meaningful differences in the quality of different detection methods, but rather the accuracy of these methods in predicting manual annotations.

Landmark detection versus face processing application Consequent to this visualization, and possibly the most important concern, is that facial landmark detectors are not used on their own but rather by face processing systems that require landmarks for a variety of purposes. Because landmark detectors are tested independently of these face processing systems, it is not clear if and how improving their accuracy affects the performance of the systems that use them. In fact, as we show in our experiments (Sect. 6), older landmark detectors which were outperformed on landmark detection benchmarks by

newer methods, sometimes provide better bottom-line results than their newer alternatives.

6 Experimental Results

We extensively test our approach and its components, reporting a wide range of quantitative and qualitative results. Following the concerns raised in Sect. 5, we propose a novel approach for evaluating face alignment and expression estimation methods—that is, by measuring their effect on bottom-line performances of face recognition (Sect. 6.1) and emotion classification benchmarks (Sect. 6.2). Doing so also places different landmark detection techniques (with varying numbers of detected landmarks) and direct approaches such as ours on even grounds, allowing for direct comparison between these methods. Although landmark prediction is not a focus of this paper, we additionally provide extensive landmark prediction results in Sect. 6.3.

We note that FaceShapeNet (Sect. 3.2) is not a contribution of this paper. The 3D shapes estimated by that network have been extensively tested by Tran et al. (2017) and shown to provide state-of-the-art 3D reconstruction accuracy as well as robustness to extreme viewing conditions. We refer to their paper for more details on these results. Their method, however, does not estimate viewpoint or adjust the 3D face to account for expressions. We therefore focus on evaluating viewpoint and expression accuracies.

6.1 Evaluating Face Alignment

Facial landmarks are predominantly used for 2D and 3D face alignment in face recognition systems (Chang et al. 2017). Rather than measure the accuracy of the landmark positions, we therefore propose to test the effects different landmark detectors (or, more generally, different face alignment techniques) have on face recognition accuracy. The rationale is that a face recognition system which uses aligned face images would be affected by the quality of the method used for this alignment. Better alignment should therefore translate to better face recognition results, regardless of the specific method used for alignment.

Importantly, the purpose of these experiments is *not* to demonstrate state-of-the-art face recognition accuracy. In particular, we do not optimize a face recognition pipeline in order to set new recognition accuracy rates. Instead, we use face recognition benchmarks to compare the effects of different alignment methods.

Face recognition for evaluating face alignment We test our improved FPN of Sect. 3.3 independently of the other two networks, using it to estimate the viewpoint of a face in an input image and then align the face in 2D and 3D. As baseline methods, we use the following popular, state-of-the-art, facial landmark detectors: dlib (King 2009), CLNF (Baltrusaitis et al. 2013), OpenFace (Baltrušaitis et al. 2016), DCLM (Zadeh et al. 2016), RCPR (Artizzu et al. 2013), and 3DDFA (Zhu et al. 2016b).

We clarify that deep networks, both our FPN and some of the baseline detectors (Baltrušaitis et al. 2016; Zadeh et al. 2016; Zhu et al. 2016b), require large quantities of training data. Training FPN requires 2.6 million images for the AlexNet-structure and 122K for the ResNet variant. These numbers are greater than those used to develop older methods such as dlib (King 2009). These differences should be considered when comparing results of these methods.

Face recognition benchmarks Our tests use two of the most recent benchmarks for face recognition: IARPA Janus Benchmark A (Klare et al. 2015) and B (Whitelam et al. 2017) (IJB-A and IJB-B, resp.). Importantly, these benchmarks were designed with the specific intention of elevating the difficulty of face recognition. This heightened challenge is reflected by, among other factors, an unprecedented amount of extreme out-of-plane rotated faces, including many appearing in near-profile views (Masi et al. 2016b). Consequently, these two benchmarks raise the bar well above other facial landmark detection benchmarks.

Face recognition pipeline All face alignment methods were tested using the same face recognition pipeline, similar to the one proposed by Masi et al. (2017, 2016b). We use their system partly because its code and trained models are publicly available, allowing for reproduction of our results.

More importantly, however, their system explicitly aligns faces to multiple viewpoints, in both 2D and 3D. These steps are highly dependent on viewpoint estimation quality and so their recognition accuracy reflects viewpoint accuracy. In practice, we used their 2D (similarity transform) and 3D (face rendering) code directly, only changing the viewpoint estimation step. Our tests compare different landmark detectors used to recover the 6DoF head pose required by their warping and rendering method (converting facial landmarks as described for our FPN training data in Sect. 3.3), with the 6DoF directly regressed by our FPN.

Their system uses a single ResNet101 architecture (He et al. 2016), trained on both real face images and synthetic, rendered views. We found that better face recognition results can be obtained by fine-tuning their network using $L2$-constrained Softmax Loss (Ranjan et al. 2017), rather than their original Softmax (Masi et al. 2017, 2016b). This fine-tuning is performed using the MS-Celeb face set (Liu et al. 2015) as the training set. Aside from this change, we use the same recognition pipeline from (Masi et al. 2017), and we refer to that paper for details.

Bounding box detection We emphasize that we tested all methods with an identical pipeline, only changing alignment

Table 2 Verification and identification on IJB-A and IJB-B, comparing landmark detection—based face alignment methods. Three baseline IJB-A results are also provided as reference at the top of the table

Method	Eval.						
	TAR@FAR			Identification rate (%)			
	.01%	0.1%	1.0%	Rank-1	Rank-5	Rank-10	Rank-20
	IJB-A (Klare et al. 2015)						
Crosswhite et al. (Crosswhite et al. 2017)	–	–	93.9	92.8	–	98.6	–
Ranjan et al. (Ranjan et al. 2017)	90.9	94.3	97.0	97.3	–	98.8	–
Masi et al. (Masi et al. 2017)	56.4	75.0	88.8	92.5	96.6	97.4	98.0
RCPR (Artizzu et al. 2013)	64.9	75.4	83.5	86.6	90.9	92.2	93.7
Dlib (King 2009)	70.5	80.4	86.8	89.2	91.9	93.0	94.2
CLNF (Baltrusaitis et al. 2013)	68.9	75.1	82.9	86.3	90.5	91.9	93.3
OpenFace (Baltrušaitis et al. 2016)	58.7	68.9	80.6	84.3	89.8	91.4	93.2
DCLM (Zadeh et al. 2016)	64.5	73.8	83.7	86.3	90.7	92.2	93.7
3DDFA (Zhu et al. 2016b)	74.8	82.8	89.0	90.3	92.8	93.5	94.4
Our FPN	77.5	85.2	90.1	91.4	93.0	93.8	94.8
Our improved FPN	**78.5**	**86.0**	**90.8**	**91.6**	**93.4**	**94.0**	**94.8**
	IJB-B (Whitelam et al. 2017)						
GOTs (Whitelam et al. 2017)*	16.0	33.0	60.0	42.0	57.0	62.0	68.0
VGG Face (Whitelam et al. 2017)*	55.0	72.0	86.0	78.0	86.0	89.0	92.0
RCPR (Artizzu et al. 2013)	71.2	83.8	93.3	83.6	90.9	93.2	95.0
Dlib (King 2009)	78.1	88.2	94.8	88.0	93.2	94.9	96.3
CLNF (Baltrusaitis et al. 2013)	74.1	85.2	93.4	84.5	90.9	93.0	94.8
OpenFace (Baltrušaitis et al. 2016)	54.8	71.6	87.0	74.3	84.1	87.8	90.9
DCLM (Zadeh et al. 2016)	67.6	81.0	92.0	81.8	89.7	92.0	94.1
3DDFA (Zhu et al. 2016b)	78.5	89.1	95.6	89.0	94.1	95.5	96.9
Our FPN	**83.2**	**91.6**	96.5	91.1	95.3	96.5	**97.5**
Our improved FPN	**83.2**	**91.6**	**96.6**	**91.6**	**95.6**	**96.7**	**97.5**

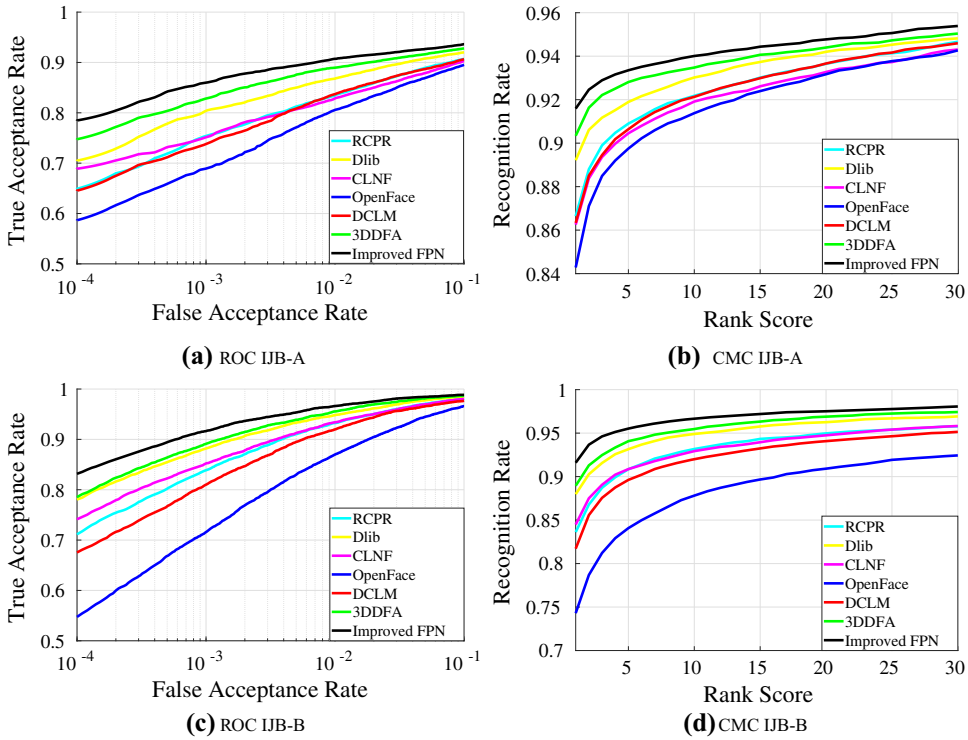
Values in bold indicates the best performance
*Numbers estimated from the ROC and CMC in (Whitelam et al. 2017)

methods; different results vary only in the method used to estimate facial pose. The only other difference was in the facial bounding box detector.

