3D Face Reconstruction with Dense Landmarks

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson,
Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin,
Toby Sharp, Ivan Stojiljković, Tom Cashman, and Julien Valentin

Microsoft

Abstract. Landmarks often play a key role in face analysis, but many aspects of identity or expression cannot be represented by sparse landmarks alone. Thus, in order to reconstruct faces more accurately, landmarks are often combined with additional signals like depth images or techniques like differentiable rendering. Can we keep things simple by just using more landmarks? In answer, we present the first method that accurately predicts 10× as many landmarks as usual, covering the whole head, including the eyes and teeth. This is accomplished using synthetic training data, which guarantees perfect landmark annotations. By fitting a morphable model to these dense landmarks, we achieve state-of-the-art results for monocular 3D face reconstruction in the wild. We show that dense landmarks are an ideal signal for integrating face shape information across frames by demonstrating accurate and expressive facial performance capture in both monocular and multi-view scenarios. Finally, our method is highly efficient: we can predict dense landmarks and fit our 3D face model at over 150FPS on a single CPU thread. Please see our website: `https://microsoft.github.io/DenseLandmarks/`.

Keywords: Dense correspondences, 3D morphable model, face alignment, landmarks, synthetic data

1 Introduction

Landmarks are points in correspondence across all faces, like the tip of the nose or the corner of the eye. They often play a role in face-related computer vision, e.g., being used to extract facial regions of interest [34], or helping to constrain 3D model fitting [26, 79]. Unfortunately, many aspects of facial identity or expression cannot be encoded by a typical sparse set of 68 landmarks alone. For example, without landmarks on the cheeks, we cannot tell whether or not someone has high cheek-bones. Likewise, without landmarks around the outer eye region, we cannot tell if someone is softly closing their eyes, or scrunching up their face.

In order to reconstruct faces more accurately, previous work has therefore used additional signals beyond color images, such as depth images [64] or optical flow [13]. However, these signals may not be available or reliable to compute. Instead, given color images alone, others have approached the problem using analysis-by-synthesis: minimizing a photometric error [26] between a generative 3D face model and an observed image using differentiable rendering [18, 27].

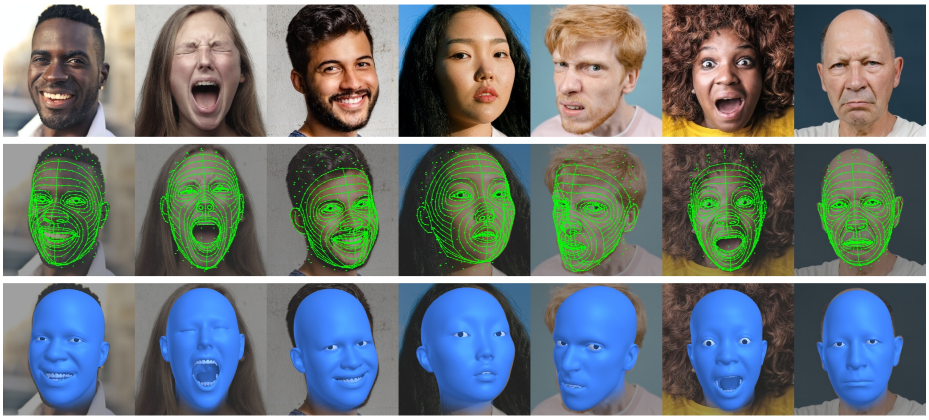


Fig. 1. Given a single image (top), we first robustly and accurately predict 703 landmarks (middle). To aid visualization, we draw lines between landmarks. We then fit our 3D morphable face model to these landmarks to reconstruct faces in 3D (bottom).

Unfortunately, these approaches are limited by the approximations that must be made in order for differentiable rendering to be computationally feasible. In reality, faces are not purely Lambertian [23], and many important illumination effects are not explained using spherical harmonics alone [18], e.g., ambient occlusion or shadows cast by the nose.

Faced with this complexity, wouldn't it be great if we could just use more landmarks? We present the first method that predicts over 700 landmarks both accurately and robustly. Instead of only the frontal "hockey-mask" portion of the face, our landmarks cover the entire head, including the ears, eyeballs, and teeth. As shown in Figure 1, these landmarks provide a rich signal for both facial identity and expression. Even with as few as 68, it is hard for humans to precisely annotate landmarks that are not aligned with a salient image feature. That is why we use synthetic training data which guarantees consistent annotations. Furthermore, instead of representing each landmark as just a 2D coordinate, we predict each one as a random variable: a 2D circular Gaussian with position and uncertainty [38]. This allows our predictor to express uncertainty about certain landmarks, e.g., occluded landmarks on the back of the head.

Since our dense landmarks represent points of correspondence across all faces, we can perform 3D face reconstruction by fitting a morphable face model [6] to them. Although previous approaches have fit models to landmarks in a similar way [77], we are the first to show that landmarks are the only signal required to achieve state-of-the-art results for monocular face reconstruction in the wild.

The probabilistic nature of our predictions also makes them ideal for fitting a 3D model over a temporal sequence, or across multiple views. An optimizer can discount uncertain landmarks and rely on more certain ones. We demonstrate this with accurate and expressive results for both multi-view and monocular facial performance capture. Finally, we show that predicting dense landmarks

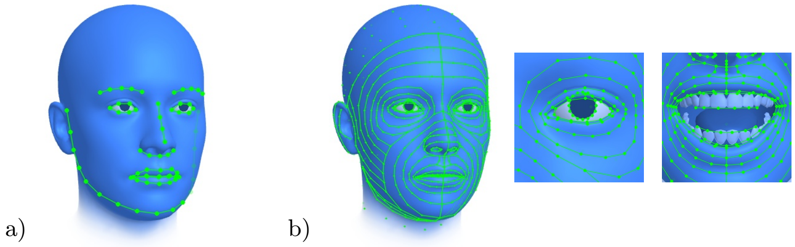


Fig. 2. Compared to a typical sparse set of 68 facial landmarks (a), our dense landmarks (b) cover the entire head in great detail, including ears, eyes, and teeth. These dense landmarks are better at encoding facial identity and subtle expressions.

and then fitting a model can be highly efficient by demonstrating real-time facial performance capture at over 150FPS on a single CPU thread.

In summary, our main contribution is to show that you can achieve more with less. You don't need parametric appearance models, illumination models, or differentiable rendering for accurate 3D face reconstruction. All you need is a sufficiently large quantity of accurate 2D landmarks and a 3D model to fit to them. In addition, we show that combining probabilistic landmarks and model fitting lets us intelligently aggregate face shape information across multiple images by demonstrating robust and expressive results for both multi-view and monocular facial performance capture.

