Learning Dense Facial Correspondences in Unconstrained Images

Ronald Yu [*1,3], Shunsuke Saito [†1,3], Haoxiang Li [‡2], Duygu Ceylan [§2], and Hao Li[¶1,3,4]

[1]University of Southern California
[2]Adobe Research
[3]Pinscreen
[4]USC Institute for Creative Technologies

Abstract

We present a minimalistic but effective neural network that computes dense facial correspondences in highly unconstrained RGB images. Our network learns a per-pixel flow and a matchability mask between 2D input photographs of a person and the projection of a textured 3D face model. To train such a network, we generate a massive dataset of synthetic faces with dense labels using renderings of a morphable face model with variations in pose, expressions, lighting, and occlusions. We found that a training refinement using real photographs is required to drastically improve the ability to handle real images. When combined with a facial detection and 3D face fitting step, we show that our approach outperforms the state-of-the-art face alignment methods in terms of accuracy and speed. By directly estimating dense correspondences, we do not rely on the full visibility of sparse facial landmarks and are not limited to the model space of regression-based approaches. We also assess our method on video frames and demonstrate successful per-frame processing under extreme pose variations, occlusions, and lighting conditions. Compared to existing 3D facial tracking techniques, our fitting does not rely on previous frames or frontal facial initialization and is robust to imperfect face detections.

1. Introduction

By introducing 3D facial alignment techniques that can process images in the wild, it is possible to improve the performance of facial recognition methods [8, 44, 16, 24]; compelling 3D face models of a person can be generated for gaming and virtual reality applications [7, 26, 37, 14, 50, 43]; and an accurate tracking model can be initialized for real-time facial performance capture and animation [52, 12, 55, 49]. Most of the techniques rely on a robust detection of sparse facial landmarks (eyes, nose, lips, etc.) and tend to perform best when most features are visible, front-facing, and free from occlusions or challenging lighting conditions. In many applications, such as video facial analytics or driver attention monitoring, these assumptions do not hold since the subjects are recorded in a fully unconstrained environment.

With the recent advancement of deep learning techniques, highly robust regression methods have emerged that can successfully fit a 3D face model for extremely difficult cases, such as side views of a face or occlusions by hair. State-of-the-art methods are based on regression [68, 31] and directly regress the shape parameters of a 3D morphable model (3DMM) [7] and expression coefficients [13] from an image using cascaded network structures. While achieving impressive accuracies on several challenging benchmark datasets, they still tend to perform poorly in extreme real-case scenarios, as demonstrated in this paper. Not only is such approach limited to variations defined by the face model space, but its performance relies on a perfectly tight facial bounding box detection. Despite significant progress, even the cutting edge face detectors [25] cannot guarantee a clean face localization and cropping for extreme images.

Instead of a regression method, we propose an alternative deep learning approach that estimates dense pixel-wise correspondences between the input image and a 3DMM model along with a *matchability mask*, which defines which pixels belong to the face and have valid correspondences. We perform dense correspondence estimation by predicting a per-pixel 2D flow vector between the input image and a synthetic rendering of a 3DMM. Once the correspondences are established, we fit a 3DMM to the input using available correspondences. Compared to sparse landmark detection techniques, dense correspondences provide more robust constraints, since any part of the face can be used for

[*]ronaldyu@usc.edu
[†]shunsuke.saito16@gmail.com
[‡]haoxli@adobe.com
[§]ceylan@adobe.com
[¶]hao@hao-li.com

matching, and our predicted matchability mask helps to distinguish non-visible parts of the face. Furthermore, our 2D flow computation is less sensitive to clean bounding box estimations during an initial face detection as oppose to existing regression approaches.

Inspired by the recent work of [66], we train a simple encoder-decoder network using synthetically generated 3DMMs with variations in pose, shape, appearance, and lighting, and simulate occlusions using random box renderings. Since every face of the 3DMM have consistent mesh topologies, the dense labels are automatically present. We further refine the training using real photographs with corresponding face models obtained from the regression technique of [68], and predict the flow between the input image and a statistical mean face. We show that by combining our dense correspondence computation with a subsequent face fitting step, we can perform comparably with the current state-of-the-art face alignment techniques on difficult images in terms of accuracy and are significantly faster on public datasets. Furthermore, we demonstrate highly effective pose estimations and 3D face fittings on extremely challenging images and videos.

2. Related Work

2D Face Alignment. Facial alignment for images and videos has attracted a lot of attention from the research community due to its wide range of applications. 2D facial alignment approaches aim to localize a set of fiducial points in the face. Classical approaches include the Active Appearance Models (AAM) [17, 42, 51, 56] and Constrained Local Models (CLM) [19, 52, 2]. Another common approach is to learn regression functions that map hand-crafted image features to 2D landmark positions directly [57, 59, 15, 3, 46, 32, 36, 60, 67]. With the recent success of deep learning methods, several approaches have replaced the use of hand-crafted features with a convolutional neural network [53, 64, 63, 45]. While such purely 2D methods have shown impressive results especially for frontal and non-occluded faces, modeling of occlusions has been mostly avoided. More recent approaches [11, 29, 21, 61, 49] have attacked this problem by introducing occlusion variation in the training data. Handling large pose variations under difficult illumination conditions, however, still remains challenging for 2D methods. To handle large pose variations, several approaches have proposed to use multiple shape models for different views [18, 70, 62]. However, due to the requirement to test all these possible views, such methods are computationally very expensive.