Facial landmark detectors can be sensitive to the face detector they are used with. We therefore report results obtained when running landmark detectors with the best bounding boxes we were able to determine for each method. Specifically, FPN was tested with the bounding boxes returned by the detector of Yang and Nevatia (2016), as described in Sect. 3.3, including the rescale factor of $\times 1.25$. We found that most facial landmark detectors performed best when applied with the same face detector but without the 25% increase. Finally, 3DDFA (Zhu et al. 2016b) was tested with the same face detector followed by the face box expansion code provided by its authors.

Face verification and identification results Face verification and identification results on both IJB-A and IJB-B are provided in Table 2. We report multiple recognition metrics for both verification and identification. For verification, these measure the recall (true acceptance rate) at three cutoff points of the false alarm rate (TAR-{1%,0.1%,0.01%}). For identification we provide recognition rates at four ranks from the CMC (cumulative matching characteristic). The overall performances in terms of ROC and CMC curves are shown in Fig. 9. For clarity, the figure only reports results for the ResNet101 version of our FPN. The table also provides, as reference, three state-of-the-art IJB-A results (Crosswhite et al. 2017; Masi et al. 2017; Ranjan et al. 2017) and baseline results from (Whitelam et al. 2017) for IJB-B. To our knowledge, we are the first to report verification and identification accuracies on IJB-B.

Faces aligned with our original FPN and the ResNet101 FPN (improved FPN) lead to better recognition accuracy, even compared to the most recent, state-of-the-art facial landmark detection method of (Zadeh et al. 2016). Remarkably, faces aligned with FPN provide substantially better recognition accuracy than those aligned with the facial landmark detector of Baltrušaitis et al. (2016), which is the method used



(a) ROC IJB-A **(b)** CMC IJB-A **(c)** ROC IJB-B **(d)** CMC IJB-B

Fig. 9 Verification and identification results on IJB-A and IJB-B. ROC and CMC curves accompanying the results reported in Table 2

Table 3 Runtime for estimating 6DoF facial viewpoint. Times reported in seconds for the AlexNet FPN, tested on a GPU. On the CPU, FPN runtime was 0.07 s. 3DDFA used the AFW collection for training. Code provided for 3DDFA (Zhu et al. 2016b) did not allow testing on the GPU; their paper reports GPU runtime to be 0.076 s

RCPR	Dlib	CLNF	OpenFace	DCLM	3DDFA	Our FPN
0.19	0.009	0.38	0.31	15.83	0.6	0.005

to produce the viewpoint labels for training our FPN. This result supports the claims made in Sect. 3.5 on the robustness of deep networks to label noise.

Face alignment runtime An important factor when designing face processing pipelines is the speed required by different components in these systems. We therefore measure the run time required by different face alignment methods. Our tests were performed on a machine with an NVIDIA, GeForce GTX TITAN X, 12GB RAM, and an Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz, 132GB RAM. The only exception was 3DDFA (Zhu et al. 2016b), which required a Windows system and was tested using an Intel(R) Core(TM) i7-4820K CPU @ 3.70GHz (8 CPUs), 16GB RAM, running MS-Windows 8 Pro, 64-bit.

Table 3 reports mean, per-image runtime for landmark detection compared with our AlexNet-based FPN. FPN is an order of magnitude faster than nearly all other alignment methods. Dlib (King 2009) is slightly slower than FPN, but far less accurate in the face recognition tests (Table 2).

Also noteworthy is that although our FPN produced better recognition results, of all the other methods, those of dlib were nearly the best. This is significant, as dlib is one of the oldest methods. Later methods were shown to excel on landmark detection benchmarks. The fact that the older dlib produces better face recognition results despite no longer being state-of-the-art on landmark detection benchmarks, supports our concerns regarding these landmark detection benchmarks and the common practice of measuring detection accuracy independently from bottom-line performance of the methods which use these detectors (Sect. 5).

6.2 Evaluating Expression Estimation

We next compare different means of estimating non-rigid facial deformations due to expressions. It is important to note that few previous methods for 3DMM face shape estimation provided quantitative results; instead, most relied primarily on qualitative examples. Although we similarly offer qualitative results of our face modeling, including expression estimation (Sect. 6.4), we also provide quantitative results.

We propose to compare different methods designed to capture facial expressions: facial landmark detection methods typically used for this purpose and our FEN. Because we are unaware of any benchmark designed for estimating facial expression accuracy, we propose instead to use emotion classification benchmarks for this purpose. Specifically, we use estimated facial expressions as feature vectors and test how

well these feature vectors capture the seven facial emotions in emotion classification benchmarks.

We again note that our goal is not to break classification accuracy records on these benchmarks. Although some emotion classification methods use landmark detection for this purpose, others take a direct, landmark-free approach and produce deep emotion features directly from image intensities (Levi and Hassner 2015; Surace et al. 2017). Our goal is not to outperform these methods, but only to use existing emotion classification benchmarks as quantitative tests of facial expression estimation methods.

Benchmark settings We use two emotion classification benchmarks containing face images labeled for discrete emotion classes. For each image we estimate its expression coefficients, either directly using our FEN or 3DDFA, or by using detected landmarks to solve Eq. (5) for the expression coefficients, as described in Sect. 3.4. We then attempt to classify the emotion of a test image using the same, simple classification pipeline applied to the different 29D expression estimates.

Our tests use the Extended Cohn-Kanade (CK+) dataset (Lucey et al. 2010) and the Emotion Recognition in the Wild Challenge (EmotiW-17) dataset Dhall et al. (2017). The CK+ dataset is a constrained set, with frontal faces taken in the controlled conditions. EmotiW-17, on the other hand, contains highly challenging video frames collected from 54 movie DVDs (Dhall et al. 2012).

The CK+ dataset contains 327 face video clips labeled for seven emotion classes: anger (An), contempt (Co), disgust (Di), fear (Fe), happy (Ha), sadness (Sa), surprise (Su). From each clip, we take the peak frame (the end of video)—the frame assigned with an emotion label—and use the expression estimated from it for emotion classification. Following the protocol used by Lucey et al. (2010), we ran a leave-one-subject-out test protocol to assess performance.

The EmotiW-17 dataset, on the other hand, contains 383 face video clips. These clips are labeled for the seven emotion classes: anger (An), disgust (Di), fear (Fe), happy (Ha), neutral (Ne), sadness (Sa), surprise (Su). We estimate 29D expression representations for every frame and apply element-wise average pooling over the estimated expression coefficients of all frames in each video.

Finally, we also evaluate the robustness of different methods to scale changes. Specifically, we tested all methods on multiple versions of the CK+ and EmotiW-17 benchmarks, each version with all images scaled down by factors of $\times 0.8$, 0.6, 0.4, and 0.2 of their sizes.

Emotion classification pipeline The same, simple classification method was used in all our tests. We preferred a simple classification method, rather than a state-of-the-art technique, in order to avoid masking the accuracy of the landmark detector/emotion estimation with an elaborate classifier. We therefore use a simple kNN classifier with $K = 5$, without optimizing for K. Nevertheless, we additionally report results with a SVM (RBF kernel) to show the consistent improvement of our method irrespective of the classifier used. Of course, all reported results are far from the state-of-the-art on this set. As previously noted, our goal is not to outperform state-of-the-art emotion classification, but rather to compare methods for expression coefficient estimation.

Baseline methods We compare our approach to widely used, state-of-the-art face landmark detectors: Dlib (King 2009), CLNF (Baltrusaitis et al. 2013), OpenFace (Baltrušaitis et al. 2016), DCLM (Zadeh et al. 2016), RCPR Artizzu et al. (2013), and 3DDFA (Zhu et al. 2016b). Of these, 3DDFA is the only one that, similar to our FEN, directly estimates 29D expression coefficients vectors. For all other methods, following landmark detection, expression coefficients were estimated using Eq. (5).