2 Related work

Reconstructing faces in 3D from images is a mature field at the intersection of vision and graphics. We focus our literature review on methods that are closer to our own, and refer the reader to Morales et al. [47] for an extensive survey.

Regression-based 3D face reconstruction DNN-based regression has been extensively used as a tool for 3D face reconstruction. Techniques fall into two broad categories: supervised, and self-supervised. Approaches either use 3D Morphable Models (3DMMs) [7, 28, 41], or eschew linear models and instead learn a non-linear one as part of the training process [66].

Fully supervised techniques either use parameter values from a 3DMM that is fit to the data via optimization as labels [14, 68, 73], or known face geometry is posed by sampling from a 3DMM and rendered to create synthetic datasets [21, 27, 51, 57]. Self-supervised approaches commonly use landmark reprojection error and/or perceptual loss via differentiable rendering [17, 23, 27, 31, 32, 45, 52, 55, 62, 63, 66, 67]. Other techniques augment this with 3D or multiview constraints [20, 44, 58, 61, 74, 75]. While this is similar to our technique, we only use a DNN to regress landmark positions which are then used to optimize 3DMM parameters, as in the large body of hybrid model-fitting methods [8, 33].

Optimization-based 3D face reconstruction Traditionally, markerless reconstruction of face geometry is achieved with multi-view stereo [4, 56], followed by optical flow based alignment, and then optimisation using geometric and temporal priors [5, 9, 50]. While such methods produce detailed results, each step takes hours to complete. They also suffer from drift and other issues due to their reliance on optical flow and multi-view stereo [15]. While our method cannot reconstruct faces in such fine detail, it accurately recovers the low-frequency shape of the face, and aligns it with a common topology. This enriches the raw data with semantics, making it useful for other tasks.

If only a single image is available, dense photometric [18, 65], depth [64], or optical flow [13] constraints are commonly used to recover face shape and motion. However, these methods still rely on sparse landmarks for initializing the optimization close to the dense constraint's basin of convergence, and coping with fast head motion [79]. In contrast, we argue that dense landmarks alone are sufficient for accurately recovering the overall shape of the face.

Dense landmark prediction While sparse landmark prediction is a mainstay of the field [12], few methods directly predict dense landmarks or correspondences. This is because annotating a face with dense landmarks is a highly ambiguous task, so either synthetic data [71], pseudo-labels made with model-fitting [16, 25, 78], or semi-automatic refinement of training data [36, 37] are used. Another issue with predicting dense landmarks is that heatmaps, the *de facto* technique for predicting landmarks [11, 12], rise in computational complexity with the number of landmarks. While a few previous methods have predicted dense frontal-face landmarks via cascade regression [36] or direct regression [16, 29, 37], we are the first to accurately and robustly predict over 700 landmarks covering the whole head, including eyes and teeth.

Some methods choose to predict dense correspondences as an image instead, where each pixel corresponds to a fixed point in a UV-unwrapping of the face [1, 25] or body [30, 60]. Such parameterization suffers from several drawbacks. How does one handle self-occluded portions of the face, e.g., the back of the head? Furthermore, what occurs at UV-island boundaries? If a pixel is half-nose and half-cheek, to which does it correspond? Instead, we choose to discretize the face into dense landmarks. This lets us predict parts of the face that are self-occluded, or lie outside image bounds. Having a fixed set of correspondences also benefits the model-fitter, making it more amenable to running in real-time.

3 Method

In recent years, methods for 3D face reconstruction have become more and more complicated, involving differentiable rendering and complex neural network training strategies. We show instead that success can be found by keeping things simple. Our approach consists of two stages: First we predict probabilistic dense 2D landmarks L using a traditional convolutional neural network (CNN). Then, we fit a 3D face model, parameterized by $\mathbf{\Phi}$, to the 2D landmarks by

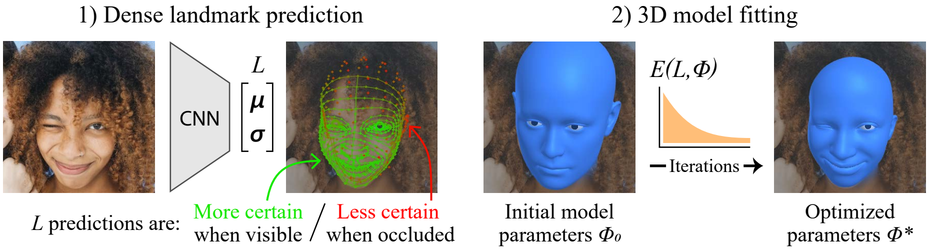

1) Dense landmark prediction
2) 3D model fitting
CNN
L
μ
σ
E(L,Φ)
Iterations
L predictions are:
More certain when visible
Less certain when occluded
Initial model parameters Φ₀
Optimized parameters Φ*

Fig. 3. Given an image, we first predict probabilistic dense landmarks L, each with position μ and certainty σ. Then, we fit our 3D face model to L, minimizing an energy E by optimizing model parameters $\mathbf{\Phi}$.

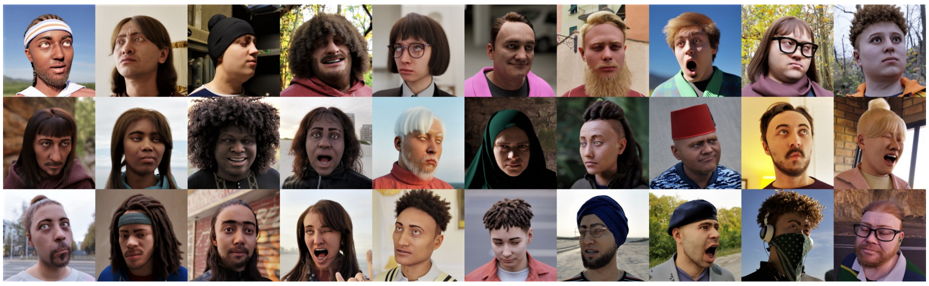


Fig. 4. Examples of our synthetic training data. Without the perfectly consistent annotations provided by synthetic data, dense landmark prediction would not be possible.

minimizing an energy function $E(\mathbf{\Phi}; L)$. Images themselves are not part of this optimization; the only data used are 2D landmarks.

The main difference between our work and previous approaches is the number and quality of landmarks. No one before has predicted so many 2D landmarks, so accurately. This lets us achieve accurate 3D face reconstruction results by fitting a 3D model to these landmarks alone.

