How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)

Adrian Bulat and Georgios Tzimiropoulos
Computer Vision Laboratory, The University of Nottingham
Nottingham, United Kingdom
{adrian.bulat, yorgos.tzimiropoulos}@nottingham.ac.uk

Abstract

This paper investigates how far a very deep neural network is from attaining close to saturating performance on existing 2D and 3D face alignment datasets. To this end, we make the following 5 contributions: (a) we construct, for the first time, a very strong baseline by combining a state-of-the-art architecture for landmark localization with a state-of-the-art residual block, train it on a very large yet synthetically expanded 2D facial landmark dataset and finally evaluate it on all other 2D facial landmark datasets. (b) We create a guided by 2D landmarks network which converts 2D landmark annotations to 3D and unifies all existing datasets, leading to the creation of LS3D-W, the largest and most challenging 3D facial landmark dataset to date (~230,000 images). (c) Following that, we train a neural network for 3D face alignment and evaluate it on the newly introduced LS3D-W. (d) We further look into the effect of all "traditional" factors affecting face alignment performance like large pose, initialization and resolution, and introduce a "new" one, namely the size of the network. (e) We show that both 2D and 3D face alignment networks achieve performance of remarkable accuracy which is probably close to saturating the datasets used. Training and testing code as well as the dataset can be downloaded from `https://www.adrianbulat.com/face-alignment/`

1. Introduction

With the advent of Deep Learning and the development of large annotated datasets, recent work has shown results of unprecedented accuracy even on the most challenging computer vision tasks. In this work, we focus on landmark localization, in particular, on facial landmark localization, also known as face alignment, arguably one of the most heavily researched topics in computer vision over the last decades. Very recent work on landmark localization using Convolutional Neural Networks (CNNs) has pushed the boundaries in other domains like human pose estimation [39, 38, 24, 17, 27, 42, 23, 5], yet it remains unclear what has been achieved so far for the case of face alignment. The aim of this work is to address this gap in literature.

Historically, different techniques have been used for landmark localization depending on the task in hand. For example, work in human pose estimation, prior to the advent of neural networks, was primarily based on pictorial structures [12] and sophisticated extensions [44, 25, 36, 32, 26] due to their ability to model large appearance changes and accommodate a wide spectrum of human poses. Such methods though have not been shown capable of achieving the high degree of accuracy exhibited by cascaded regression methods for the task of face alignment [11, 8, 43, 50, 41]. On the other hand, the performance of cascaded regression methods is known to deteriorate for cases of inaccurate initialisation, and large (and unfamiliar) facial poses when there is a significant number of self-occluded landmarks or large in-plane rotations.

More recently, fully Convolutional Neural Network architectures based on heatmap regression have revolutionized human pose estimation [39, 38, 24, 17, 27, 42, 23, 5] producing results of remarkable accuracy even for the most challenging datasets [1]. Thanks to their end-to-end training and little need for hand engineering, such methods can be readily applied to the problem of face alignment. Following this path, our main contribution is to construct and train such a powerful network for face alignment and investigate for the first time how far it is from attaining close to saturating performance on all existing 2D face alignment datasets and a newly introduced large scale 3D dataset. More specifically, **our contributions** are:

1. We construct, for the first time, a very strong baseline by combining a state-of-the-art architecture for landmark localization with a state-of-the-art residual block and train it on a very large yet synthetically expanded 2D facial landmark dataset. Then, we evaluate it on *all other* 2D datasets (~230,000 images), investigating how

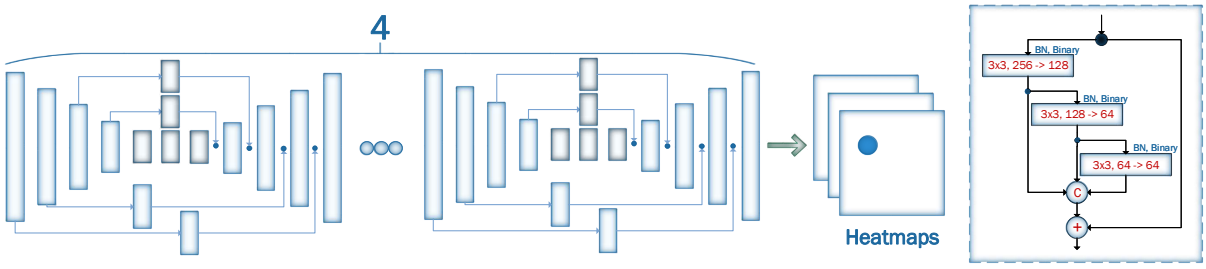


Figure 1: The Face Alignment Network (FAN) constructed by stacking four HGs in which all bottleneck blocks (depicted as rectangles) were replaced with the hierarchical, parallel and multi-scale block proposed of [7].

far are we from solving 2D face alignment.