3D Face Alignment. In the context of 3D facial alignment, earlier works have focused on optimization based methods that minimize the difference between the input image and the model appearance [7, 47]. As in the case of 2D facial alignment, an alternative approach is to regress the parameters of a 3D face model based on image features around landmark points [12, 28, 30]. More recent methods have focused on performing this regression with neural networks, specifically with cascaded [68, 31, 40] or very deep network structures [1, 35]. While such methods achieve impressive results on challenging datasets, the main drawback is being limited to the shape space represented by the utilized 3D morphable model. We present an alternative approach of predicting dense correspondences between the input image and the face model, which can potentially be propagated to any 3D morphable model with little effort. We provide extensive comparisons between our alternative approach and the recent 3D facial alignment methods and show that our method outperforms both in terms of accuracy and speed (see Section 4).

Dense Correspondence Estimation. The problem of dense correspondence estimation has been mainly investigated in the form of *optical flow estimation* for tracking purposes [23, 6, 4, 10]. To compute dense correspondences between different scenes and different instances of an object category, energy minimization approaches that match hand-crafted features with additional smoothness priors have been proposed [5, 39, 33, 9]. Such approaches have been further improved by either jointly solving for co-segmentation [54] or analyzing collections of images [65]. Last but not least, recent methods have explored the power of neural networks for predicting dense optical flow [58, 20, 27] and correspondences [66]. These methods have shown impressive results that motivate us to explore the power of predicting dense flow to tackle the problem of 3D facial alignment. [22] has also explored using dense correspondences for the purpose of 3D facial landmark alignment. We show that dense correspondences provide robust constraints for 3D face fitting under large pose and illumination conditions since they enable any part of the face to be utilized for matching.

3. Proposed Method

3.1. Overview

Our approach takes a source image and outputs per-pixel correspondences between the source image and a 3D morphable model (3DMM). Since correspondences are well-defined only on regions of the face that are visible in the source image, we also output a *matchability mask* that predicts the probability of each correspondence being valid or not. We perform dense correspondence estimation by predicting a per-pixel 2D flow between the source image and a synthetic rendering of the 3DMM depicting a frontal mean face. Both 2D flow and the matchability are predicted by a convolutional neural network (Section 3.3). Valid 2D correspondences are easily translated to 2D-3D correspondences since each pixel in the synthetic rendering is directly associated with the 3DMM. Such 2D-3D correspondences are then used to guide the alignment of the 3DMM to the source image (Section 3.4).This pipeline is illustrated in Figure 1.

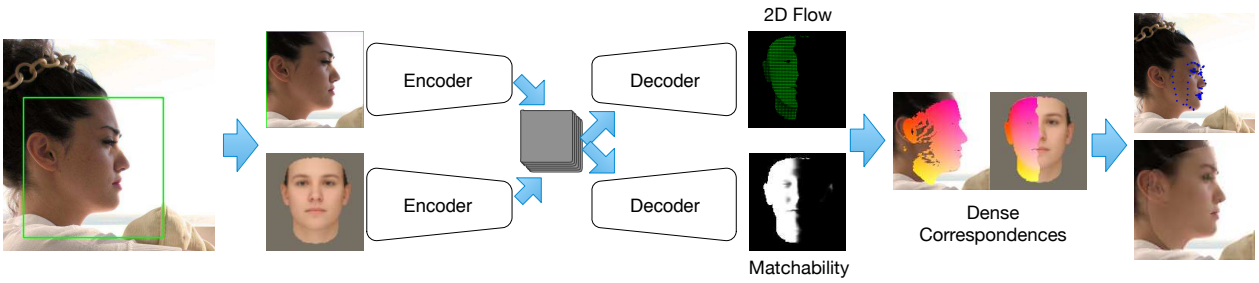

Figure 1. Given an input image with a detected face, we propose an encoder-decoder architecture that predicts Dense Facial Correspondences between the input image and a 3D morphable face model. These correspondences are estimated as 2D flow between the input image and a synthetic rendering of a frontal, mean face. We also predict a matchability mask which indicates which correspondences are valid or not. Using such correspondences, we can perform 3D face alignment even in very challenging cases of large pose and illumination variation.

We next discuss each stage in more detail.

3.2. 3D Morphable Model

We use the 3D morphable model (3DMM) proposed by Blanz and Vetter [7]. Each 3D face, S, is represented as:

$$S = \bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp}, \quad (1)$$

where $\bar{S}$ is the mean 3D face, A_{id} and A_{exp} are the basis for the identity and the expression respectively. α_{id} and α_{exp} denote the parameters for the identity and the expression basis. Moreover, in order to project a 3D face S to a 2D image we use perspective projection:

$$S_{2D} = \Pi_f(\mathbf{R}(\bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_e xp) + \mathbf{t}), \quad (2)$$

where Π_f is the projection operator that depends on the focal length, f, and the principal point defined to be the center of the image. $\mathbf{R}$ and $\mathbf{t}$ denote the rotation and the translation components of the pose. Thus, aligning the 3DMM with an image is equivalent to finding the set of parameters $(f, \mathbf{R}, \mathbf{t}, \alpha_{id}, \alpha_{exp})$ that minimizes an alignment error as described in Section 3.4.