3.1 Landmark prediction

Synthetic training data. Our results are only possible because we use synthetic training data. While a human can consistently label face images with e.g., 68 landmarks, it would be almost impossible for them to annotate an image with dense landmarks. How would it be possible to consistently annotate occluded landmarks on the back of the head, or multiple landmarks over a largely featureless patch of skin e.g., the forehead? In previous work, pseudo-labelled real images with dense correspondences are obtained by fitting a 3DMM to images [1], but the resulting label consistency heavily depends on the quality of the 3D fitting. Using synthetic data has the advantage of guaranteeing perfectly consistent labels. We rendered a training dataset of 100k images using the method of Wood et al. [71] with some minor modifications: we include expression-dependent

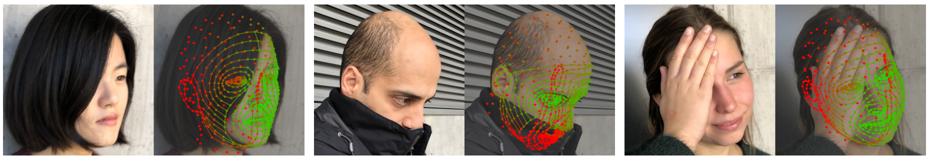


Fig. 5. When parts of the face are occluded by e.g. hair or clothing, the corresponding landmarks are predicted with high uncertainty (red), compared to those visible (green).

wrinkle texture maps for more realistic skin appearance, and additional clothing, accessory, and hair assets. See Figure 4 for some examples.

Probabilistic landmark regression. We predict each landmark as a random variable with the probability density function of a circular 2D Gaussian. So $L_i = \{\boldsymbol{\mu}_i, \sigma_i\}$, where $\boldsymbol{\mu}_i = [x_i, y_i]$ is the expected position of that landmark, and σ_i (the standard deviation) is a measure of uncertainty. Our training data includes labels for landmark positions $\boldsymbol{\mu}'_i = [x'_i, y'_i]$, but not for σ. The network learns to output σ in an unsupervised fashion to show that it is certain about some landmarks, e.g., visible landmarks on the front of the face, and uncertain about others, e.g., landmarks hidden behind hair (see Figure 5). This is achieved by training the network with a Gaussian negative log likelihood (GNLL) loss [38]:

$$\text{Loss}(L) = \sum_{i=1}^{|L|} \lambda_i \left(\underbrace{\log\left(\sigma_i^2\right)}_{\text{Loss}_\sigma} + \underbrace{\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}'_i\|^2}{2\sigma_i^2}}_{\text{Loss}_\mu} \right) \tag{1}$$

Loss_σ penalizes the network for being too uncertain, and Loss_μ penalizes the network for being inaccurate. λ_i is a per-landmark weight that focuses the loss on certain parts of the face. This is the only loss used during training.

The probabilistic nature of our landmark predictions is important for accuracy. A network trained with the GNLL loss is more accurate than a network trained with L2 loss on positions only. Perhaps this is the result of the CNN being able to discount challenging landmarks (e.g., fully occluded ones), and spend more capacity on making precise predictions about visible landmarks.

Landmarks are commonly predicted via heatmaps [11]. However, generating heatmaps is computationally expensive [42]; it would not be feasible to output over 700 heatmaps in real-time. Heatmaps also prevent us predicting landmarks outside image bounds. Instead, we keep things simple, and directly regress position and uncertainty using a traditional CNN. We are able to take any off-the-shelf architecture, and alter the final fully-connected layer to output three values per-landmark: two for position and one for uncertainty. Since this final layer represents a small percentage of total CNN compute, our method scales well with landmark quantity.

Training details. Landmark coordinates are normalized from $[0, S]$ to $[-1, 1]$, for a square image of size $S \times S$. Rather than directly outputting σ, we predict

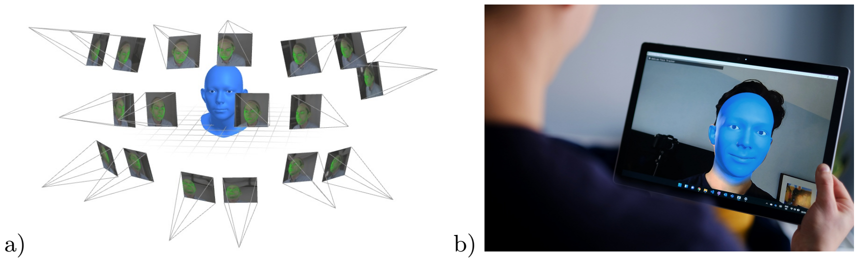


Fig. 6. We implemented two versions of our approach: one for processing multi-view recordings offline (a), and one for real-time facial performance capture (b).

$\log \sigma$, and take its exponential to ensure σ is positive. Using PyTorch [48], we train ResNet [35] and MobileNet V2 [54] models from the timm [70] library using AdamW [46] with automatically determined learning rate [22]. We use data augmentation to help our synthetic data cross the domain gap [71].

3.2 3D model fitting

Given probabilistic dense 2D landmarks L, our goal is to find optimal model parameters $\mathbf{\Phi}^*$ that minimize the following energy:

$$E(\mathbf{\Phi}; L) = \underbrace{E_{\text{landmarks}}}_{\text{Data term}} + \underbrace{E_{\text{identity}} + E_{\text{expression}} + E_{\text{joints}} + E_{\text{temporal}} + E_{\text{intersect}}}_{\text{Regularizers}}$$

$E_{\text{landmarks}}$ is the only term that encourages the 3D model to explain the observed 2D landmarks. The other terms use prior knowledge to regularize the fit.

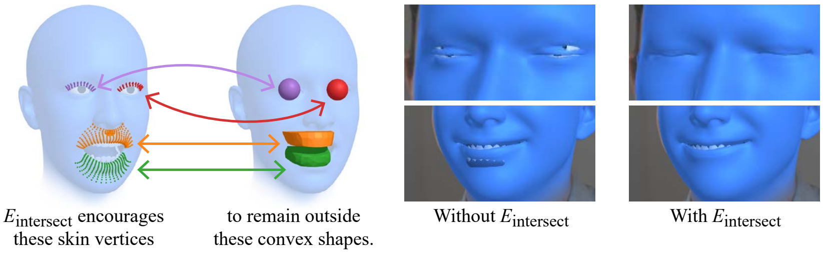
Part of the beauty of our approach is how naturally it scales to multiple images and cameras. In this section we present the general form of our method, suitable for F frames over C cameras, i.e., multi-view performance capture.