2. In order to overcome the scarcity of 3D face alignment datasets, we further propose a guided-by-2D landmarks CNN which converts 2D annotations to 3D [1] and use it to create LS3D-W, the largest and most challenging 3D facial landmark dataset to date (~230,000 images), obtained from unifying almost all existing datasets to date.
3. Following that, we train a 3D face alignment network and then evaluate it on the newly introduced large scale 3D facial landmark dataset, investigating how far are we from solving 3D face alignment.
4. We further look into the effect of all "traditional" factors affecting face alignment performance like large poses, initialization and resolution, and introduce a "new" one, namely the size of the network.
5. We show that both 2D and 3D face alignment networks achieve performance of remarkable accuracy which is probably close to saturating the datasets used.

2. Closely related work

This Section reviews related work on face alignment and landmark localization. Datasets are described in detail in the next Section.

2D face alignment. Prior to the advent of Deep Learning, methods based on cascaded regression had emerged as the state-of-the-art in 2D face alignment, see for example [8, 43, 50, 41]. Such methods are now considered to have largely "solved" the 2D face alignment problem for faces with controlled pose variation like the ones of LFPW [2], Helen [22] and 300-W [30].

We will keep the main result from these works, namely their performance on the frontal dataset of LFPW [2]. This performance will be used as a measure of comparison of how well the methods described in this paper perform assuming that a method achieving a similar error curve on a different dataset is close to saturating that dataset.

CNNs for face alignment. By no means we are the first to use CNNs for face alignment. The method of [35] uses a CNN cascade to regress the facial landmark locations. The work in [47] proposes multi-task learning for joint facial landmark localization and attribute classification. More recently, the method of [40] extends [43] within recurrent neural networks. All these methods have been mainly shown effective for the near-frontal faces of 300-W [30].

Recent works on large pose and 3D face alignment includes [20, 50] which perform face alignment by fitting a 3D Morphable Model (3DMM) to a 2D facial image. The work in [20] proposes to fit a dense 3DMM using a cascade of CNNs. The approach of [50] fits a 3DMM in an iterative manner through a single CNN which is augmented by additional input channels (besides RGB) representing shape features at each iteration. More recent works that are closer to the methods presented in this paper are [4] and [6]. Nevertheless, [4] is evaluated on [20] which is a relatively small dataset (3900 images for training and 1200 for testing) and [6] on [19] which is of moderate size (16,2000 images for training and 4,900 for testing), includes mainly images collected in the lab and does not cover the full spectrum of facial poses. Hence, the results of [4] and [6] are not conclusive in regards to the main questions posed in our paper.

Landmark localization. A detailed review of state-of-the-art methods on landmark localization for human pose estimation is beyond the scope of this work, please see [39, 38, 24, 17, 27, 42, 23, 5]. For the needs of this work, we built a powerful CNN for 2D and 3D face alignment based on two components: (a) the state-of-the-art HourGlass (HG) network of [23], and (b) the hierarchical, parallel & multi-scale block recently proposed in [7]. In particular, we replaced the bottleneck block [15] used in [23] with the block proposed in [7].

Transferring landmark annotations. There are a few works that have attempted to unify facial alignment datasets by transferring landmark annotations, typically through exploiting common landmarks across datasets [49, 34, 46]. Such methods have been primarily shown to be successful when landmarks are transferred from more challenging to less challenging images, for example in [49] the target

[1] The 3D annotations are actually the 2D projections of the 3D facial landmarks but for simplicity we will just call them 3D. In the supplementary material, we present a method for extending them to full 3D.

dataset is LFW [16] or [34] provides annotations only for the relatively easy images of AFLW [21]. Hence, the community primarily relies on the unification performed manually by the 300-W challenge [29] which contains less than 5,000 near frontal images annotated from a 2D perspective.

Using 300-W-LP [50] as a basis, this paper presents the first attempt to provide 3D annotations for *all other* datasets, namely AFLW-2000 [50] (2,000 images), 300-W test set [28] (600 images), 300-VW [33] (218,595 frames), and Menpo training set (9,000 images). To this end, we propose a guided-by-2D landmarks CNN which converts 2D annotations to 3D and unifies all aforementioned datasets.

3. Datasets

In this Section, we provide a description of how existing 2D and 3D datasets were used for training and testing for the purposes of our experiments. We note that the 3D annotations preserve correspondence across pose as opposed to the 2D ones and, in general, they should be preferred. We emphasize that the 3D annotations are actually the 2D projections of the 3D facial landmark coordinates but for simplicity we will just call them 3D. In the supplementary material, we present a method for extending these annotations to full 3D. Finally, we emphasize that we performed cross-database experiments only.

Dataset	Size	pose	annot.	synt.
300-W	4,000	$[-45^o, 45^o]$	2D	No
300W-LP-2D	61,225	$[-90^o, 90^o]$	2D	Yes
300W-LP-3D	61,225	$[-90^o, 90^o]$	3D	Yes
AFLW2000-3D	2,000	$[-90^o, 90^o]$	3D	No
300-VW	218,595	$[-45^o, 45^o]$	2D	No
LS3D-W (ours)	230,000	$[-90^o, 90^o]$	3D	No

Table 1: Summary of the most popular face alignment datasets and their main characteristics.