Emotion classification results Fig. 10 reports the emotion classification confusion matrix on the original CK+ data set (unscaled) for our FEN (Fig. 10c), comparing it to the other two best performing methods: 3DDFA (Fig. 10b) and DCLM (Fig. 10a). Our expression coefficients were able to capture emotions of surprise (Su), happy (Ha), and disgust (Di), but were less able to represent more subtle facial emotions: anger (An), contempt (Co), fear (Fe), and sadness (Sa). These same classes were not handled well by the other methods. Overall, however, our expressions were noticeably better at capturing all emotion classes.

Figure 11 shows the emotion classification confusion matrix on the original EmotiW-17 data set (unscaled). Our expression coefficients were able to capture neutral (Ne), happy (Ha), sad (Sa), and angry (An), but were less able to represent the emotions which were less clearly defined by expressions in the benchmark: disgust (Di), fear (Fe), and surprise (Su), which from our observations, often appear very similar to angry (An). Note that these confusion matrices reflect our experiments with a kNN classifier.

Sensitivity to scale changes Figures 12 and 13 additionally provide emotion classification performances of all methods on both CK+ and EmotiW-17 images, but with increasing scale changes, for both the kNN and SVM classifiers. The plot shows the sensitivity of each tested method with respect to the input resolution. The x-axis reports the downsizing applied to the images, with a factor proportional to the scale. That is, scale $= 1$ represents original image sizes (640×490 for CK+; 730×576 for EmotiW-17), while the lowest scale of 0.2 refers to 128×98 images in CK+ and 146×115 in EmotiW-17. Note that whenever deep networks were used, images were subsequently rescaled to the size of the input layer: 224×224.

Figures 12 and 13 both show that our approach is not only the most accurate, but also the most stable across

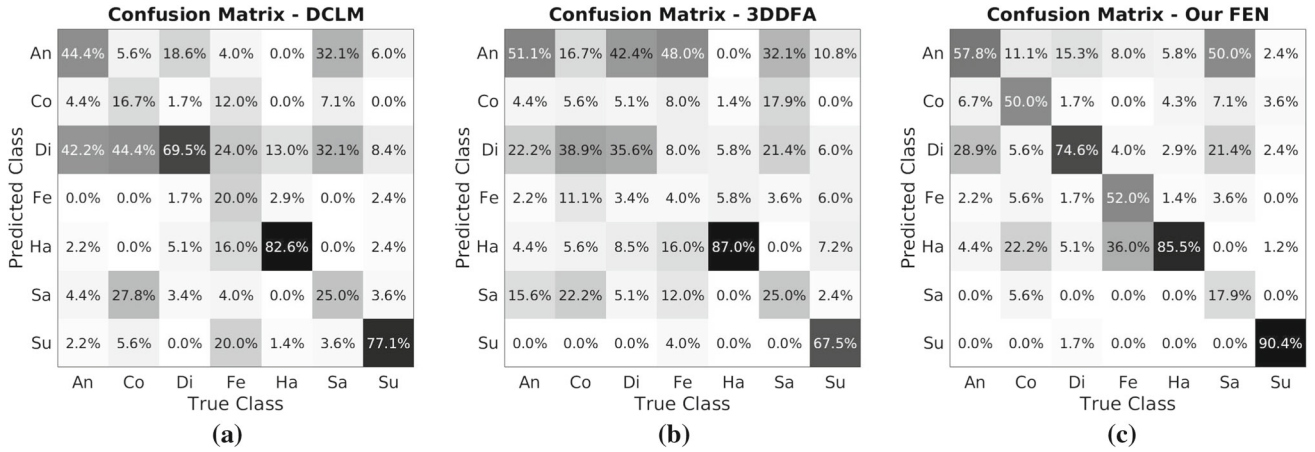


Fig. 10 Confusion Matrices for emotion classification on the CK+ benchmark. Classification confusions on the seven emotion classes in CK+, for the original (unscaled) images. Results provided for the top performing three methods: **a** DCLM landmarks (Zadeh et al. 2016) and Expression Fitting, **b** the deep, direct method of (Zhu et al. 2016b), **c** our FEN

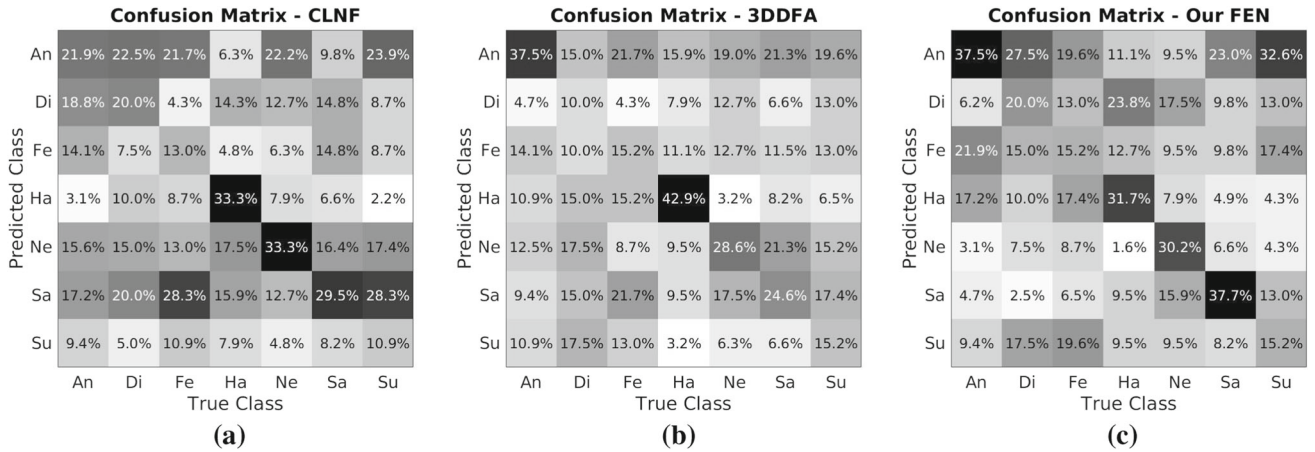


Fig. 11 Confusion Matrices for emotion classification on the EmotiW-17 benchmark. Classification confusions on the seven emotion classes in EmotiW-17, for the original (unscaled) images. Results provided for the top performing three methods: **a** CLNF landmarks (Baltrusaitis et al. 2013) and Expression Fitting, **b** the deep, direct method of Zhu et al. (2016b), **c** our FEN

scale changes. This robustness indicates that facial landmark detection methods are more sensitive to scale changes; the same face at different scales results in different landmark detections. This is an important factor that is not explicitly examined by facial detection benchmarks. The figures also show that SVM generally performs better than kNN but, importantly, the improvement of our approach over landmark-based methods is consistent, irrespective of the classifier.

Qualitative expression results Figure 14 provides rendered views of the expressions estimated for CK+ images labeled as presenting *fear*, *surprise*, and *contempt* facial emotions. These examples illustrate estimates obtained on the original images ($\times 1$) and the images scaled down to the lowest resolution ($\times 0.2$). The figure also provides results for the two methods which performed best (aside from our FEN): DCLM (Zadeh et al. 2016) and 3DDFA (Zhu et al. 2016b). These rendered images support our quantitative results, and show that DCLM produces visibly different expression coefficients for images at different scales (e.g., DCLM results for *surprise*).

Figure 15 provides additional qualitative results on unconstrained, EmotiW-17 images. Here the two best performing methods (other than our own) were CLNF (Baltrusaitis et al. 2013) and 3DDFA (Zhu et al. 2016b). CLNF is noticeably affected by scale changes (e.g., *neutral*).

Expression estimation runtime Table 4 lists runtimes for the various methods tested in our experiments. Our method is the fastest. We note that expression fitting methods relying on landmarks involve three separate steps: (i) landmark detec-

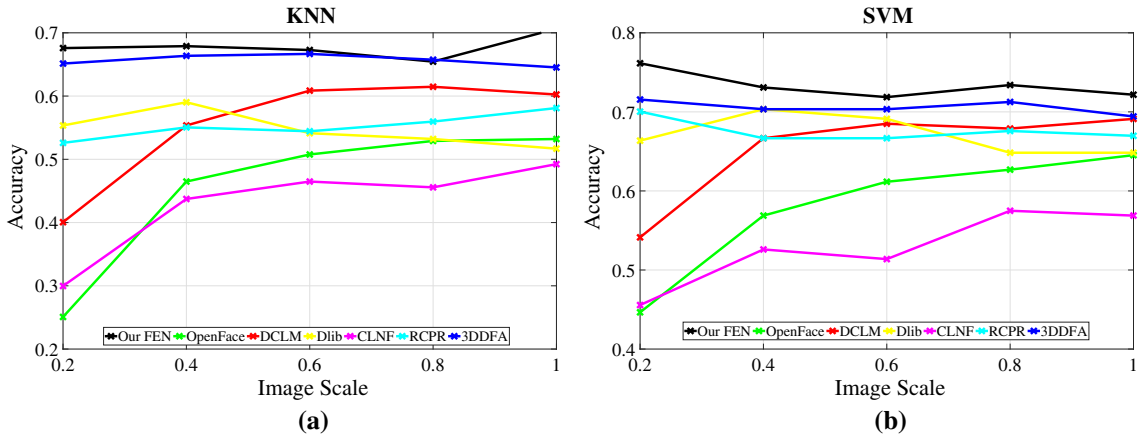


Fig. 12 Emotion classification over scales on the CK+ benchmark. Curves report emotion classification accuracy over different scales of the input images. Lower scale indicates lower resolution. Original resolution is 640 × 490. **a** Reports results with a simple kNN classifier. **b** Same as **a**, now using a SVM (RBF kernel) classifier