3.3. Dense Correspondence Prediction

Given a source image, I_s, and a rendering of the 3DMM showing a frontal, mean 3D face, which we call the target image I_t, we propose a network architecture to predict a per-pixel 2D flow from I_s to I_t along with a matchability mask. For each pixel location $\mathbf{p}_s = (x, y)$ in I_s, the 2D flow $\mathbf{F}_{s,t}(x, y) = (\Delta x, \Delta y)$ maps $\mathbf{p}_s$ to the location $\mathbf{q}_t = (x + \Delta x, y + \Delta y)$ in I_t, predicting that $\mathbf{p}_s$ and $\mathbf{q}_t$ are semantic correspondences. Since the 2D flow is well defined only for parts of the face visible in I_s and I_t, we also predict a matchability score $m_{s,t}(\mathbf{p}_s) \in [0, 1]$, where $m_{s,t}(\mathbf{p}_s) = 1$ if p_s has a valid correspondence in I_t. We note that we first perform face detection in the source image and predict the flow on the cropped image based on the detection result. A visualization of our network output can be seen in Figure 2

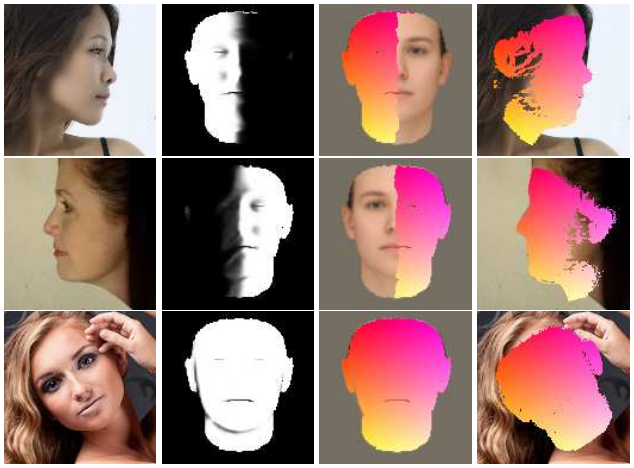

Figure 2. Sample visualizations of dense correspondences between our input and our template image for two inputs. The first image is the input. The second image is the matchability map that indicates which pixels on the template are matchable. Note that we do not explicitly segment out external occlusions and occluded areas are also considered matchable as long as they reside in the face shape. The two color maps on the right show the dense correspondence between the two images. The holes found at the edges of the dense correspondences in the third column indicate that the pixel cannot be matched and does not have a flow to our frontal facing template.

Our network architecture for predicting the 2D flow and the matchability generally follows the architecture recently proposed by Zhou et al. [66]. Specifically, we use two encoder branches that take the source and target images as input respectively. The output of these encoders are concatenated and provided as input to two decoder branches. The *flow decoder* outputs a 2-channel feature map with the same size as the source image, where the channels specify the 2D flow $(\Delta x, \Delta y)$ for each pixel. The *matchability decoder*, on the other hand, outputs a single channel feature map with the same size as the source image representing the probability of each pixel being matchable.

The encoders consist of 8 convolutional layers, each

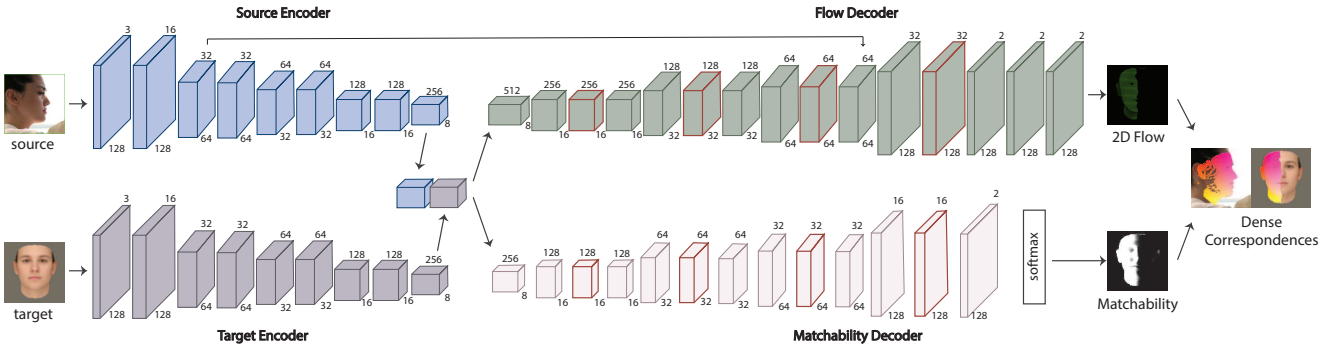

Figure 3. Our network architecture.

followed by a ReLU layer. Every second convolution layer has a stride of 2 in order to decrease the spatial dimension by half. Each decoder begins with four triplets of convolution layers. In each triplet, the third layer is a deconvolutional layer with a stride of 2. The first and third layer of each triplet is followed by an ReLU layer. For the flow decoder, the triplets are followed by three additional convolutional layers with an ReLU in between. For the matchability decoder, the triplets are followed by a single convolutional layer and a softmax function that classifies each pixel as matchable or not matchable. Details of the network architecture can be seen in Figure 3.

Loss function. Given a source and a target image I_s and I_t, for each pixel (x, y) in I_s we denote by $\mathbf{F}(x, y)$ and $m(x, y)$ the 2D flow and matchability predictions and their ground truths by $\widetilde{\mathbf{F}}(x, y)$ and $\widetilde{m}(x, y)$ respectively. We train the network to minimize the following loss $\mathbf{L}(I_s, I_t)$:

$$\begin{aligned}\mathbf{L}(I_s, I_t) &= \sum_{x,y} \widetilde{m}(x, y)||\mathbf{F}(x, y) - \widetilde{\mathbf{F}}(x, y)||^2 \\ &+ \lambda \sum_{x,y} \mathbf{L}_C(m(x, y), \widetilde{m}(x, y)),\end{aligned}$$

where $\mathbf{L}_C$ denotes the cross-entropy loss and λ is a hyper-parameter.

Training procedure In order to train the proposed network architecture, we need access to images where ground truth correspondences with a 3DMM are available. We use the recently released large-pose 300W (300W-LP) dataset [68] which provides the parameters of a 3DMM fitting each image in the dataset. Given the pose and the 3DMM parameters, we project the 3D face to the input image using Equation 2. For each pixel p_s in the input image, we identify the $uv-$coordinate of the 3DMM surface point that projects to it. Then, we find the pixel q_t in the rendering of the frontal, mean face that has the most similar $uv-$ coordinate. If the distance between the $uv-$coordinates is less than a threshold (0.015 in our experiments), we define a ground truth correspondence between the pixel p_s in the input image and q_t in the rendering of the frontal, mean template.