3.1. Training datasets

For training and validation, we used 300-W-LP [50], a synthetically expanded version of 300-W [29]. 300-W-LP provides both 2D and 3D landmarks allowing for training models and conducting experiments using both types of annotations. For some 2D experiments, we also used the original 300-W dataset [29] for fine tuning, only. This is because the 2D landmarks of 300-W-LP are not entirely compatible with the 2D landmarks of the test sets used in our experiments (i.e. 300-W test set, [28], 300-VW [33] and Menpo [45]), but the original annotations from 300-W are.

300-W. 300-W [29] is currently the most widely-used *in-the-wild* dataset for 2D face alignment. The dataset itself is a concatenation of a series of smaller datasets: LFPW [3], HELEN [22], AFW [51] and iBUG [30], where each image was re-annotated in a consistent manner using the 68 2D landmark configuration of Multi-PIE [13]. The dataset contains in total ~4,000 near frontal facial images.

300W-LP-2D and 300W-LP-3D. 300-W-LP is a synthetically generated dataset obtained by rendering the faces of 300-W into larger poses, ranging from -90^0 to 90^0, using the profiling method of [50]. The dataset contains 61,225 images providing both 2D (300W-LP-2D) and 3D landmark annotations (300W-LP-3D).

3.2. Test datasets

This Section describes the test sets used for our 2D and 3D experiments. Observe that there is a large number of 2D datasets/annotations which are however problematic for moderately large poses (2D landmarks lose correspondence) and that the only in-the-wild 3D test set is AFLW2000-3D [50] [2]. We address this significant gap in 3D face alignment datasets in Section 6.

3.2.1 2D datasets

300-W test set. The 300-W test set consists of the 600 images used for the evaluation purposes of the 300-W Challenge [28]. The images are split in two categories: *Indoor* and *Outdoor*. All images were annotated with the same 68 2D landmarks as the ones used in the 300-W data set.

300-VW. 300-VW[33] is a large-scale face tracking dataset, containing 114 videos and in total 218,595 frames. From the total of 114 videos, 64 are used for testing and 50 for training. The test videos are further separated into three categories (A, B, and C) with the last one being the most challenging. It is worth noting that some videos (especially from category C) contain very low resolution/poor quality faces. Due to the semi-automatic annotation approach (see [33] for more details), in some cases, the annotations for these videos are not so accurate (see Fig. 3). Another source of annotation error is caused by facial pose, i.e. large poses are also not accurately annotated (see Fig. 3).

Menpo. Menpo is a recently introduced dataset [45] containing landmark annotations for about 9,000 faces from FDDB [18] and ALFW. Frontal faces were annotated in terms of 68 landmarks using the same annotation policy as the one of 300-W but profile faces in terms of 39 different landmarks which are not in correspondence with the landmarks from the 68-point mark-up.

3.2.2 3D datasets

AFLW2000-3D. AFLW2000-3D [50] is a dataset constructed by re-annotating the first 2000 images from AFLW [21] using 68 3D landmarks in a consistent manner with the

[2] The data from [19] includes mainly images collected in the lab and do not cover the full spectrum of facial poses.

ones from 300W-LP-3D. The faces of this dataset contain large-pose variations (yaw from -90^o to 90^o), with various expressions and illumination conditions. However, some annotations, especially for larger poses or occluded faces are not so accurate (see Fig. 6).

3.3. Metrics

Traditionally, the metric used for face alignment is the point-to-point Euclidean distance normalized by the interocular distance [10, 29, 33]. However, as noted in [51], this error metric is biased for profile faces for which the interocular distance can be very small. Hence, we normalize by the bounding box size. In particular, we used the Normalized Mean Error defined as:

$$\text{NME} = \frac{1}{N}\sum_{k=1}^{N}\frac{\|\mathbf{x}_k - \mathbf{y}_k\|_2}{d}, \tag{1}$$

where $\mathbf{x}$ denotes the ground truth landmarks for a given face, $\mathbf{y}$ the corresponding prediction and d is the square-root of the ground truth bounding box, computed as $d = \sqrt{w_{bbox} * h_{bbox}}$. Although we conducted both 2D and 3D experiments, we opted to use the same bounding box definition for both experiments; in particular we used the bounding box calculated from the 2D landmarks. This way, we can readily compare the accuracy achieved in 2D and 3D.

4. Method

This Section describes FAN, the network used for 2D and 3D face alignment. It also describes 2D-to-3D FAN, the network used for constructing the very large scale 3D face alignment dataset (LS3D-W) containing more than 230,000 3D landmark annotations.

4.1. 2D and 3D Face Alignment Networks

We coin the network used for our experiments simply Face Alignment Network (FAN). To our knowledge, it is the first time that such a powerful network is trained and evaluated for large scale 2D/3D face alignment experiments.