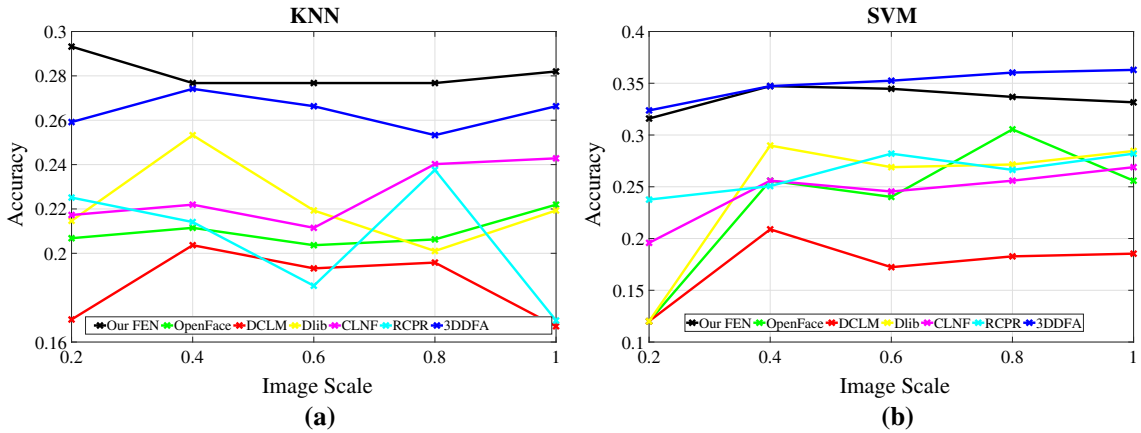


Fig. 13 Emotion classification over scales on the EmotiW-17 benchmark. Curves report emotion classification accuracy over different scales of the input images. Lower scale indicates lower resolution. Original resolution is 720×576. **a** reports results with a simple kNN classifier. **b** Same as **a**, now using a SVM (RBF kernel) classifier

tion, (ii) pose estimation, and (iii) expression fitting. By doing so, the total processing time is a sum of multiple factors. Thus, although some landmark detection methods (e.g., dlib) are very efficient in extracting landmarks (0.009 s), they still need to address the optimization problem of Eq. (5), leading to a runtime which is slower than the proposed method.

We again note that code for the 3DDFA method of Zhu et al. (2016b) was released for use on the CPU. In their paper they report GPU performances which are much faster (0.076 s for landmark detection). Runtimes were measured on the same machines used in Sect. 6.1.

6.3 Landmark Detection Accuracy

Section 4 explains how facial landmark estimates can be obtained from our 3D models. Although our goal is 3D face modeling and *not* breaking published state-of-the-art performance on face landmark detection benchmarks, it is instructional to consider the accuracy of landmarks estimated as a by-products of our FAME approach. We therefore tested our method on landmark detection benchmarks for 2D landmarks—the 300W (Sagonas et al. 2016) and AFLW-PIFA (Jourabloo and Liu 2016) benchmarks—and 3D landmarks, the AFLW 2000-3D dataset (Zhu et al. 2016b).

300W 300W (Sagonas et al. 2016) contains multiple face alignment sets with 68 landmark annotations: AFW, LFPW, HELEN, and iBUG.

AFLW-PIFA This dataset (Jourabloo and Liu 2015) offers 5200 images sampled from AFLW (Köstinger et al. 2011) with a balanced distribution of yaw angles and left versus

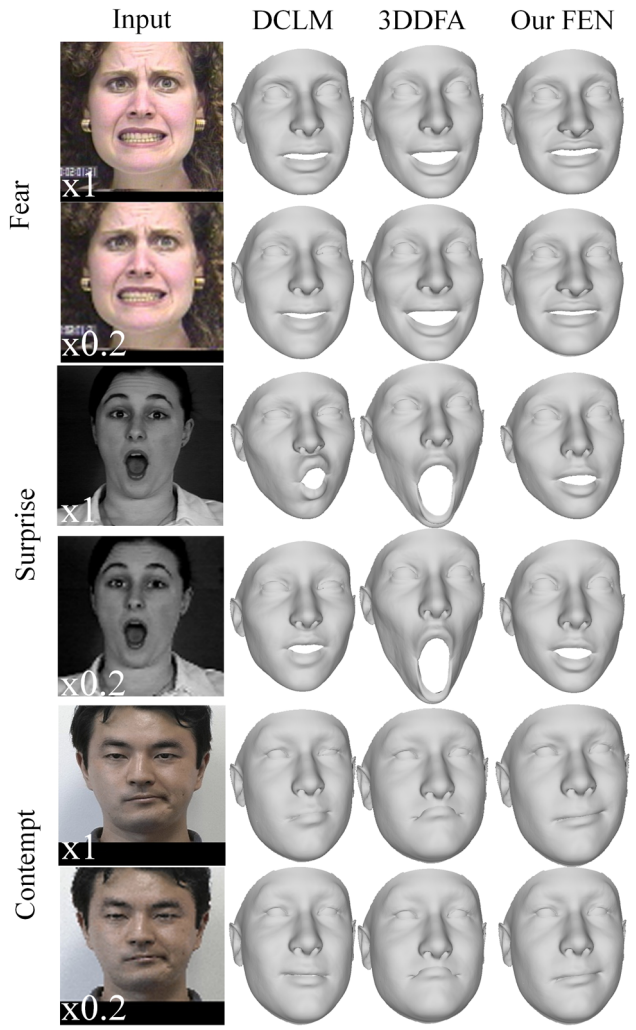


Fig. 14 Qualitative expression estimation on CK+. Example results for *fear*, *surprise*, and *contempt* facial emotions. Results provided for the original images (×1) and the images scaled down to the lowest resolution (×0.2). Results provided for the three best performing methods: DCLM (Zadeh et al. 2016), 3DDFA (Zhu et al. 2016b), and our FEN. Compared with FEN, 3DDFA overestimates expressions and DCLM is visibly less stable across scale changes. Note: Results rendered with a generic shape and fixed viewpoint

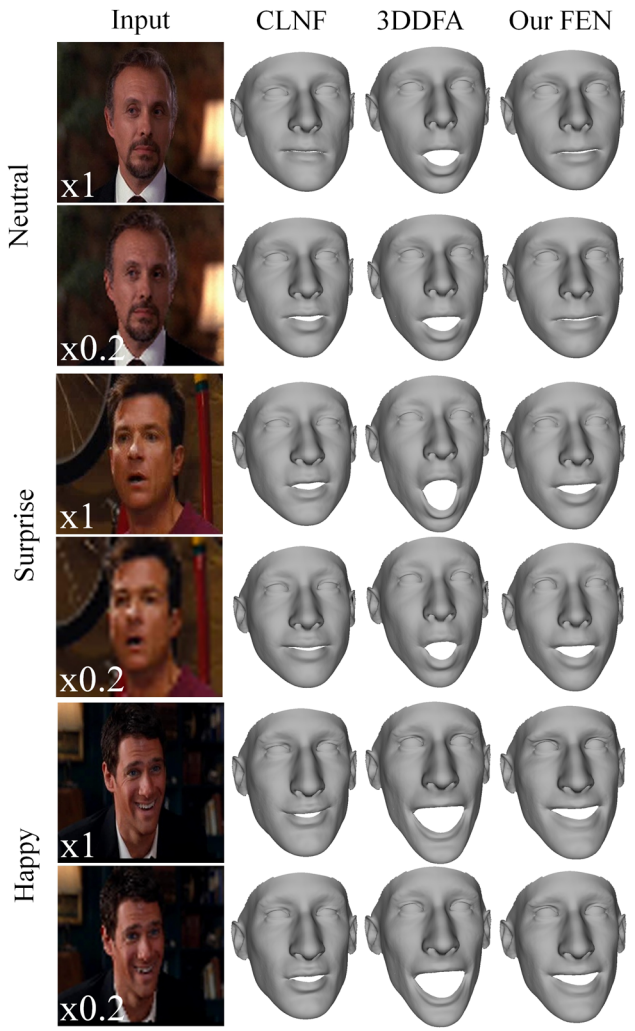


Fig. 15 Qualitative expression estimation on EmotiW-17. Example results for *neutral*, *surprise*, and *happy* facial emotions. Results provided for the original images (×1) and the images scaled down to the lowest resolution (×0.2). Baseline results for the three best performing methods: CLNF (Baltrusaitis et al. 2013), 3DDFA (Zhu et al. 2016b), and our FEN. Compared with FEN, 3DDFA overestimates expressions and CLNF is less stable across scale changes. Note: Results rendered with a generic shape and fixed viewpoint

right viewpoints. Each image is labeled with up to 21 landmarks with a visibility label for each landmark. Jourabloo and Liu (2016), offer 13 additional landmarks for these images, for a total of 34 landmarks per image.

AFLW2000-3D 3D face modeling and alignment is widely evaluated on the set offered by Zhu et al. (2016b). The data set contains ground truth 3D faces and corresponding 68 landmarks for the first 2000 AFLW samples.