We observe that training the network from scratch directly by feeding "in-the-wild" real face images does not converge. We assume this is due to the large appearance variations as well as the noisy ground-truth annotations. To overcome this challenge, we propose to use a large scale synthetic data in a pre-training process where we learn dense correspondences between random pairs of synthetic faces. Specifically, using the 3DMM we generate random pairs of synthetic renderings showing faces with varying identity, expression, pose, lighting, and occlusion (see Figure 4). Since both images in the pair are generated from the 3DMM, we have direct access to perfect ground truth correspondences and matchability masks. Note that although our framework is robust to external occlusions and able to detect self-occlusions in our matchability mask, we do not explicitly segment out external occlusions such as in [49] due to limitations in our training data. During this pre-training stage, since we provide both of the encoders with synthetic renderings we share their weights. After convergence, the network accurately estimates dense correspondences between two synthetic faces, even with extreme pose and lighting.

We next fine-tune our network on the 300W-LP dataset [1]. We fix the input to the target encoder branch to be the frontal, mean face rendering while the input to the source branch are the real face images. Thus, in this stage the two encoder branches no longer share weights.

Although the input to one of the encoder branches is fixed, we note that this input branch is essential to our framework. If we only use a one-encoder-branch network architecture during the pre-training stage, training with synthetic data may adversely lead to overfitting to the appearance of our synthetic images. To address this, we have two input branches take an image pair to predict the flow to guide the network to learn the correspondences instead of memorizing the appearance. We then fine-tune the network with real data while keeping one input fixed. In our experiments, we observed that the training does not converge

[1] 300W-LP also contains synthetic faces, but with more realistic texture and background compared with our synthetic faces.

when discarding the constant branch since the task would not align well to the pre-training setting. For this reason, we keep a constant encoder branch in our final model.

3.4. 3DMM Alignment

Once our network is trained, during test time, given a single input image we first predict dense correspondences and the matchability mask with respect to the rendering of the frontal, mean face template. We filter our correspondences with low matchability scores and translate the remaining correspondences to 2D-3D correspondences between the input image and the 3DMM. We use such 2D-3D correspondences to fit the 3DMM to the input image to further refine our results.

Given a set of (p^i, q^i) correspondences where p^i denotes a pixel in the input image and q^i denotes its corresponding vertex in the 3DMM, we minimize the following energy function defined over the parameters $\mathcal{X} = (f, \mathbf{R}, \mathbf{t}, \alpha_{id}, \alpha_{exp})$:

$$\begin{aligned} E(\mathcal{X}) &= E_{data}(\mathcal{X}) + E_{reg}(\mathcal{X}), \qquad (3) \\ E_{data}(\mathcal{X}) &= \sum_i w_i \|p^i - \Pi_f(\mathbf{R}S_{q^i} + \mathbf{t})\|^2, \\ E_{reg}(\mathcal{X}) &= w_{id} \sum_{id,i} (\frac{\alpha_{id,i}}{\sigma_{id,i}})^2 + w_{exp} \sum_{exp,i} (\frac{\alpha_{exp,i}}{\sigma_{exp,i}})^2, \end{aligned}$$

where E_{data} measures the error between the projected 3D face points and their corresponding 2D points and E_{reg} is a statistical prior over the identity and the expression blendshapes [50] and we set $w_{id} = 2.5 \times 10^{-5}, w_{exp} = 1000$. $S_{q^i} = (\bar{S} + A_{id}\alpha_{id} + A_{exp}\alpha_{exp})_{q^i}$ denotes the position of the vertex q_i in the 3D face defined by the parameters $(\alpha_{id}, \alpha_{exp})$. w_i defines the weight of each correspondence and all weights are initialized to be equal. We iteratively solve for the parameters listed in Eq. 3 to minimize the L2 distance between the mesh's projected vertices and their estimated pixel location based on the dense correspondence at each iteration. This standard formulation is also used in many previous papers such as [55], [49], and [12]. Once the parameters of the 3DMM that aligns best with the input image are computed, we recover any missing correspondence and refine our predictions.

4. Experiments

4.1. Training Data

Our network operates on images of size 128×128. We train our network first on a large set of synthetic renderings of a 3D morphable model. To generate these renderings, we randomly sample different facial textures from the Chicago face database [41], we sample the identity and expression parameters as well as the rotation and translation from a Gaussian distribution. We also randomly sample spherical harmonics values from a database of lighting environments and apply it to the face to generate a total of 200k renderings on gray background. We also composite an additional 200k renderings with random background images downloaded from the COCO dataset [38](see Figure 4). We use a total of 100k random pairs of source and target images with gray background and 100k random pairs of source and target images with real background for training. We use a batch size of 12 and learning rate of $1e{-}4$ for roughly 2 epochs. One epoch takes roughly 6 hours.

Once the network converges, we fine-tune it with the images from the 300W-LP dataset [68]. We also perturb images from the 300W-LP dataset with 2D image-plane scale, translation, and rotation, and we also synthesize occlusions by drawing rectangles similar to the method of [49]. An example of the rectangles we synthesize to simulate occlusion is seen in Figure 4 . We first train with a learning rate of $1e{-}4$ for about two epochs and then drop the learning rate by a factor of 10 and train for another epoch.

Figure 4. We train our network first on a large set of synthetic data with variations in shape, expression, facial texture, and illumination. We further composite some of these renderings with real background images. We later synthesize occlusions onto real background images as seen in the image on the far right.

