Cascade Multi-view Hourglass Model for Robust 3D Face Alignment

Jiankang Deng[1], Yuxiang Zhou[1], Shiyang Cheng[1], and Stefanos Zaferiou[1,2]
[1] Department of Computing, Imperial College London, UK
[2] Centre for Machine Vision and Signal Analysis, University of Oulu, Finland
{*j.deng16, yuxiang.zhou10, shiyang.cheng11, s.zafeiriou*}*@imperial.ac.uk*

***Abstract*— Estimating the 3D facial landmarks from a 2D image remains a challenging problem. Even though state-of-the-art 2D alignment methods are able to predict accurate landmarks for semi-frontal faces, the majority of them fail to provide semantically consistent landmarks for profile faces. A de facto solution to this problem is through 3D face alignment that preserves correspondence across different poses. In this paper, we proposed a Cascade Multi-view Hourglass Model for 3D face alignment, where the first Hourglass model is explored to jointly predict semi-frontal and profile 2D facial landmarks, after removing spatial transformations, another Hourglass model is employed to estimate the 3D facial shapes. To improve the capacity without sacrificing the computational complexity, the original residual bottleneck block in the Hourglass model is replaced by a parallel, multi-scale inception-resnet block. Extensive experiments on two challenging 3D face alignment datasets, AFLW2000-3D and Menpo-3D, show the robustness of the proposed method under continuous pose changes.**

I. Introduction

Facial landmark localisation [14], [25], [26], [18], [1], [3] and tracking [2], [12], [37], [27] under the unconstrained environment have recently received considerable attention due to the wide applications such as human-computer interaction, video surveillance and entertainment.

The current state-of-the-art 2D face alignment benchmarks [41] revolves around Deep Convolutional Neural Networks(DCNNs) [22], [31], [33], [16] equipped with resolution preserved structure, alleged Hourglass architecture [29], [9], [8], [38], [15]. Even though the performance of 2D face alignment is almost saturated on the public benchmarks [9], 2D facial landmark annotations are not always semantically consistent and hardly preserve the 3D structure of the human face. This is particularly evident for the landmarks on the facial contour under large pose variations. By contrast, 3D annotations preserve correspondence across poses. In this paper, the 3D annotations refer to the 2D projections of the 3D facial landmarks. In [9], 3D facial landmarks projections are extended to full 3D facial landmarks by adding an extra regression network to predict the depth.

3D Face alignment under unconstrained conditions is very challenging as facial appearance can change dramatically due to extreme pose, camera defocus, low resolution and occlusion. Besides, self-occlusion caused by the large pose variations makes the boundary features unreliable. To keep correspondence with the 3D structure of the human face, one might only predict the occluded landmarks by contextual information.

978-1-5386-2335-0/18/$31.00

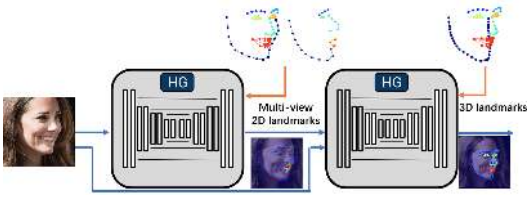


Fig. 1. Cascade Multi-view Hourglass Model for 3D face alignment.

In this paper, we propose a Cascade Multi-view Hourglass Model (CMHM) for 3D face alignment. As illustrated in Fig. 1, two Hourglass models are cascaded with supervision signals from 2D and 3D facial landmarks. We address this alignment problem with the goal of improving the accuracy under large pose variations. More specifically, we have made three contributions:

1) We replace the residual bottleneck block in the Hourglass model by a parallel and multi-scale inception-resnet block, which improves the capacity and keeps the computational complexity of the Hourglass model.
2) We capitalise on the correspondences between the frontal and profile facial shapes and formulate a novel Multi-view Hourglass Model (MHM) which jointly estimates both semi-frontal and profile 2D facial landmarks.
3) We employ a cascade strategy where Multi-view Hourglass Model is first applied to find the 2D facial landmarks. After removing the similarity transformation, another Hourglass model is performed to estimate the 3D facial landmarks.

Based on these improvements, the proposed method achieves state-of-the-art performances on the latest 3D benchmarks, AFLW2000-3D [45] and Menpo-3D [40].

II. Related Works

To better understand the problem of 2D and 3D face alignment, we review some related deep learning based methods.

In [32], [44], [24], DCNNs were employed in 2D face alignment, and the resolution loss within the pooling step was compensated by the image enlargement in a global to local way. Zhang et al. [42] adopted the similar coarse-to-fine framework with auto-encoder networks. Ranjan et al. [30] combined outputs of multi-resolution convolutional layers to predict the landmark locations.

After the presence of the fully-convolutional network (FCN) [24], direct landmark coordinate prediction changed to the landmark response map prediction. Lai et al. [23], Xiao et al. [34] and Bulat et al. [6] employed the convolutional and de-convolutional network to generate the response map for each facial landmark, and added a refinement step by utilising a network that performs regression.

In the area of articulated human pose estimation, Alejandro et al. [29] proposed a novel stacked hourglass model, which repeated bottom-up and top-down processing in conjunction with intermediate supervision and obtained state-of-the-art results. Yang et al. [38] won the Menpo Challenge by improving initialisation and stacking multiple Hourglass models. Deng et al. [15] proposed a joint multi-view Hourglass model for 2D face alignment under large pose variation. Bulat et al. [8] further explored binarized Hourglass-like convolutional network for face alignment with limited resources.