Importantly, ground truth annotations in 300W and AFLW-PIFA are provided only for visible face contours. Thus, the same landmarks in different images reflect different facial details (Sect. 5). Landmark annotations in AFLW 2000-3D (Zhu et al. 2016b), on the other hand, reflect the same facial features, though these features may be self-occluded in different images and their locations were therefore guessed by the annotators. Regardless, landmark annotations in all these benchmarks are conceptually different.

Evaluation metrics Alignment accuracy is evaluated by the normalized mean error (NME): the average of landmark error normalized by the interocular distance on 300W (Sagonas et al. 2016), and by the bounding box size on AFLW-PIFA and AFLW2000-3D datasets (Jourabloo and Liu 2015; Yu et al. 2013).

300W results with ground truth bounding boxes We follow the standard protocol of Zhu et al. (2015a), where the train-

Table 4 Expression estimation runtime. Runtime for expression fitting for recent methods. Landmark-based methods need to address for landmark extraction and then optimization fitting at test-time; where as deep methods are solving the entire problem in a single step

Time	Landmark-based					Deep, direct	
	Dlib	DCLM	OpenFace	CLNF	RCPR	3DDFA	Us
Landmarks	0.009	15.83	0.31	0.38	0.19	–	–
Pose fitting	0.29	0.29	0.29	0.29	0.29	–	–
Expr. fitting	0.30	0.30	0.30	0.30	0.30	–	–
Total	0.599	16.42	0.90	0.97	0.78	0.6	0.088

Time for pose and expression fitting are shared by all five landmark–based methods. Time reported in seconds per image

Table 5 The NME (%) of 68 point detection results on 300W with ground truth bounding boxes provided by 300W. We use the typical split: common (HELEN and LFPW), challenging (iBUG), and full (HELEN, LFPW, and iBUG)

Method	Comm.	Chall.	Full	Sec./im.
TSPM (Zhu and Ramanan 2012)	8.22	18.33	10.20	–
ESR (Cao et al. 2014)	5.28	17.00	7.58	–
CFAN (Zhang et al. 2014)	5.50	16.78	7.69	–
RCPR (Artizzu et al. 2013)	6.18	17.26	8.35	0.19
SDM (Xiong and De la Torre 2013)	5.57	15.40	7.50	–
LBF (Ren et al. 2014)	4.95	11.98	6.32	–
Dlib* (King 2009)	5.41	20.31	8.33	0.009
CLNF* (Baltrusaitis et al. 2013)	5.64	17.08	7.88	0.19
OpenFace* (Baltrušaitis et al. 2016)	4.57	14.41	6.50	0.28
TCNN (Wu et al. 2017)	4.10	11.86	5.62	–
PCD-CNN (Kumar and Chellappa 2018)	3.67	7.62	4.44	–
DCLM (Zadeh et al. 2016)	3.42	7.66	4.25	15.83
3DDFA (Zhu et al. 2016b)	6.15	10.59	7.01	0.6
3DDFA+SDM (Zhu et al. 2016b)	5.53	9.56	6.31	–
FPN (AlexNet), FEN (ResNet101), FaceShapeNet (ResNet101)				
FPN	8.34	13.78	9.40	**0.005**
FPN+FEN+FaceShapeNet	5.80	11.39	6.89	0.029
FPN+FEN+FaceShapeNet+ref.	5.03	10.59	6.12	0.20
FPN (ResNet101), FEN (ResNet101), FaceShapeNet (ResNet101)				
FPN	5.78	9.82	6.57	0.088
FPN+FEN+FaceShapeNet	3.93	7.57	4.64	0.112
FPN+FEN+FaceShapeNet+ref.	**3.34**	**6.56**	**3.97**	0.283

Values in bold indicates the best performance
*These methods were tested by us using codes provided by their authors

ing part of LFPW, HELEN and the entire AFW are used for fine-tuning our FAME networks, and perform testing on three parts: the test samples from LFPW and HELEN as the *common* subset, the 135-image iBUG as the *challenging* subset, and the union of them as the *full* set with 689 images in total.

Table 5 compares our landmark detection errors with those of recent state-of-the-art methods. We use the ground truth face bounding boxes provided by the benchmark. Note that all the baseline results provided in Table 5 except for dlib, CLNF, and OpenFace were reported by their authors in the original publications.

Our results (*FPN (AlexNet) + FEN (ResNet101) + FaceShapeNet (ResNet101) + refinement*) achieve comparable results with the recent state-of-the-art (Zhu et al. 2016b). Although more accurate detections are reported by Zadeh et al. (2016), their method is *three orders of magnitude* slower than our own. Furthermore, as our face recognition results show in Sect. 6.1, better landmark detection accuracy does not always imply better bottom-line performance of the face processing pipeline.

Finally, we note the effects of better approximating the 3D face shape, as evident in our bottom three results. Landmarks estimated using FPN alone are not particularly accurate. By adding shape and expression estimation (FPN + FEN + FaceShapeNet), predictions are substantially improved. Landmark refinement (Sect. 4.2) provides an additional drop in landmark localization errors compared with the manual annotations.

Table 6 The NME (%) of 68 point detection results on 300W with the bounding box provided by the face detector of Yang and Nevatia (2016). We use the typical splits: common (HELEN and LFPW), challenging (iBUG), and full (HELEN, LFPW, and iBUG)

Method	Comm.	Chall.	Full	Sec./im.
RCPR (Artizzu et al. 2013)	8.58	22.33	11.27	0.19
Dlib (King 2009)	4.50	17.23	6.99	0.009
CLNF (Baltrusaitis et al. 2013)	8.19	21.09	10.72	0.38
OpenFace (Baltrušaitis et al. 2016)	4.81	15.45	6.90	0.31
DCLM (Zadeh et al. 2016)	4.10	13.74	5.99	15.83
3DDFA (Zhu et al. 2016b)	10.64	12.87	11.08	0.6
FPN (AlexNet), FEN (ResNet101), FaceShapeNet (ResNet101)				
FPN	8.17	13.14	9.14	**0.005**
FPN+FEN+FaceShapeNet	5.77	10.84	6.76	0.029
FPN+FEN+FaceShapeNet+ref.	5.51	10.33	6.45	0.20
FPN (ResNet101), FEN (ResNet101), FaceShapeNet (ResNet101)				
FPN	6.35	11.38	7.33	0.088
FPN+FEN+FaceShapeNet	4.79	9.67	5.75	0.112
FPN+FEN+FaceShapeNet+ref.	**3.97**	**8.51**	**4.86**	0.283

Values in bold indicates the best performance

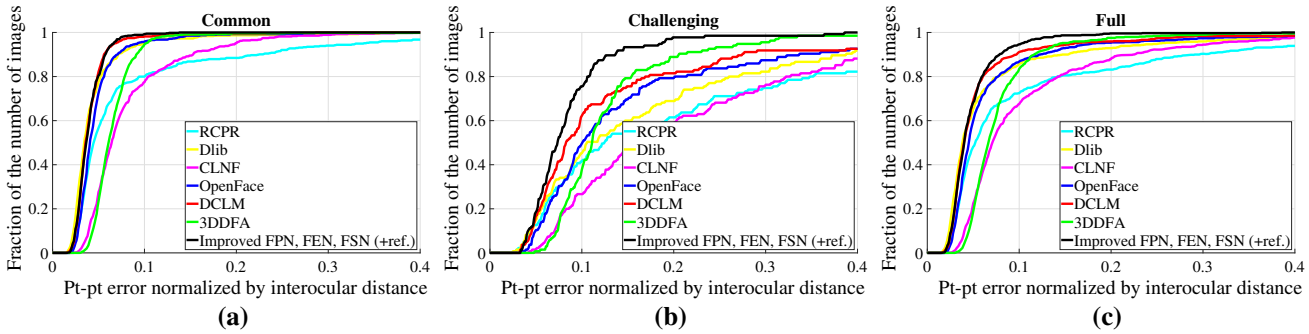


Fig. 16 Comparisons of CED curves on 300W with the face bounding boxes detected by (Yang and Nevatia 2016). Results provided for the Common (HELEN and LFPW), Challenging (iBUG), and Full (HELEN, LFPW, and iBUG) splits

We report additional improvement in detection accuracy by adopting the deeper ResNet101 version of our FPN and training it with more profile faces (Sect. 3.3). Some of this improvement may be due to the fact that our method estimates positions for fixed, physical facial positions, including positions that are occluded from view, whereas 300W measures accuracy of contour points which change depending on viewpoint (Sect. 5). Refinement moves our landmarks towards the visible contour (see Fig. 6e versus f) and reduces some of these errors.

300W results with detected bounding boxes The 300W benchmark provides ground truth face bounding boxes for all of its images. In practical scenarios, such bounding boxes would likely not be available, and a face detection method would be used to obtain these bounding boxes. We therefore tested performances on 300W images using *detected bounding boxes*. Our FAME networks are fine-tuned on the training parts of LFPW, HELEN, and the entire AFW using the same face detector. The results are provided in Table 6. Figure 16 additionally offers cumulative error distribution (CED) curves on 300W with detected bounding boxes.