4.2. Qualitative Evaluations

We evaluate the performance of our method both for 2D and 3D facial alignment on the recently released AFLW2000 [68] dataset of challenging and large pose images and show qualitative results in Figure 5. We provide comparisons with the recent methods that tackle the problem of face alignment under large pose variations [68, 31] as well as state-of-the-art face trackers including Kazemi et al. [32] and the TVS implementation of Saragih et al. [51]. Both of them have been widely deployed in the industry. Furthermore, we demonstrate the performance of combining our method with 2D face alignment method by initializing the face tracker with our predictions of 2D facial landmarks. We observe that our method is more robust to heavy occlusions, large variations in illumination, translation, and image-axis rotation. Our method can also serve as a better starting point for 2D face alignment method such as Saragih et al. [51] to significantly improve its performance.

We also report evaluations on extremely challenging images and video sequences captured in the wild in the supplemental materials.

4.3. Quantitative Evaluation

In addition to qualitative results, we perform quantitative evaluations by measuring the accuracy of the 2D facial landmarks. Once a 3D face model is aligned to an input image

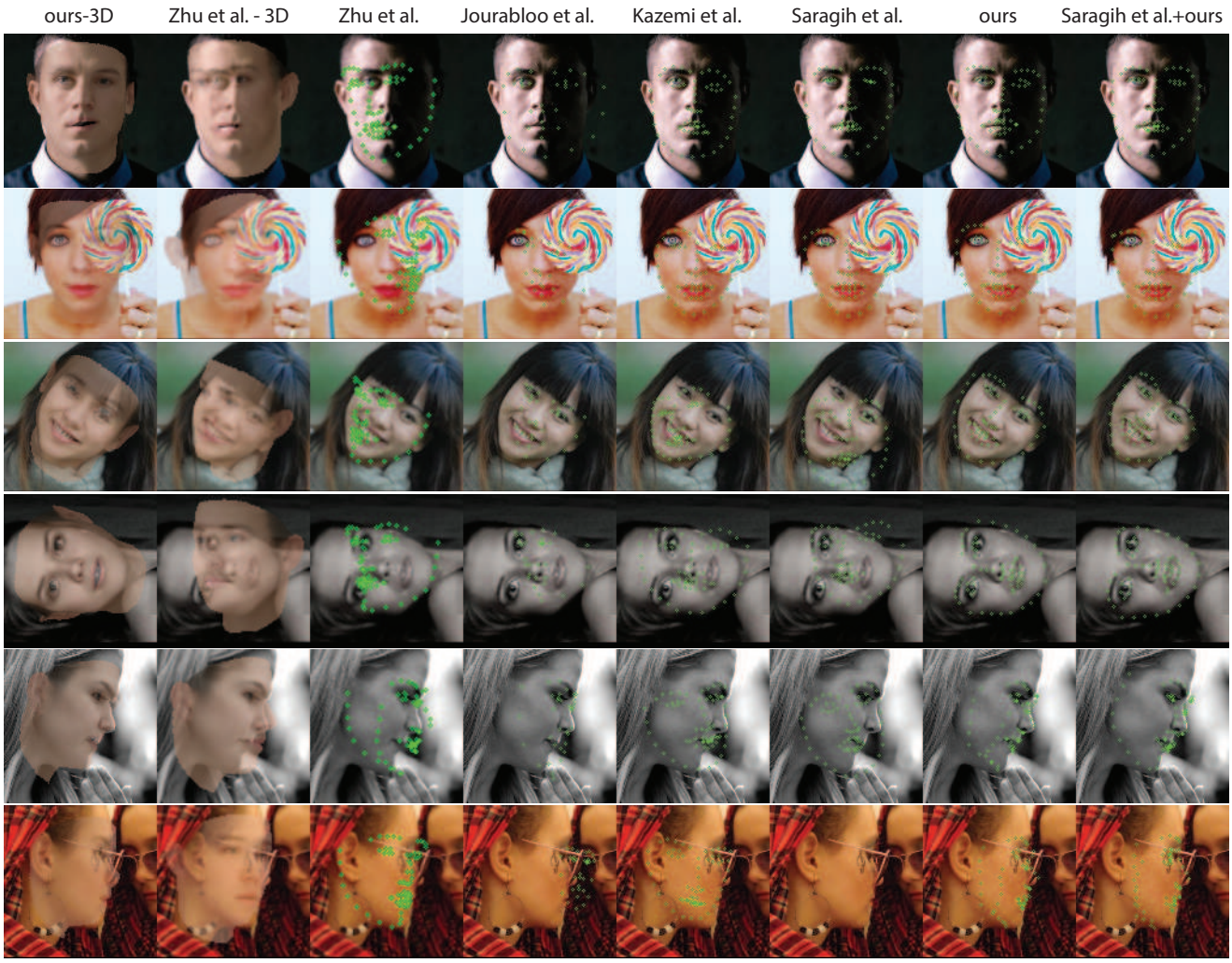

Figure 5. We provide visual 2D and 3D facial alignment results on the AFLW2000 dataset [68] using our method, the method of Zhu et al. [68], Jourabloo et al. [31], Kazemi et al. [32], and Saragih et al. [51]. We also show the results obtained by [51] when initialized with our predictions.

using the estimated dense correspondences, we can identify the 2D facial landmarks from the annotated vertices on the 3D model. We then measure the normalized mean error (NMS) [68] between the ground truth and predicted facial landmarks.

We evaluate our method on several challenging datasets such as the 68 landmarks on AFLW2000 [68], 21 landmarks on AFLW [34], and 21 landmarks on AFLW-PIFA [30] [2]. Since we did not have much training data with real images, we included images from the 300W Challenge dataset [48] and their synthesized side views in our training set, so we did not evaluate our method on this dataset. We compare our performance to state-of-the-art face alignment methods including Zhu et al. [68] and Jourabloo et al. [31].