We construct FAN based on one of the state-of-the-art architectures for human pose estimation, namely the *Hour-Glass* (HG) network of [23]. In particularly, we used a stack of four HG networks (see Fig. 1). While [23] uses the bottleneck block of [14] as the main building block for the HG, we go one step further and replace the bottleneck block with the recently introduced hierarchical, parallel and multi-scale block of [7]. As it was shown in [7], this block outperforms the original bottleneck of [14] when the same number of network parameter were used. Finally, we used 300W-LP-2D and 300W-LP-3D to train 2D-FAN and 3D-FAN, respectively.

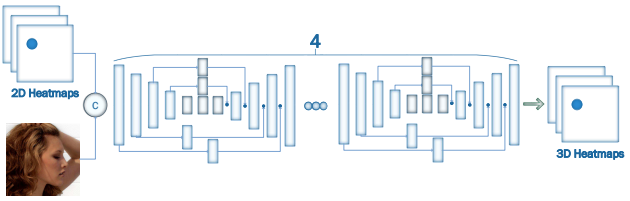


Figure 2: The 2D-to-3D-FAN network used for the creation of the LS3D-W dataset. The network takes as input the RGB image and the 2D landmarks and outputs the corresponding 2D projections of the 3D landmarks.

4.2. 2D-to-3D Face Alignment Network

Our aim is to create the very first very large scale dataset of 3D facial landmarks for which annotations are scarce. To this end, we followed a guided-based approach in which a FAN for predicting 3D landmarks is guided by 2D landmarks. In particular, we created a 3D-FAN in which the input RGB channels have been augmented with 68 additional channels, one for each 2D landmark, containing a 2D Gaussian with std = 1px centered at each landmark's location. We call this network 2D-to-3D FAN. Given the 2D facial landmarks for an image, 2D-to-3D FAN converts them to 3D. To train 2D-to-3D FAN, we used 300-W-LP which provides both 2D and 3D annotations for the same image. We emphasize again that the 3D annotations are actually the 2D projections of the 3D coordinates but for simplicity we call them 3D. Please see supplementary material for extending these annotations to full 3D.

4.3. Training

For all of our experiments, we independently trained three distinct networks: 2D-FAN, 3D-FAN, and 2D-to-3D-FAN. For the first two networks, we set the initial learning rate to 10^{-4} and used a minibatch of 10. During the process, we dropped the learning rate to 10^{-5} after 15 epochs and to 10^{-6} after another 15, training for a total of 40 epochs. We also applied random augmentation: flipping, rotation (from -50^o to 50^o), color jittering, scale noise (from 0.8 to 1.2) and random occlusion. The 2D-to-3D-FAN model was trained by following a similar procedure increasing the amount of augmentation even further: rotation (from -70^o to 70^o) and scale (from 0.7 to 1.3). Additionally, the learning rate initially was set to 10^{-3}. All networks were implemented in Torch7 [9] and trained using rmsprop [37].

5. 2D face alignment

This Section evaluates 2D-FAN (trained on 300-W-LP-2D), on 300-W test set, 300-VW (both training and test sets), and Menpo (frontal subset). Overall, 2D-FAN is evaluated on more than 220,000 images. Prior to reporting our results, the following points need to be emphasized:

1. 300-W-LP-2D contains a wide range of poses (yaw angles in $[-90°, 90°]$, yet it is still a synthetically generated dataset as this wide spectrum of poses were produced by warping the nearly frontal images of the 300-W dataset. It is evident that this lack of real data largely increases the difficulty of the experiment.
2. The 2D landmarks of 300-W-LP-2D that 2D-FAN was trained on are slightly different from the 2D landmarks of the 300-W test set, 300-VW and Menpo. To alleviate this, the 2D-FAN was further fine-tuned on the original 300-W training set for a few epochs. Although this seems to resolve the issue, this discrepancy obviously increases the difficulty of the experiment.
3. We compare the performance of 2D-FAN on all the aforementioned datasets with that of an unconventional baseline: the performance of a recent state-of-the-art method, namely MDM [40] on LFPW test set, initialized with the ground truth bounding boxes. We call this result MDM-on-LFPW. As there is very little performance progress made on the frontal dataset of LFPW over the past years, we assume that a state-of-the-art method like MDM (nearly) saturates it. Hence, we use the produced error curve to compare how well our method does on the much more challenging aforementioned test sets.

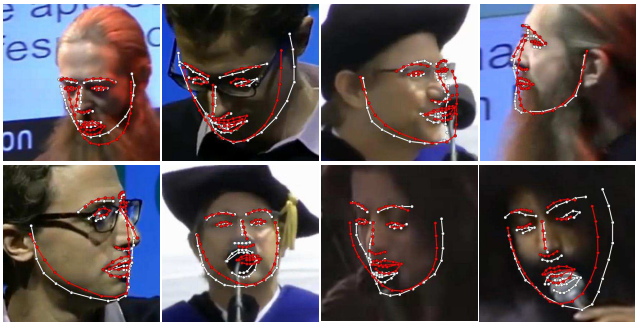


Figure 3: Fittings with the highest error from 300-VW (NME 6.8-7%). Red: ground truth. White: our predictions. In most cases, our predictions are more accurate than the ground truth.