The same face detection method of Yang and Nevatia (2016) was used with all landmark detectors. Because landmark detectors can be sensitive to the choice of face detection method, we attempted to optimize performances for these baseline methods by scaling the face detection bounding box, using the best scaling for each method. Note that we report baseline results in Table 6 only for methods for which we could find code available. Our results are, however, consistent with those reported by the different authors, appearing in Table 5.

There are a number of noteworthy observations from our results in Table 6. First, the accuracy of all methods tested here dropped somewhat compared to the precision reported using ground truth bounding boxes (Table 5). All variations of our approach, however, show only small drops in accuracy. We attribute this robustness to face bounding boxes to the scale changes which we synthetically introduced to the data set used to train our FPN (Sect. 3.3). By making our network

Table 7 The NME (%) of 68 point detection results on AFLW2000-3D for different ranges of yaw angles.

Method	[0, 30]	(30, 60]	(60, 90]	mean	std
RCPR (Artizzu et al. 2013)	4.26	5.96	13.18	7.80	4.74
ESR (Cao et al. 2014)	4.60	6.70	12.67	7.99	4.19
SDM (Xiong and De la Torre 2013)	3.67	4.94	9.76	6.12	3.21
3DDFA (Zhu et al. 2016b)	3.78	4.54	7.93	5.42	2.21
3DDFA+SDM (Zhu et al. 2016b)	3.43	4.24	7.17	4.94	1.97
3DSTN (AlexNet) (Bhagavatula et al. 2017)	3.71	5.33	7.19	5.41	1.74
3DSTN (VGG-16) (Bhagavatula et al. 2017)	3.15	4.33	**5.98**	**4.49**	**1.42**
FPN (AlexNet), FEN (ResNet101), FaceShapeNet (ResNet101)					
FPN	5.38	10.31	19.08	11.59	6.94
FPN+FEN+FaceShapeNet	4.20	5.12	8.14	5.82	2.06
FPN+FEN+FaceShapeNet+refinement	3.35	4.15	7.05	4.85	1.94
FPN (ResNet101), FEN (ResNet101), FaceShapeNet (ResNet101)					
FPN	3.78	4.23	7.25	5.09	1.89
FPN+FEN+FaceShapeNet	3.46	4.13	7.09	4.89	1.93
FPN+FEN+FaceShapeNet+refinement	**3.11**	**3.84**	6.60	4.52	1.84

Values in bold indicates the best performance

robust to scale changes, it better handles differences in the tightness of the detected facial bounding boxes.

Another remarkable result is that in a realistic use-case where a face detector is used to obtain face bounding boxes, dlib (King 2009) appears to perform very well, despite its age and despite being the fastest landmark detector we tested. This result is consistent with those in Sect. 6.1 where dlib also performed well in aligning faces for recognition.

Finally, our method (*FPN + FEN + FaceShapeNet + refinement*) obtains the most accurate landmark detection results. Even without the landmark-specific refinement step, the accuracy of our approach (*FPN + FEN + FaceShapeNet*) is comparable to the existing state-of-the-art, DCLM (Zadeh et al. 2016), outperforming it on both the *challenging* and *full* splits despite being at least an order of magnitude faster.

Landmark detection runtime Both Tables 5 and 6 also report the runtimes for the various methods we tested. These runtimes were all measured by us on the same machine (see Sect. 6.1); missing results in Table 5 represent methods for which we were unable to run the code ourselves.

It is worth noting that our full FAME approach is only slower than RCPR (Artizzu et al. 2013) and the very fast dlib (King 2009). These two methods, however, provide less accurate rigid alignment than even our (much faster) FPN alone (Sect. 6.1) and are less accurate than our full approach in landmark localization (Tables 5 and 6).

AFLW 2000-3D results Because we estimate 3D face shape, we can also report landmark detection accuracy with 3D landmarks on the AFLW 2000-3D benchmark of Zhu et al. (2016b). Results are provided in Table 7 for three categories of the absolute yaw angles: [0, 30], (30, 60], and (60, 90]. Following Bhagavatula et al. (2017), we use the bounding box associated with 68 landmarks. The NME is computed using the bounding box size. The cumulative error distribution (CED) curves are reported in Fig. 17. All baselines except for 3DSTN (Bhagavatula et al. 2017) were reported by their respective authors.

The very recent method of Bhagavatula et al. (2017) appears to perform better than most other baselines in most of the tests. Our approach (*FPN + FEN + FaceShapeNet + refinement*) is the most accurate in the ranges [0.30] (30, 60], coming in second in (60, 90], outperforming other methods designed for 3D landmark detection (e.g., 3DDFA (Zhu et al. 2016b)). Even without the landmark-specific refinement step, our method is the most accurate in (30, 60].

AFLW-PIFA results We report landmark detection results on AFLW, strictly following the PIFA protocol suggested by Jourabloo and Liu (2015).[2] PIFA provides 5200 images where the numbers of images with absolute yaw angle viewpoints within [0,30), [30,60), [60,90] are approximatively one third each. Finally, 3901 of these images are used for training and 1299 for testing. Note that Kumar et al. (2017) used many more images to train (23,386) and a different testing partition with 1000 images. Their results are not directly comparable to all others.

All AFLW-PIFA images are labeled with up to 21 landmarks (Jourabloo and Liu 2015) and 34 landmarks (Jourabloo and Liu 2016). Results based on our best method using FPN, FEN, and FaceShapeNet (all with a ResNet101 architecture) are provided in Table 8. For fair comparison, we report results for both 34 landmarks and 21 landmarks.

[2] The train/test partitions of PIFA are available at http://cvlab.cse.msu.edu/project-pifa.html.

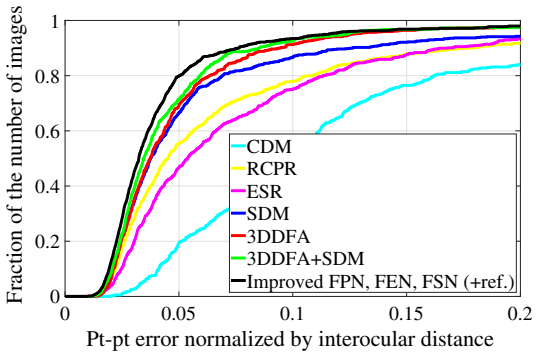


Fig. 17 Comparisons of CED curves on AFLW2000-3D. To balance the yaw distributions, we randomly sample 699 faces from AFLW 2000-3D, split evenly among the 3 yaw categories and compute the CED curve. This is done 10 times and the average of the resulting CED curves are reported

Table 8 NME (%) results on the AFLW dataset under the PIFA protocol (Jourabloo and Liu 2015, 2016)

Method	34 land.	21 land.
RCPR (Artizzu et al. 2013)	6.26	7.15
CFSS (Zhu et al. 2015a)	6.75	–
PIFA (Jourabloo and Liu 2015)	8.04	6.52
CCL (Zhu et al. 2016a)	5.81	–
PAWF (Jourabloo and Liu 2016)	4.72	–
DeFA (Liu et al. 2017)	**3.86**	–
KEPLER (Kumar et al. 2017)*	-	**2.98**
FPN	4.79	4.75
FPN+FEN+FaceShapeNet	4.20	4.09
FPN+FEN+FaceShapeNet+ref.	4.03	3.9

Values in bold indicates the best performance
* Followed a non-standard training/testing setting on the AFLW dataset

Discussion: FAME versus OpenFace The face recognition results reported in Sect. 6.1 demonstrate that our FPN better aligns faces for face recognition than OpenFace Baltrušaitis et al. (2016), the method used to produce pose labels for training our FPN. In this section, Tables 5 and 6 directly compare the landmark detection accuracy of our FAME approach with the accuracy reported by OpenFace.

Alone, the landmark accuracy of FPN is comparable with OpenFace—FPN matching the accuracy of the method used to train it—though FPN is nearly an order of magnitude faster and, unlike OpenFace, was not designed or optimized for landmark detection. This is not entirely surprising, as FPN is trained to solve a 6D regression problem (six floating point numbers), whereas landmark detection methods such as OpenFace try to solve a harder 2×49 or 2×68 dimensional regression task (49 or 68 integer 2D image coordinates). With a simpler regression problem, we can therefore train FPN to achieve better accuracy than the method used to produce its labels. Importantly, even without the refinement step which optimizes for landmark localization, our FAME approach clearly outperforms OpenFace.

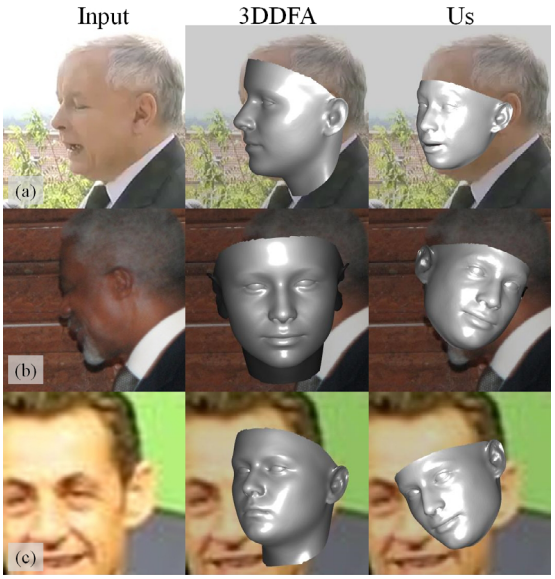


Fig. 18 Limitations of our approach. Results obtained with the state-of-the-art 3DDFA of Zhu et al. (2016b) and our full reconstruction (Us, denoting *FPN* + *FEN* + *FaceShapeNet*) for three faces in the IJB-A benchmark. See text for more details

6.4 Qualitative Results

We visualize the results of our FAME approach (*FPN* + *FEN* + *FaceShapeNet*) by rendering 3D face shape estimates, super-positioned over the original input images. We use images from the IJB-A benchmark for this purpose.