3D face model. We use the face model described in [71], comprising $N = 7{,}667$ vertices and $K = 4$ skeletal joints (the head, neck, and two eyes). Vertex positions are determined by the mesh-generating function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}) : \mathbb{R}^{|\boldsymbol{\beta}|+|\boldsymbol{\psi}|+|\boldsymbol{\theta}|} \to \mathbb{R}^{3N}$ which takes parameters $\boldsymbol{\beta} \in \mathbb{R}^{|\boldsymbol{\beta}|}$ for identity, $\boldsymbol{\psi} \in \mathbb{R}^{|\boldsymbol{\psi}|}$ for expression, and $\boldsymbol{\theta} \in \mathbb{R}^{3K+3}$ for skeletal pose (including root joint translation).

$$\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta}) = \mathcal{L}(\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\psi}), \boldsymbol{\theta}, \mathcal{J}(\boldsymbol{\beta}); \mathbf{W})$$

where $\mathcal{L}(\mathbf{V}, \boldsymbol{\theta}, \mathbf{J}; \mathbf{W})$ is a standard linear blend skinning (LBS) function [40] that rotates vertex positions $\mathbf{V} \in \mathbb{R}^{3N}$ about joint locations $\mathbf{J} \in \mathbb{R}^{3K}$ by local joint rotations in $\boldsymbol{\theta}$, with per-vertex weights $\mathbf{W} \in \mathbb{R}^{K \times N}$. The face mesh and joint locations in the bind pose are determined by $\mathcal{T}(\boldsymbol{\beta}, \boldsymbol{\psi}) : \mathbb{R}^{|\boldsymbol{\beta}|+|\boldsymbol{\psi}|} \to \mathbb{R}^{3N}$ and $\mathcal{J}(\boldsymbol{\beta}) : \mathbb{R}^{|\boldsymbol{\beta}|} \to \mathbb{R}^{3K}$ respectively. See Wood et al. [71] for more details.

Cameras are described by a world-to-camera rigid transform $\mathbf{X} \in \mathbb{R}^{3 \times 4} = [\mathbf{R}|\mathbf{T}]$ comprising rotation and translation, and a pinhole camera projection matrix $\mathbf{\Pi} \in \mathbb{R}^{3 \times 3}$. Thus, the image-space projection of the j^{th} landmark in the i^{th} camera is $\mathbf{x}_{i,j} = \mathbf{\Pi}_i \mathbf{X}_i \mathcal{M}_j$. In the monocular case, $\mathbf{X}$ can be ignored.



$E_{\text{intersect}}$ encourages these skin vertices

to remain outside these convex shapes.

Without $E_{\text{intersect}}$

With $E_{\text{intersect}}$

Fig. 7. We encourage the optimizer to avoid face mesh self-intersections by penalizing skin vertices that enter the convex hulls of the eyeballs or teeth parts.

Parameters $\mathbf{\Phi}$ are optimized to minimize E. The main parameters of interest control the face, but we also optimize camera parameters if they are unknown.

$$\mathbf{\Phi} = \{\underbrace{\boldsymbol{\beta}, \mathbf{\Psi}_{F\times|\boldsymbol{\psi}|}, \mathbf{\Theta}_{F\times|\boldsymbol{\theta}|}}_{\text{Face}}; \underbrace{\mathbf{R}_{C\times 3}, \mathbf{T}_{C\times 3}, \mathbf{f}_{C}}_{\text{Camera(s)}}\}$$

Facial identity $\boldsymbol{\beta}$ is shared over a sequence of F frames, but expression $\mathbf{\Psi}$ and pose $\mathbf{\Theta}$ vary per frame. For each of our C cameras we have six degrees of freedom for rotation $\mathbf{R}$ and translation $\mathbf{T}$, and a single focal length parameter f. In the monocular case, we only optimize focal length.

$E_{\textbf{landmarks}}$ encourages the 3D model to explain the predicted 2D landmarks:

$$E_{\text{landmarks}} = \sum_{i,j,k}^{F,C,|L|} \frac{\|\mathbf{x}_{ijk} - \boldsymbol{\mu}_{ijk}\|^2}{2\sigma_{ijk}^2} \tag{2}$$

where, for the k^{th} landmark seen by the j^{th} camera in the i^{th} frame, $[\boldsymbol{\mu}_{ijk}, \sigma_{ijk}]$ is the 2D location and uncertainty predicted by our dense landmark CNN, and $\mathbf{x}_{ijk} = \mathbf{\Pi}_j \mathbf{X}_j \mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\psi}_i, \boldsymbol{\theta}_i)_k$ is the 2D projection of that landmark on our 3D model. The similarity of Equation 2 to Loss_μ in Equation 1 is no accident: treating landmarks as 2D random variables during both prediction and model-fitting allows our approach to elegantly handle uncertainty, taking advantage of landmarks the CNN is confident in, and discounting those it is uncertain about.

$E_{\textbf{identity}}$ penalizes unlikely face shape by maximizing the relative log-likelihood of shape parameters $\boldsymbol{\beta}$ under a multivariate Gaussian Mixture Model (GMM) of G components fit to a library of 3D head scans [71]. $E_{\text{identity}} = -\log(p(\boldsymbol{\beta}))$ where $p(\boldsymbol{\beta}) = \sum_{i=1}^{G} \gamma_i \, \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\nu}_i, \, \mathbf{\Sigma}_i)$. $\boldsymbol{\nu}_i$ and $\mathbf{\Sigma}_i$ are the mean and covariance matrix of the i^{th} component, and γ_i is the weight of that component.

$E_{\textbf{expression}} = \|\boldsymbol{\psi}\|^2$ and $E_{\textbf{joints}} = \|\boldsymbol{\theta}_{i:i\in[2,K]}\|^2$ encourage the optimizer to explain the data with as little expression and joint rotation as possible. We do not penalize global translation or rotation by ignoring the root joint $\boldsymbol{\theta}_1$.

$E_{\textbf{temporal}} = \sum_{i=2,j,k}^{F,C,|L|} \|\mathbf{x}_{i,j,k} - \mathbf{x}_{i-1,j,k}\|^2$ reduces jitter by encouraging face mesh vertices $\mathbf{x}$ to remain still between neighboring frames $i-1$ and i.

$E_{\textbf{intersect}}$ encourages the optimizer to find solutions without intersections between the skin and eyeballs or teeth (Figure 7). Please refer to the supplementary material for further details.

3.3 Implementation

We implemented two versions of our system: one for processing multi-camera recordings offline, and one for real-time facial performance capture.

Our **offline** system produces the best quality results without constraints on compute. We predict 703 landmarks with a ResNet 101 [35]. To extract a facial Region-of-Interest (ROI) from an image we run a full-head probabilistic landmark CNN on multi-scale sliding windows, and select the window with the lowest uncertainty. When fitting our 3DMM, we use PyTorch [48] to minimize $E(\mathbf{\Phi})$ with L-BFGS [43], optimizing all parameters across all frames simultaneously.

For our **real-time** system, we trained a lightweight dense landmark model with MobileNet V2 architecture [54]. To compensate for a reduction in network capacity, we predict 320 landmarks rather than 703, and modify the ROI strategy: aligning the face so it appears upright with the eyes a fixed distance apart. This makes the CNN's job easier for frontal faces at the expense of profile ones.