The cumulative error curves for our 2D experiments on 300-VW, 300-W test set and Menpo are shown in Fig. 8. We additionally report the performance of MDM on all datasets initialized by ground truth bounding boxes, ICCR, the state-of-the-art face tracker of [31], on 300-VW (the only tracking dataset), and our unconventional baseline (called MDM-on-LFPW). Comparison with a number of methods in terms of AUC are also provided in Table 2.

With the exception of Category C of 300-VW, it is evident that 2D-FAN achieves literally the same performance on all datasets, outperforming MDM and ICCR, and, notably, matching the performance of MDM-on-LFPW. Out of 7,200 images (from Menpo and 300-W test set), there are in total only 18 failure cases, which represent 0.25% of the images (we consider a failure a fitting with NME $> 7\%$). After removing these cases, the 8 fittings with the highest error for each dataset are shown in Fig. 4.

Figure 4: Fittings with the highest error from 300-W test set (first row) and Menpo (second row) (NME 6.5-7%). Red: ground truth. White: our predictions. In most cases, our predictions are more accurate than the ground truth.

Regarding the Category C of 300-VW, we found that the main reason for this performance drop is the quality of the annotations which were obtained in a semi-automatic manner. After removing all failure cases (101 frames representing 0.38% of the total number of frames), Fig. 3 shows the quality of our predictions vs the ground truth landmarks for the 8 fittings with the highest error for this dataset. It is evident that in most cases our predictions are more accurate.
Conclusion: Given that 2D-FAN matches the performance of MDM-on-LFPW, we conclude that 2D-FAN achieves near saturating performance on the above 2D datasets. Notably, this result was obtained by training 2D-FAN primarily on synthetic data, and there was a mismatch between training and testing landmark annotations.

6. Large Scale 3D Faces in-the-Wild dataset

Motivated by the scarcity of 3D face alignment annotations and the remarkable performance of 2D-FAN, we opted to create a large scale 3D face alignment dataset by converting all existing 2D face alignment annotations to 3D. To this end, we trained a 2D-to-3D FAN as described in Subsection 4.2 and guided it using the predictions of 2D-FAN, creating 3D landmarks for: 300-W test set, 300-VW (both training and all 3 testing datasets), Menpo (the whole dataset).

Evaluating 2D-to-3D is difficult: the only available 3D face alignment in-the-wild dataset (not used for training) is AFLW2000-3D [50]. Hence, we applied our pipeline (consisting of applying 2D-FAN for producing the 2D landmarks and then 2D-to-3D FAN for converting them to 3D) on AFLW2000-3D and then calculated the error, shown in Fig. 5 (note that for normalization purposes, 2D bounding

Dataset	**2D-FAN(Ours)**	MDM[40]	iCCR[31]	TCDCN[47]	CFSS[48]
300VW-A	**72.1%**	70.2 %	65.9%	-	-
300VW-B	**71.2%**	67.9 %	65.5%	-	-
300VW-C	**64.1%**	54.6%	58.1%	-	-
Menpo	**67.5%**	67.1%	-	47.9%	60.5%
300W	**66.9%**	58.1%	-	41.7%	55.9%

Table 2: AUC (calculated for a threshold of 7%) on all major 2D face alignment datasets. MDM, CFSS and TCDCN were evaluated using ground truth bounding boxes and the openly available code.

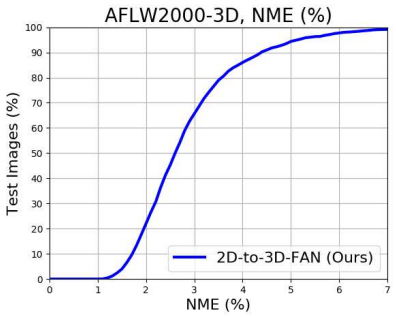


Figure 5: NME on AFLW2000-3D, between the original annotations of [50] and the ones generated by 2D-to-3D-FAN. The error is mainly introduced by the automatic annotation process of [50]. See Fig. 6 for visual examples.

box annotations are still used). The results show that there is discrepancy between our 3D landmarks and the ones provided by [50]. After removing a few failure cases (19 in total, which represent 0.9% of the data), Fig. 6 shows 8 images with the highest error between our 3D landmarks and the ones of [50]. It is evident, that this discrepancy is mainly caused from the semi-automatic annotation pipeline of [50] which does not produce accurate landmarks especially for images with difficult poses.