Figure 18 provides a few examples of the limitations of our approach, comparing them with the results obtained by the state-of-the-art 3DDFA of Zhu et al. (2016b). Because profile views are underrepresented in the VGG Face set used to train our FPN in Sect. 3.3, our network is conservative in the poses it estimates, preferring lower than 90° yaw angles (Fig. 18a). 3DDFA was explicitly designed to handle such profile views and so better handles such images. Both methods fail on images with other extreme viewpoints: larger than profile yaw rotations of the head (Fig. 18b) and extreme scales (Fig. 18c).

Finally, Fig. 19 provides a wide range of example 3D reconstructions, produced using IJB-A images. These results were selected to represent varying expressions, viewpoints, ethnicities, genders, facial expressions, occlusions, and different image qualities. We provide baseline results for 3DDFA (Zhu et al. 2016b) and the FaceShapeNet method of Tran et al. (2017) (adjusted for viewpoint using FPN, Sect. 3.3). Our estimated 3D shapes clearly capture subject-specific facial attributes (e.g., ethnicity and gender). Expressions are also very evident on the reconstructed shapes.

Fig. 19 Qualitative 3D reconstruction results. Rendered 3D reconstruction results for IJB-A images representing a wide range of viewing settings. For each image we provide results obtained by 3DDFA (Zhu et al. 2016b), the FaceShapeNet of Tran et al. (2017) (adjusted for viewpoint using FPN, Sect. 3.3), and our full approach (Us, denoting *FPN* + *FEN* + *FaceShapeNet*). These results should be considered by how well they capture the unique 3D shape of each individual, the viewpoint, and the facial expression. See text for more details

Our results are especially remarkable considering the extreme conditions in some of these images: Severe occlusions Fig. 19a, f, h, q, very low-quality photos in Fig. 19c, e, q, and a wide range of scales (see Fig. 19b versus l, n).

7 Conclusions

Over the past decade, facial landmark detection methods have played a tremendous part in advancing the capabilities of face processing applications. Despite these contributions, landmark detection methods and the benchmarks that measure their performances have their limits. We show that deep learning can be leveraged to perform tasks that, until recently, required the use of these facial landmark detectors. In particular, we show how face shape, viewpoint, and expression can be estimated directly from image intensities, without the use of facial landmarks. Moreover, facial landmarks can be obtained as by-products of our deep 3D face modeling process.

By proposing an alternative to facial landmark detection, we must also provide novel alternatives for evaluating the effectiveness of landmark-free methods such as our own. We therefore compare our method with facial landmark detectors by considering the effect these methods have on the bottom line performances of the methods that use them: face recognition for rigid 2D and 3D face alignment, and emotion classification for non-rigid, expression estimation. Of course, these tests are not meant to be exhaustive: This evaluation paradigm can potentially be extended to other benchmarks, representing other face processing tasks.

In addition to extending our tests to other face processing applications, another potential direction for future work is improvement of our proposed FAME framework. Specifically, notice that our FPN is trained to estimate pose for a generic face shape, whereas in practice, the 3D face shape that we project is subject- and expression-adjusted to the input face. This discrepancy can lead to misalignment errors, even if small ones. These errors may be mitigated by combining the three networks into a single, jointly learned, FAME network.

Acknowledgements This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600011. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

References

Artizzu, X. P., Perona, P., & Dollár, P. (2013). Robust face landmark estimation under occlusion. In *Proceedings of the international conference on computer vision*.

Asthana, A., Zafeiriou, S., Cheng, S., & Pantic, M. (2014). Incremental face alignment in the wild. In *Proceedings of the conference on computer vision pattern recognition*.

Baltrusaitis, T., Robinson, P., & Morency, L. P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the conference on computer vision pattern recognition workshops*.

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). Openface: An open source facial behavior analysis toolkit. In *Winter conference on appllications of computer vision*.

Bansal, A., Russell, B., & Gupta, A. (2016). Marr revisited: 2D-3D alignment via surface normal prediction. In *Proceedings of the conference on computer vision pattern recognition*.

Bas, A., Smith, W. A. P., Bolkart, T., & Wuhrer, S. (2016). Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences. In *ACCV workshops*.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(12), 2930–2940.

Bhagavatula, C., Zhu, C., Luu, K., & Savvides, M. (2017). Faster than real-time facial alignment: A 3D spatial transformer network approach in unconstrained poses. In *Proceedings of the international conference on computer vision*.

Blanz, V., & Vetter, T. (1999). Morphable model for the synthesis of 3D faces. In *Proceedings of ACM SIGGRAPH conference on computer graphics*.

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(9), 1063–1074.

Blanz, V., Romdhani, S., & Vetter, T. (2002). Face identification across different poses and illuminations with a 3d morphable model. In *International conference on automatic face and gesture recognition*.

Blanz, V., Scherbaum, K., Vetter, T., & Seidel, H. P. (2004). Exchanging faces in images. *Computer Graphics Forum*, *23*(3), 669–676.

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., & Zafeiriou, S. (2017). 3D face morphable models "in-the-wild". In *Proceedings of conference on computer vision pattern recognition*.

Bulat, A., & Tzimiropoulos, G. (2017a). Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the international conference on computer vision*.

Bulat, A., & Tzimiropoulos, G. (2017b). How far are we from solving the 2d and 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proceedings of the international conference on computer vision*.

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision*, *107*(2), 177–190.

Chang, F. J., Tran, A., Hassner, T., Masi, I., Nevatia, R., & Medioni, G. (2017) Faceposenet: Making a case for landmark-free face alignment. In *Proceedings of international conference on computer vision workshops*.

Chang, F. J., Tran, A. T., Hassner, T., Masi, I., Nevatia, R., & Medioni, G. (2018) Expnet: Landmark-free, deep, 3D facial expressions. In *International conference on automatic face and gesture recognition*.

Chu, B., Romdhani, S., & Chen, L. (2014). 3D-aided face recognition robust to expression and pose variations. In *Proceedings of conference on computer vision pattern recognition*.

Crosswhite, N., Byrne, J., Stauffer, C., Parkhi, O., Cao, Q., & Zisserman, A. (2017). Template adaptation for face verification and identification. In *International conference on automatic face and gesture recognition*.

Dantone, M., Gall, J., Fanelli, G., & Van Gool, L. (2012). Real-time facial feature detection using conditional regression forests. In *Proceedings of conference on computer vision pattern recognition*.

Dhall, A., Goecke, R., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, *19*(3), 34–41.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., & Gedeon, T. (2017). From individual to group-level emotion recognition: Emotiw 5.0. In *ACM ICMI*.

Dhall, A., Murthy, O. R., Goecke, R., Joshi, J., & Gedeon, T. (2015). Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: *ACM ICMI*.

Dong, X., Yan, Y., Ouyang, W., & Yang, Y. (2018a). Style aggregated network for facial landmark detection. In *Proceedings of conference on computer vision pattern recognition*.

Dong, X., Yu, S. I., Weng, X., Wei, S. E., Yang, Y., & Sheikh, Y. (2018b). Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of conference on computer vision pattern recognition*.

Dong, X., Zheng, L., Ma, F., Yang, Y., & Meng, D. (2018c). Few-example object detection with model communication. *IEEE Transactions on Pattern Analysis & Machine Intelligence*. https://doi.org/10.1109/TPAMI.2018.2844853.

Dou, P., Shah, S. K., & Kakadiaris, I. A. (2017). End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of conference on computer vision pattern recognition*.

Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, *9*(12), 2170–2179.

Everingham, M., Sivic, J., & Zisserman, A. (2006). "Hello! My name is... Buffy"—Automatic naming of characters in TV video. In *Proceedings of British machine vision conference*.

Fabian Benitez-Quiroz, C., Srinivasan, R., & Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of conference on computer vision pattern recognition*.

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

Hassner, T. (2013). Viewing real-world faces in 3D. In *Proceedings of the international conference on computer vision*. Available www.openu.ac.il/home/hassner/projects/poses.

Hassner, T., & Basri, R. (2006). Example based 3D reconstruction from single 2D images. In *Proceedings of conference on computer vision pattern recognition workshops*.

Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). Effective face frontalization in unconstrained images. In *Proceedings of conference on computer vision pattern recognition*.

Hassner, T., Masi, I., Kim, J., Choi, J., Harel, S., Natarajan, P., & Medioni, G. (2016). Pooling faces: Template based face recognition with pooled face images. In *Proceedings of conference on computer vision pattern recognition workshops*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of conference on computer vision pattern recognition*.

Huang, G. B., Jain, V., & Learned-Miller, E. (2007). Unsupervised joint alignment of complex images. In *Proceedings of the international conference on computer vision*.