Real-time model fitting. We use the Levenberg-Marquardt algorithm to optimize our model-fitting energy. Camera and identity parameters are only fit occasionally. For the majority of frames we fit pose and expression parameters only. We rewrite the energy E in terms of the vector of residuals, $\mathbf{r}$, as $E(\mathbf{\Phi}) = \|\mathbf{r}(\mathbf{\Phi})\|^2 = \sum_i r_i(\mathbf{\Phi})^2$. Then at each iteration k of our optimization, we can compute $\mathbf{r}(\mathbf{\Phi}_k)$ and the Jacobian, $J(\mathbf{\Phi}_k) = \frac{\partial \mathbf{r}(\mathbf{\Phi})}{\partial \mathbf{\Phi}}|^{\mathbf{\Phi}=\mathbf{\Phi}_k}$, and use these to solve the symmetric, positive-semi-definite linear system, $(J^T J + \lambda \text{diag}(J^T J))\boldsymbol{\delta}_k = -J^T \mathbf{r}$ via Cholesky decomposition. We then apply the update rule, $\mathbf{\Phi}_{k+1} = \mathbf{\Phi}_k + \boldsymbol{\delta}_k$.

In practice we do not actually form the residual vector $\mathbf{r}$ nor the Jacobian matrix J. Instead, for performance reasons, we directly compute the quantities $J^T J$ and $J^T \mathbf{r}$ as we visit each term $r_i(\mathbf{\Phi}_k)$ of the energy. Most of the computational cost is incurred in evaluating these products for the landmark data term, as expected. However, the Jacobian of landmark term residuals is not fully dense. Each individual landmark depends on its own subset of expression parameters, and is invariant to other expression parameters. We performed a static analysis of the sparsity of each landmark term with respect to parameters, $\partial r_i / \partial \Phi_j$, and we use this set of i, j indices to reduce the cost of our outer products from $O(|\mathbf{\Phi}|^2)$ to $O(m_i^2)$, where m_i is the sparsified dimensionality of $\partial r_i / \partial \mathbf{\Phi}$. We further enhance the sparsity by ignoring any components of the Jacobian with an absolute value below a certain empirically-determined threshold.

By exploiting sparsity in this way, the landmark term residuals and their derivatives become very cheap to evaluate. This formulation avoids the correspondence problem usually seen with depth images [59], which requires a more expensive optimization. In addition, adding more landmarks does not significantly increase the cost of optimization. It therefore becomes possible to implement a very detailed and well-regularized fitter with a relatively small compute

	Common	Challenging	Private
Method	NME	NME	$FR_{10\%}$
LAB [72]	🟠 2.98	5.19	0.83
AWING [69]	🟡 **2.72**	🟡 **4.52**	⚪ 0.33
ODN [76]	3.56	6.67	-
3FabRec [10]	3.36	5.74	🟡 **0.17**
Wood et al. [71]	3.09	4.86	🟠 0.50
LUVLi [39]	⚪ 2.76	5.16	-
ours (L2)	3.30	🟠 5.12	⚪ 0.33
ours (GNLL)	3.03	⚪ 4.80	🟡 **0.17**

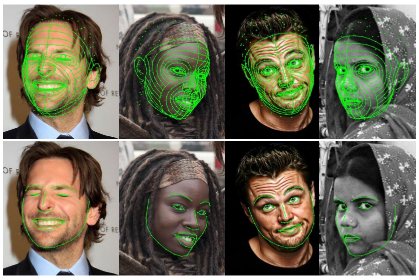


Fig. 8. Left: results on 300W dataset, lower is better. Note competitive performance of our model (despite being evaluated across-dataset) and importance of GNLL loss. Right: sample predictions (top row) with label-translated results (bottom row).

burden, simply by adding a sufficient number of landmarks. The cost of the Cholesky solve for the update $\boldsymbol{\delta}_k$ is independent of the number of landmarks.

4 Evaluation

4.1 Landmark accuracy

We measure the accuracy of a ResNet 101 dense landmark model on the **300W** [53] dataset. For benchmark purposes only, we employ label translation [71] to deal with systematic inconsistencies between our 703 predicted dense landmarks and the 68 sparse landmarks labelled as ground truth (see Figure 8). While previous work [71] used label translation to evaluate a synthetically-trained sparse landmark predictor, we use it to evaluate a dense landmark predictor.

We use the standard normalized mean error (NME) and failure rate ($FR_{10\%}$) error metrics [53]. Our model's results in Figure 8 are competitive with the state of the art, despite being trained with synthetic data alone. Note: these results provide a conservative estimate of our method's accuracy as the translation network may introduce error, especially for rarely seen expressions.

Ablation study We measured the importance of predicting each landmark as a random variable rather than as a 2D coordinate. We trained two landmark prediction models, one with our proposed GNLL loss (Equation 1), and one with a simpler L2 loss on landmark coordinate only. Results in Figure 8 confirm that including uncertainty in landmark regression results in better accuracy.

Qualitative comparisons are shown in Figure 9 between our real-time dense landmark model (MobileNet V2) and MediaPipe Attention Mesh [29], a publicly available dense landmark method designed for mobile devices. Our method is more robust, perhaps due to the consistency and diversity of our synthetic training data. See the supplementary material for additional qualitative results, including landmark predictions on the Challenging subset of 300W.

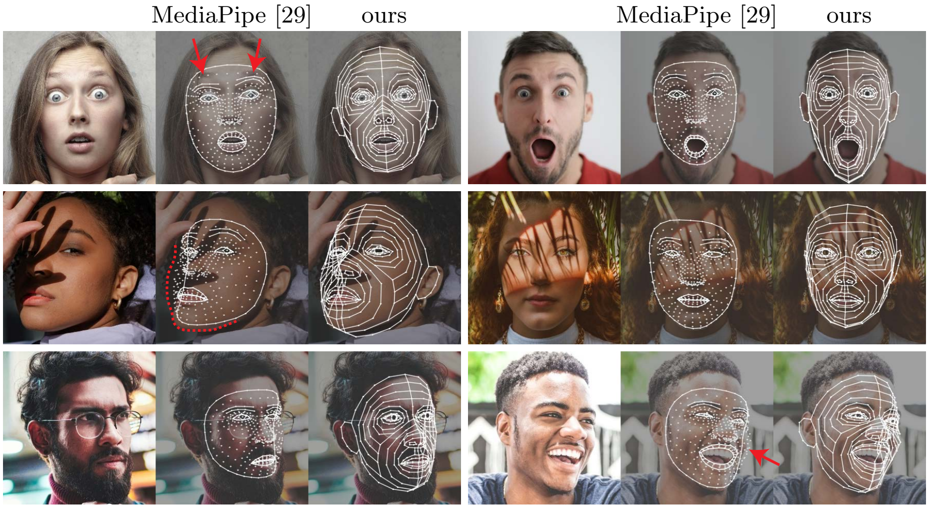


Fig. 9. We compare our real-time landmark CNN (MobileNet V2) with MediaPipe Attention Mesh [29], a publicly available method for dense landmark prediction. Our approach is more robust to challenging expressions and illumination.