In Table 3, we report our results on the visible landmarks on the complete AFLW [34] along with the accuracy achieved by Zhu et al. [68], which has been shown to outperform previous existing methods on this dataset.

Method	0 to 30	30 to 60	60 to 90
RCPR [11]	4.26	5.96	13.18
ESR [15]	5.60	6.70	12.67
SDM [60]	**3.67**	**4.94**	9.76
Zhu et al. [68]	3.78	**4.54**	**7.93**
Our Method	**3.62**	6.06	**9.56**

Table 1. Performance evaluation on AFLW2000 (68 landmarks): we report the NMS for faces in small ($[0, 30]$), medium($[30, 60]$), and large($[60, 90]$) pose with respect to the yaw angles. The top two results in each category are highlighted in bold.

In Table 1 we show our performance with NMS overall of the ground truth 68 landmarks in AFLW2000. One issue of this setting is that the exact locations of invisible and contour landmarks is unclear [69] and subjective. More-

[2] On AFLW-PIFA, the ground-truth annotations have 34 landmarks. But we are only clear about their definitions on a 3D face for 21 out of them.

over, their 2D projections may vary with different fitting and projection methods. Zhu et al. [69] propose "landmark marching" to address this problem. But still the definitions of contour landmarks are arguable. In Figure 6, we observe that a decent-quality fitting can have a high NMS due to the subjective nature of the contour and invisible landmarks. A strong motivation for detecting the contour and invisible landmarks is to help 3D face fitting. However, in our method, we are able to get accurate 3D face fitting from dense correspondences without relying on invisible landmarks and the need of defining contour landmarks.

To better understand the performance of our method, we exclude the contour and invisible landmarks and evaluate the NMS over only the inner and visible landmarks. For each face image, the ground-truth visibility of a landmark can be obtained from its ground-truth 3D face provided in the AFLW2000 dataset. We report our results in Table 2. We observe that our method consistently achieves comparable performance with the state-of-the-art for faces in small, medium, and large pose.

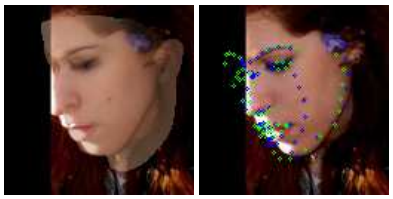

Figure 6. The problem of evaluation against invisible and contour landmarks: we show a typical large-pose face image in AFLW2000 and its landmarks. On the left we show our 3D face fitting result. On the right we show the projected landmarks of our fitting are in green and ground truth landmarks are in blue. While the fitting is decent, the NMS over 68 landmarks is as high as 9.53.

Pose (Yaw Angle)	Zhu et al [68]	Our Method
Small [0°-30°]	4.30	3.14
Medium [30°-60°]	4.41	3.84
Large [>60°]	6.68	5.53
All Images	4.60	3.58

Table 2. Performance evaluation on AFLW2000 for visible inner landmarks: we report the NMS for faces in small ($[0, 30]$), medium($[30, 60]$), large($[60, 90]$) pose with respect to the yaw angles, and across all the images.

We also compare our method on the AFLW-PIFA [30] dataset with another large-pose face alignment method from Jourabloo et al. [31]. In Table 4, we see that our method again achieves comparable performance.

Pose (Yaw Angle)	Zhu et al [68]	Our Method
Small [0°-30°]	5.00	5.94
Medium [30°-60°]	5.06	6.48
Large [>60°]	6.74	7.96

Table 3. Performance evaluation on AFLW[68]. We report NMS across all visible landmarks

Jourabloo et al. [31]	Our Method
4.72	5.42

Table 4. Performance evaluation on AFLW-PIFA [31]. We report NMS error across the original 21 AFLW landmarks.

Jourabloo et al. [31]	Zhu et al. [68]	Our Method
1666	75.7	9.35

Table 5. Comparisons of Runtimes in Milliseconds

Additional quantitative evaluations of our 3D model fitting and dense correspondences can be found in the supplemental materials.

4.4. Runtime

In addition to comparably robust and accurate face alignment, our method is one to several orders of magnitude more efficient compared with other state-of-the-art methods on large pose face alignment. Our method takes only a single iteration at test time, providing us with a large advantage in terms of efficiency. In Table 5, we summarize the runtime speed of the competing methods for large pose face alignment. Without an iterative process, it takes only 9ms on an NVIDIA Titan X GPU for our network to estimate the dense correspondences, which is one to several orders of faster than others. A 3D face-fitting post-process would take up to an additional 10 ms on the CPU, meaning that in total our pipeline can obtain a 3D face-fitting from an input image within 19 ms.

4.5. Limitations

We see in Figures 7 to 9 the limits of what our network can achieve. Namely, with respect to side faces and occlusion, we are robust enough to obtain good results when one key feature (i.e. face, mouth, nose) disappears, but performance drop as more features disappear due to extreme pose or occlusion.

5. Discussion

We have presented an alternative deep learning solution for 3D facial fitting to some top performing regression based techniques [68, 31]. Our experiments show that it is possible to reliably estimate dense 2D facial correspondences from RGB images by training a convolutional neural network with encoder-decoder architecture using a combination of real photographs and synthetic renderings with 3DMM variations, perturbations, and simulated occlusions and lighting.

With the same amount of real-world training data we are 28.5% more accurate on inner visible landmarks than Zhu et al. [68] for the AFLW2000 dataset and 14.8% less accurate to Jourabloo et al. [31] for the PIFA dataset. Generally our results can be considered comparable as other methods outperform us in certain cases (e.g their fitting

Figure 7. We observe a gradual decrease in performance for pose estimation as the pose becomes more extreme and key features (i.e. nose, eye, mouth) disappear. Our algorithm completely misses on the fourth frame when the nose is completely occluded by the rest of the face. The estimated yaw angle of the previous frame is 70°.