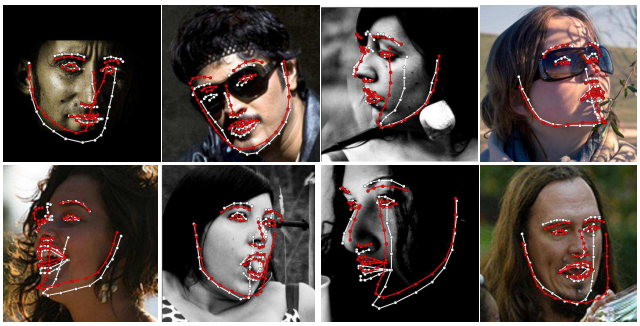


Figure 6: Fittings with the highest error from AFLW2000-3D (NME 7-8%). Red: ground truth from [50]. White: predictions of 2D-to-3D-FAN. In most cases, our predictions are more accurate than the ground truth.

By additionally including AFLW2000-3D into the aforementioned datasets, overall, ~230,000 images were annotated in terms of 3D landmarks leading to the creation of the Large Scale 3D Faces in-the-Wild dataset (LS3D-W), the largest 3D face alignment dataset to date.

7. 3D face alignment

This Section evaluates 3D-FAN trained on 300-W-LP-3D, on LS3D-W (described in the previous Section) i.e. on the 3D landmarks of the 300-W test set, 300-VW (both training and test sets), and Menpo (the whole dataset) and AFLW2000-3D (re-annotated). Overall, 3D-FAN is evaluated on ~230,000 images. Note that compared to the 2D experiments reported in Section 5, more images in large poses have been used as our 3D experiments also include AFLW2000-3D and the profile images of Menpo (~2000 more images in total).

The results of our 3D face alignment experiments on 300-W test set, 300-VW, Menpo and AFLW2000-3D are shown in Fig. 9. We additionally report the performance of the state-of-the-art method of 3DDFA (trained on the same dataset as 3D-FAN) on all datasets.

Conclusion: 3D-FAN essentially produces the same accuracy on all datasets largely outperforming 3DDFA. This accuracy is slightly increased compared to the one achieved by 2D-FAN, especially for the part of the error curve for which the error is less than 2% something which is not surprising as now the training and testing datasets are annotated using the same mark-up.

8. Ablation studies

To further investigate the performance of 3D-FAN under challenging conditions, we firstly created a dataset of 7,200 images from LS3D-W so that there is an equal number of images in yaw angles $[0^o - 30^o]$, $[30^o - 60^o]$ and $[60^o - 90^o]$. We call this dataset LS3D-W Balanced. Then, we conducted the following experiments:

Performance across pose. We report the performance of 3D-FAN on LS3D-W Balanced for each pose separately in terms of the Area Under the Curve (AUC) (calculated for a threshold of 7%) in Table 3. We observe only a slight degradation of performance for very large poses ($[60^o - 90^o]$). We believe that this is to some extent to be expected as 3D-FAN was largely trained with synthetic data

Yaw	#images	3D-FAN (Ours)
$[0^o - 30^o]$	2400	73.5%
$[30^o - 60^o]$	2400	74.6%
$[60^o - 90^o]$	2400	68.8%

Table 3: AUC (calculated for a threshold of 7%) on the LS3D-W Balanced for different yaw angles.

for these poses (300-W-LP-3D). This data was produced by warping frontal images (i.e. the ones of 300-W) to very large poses which causes face distortion especially for the face region close to the ears.
Conclusion: Facial pose is not a major issue for 3D-FAN.
Performance across resolution. We repeated the previous

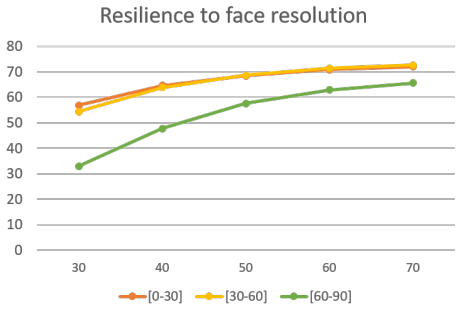


Figure 7: AUC on the LS3D-W Balanced for different face resolutions. Up to 30px, performance remains high.

experiment but for different face resolutions (resolution is reduced relative to the face size defined by the tight bounding box) and report the performance of 3D-FAN in terms of AUC in Fig. 7. Note that we did not retrain 3D-FAN to particularly work for such low resolutions. We observe significant performance drop for all poses only when the face size is as low as 30 pixels.
Conclusion: Resolution is not a major issue for 3D-FAN.

Noise	$[0^o - 30^o]$	$[30^o - 60^o]$	$[60^o - 90^o]$
0%	74.5%	75.2%	69.8%
10%	73.5%	74.6%	68.8%
20%	70.8%	71.7%	66.1%
30%	63.8%	63.5%	57.2%

Table 4: AUC on the LS3D-W Balanced for different levels of initialization noise. The network was trained with a noise level of up to 20%.

Performance across noisy initializations. For all reported results so far, we used 10% of noise added to the ground truth bounding boxes. Note that 3D-FAN was trained with noise level of 20% percent. Herein, we repeated the previous experiment but for different noise levels and report the performance of 3D-FAN in terms of AUC in Table 4. We observe only small performance decrease for noise level equal to 30% which is greater than the level of noise that the network was trained with.
Conclusion: Initialization is not a major issue for 3D-FAN.