4.2 3D face reconstruction

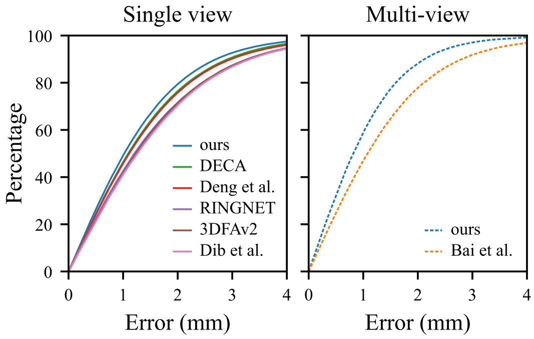
Quantitatively, we compare our offline approach with recent methods on two benchmarks: the NoW Challenge [55] and the MICC dataset [2].

The NoW Challenge [55] provides a standard evaluation protocol for measuring the accuracy and robustness of 3D face reconstruction in the wild. It consists of 2054 face images of 100 subjects along with a 3D head scan for each subject which serves as ground truth. We undertake the challenge in two ways: *single view*, where we fit our face model to each image separately, and *multi-view*, where we fit a per-subject face model to all image of a particular subject. As shown in Figure 10, we achieve state of the art results.

The MICC dataset [2] consists of 3D face scans and videos of 53 subjects. The videos were recorded in three environments: a “cooperative” laboratory environment, an indoor environment, and an outdoor environment. We follow Deng et al. [17], and evaluate our method in two ways: *single view*, where we estimate one face shape per frame in a video, and average the resulting face meshes, and *multi-view*, where we fit a single face model to all frames in a video jointly. As shown in Table 1, we achieve state of the art results.

Note that many previous methods are incapable of aggregating face shape information across multiple views. The fact ours can benefit from multiple views highlights the flexibility of our hybrid model-fitting approach.

Ablation studies We conducted an experiment to measure the importance of landmark quantity for 3D face reconstruction. We trained three landmark CNNs,



Method	Error (mm)		
Single view	Median	Mean	Std
Deng et al. [17]	1.11	1.41	1.21
RingNet [55]	1.21	1.53	1.31
3DFAv2 [31]	1.23	1.57	1.39
DECA [24]	1.09	1.38	1.18
Dib et al. [19]	1.26	1.57	1.31
ours	**1.02**	**1.28**	**1.08**
Multi-view			
Bai et al. [3]	1.08	1.35	1.15
ours	**0.81**	**1.01**	**0.84**

Fig. 10. Results for the NoW Challenge [55]. We outperform the state of the art on both single- and multi-view 3D face reconstruction.

Table 1. Results on the MICC dataset [2], following the single and multi-frame evaluation protocol of Deng et al. [17]. We achieve state-of-the-art results.

Method	Error (mm), mean		
Single view	Coop.	Indoor	Outdoor
Tran et al. [68]	1.97	2.03	1.93
Genova et al. [27]	1.78	1.78	1.76
Deng et al. [17]	1.66	1.66	1.69
ours	**1.64**	**1.62**	**1.61**

Method	Error (mm), mean		
Multi-view	Coop.	Indoor	Outdoor
Piotraschke and Blanz [49]	1.68	1.67	1.72
Deng et al. [17]	1.60	1.61	1.63
ours	**1.43**	**1.42**	**1.42**

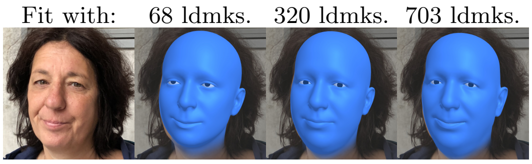
predicting 703, 320, and 68 landmarks respectively, and used these on the NoW Challenge (validation set). As shown in Figure 11, fitting with more landmarks results in more accurate 3D face reconstruction.

In addition, we investigated the importance of using landmark uncertainty σ in model fitting. We fit our model to 703 landmark predictions on the NoW validation set, but using fixed rather than predicted σ. Figure 11 (bottom row of table) shows that fitting without σ leads to worse results.

Qualitative comparisons between our work and several publicly available methods [17, 24, 31, 55, 58] can be found in Figure 13.

4.3 Facial performance capture

Multi-view Good synthetic training data requires a database of facial expression parameters from which to sample. We acquired such a database by conducting markerless facial performance capture for 108 subjects. We recorded each subject in our 17-camera studio, and processed each recording with our offline multi-view model fitter. For a 520 frame sequence it takes 3 minutes to predict dense landmarks for all images, and a further 9 minutes to optimize face model parameters. See Figure 12 for some of the 125,000 frames of expression data captured with our system. As the system which is used to create the database



Number of Landmarks	Error (mm) Median	Mean	Std
68	1.10	1.38	1.16
320	1.00	1.24	1.02
703	**0.95**	**1.17**	**0.97**
703 (without σ)	1.02	1.26	1.03

Fig. 11. Ablation studies on the NoW [55] validation set confirm that denser is better: model fitting with more landmarks leads to more accurate results. In addition, we see that fitting without using σ leads to worse results.

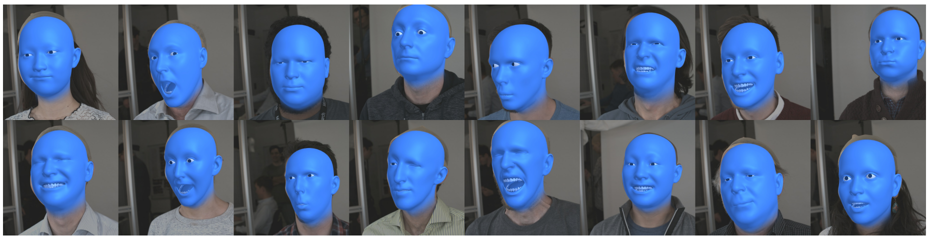


Fig. 12. We demonstrate the robustness and reliability of our method by using it to collect a massive database of 125,000 facial expressions, fully automatically.

is then subsequently re-trained with it, we produced several databases in this manner until no further improvement was seen. We do not reconstruct faces in fine detail like previous multi-view stereo approaches [5, 9, 50]. However, while previous work can track a detailed 3D mesh over a performance, our approach reconstructs the performance with richer semantics: identity and expression parameters for our generative model. In many cases it is sufficient to reconstruct the low-frequency shape of the face accurately, without fine details.