Huber, P., Hu, G., Tena, R., Mortazavian, P., Koppen, W., Christmas, W., Rtsch, M., & Kittler, J. (2016). A multiresolution 3D morphable face model and fitting framework. In *VISAPP*.

Jackson, A. S., Bulat, A., Argyriou, V., & Tzimiropoulos, G. (2017). Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the international conference on computer vision*

Jeni, L. A., Cohn, J. F., & Kanade, T. (2015). Dense 3D face alignment from 2D videos in real-time. In *International conference on automatic face and gesture recognition*.

Jourabloo, A., & Liu, X. (2015). Pose-invariant 3d face alignment. In *Proceedings of conference on computer vision pattern recognition*.

Jourabloo, A., & Liu, X. (2016). Large-pose face alignment via cnn-based dense 3D model fitting. In *Proceedings of conference on computer vision pattern recognition*.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. In *Proceedings of conference on computer vision pattern recognition*.

Kemelmacher-Shlizerman, I., & Basri, R. (2011). 3D face reconstruction from a single image using a single reference face shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 394–405.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., & Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A. In *Proceedings of conference on computer vision pattern recognition*.

Kosti, R., Alvarez, J. M., Recasens, A., & Lapedriza, A. (2017). Emotion recognition in context. In *Proceedings of conference on computer vision pattern recognition*.

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proceedings of the international conference on computer vision workshops*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Neural information processing systems*.

Kumar, A., Alavi, A., & Chellappa, R. (2017). Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *Automatic face and gesture recognition*.

Kumar, A., & Chellappa, R. (2018). Disentangling 3D pose in a dendritic cnn for unconstrained 2d face alignment. In *Proceedings of conference on computer vision pattern recognition*.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. (2012). Interactive facial feature localization. In *European conference on computer vision*.

Levi, G., & Hassner, T. (2015). Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *ACM ICMI*.

Li, C., Zhou, K., & Lin, S. (2014). Intrinsic face image decomposition with human face priors. In *European conference on computer vision*.

Liu, Y., Jourabloo, A., Ren, W., & Liu, X. (2017). Dense face alignment. In *Proceedings of conference on computer vision pattern recognition*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the international conference on computer vision*.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of conference on computer vision pattern recognition workshops*.

Masi, I., Ferrari, C., Del Bimbo, A., & Medioni, G. (2014). Pose independent face recognition by localizing local binary patterns via deformation components. In *International conference on pattern recognition* (pp. 4477–4482). IEEE.

Masi, I., Chang, F. J., Choi, J., Harel, S., Kim, J., Kim, K., et al. (2018a). Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 379–393.

Masi, I., Hassner, T., Tran, A. T., & Medioni, G. (2017). Rapid synthesis of massive face sets for improved face recognition. In *International conference on automatic face and gesture recognition* (pp. 604–611). IEEE.

Masi, I., Rawls, S., Medioni, G., & Natarajan, P. (2016a). Pose-aware face recognition in the wild. In *Proceedings of conference on computer vision pattern recognition*.

Masi, I., Tran, A., Hassner, T., Leksut, J. T., & Medioni, G. (2016b). Do we really need to collect millions of faces for effective face recognition?. In *European conference computer vision*. Available www.openu.ac.il/home/hassner/projects/augmented_faces.

Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2018b). Deep face recognition: A survey. In *Conference on graphics, patterns and images*.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of British machine vision conference*.

Paysan, P., Knothe, R., Amberg, B., Romhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *International conference on advanced video and signal based surveillance*.

Poirson, P., Ammirato, P., Fu, C. Y., Liu, W., Kosecka, J., & Berg, A. C. (2016). Fast single shot detection and pose estimation. In *3DV*.

Ranjan, R., Castillo, C. D., & Chellappa, R. (2017). L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507.

Ren, S., Cao, X., Wei, Y., & Sun, J. (2014). Face alignment at 3000 fps via regressing local binary features. In *Proceedings of conference on computer vision pattern recognition*.

Richardson, E., Sela, M., & Kimmel, R. (2016). 3d face reconstruction by learning from synthetic data. In *3DV*.

Richardson, E., Sela, M., Or-El, R., & Kimmel, R. (2017). Learning detailed face reconstruction from a single image. In *Proceedings of conference on computer vision pattern recognition*.

Romdhani, S., & Vetter, T. (2003). Efficient, robust and accurate fitting of a 3D morphable model. In *Proceedings of the international conference on computer vision*.

Romdhani, S., & Vetter, T. (2005). Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proceedings of conference on computer vision pattern recognition*.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of conference on computer vision pattern recognition workshops*.

Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2016). 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, *47*, 3–18.

Sela, M., Richardson, E., & Kimmel, R. (2017). Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the international conference on computer vision*.

Sengupta, S., Kanazawa, A., Castillo, C. D., & Jacobs, D. (2018). SfSNet: Learning shape, reflectance and illuminance of faces in the wild. In *Proceedings of conference on computer vision pattern recognition*.

Su, H., Qi, C. R., Li, Y., & Guibas, L. J. (2015). Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proceedings of the international conference on computer vision*.

Surace, L., Patacchiola, M., Battini Sönmez, E., Spataro, W., & Cangelosi, A. (2017). Emotion recognition in the wild using deep neural networks and Bayesian classifiers. In *ACM ICMI*.

Tang, H., Hu, Y., Fu, Y., Hasegawa-Johnson, M., & Huang, T. S. (2008). Real-time conversion from a single 2d face image to a 3D text-driven emotive audio-visual avatar. In *International conference on multimedia and expo*.

Tewari, A., Zollhfer, M., Garrido, P., Florian Bernard, H. K., Prez, P., & Theobalt, C. (2018). Self-supervised multi-level face model learning for monocular reconstruction at over 250 Hz. In *Proceedings of conference on computer vision pattern recognition*.

Tran, A., Hassner, T., Masi, I., & Medioni, G. (2017). Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of conference on computer vision pattern recognition*.

Tran, A. T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., & Medioni, G. (2018) Extreme 3D face reconstruction: Looking past occlusions. In *Proceedings of conference on computer vision pattern recognition*.

Vetter, T., & Blanz, V. (1998). Estimating coloured 3D face models from single images: An example based approach. In *European conference on computer vision*.

Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A. K., Duncan, J. A., & Allen, K., et al. (2017). Iarpa janus benchmark-b face dataset. In *Proceedings of conference on computer vision pattern recognition workshops*.

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *Proceedings of conference on computer vision pattern recognition*.

Wu, Y., Hassner, T., Kim, K., Medioni, G., & Natarajan, P. (2017). Facial landmark detection with tweaked convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 3067–3074.

Xiang, Y., Mottaghi, R., & Savarese, S. (2014). Beyond pascal: A benchmark for 3D object detection in the wild. In *Winter conference on applications of computer vision*.

Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., & Savarese, S. (2016). Objectnet3D: A large scale database for 3D object recognition. In *European conference on computer vision*.

Xie, L., Wang, J., Wei, Z., Wang, M., & Tian, Q. (2016). Disturblabel: Regularizing cnn on the loss layer. In *Proceedings of conference on computer vision pattern recognition* (pp. 4753–4762).

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of conference on computer vision pattern recognition*.

Yang, Z., & Nevatia, R. (2016). A multi-scale cascade fully convolutional network face detector. In *ICPR*.

Yang, F., Wang, J., Shechtman, E., Bourdev, L., & Metaxas, D. (2011). Expression flow for 3D-aware face component transfer. *ACM Transactions on Graphics*, *30*(4), 60.

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Learning face representation from scratch. arXiv preprint arXiv:1411.7923. Available http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html.

Yu, X., Huang, J., Zhang, S., Yan, W., & Metaxas, D. N. (2013). Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of the international conference on computer vision* (pp. 1944–1951). IEEE.

Zadeh, A., Baltrušaitis, T., & Morency, L. P. (2016). Deep constrained local models for facial landmark detection. arXiv preprint arXiv:1611.08657.

Zafeiriou, S., Chrysos, G. G., Roussos, A., Ververas, E., Deng, J., & Trigeorgis, G. (2017). The 3D menpo facial landmark tracking challenge. In *Proceedings of international conference on computer vision workshops*.

Zafeiriou, S., Papaioannou, A., Kotsia, I., Nicolaou, M., & Zhao, G. (2016) Facial affect "in-the-wild". In *Proceedings of conference on computer vision pattern recognition workshops* (pp. 36–47).

Zhang, J., Shan, S., Kan, M., & Chen, X. (2014). Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *European conference on computer vision*. Springer.

Zhang, K., Tan, L., Li, Z., & Qiao, Y. (2016). Gender and smile classification using deep convolutional neural networks. In *Proceedings of conference on computer vision pattern recognition workshops* (pp. 34–38).

Zhu, S., Li, C., Change Loy, C., & Tang, X. (2015a). Face alignment by coarse-to-fine shape searching. In *Proceedings of conference on computer vision pattern recognition*.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2016a). Unconstrained face alignment via cascaded compositional learning. In *Proceedings of conference on computer vision pattern recognition*.

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. (2016b). Face alignment across large poses: A 3D solution. In *Proceedings of conference on computer vision pattern recognition*.

Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015b). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of conference on computer vision pattern recognition* (pp. 787–796).

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Proceedings of conference on computer vision pattern recognition*.