#params	$[0^o - 30^o]$	$[30^o - 60^o]$	$[60^o - 90^o]$
2M	70.9%	69.9%	55.8%
4M	71.0%	70.5%	57.0%
6M	71.5%	71.1%	58.3%
12M	72.7%	72.7%	67.1%
18M	73.4%	74.2%	68.3%
24M	73.5%	74.6%	68.8%

Table 5: AUC on the LS3D-W Balanced for various network sizes. Between 12-24M parameters, performance remains almost the same.

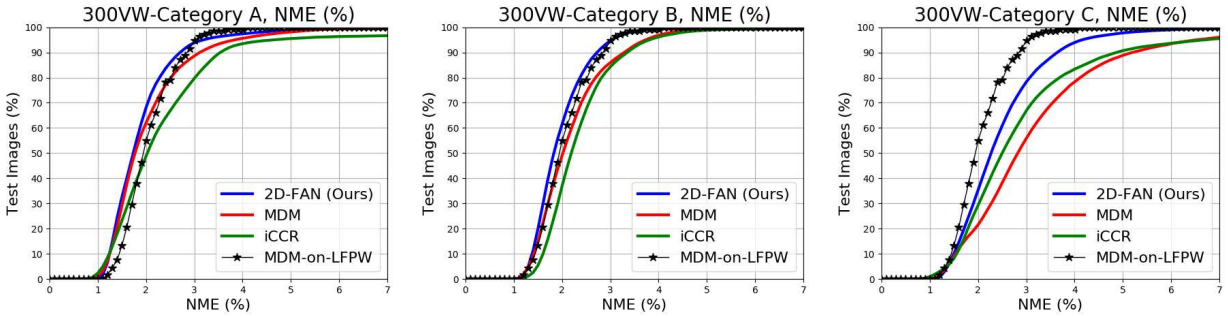
Performance across different network sizes. For all reported results so far, we used a very powerful 3D-FAN with 24M parameters. Herein, we repeated the previous experiment varying the number of network parameters and report the performance of 3D-FAN in terms of AUC in Table 5. The number of parameters is varied by firstly reducing the number of HG networks used from 4 to 1. Then, the number of parameters was dropped further by reducing the number of channels inside the building block. It is important to note that even then biggest network is able to run on 28-30 fps on a TitanX GPU while the smallest one can reach 150 fps. We observe that up to 12M, there is only a small performance drop and that the network's performance starts to drop significantly only when the number of parameters becomes as low as 6M.
Conclusion: There is a moderate performance drop vs the number of parameters of 3D-FAN. We believe that this is an interesting direction for future work.

9. Conclusions

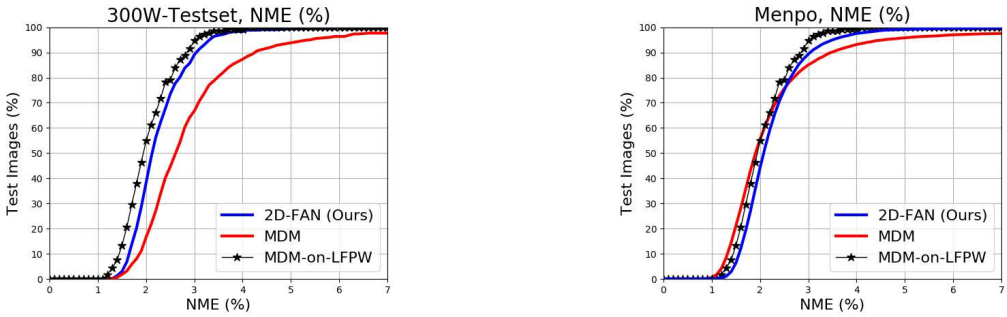
We constructed a state-of-the-art neural network for landmark localization, trained it for 2D and 3D face alignment, and evaluate it on hundreds of thousands of images. Our result show that our network nearly saturates these datasets, showing also remarkable resilience to pose, resolution, initialization, and even to the number of the network parameters used. Although some very unfamiliar poses were not explored in these datasets, there is no reason to believe, that given sufficient data, the network does not have the learning capacity to accommodate them, too.

10. Acknowledgments

Adrian Bulat was funded by a PhD scholarship from the University of Nottingham. This work was supported in part by the EPSRC project EP/M02153X/1 Facial Deformable Models of Animals.