Real-time monocular See the last two columns of Figure 13 for a comparison between our offline and real-time systems for monocular 3D model-fitting. While our offline system produces the best possible results by using a large CNN and optimizing over all frames simultaneously, our real-time system can still produce accurate and expressive results fitting frame-to-frame. Please refer to the supplementary material for more results. Running on a single CPU thread (i5-11600K), our real-time system spends 6.5ms processing a frame (150FPS), of which 4.1ms is spent predicting dense landmarks and 2.3ms is spent fitting our face model.

5 Limitations and future work

Our method depends entirely on accurate landmarks. As shown in Figure 14, if landmarks are poorly predicted, the resulting model fit suffers. We plan to address this by improving our synthetic training data. Additionally, since our model does not include tongue articulation we cannot recover tongue movement.

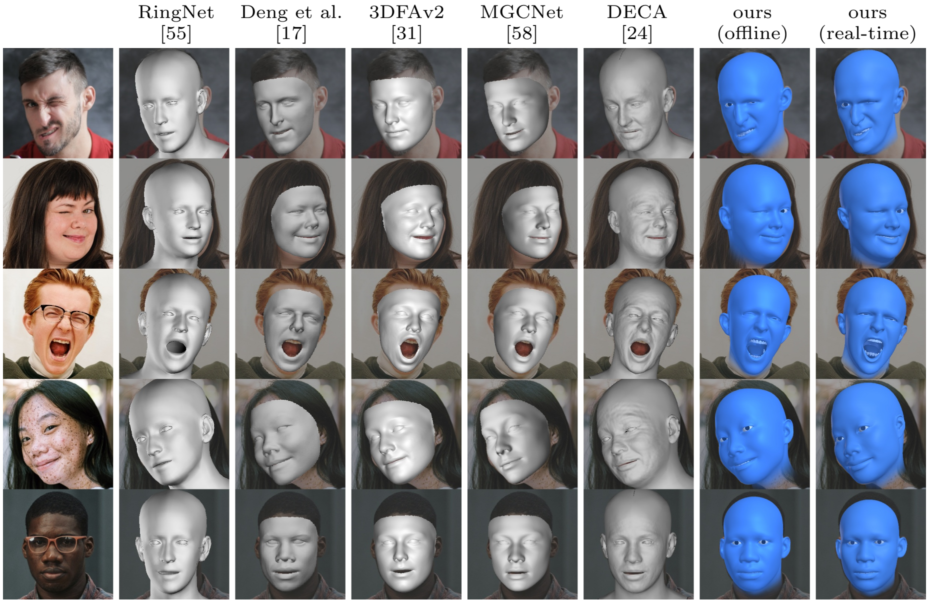


Fig. 13. Compared to previous recent monocular 3D face reconstruction methods, ours better captures gaze, expressions like winks and sneers, and subtleties of facial identity. In addition, our method can run in real time with only a minor loss of fidelity.

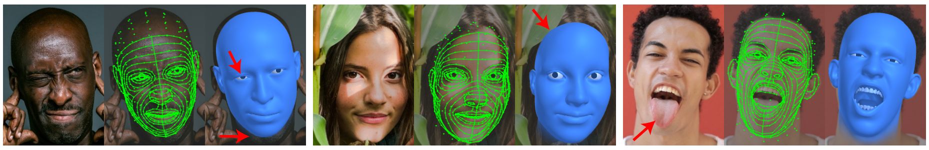


Fig. 14. Bad landmarks result in bad fits, and we are incapable of tracking the tongue.

Heatmaps have dominated landmark prediction for some time [11, 12]. We were pleasantly surprised to find that directly regressing 2D landmark coordinates with unspecialized architectures works well and eliminates the need for computationally-costly heatmap generation. In addition, we were surprised that predicting σ helps accuracy. We look forward to further investigating direct probabilistic landmark regression as an alternative to heatmaps in future work.

In conclusion, we have demonstrated that dense landmarks are an ideal signal for 3D face reconstruction. Quantitative and qualitative evaluations have shown that our approach outperforms those previous by a significant margin, and excels at multi-view and monocular facial performance capture. Finally, our approach is highly efficient, and runs at over 150FPS on a single CPU thread.

Acknowledgements Thanks to Chirag Raman and Jamie Shotton for their contributions, and Jiaolong Yang and Timo Bolkart for help with evaluation.

Bibliography

[1] Alp Güler, R., Trigeorgis, G., Antonakos, E., Snape, P., Zafeiriou, S., Kokkinos, I.: DenseReg: Fully Convolutional Dense Shape regression In-the-Wild. In: CVPR (2017)

[2] Bagdanov, A.D., Del Bimbo, A., Masi, I.: The Florence 2D/3D Hybrid Face Dataset. In: Workshop on Human Gesture and Behavior Understanding. ACM (2011)

[3] Bai, Z., Cui, Z., Liu, X., Tan, P.: Riggable 3D Face Reconstruction via In-Network Optimization. In: CVPR (2021)

[4] Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-Quality Single-Shot Capture of Facial Geometry. ACM Trans. Graph. (2010)

[5] Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R.W., Gross, M.: High-quality passive facial performance capture using anchor frames. ACM Trans. Graph. (2011)

[6] Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: Computer graphics and interactive techniques (1999)

[7] Blanz, V., Vetter, T.: Face Recognition Based on Fitting a 3D Morphable Model. TPAMI (2003)

[8] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (Oct 2016)

[9] Bradley, D., Heidrich, W., Popa, T., Sheffer, A.: High Resolution Passive Facial Performance Capture. ACM Trans. Graph. **29**(4) (2010)

[10] Browatzki, B., Wallraven, C.: 3FabRec: Fast Few-shot Face alignment by Reconstruction. In: CVPR (2020)

[11] Bulat, A., Sanchez, E., Tzimiropoulos, G.: Subpixel Heatmap Regression for Facial Landmark Localization. In: BMVC (2021)

[12] Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In: ICCV (2017)

[13] Cao, C., Chai, M., Woodford, O., Luo, L.: Stabilized real-time face tracking via a learned dynamic rigidity prior. ACM Transactions on Graphics (2018)

[14] Chandran, P., Bradley, D., Gross, M., Beeler, T.: Semantic Deep Face Models. In: International Conference on 3D Vision (3DV) (2020)

[15] Cong, M., Lan, L., Fedkiw, R.: Local geometric indexing of high resolution data for facial reconstruction from sparse markers. CoRR **abs/1903.00119** (2019), `http://arxiv.org/abs/1903.00119`

[16] Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: RetinaFace: Single-shot Multi-level Face Localisation in the Wild. In: CVPR (2020)

[17] Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3D Face Reconstruction With Weakly-Supervised Learning: From Single Image to Image Set. In: CVPR Workshops (2019)