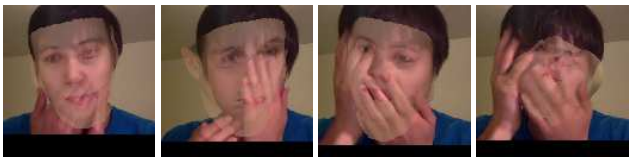

Figure 8. Like with pose, we can handle the disappearance of one key feature, and performance gradually deteriorates until our algorithm completely misses when an eye, nose, and mouth are all occluded in the fourth frame.

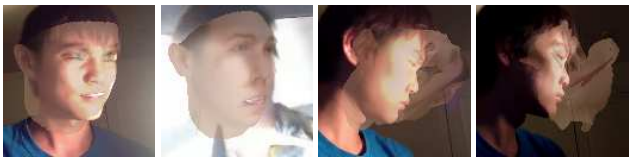

Figure 9. We see that for a frontal face, we handle intense lighting conditions both from artificial and natural light quite well. However, when extreme lighting is combined with large head pose, we see our performance suffers significantly.

among side views and wide-open mouths are slightly more accurate) while we outperform related works in other cases (e.g. we are more robust to external occlusion, illumination, and image-axis rotation). However, our approach is significantly faster (refer to Table 5) and shows increased robustness on our real test cases, such as for facial tracking in the wild, where the facial detection bounding box is not reliable and does not always provide a tight crop (see supplemental materials).

When assessing the sparse landmark positions after a 3DMM fitting step, our approach is less accurate than some cutting-edge landmark detectors such as TVS (commercial variant of [51]) or [32], but we can handle extreme conditions such as large poses, challenging lighting, and occlusions. In addition to the robustness, our method is one to several orders faster than other state-of-the-art large pose face alignment methods, and is the only one that can be real-time. Our experiments suggest that a reasonable design choice is to use our efficient and robust dense 2D flow prediction as initialization for a refined and more accurate sparse landmark detection step.

Though more efficient, our current training is limited to facial shape and appearance variations from 3DMM and photos provided by Zhu et al. [68], which indicates a similar performance to existing regression approaches. Nevertheless, since we predict 2D flows directly, we are not limited to the model space of 3DMM, and could potentially increase the dimensionality of variations, and include more expressions, facial hair, and potentially non-realistic faces such as drawings and cartoon characters. Hence, the full capabilities of our dense correspondence approach is not fully leveraged, but new training data sources need to be investigated.

Future Work. While we improve the state-of-the art in terms of efficiency and robustness, the presented framework is far from perfect. For example, although our matchability mask is able to accurately detect self-occlusion, we do not explicitly segment out external occlusions from the face region of the image due to limitations in our training data. Our accuracy is also limited by the low resolution (128x128) of the DNN input, and we would like to improve the resolution and accuracy to eliminate any need for a refinement step. Additionally, although we improve state-of-the-art in terms of robustness, there are still cases such as the ones listed in the Section 4.5 where our method fails, and we would like to extend out method to address these limitations.

We could improve our framework by introducing more training data with accurate face segmentation ground-truth, and better ground-truth fitting of the whole head, especially the back of the head and ears, would allow us to accurately track faces with even more extreme poses where close to all the sparse landmarks that are traditionally tracked in other methods are hidden. We will plan to explore new directions to generate more training data with dense facial labels using both computer graphics and machine learning techniques. Recent advancements in generative adversarial networks are promising areas for exploration. We could also extend the framework to directly infer 3D positions, eliminating the need to do post-hoc 3D fitting. If we can accurately infer dense correspondences for shapes beyond the space spanned by 3DMM, we could also model faces with more details and capture facial hair and impact general 3D reconstruction techniques such as structure from motion and multi-view stereo.

Acknowledgements We would like to thank Iman Sadeghi, Melanie Hamasaki, and Justin Kriebal for acting as capture models. This research is supported in part by Adobe, Oculus & Facebook, Huawei, Sony, the Google Faculty Research Award, the Okawa Foundation Research Grant, and the U.S. Army Research Laboratory (ARL) under contract W911NF-14-D-0005. 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 ARL 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

[1] A. T. an Trãn, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. *arXiv*, 2016. 2

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE CVPR*, pages 3444–3451, 2013. 2

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *IEEE CVPR*, pages 1859–1866, 2014. 2

[4] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92(1):1–31, 2011. 2

[5] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In *IEEE ECCV*, ECCV'10, pages 29–43, Berlin, Heidelberg, 2010. Springer-Verlag. 2

[6] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996. 2

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *ACM SIGGRAPH*, pages 187–194, 1999. 1, 2, 3

[8] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE PAMI*, 25(9):1063–1074, Sept. 2003. 1

[9] H. Bristow, J. Valmadre, and S. Lucey. Dense semantic correspondence where every pixel is a classifier. In *IEEE ICCV*, pages 4024–4031, 2015. 2

[10] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE PAMI*, 33(3):500–513, 2011. 2

[11] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *IEEE ICCV*, pages 1513–1520, 2013. 2, 6

[12] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM TOG*, 33(4):43, 2014. 1, 2, 5

[13] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: a 3d facial expression database for visual computing. *IEEE TVCG*, 20(3):413–425, 2014. 1

[14] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. Real-time facial animation with image-based dynamic avatars. *ACM TOG*, 35(4):126, 2016. 1

[15] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *IEEE IJCV*, 107(2):177–190, 2014. 2, 6