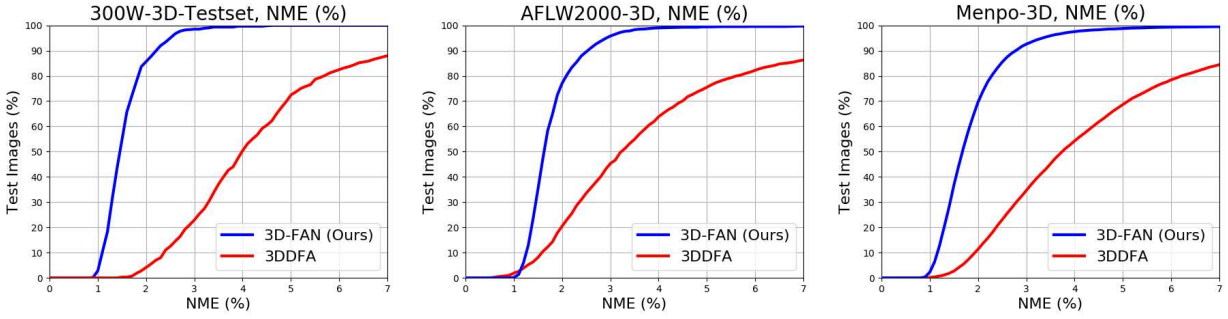


(a) 300VW-Category A (b) 300VW-Category B (c) 300VW-Category C

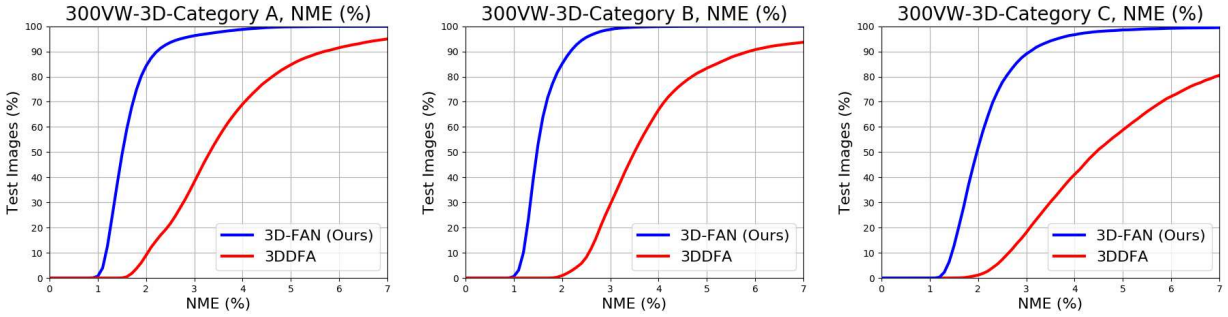


(d) 300-W Testset (Indoor and Outdoor subset). (e) Menpo (on faces annotated with 68 points).

Figure 8: 2D face alignment experiments: NME (all 68 points used) on 300-VW (a-c), 300-W Testset (d) and Menpo (e). Our model is called 2D-FAN. MDM is initialized with ground truth bounding boxes. **Note: MDM-on-LFPW is not a method but the curved produced by running MDM on LFPW test set, initialized with the ground truth bounding boxes.**



(a) 300-W-3D Test set (b) AFLW2000-3D, re-annotated using 2D-to-3D FAN. (c) Menpo-3D



(d) 300VW-3D Category A (e) 300VW-3D Category B (f) 300VW-3D Category C

Figure 9: 3D face alignment experiments: NME (all 68 points used) on the newly introduced LS3D-W dataset.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[2] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *TPAMI*, 2013.

[4] A. Bulat and G. Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016.

[5] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016.

[6] A. Bulat and G. Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *ECCV*, 2016.

[7] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *arXiv*, 2017.

[8] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[9] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *NIPS-W*, 2011.

[10] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[11] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, 2010.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *IVC*, 2010.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, 2016.

[16] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007.

[17] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *arXiv*, 2016.

[18] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*, 2010.

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

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

[21] M. Köstinger, 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 *ICCV-W*, 2011.

[22] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012.

[23] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[24] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015.

[25] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013.

[26] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *CVPR*, 2013.

[27] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. V. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.

[28] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *IVC*, 47:3–18, 2016.

[29] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *CVPR*, 2013.

[30] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR*, 2013.

[31] E. Sánchez-Lozano, B. Martinez, G. Tzimiropoulos, and M. Valstar. Cascaded continuous regression for real-time incremental face tracking. In *ECCV*, 2016.

[32] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.

[33] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCVW*, 2015.

[34] B. M. Smith and L. Zhang. Collaborative facial landmark localization for transferring annotations across datasets. In *ECCV*, 2014.

[35] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[36] Y. Tian, C. L. Zitnick, and S. G. Narasimhan. Exploring the spatial hierarchy of mixture models for human pose estimation. In *ECCV*. 2012.

[37] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.

[38] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[39] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[40] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *CVPR*, 2016.

[41] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *CVPR*, 2015.

[42] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.

[43] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[44] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[45] S. Zaferiou. The menpo facial landmark localisation challenge. In *CVPR-W*, 2017.

[46] J. Zhang, M. Kan, S. Shan, and X. Chen. Leveraging datasets with varying annotations for face alignment via deep regression network. In *ICCV*, 2015.

[47] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*. 2014.

[48] S. Zhu, C. Li, C. Change Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.

[49] S. Zhu, C. Li, C. C. Loy, and X. Tang. Transferring landmark annotations for cross-dataset face alignment. *arXiv*, 2014.

[50] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016.

[51] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*. IEEE, 2012.