[18] Dib, A., Bharaj, G., Ahn, J., Thébault, C., Gosselin, P., Romeo, M., Chevallier, L.: Practical face reconstruction via differentiable ray tracing. Computer Graphics Forum **40**(2) (2021)

[19] Dib, A., Thebault, C., Ahn, J., Gosselin, P.H., Theobalt, C., Chevallier, L.: Towards High Fidelity Monocular Face Reconstruction with Rich Reflectance using Self-supervised Learning and Ray Tracing. In: CVPR (2021)

[20] Dou, P., Kakadiaris, I.A.: Multi-view 3D face reconstruction with deep recurrent neural networks. Image and Vision Computing (2018)

[21] Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3D face reconstruction with deep neural networks. In: CVPR (2017)

[22] Falcon, W., et al.: Pytorch lightning. GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning **3**, 6 (2019)

[23] Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. ACM Transactions on Graphics (ToG) (2021)

[24] Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. ACM Transactions on Graphics (ToG) (2021)

[25] Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In: ECCV (2018)

[26] Garrido, P., Zollhöfer, M., Casas, D., Valgaerts, L., Varanasi, K., Pérez, P., Theobalt, C.: Reconstruction of Personalized 3D Face Rigs from Monocular Video. ACM Trans. Graph. **35**(3) (2016)

[27] Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised Training for 3D Morphable Model Regression. In: CVPR (2018)

[28] Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., Vetter, T.: Morphable face models-an open framework. In: Automatic Face & Gesture Recognition (FG). IEEE (2018)

[29] Grishchenko, I., Ablavatski, A., Kartynnik, Y., Raveendran, K., Grundmann, M.: Attention Mesh: High-fidelity Face Mesh Prediction in Real-time. In: CVPR Workshops (2020)

[30] Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018)

[31] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S.Z.: Towards Fast, Accurate and Stable 3D Dense Face Alignment. In: ECCV (2020)

[32] Guo, Y., Cai, J., Jiang, B., Zheng, J., et al.: Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. TPAMI (2018)

[33] Han, S., Liu, B., Cabezas, R., Twigg, C.D., Zhang, P., Petkau, J., Yu, T.H., Tai, C.J., Akbay, M., Wang, Z., et al.: Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Transactions on Graphics (TOG) **39**(4), 87–1 (2020)

[34] Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective Face Frontalization in Unconstrained Images. In: CVPR (2015)

[35] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[36] Jeni, L.A., Cohn, J.F., Kanade, T.: Dense 3D face alignment from 2D videos in real-time. In: Automatic Face and Gesture Recognition (FG). IEEE (2015)

[37] Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time facial surface geometry from monocular video on mobile GPUs. In: CVPR Workshops (2019)

[38] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems **30** (2017)

[39] Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: CVPR (2020)

[40] Lewis, J.P., Cordner, M., Fong, N.: Pose Space Deformation: A Unified Approach to Shape Interpolation and Skeleton-Driven Deformation. In: SIGGRAPH (2000)

[41] Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) (2017)

[42] Li, Y., Yang, S., Zhang, S., Wang, Z., Yang, W., Xia, S.T., Zhou, E.: Is 2D Heatmap Representation Even Necessary for Human Pose Estimation? (2021)

[43] Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Mathematical programming **45**(1), 503–528 (1989)

[44] Liu, F., Zhu, R., Zeng, D., Zhao, Q., Liu, X.: Disentangling features in 3D face shapes for joint face reconstruction and recognition. In: CVPR (2018)

[45] Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense Face Alignment. In: ICCV Workshops (2017)

[46] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019)

[47] Morales, A., Piella, G., Sukno, F.M.: Survey on 3d face reconstruction from uncalibrated images. Computer Science Review **40**, 100400 (2021)

[48] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)

[49] Piotraschke, M., Blanz, V.: Automated 3D face reconstruction from multiple images using quality measures. In: CVPR (2016)

[50] Popa, T., South-Dickinson, I., Bradley, D., Sheffer, A., Heidrich, W.: Globally Consistent Space-Time Reconstruction. Comput. Graph. Forum (2010)

[51] Richardson, E., Sela, M., Kimmel, R.: 3D face reconstruction by learning from synthetic data. In: 3DV. IEEE (2016)

[52] Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: CVPR (2017)

[53] Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces In-the-wild challenge: Database and results. Image and Vision Computing (IMAVIS) (2016)

[54] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobilenetV2: Inverted residuals and linear bottlenecks. In: CVPR (2018)

[55] Sanyal, S., Bolkart, T., Feng, H., Black, M.: Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In: CVPR (2019)

[56] Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. CVPR (2006)

[57] Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: ICCV (2017)

[58] Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In: ECCV (2020)

[59] Taylor, J., Bordeaux, L., Cashman, T., Corish, B., Keskin, C., Sharp, T., Soto, E., Sweeney, D., Valentin, J., Luff, B., et al.: Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. ACM Transactions on Graphics (ToG) (2016)

[60] Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: CVPR (2012)

[61] Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: FML: Face Model Learning from Videos. In: CVPR (2019)

[62] Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised m¡ulti-level Face Model Learning for Monocular Reconstruction at over 250 Hz. In: CVPR (2018)

[63] Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: ICCV Workshops (2017)

[64] Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., Theobalt, C.: Real-time expression transfer for facial reenactment. ACM Trans. Graph. (oct 2015)

[65] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In: CVPR (2016)

[66] Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3D face morphable model. In: CVPR (2019)

[67] Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: CVPR (2018)

[68] Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: CVPR (2017)

[69] Wang, X., Bo, L., Fuxin, L.: Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression. In: ICCV (2019)

[70] Wightman, R.: Pytorch image models. `https://github.com/rwightman/pytorch-image-models` (2019). https://doi.org/10.5281/zenodo.4414861

[71] Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Johnson, M., Estellers, V., Cashman, T.J., Shotton, J.: Fake It Till You Make It: Face analysis in the wild using synthetic data alone (2021)

[72] Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at Boundary: A Boundary-Aware Face Alignment Algorithm. In: CVPR (2018)

[73] Yi, H., Li, C., Cao, Q., Shen, X., Li, S., Wang, G., Tai, Y.W.: MMFace: A Multi-metric Regression Network for Unconstrained Face Reconstruction. In: CVPR (2019)

[74] Yoon, J.S., Shiratori, T., Yu, S.I., Park, H.S.: Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In: CVPR (2019)

[75] Zhou, Y., Deng, J., Kotsia, I., Zafeiriou, S.: Dense 3D Face Decoding over 2500FPS: Joint Texture & Shape Convolutional Mesh Decoders. In: CVPR (2019)

[76] Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust Facial Landmark Detection via Occlusion-Adaptive Deep Networks. In: CVPR (2019)

[77] Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face Alignment Across Large Poses: A 3D Solution. In: CVPR (2016)

[78] Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)

[79] Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. Computer Graphics Forum **37**(2) (2018)