[16] B. Chu, S. Romdhani, and L. Chen. 3d-aided face recognition robust to expression and pose variations. In *IEEE CVPR*, pages 1907–1914, 2014. 1

[17] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE PAMI*, 23(6):681–685, 2001. 2

[18] T. F. Cootes, G. V. Wheeler, K. N. Walker, and C. J. Taylor. View-based active appearance models. *Image and vision computing*, 20(9):657–664, 2002. 2

[19] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recogn.*, 41(10):3054–3067, 2008. 2

[20] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE ICCV*, pages 2758–2766, 2015. 2

[21] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *IEEE CVPR*, pages 1899–1906, 2014. 2

[22] R. A. Güler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, and I. Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. *arXiv preprint arXiv:1612.01202*, 2016. 2

[23] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[24] G. Hu, F. Yan, C. Chan, W. Deng, W. J. Christmas, J. Kittler, and N. M. Robertson. Face recognition using a unified 3d morphable model. In *IEEE ECCV*, pages 73–89, 2016. 1

[25] P. Hu and D. Ramanan. Finding tiny faces. *CoRR*, abs/1612.04402, 2016. 1

[26] A. E. Ichim, S. Bouaziz, and M. Pauly. Dynamic 3d avatar creation from hand-held video input. In *ACM SIGGRAPH*, pages 45:1–45:14, 2015. 1

[27] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. *arXiv*, 2016. 2

[28] L. A. Jeni, J. F. Cohn, and T. Kanade. Dense 3d face alignment from 2d videos in real-time, 2015. 2

[29] X. Jia, H. Yang, A. Lin, K.-P. Chan, and I. Patras. Structured semi-supervised forest for facial landmarks localization with face mask reasoning. In *BMVC*, 2014. 2

[30] A. Jourabloo and X. Liu. Pose-invariant 3d face alignment. In *IEEE ICCV*, pages 3694–3702, Dec 2015. 2, 6, 7

[31] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *IEEE CVPR*, pages 4188–4196, 2016. 1, 2, 5, 6, 7

[32] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE CVPR*, pages 1867–1874, 2014. 2, 5, 6, 8

[33] J. Kim, C. Liu, F. Sha, and K. Grauman. Deformable spatial pyramid matching for fast dense correspondences. In *IEEE CVPR*, pages 2307–2314, 2013. 2

[34] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE Int. Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 6

[35] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen. Facial performance capture with deep neural networks. *arXiv preprint arXiv:1609.06536*, 2016. 2

[36] D. Lee, H. Park, and C. D. Yoo. Face alignment using cascade gaussian process regression trees. In *IEEE CVPR*, pages 4204–4212, 2015. 2

[37] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4), July 2015. 1

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *IEEE ECCV*, pages 740–755, Cham, 2014. 5

[39] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE PAMI*, 33(5):978–994, 2011. 2

[40] F. Liu, D. Zeng, Q. Zhao, and X. Liu. Joint face alignment and 3d face reconstruction. In *IEEE ECCV*, pages 545–560, 2016. 2

[41] D. S. Ma, J. Correll, and B. Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4):1122–1135, 2015. 5

[42] I. Matthews and S. Baker. Active appearance models revisited. *IEEE IJCV*, 60(2):135–164, 2004. 2

[43] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2016)*, 35(6), December 2016. 1

[44] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In S. Tubaro and J.-L. Dugelay, editors, *AVSS*, pages 296–301. IEEE Computer Society, 2009. 1

[45] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas. A recurrent encoder-decoder network for sequential face alignment. In *IEEE ECCV*, pages 38–56, 2016. 2

[46] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE CVPR*, pages 1685–1692, 2014. 2

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

[48] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013. 6

[49] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from rgb input. In *IEEE ECCV*, 2016. 1, 2, 4, 5

[50] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li. Photorealistic facial texture inference using deep neural networks. In *IEEE CVPR*, 2016. 1, 5

[51] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *IEEE ICCV*, pages 1–8, 2007. 2, 5, 6, 8

[52] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011. 1, 2

[53] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE CVPR*, pages 3476–3483, 2013. 2

[54] T. Taniai, S. N. Sinha, and Y. Sato. Joint recovery of dense correspondence and cosegmentation in two images. In *IEEE CVPR*, pages 4246–4255, 2016. 2

[55] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE CVPR*, 2016. 1, 5

[56] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *IEEE ICCV*, pages 593–600, 2013. 2

[57] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *IEEE CVPR*, pages 2729–2736, 2010. 2

[58] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *IEEE ICCV*, pages 1385–1392, 2013. 2

[59] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE CVPR*, pages 532–539, 2013. 2

[60] X. Xiong and F. D. la Torre. Global supervised descent method. In *IEEE CVPR*, pages 2664–2673, June 2015. 2, 6

[61] H. Yang, X. He, X. Jia, and I. Patras. Robust face alignment under occlusion via regional predictive power estimation. *IEEE Trans. on Image Proc.*, 2015. 2

[62] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *IEEE ICCV*, pages 1944–1951, 2013. 2

[63] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *IEEE ECCV*, pages 94–108, 2014. 2

[64] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *IEEE ICCV Workshops*, pages 386–391, 2013. 2

[65] T. Zhou, Y. Jae Lee, S. X. Yu, and A. A. Efros. Flowweb: Joint image set alignment by weaving consistent, pixel-wise correspondences. In *IEEE CVPR*, pages 1191–1200, 2015. 2

[66] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *IEEE CVPR*, June 2016. 2, 3

[67] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *IEEE CVPR*, pages 4998–5006, 2015. 2

[68] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *IEEE CVPR*, June 2016. 1, 2, 4, 5, 6, 7, 8

[69] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE CVPR*, pages 787–796, 2015. 6, 7

[70] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE CVPR*, pages 2879–2886, 2012. 2