To solve the problem of large pose face alignment, 3D face fitting methodologies have been considered [19], [20], [45], which aims to fit a 3D morphable model (3DMM) [4] to a 2D image. [19] aligned faces of arbitrary poses with the assist of a sparse 3D point distribution model. The model parameter and projection matrix are estimated by the cascaded linear or nonlinear regressors. [20] extended [19] by fitting a dense 3D morphable model, employing the CNN regressor with 3D-enabled features, and estimating contour landmarks. [45] fitted a dense 3D face model to the image via CNN and synthesised large-scale training samples in profile views to solve the problem of data labelling. 3D face alignment methods model the 3D face shape with a linear subspace and achieve fitting by minimising the difference between image and model appearance.

Although 3D alignment methods can cover arbitrary poses, the accuracy of alignment is bounded by the linear parametric 3D model, and the invisible landmarks are predicted based on 3DMM fitting results on the visible appearance. By contrast, Bulat et al. [9], [7] directly utilised stacked Hourglass model trained on large-scale data to predict 3D facial landmarks and obtained state-of-the-art results.

III. Cascade Multi-view Hourglass Model

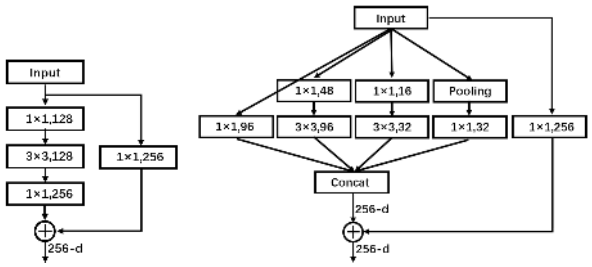
A. Inception-Resnet Hourglass

Hourglass [29] is a symmetric top-down and bottom-up fully convolutional network. The input signals are branched out before each down-sampling step and combined together before each up-sampling step to maintain the resolution information. n scale Hourglass is able to extract features from the original scale to $1/2^n$ scale without resolution loss throughout the whole network. The increasing depth of network design helps to increase contextual region, which incorporates global shape inference and increases robustness when local observation is blurred.

Hourglass [29] is constructed based on Residual blocks [16] (Fig. 2(a)), and could be represented as follows:

$$x_{n+1} = H(x_n) + F(x_n, W_n), \tag{1}$$

where x_n and x_{n+1} are the input and output of the n-th unit, $H(x_n)$ is the identity mapping, W_n is the weight, and F is the stacked convolution, batch normalisation, and ReLU non-linearity. To improve the model capacity and compress the computational complexity of the Hourglass model, we replace the bottleneck block with a parallel and multi-scale inception block, and construct the inception-resnet block [28] as shown in Fig. 2(b).



(a) Resnet block (b) Inception-Resnet block

Fig. 2. Resnet and Inception-Resnet blocks to construct Hourglass model.

B. Multi-view 2D Alignment

Based on the Hourglass model [29], we formulate the Multi-view Hourglass Model (MHM) which tries to jointly estimate both semi-frontal (68 landmarks) and profile (39 landmarks) face shapes. Unlike other methods which employ distinct models, we try to capitalise on the correspondences between the profile and frontal facial shapes. As shown in Fig. 3, for each landmark on the profile face, the nearest landmark on the frontal face is regarded as its corresponding landmark in the union set, thus we can form the union landmark set with 68 landmarks. During the training, we use the view status to select the corresponding response maps for the loss computation.

$$L = \frac{1}{N}\sum_{n=1}^{N}(v_n^* \sum_{ij} \|m_n(i,j) - m_n^*(i,j)\|_2^2), \tag{2}$$

where $m_n(i,j)$ and $m_n^*(i,j)$ represent the estimated and the ground truth response maps at pixel location (i,j) for the n-th landmark correspondingly, and $v_n \in \{0,1\}$ is the indicator to select the corresponding response map to calculate the loss.

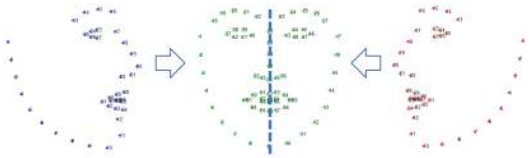


Fig. 3. Correspondence between the profile (39 landmarks) and frontal (68 landmarks) facial shapes. We define a union landmark set with 68 landmarks for Multi-view Hourglass Model.

C. Cascade 2D and 3D Alignment

Our prior knowledge shows that the variance of the regression target and capacity of the regression model are the two main factors affecting the performance of the regression

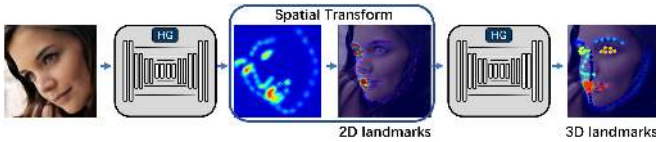


Fig. 4. 2D alignment acts as spatial transform network for 3D alignment.

task. As our 3D alignment task is formulated as a heatmap regression problem, we adopt a cascade framework to predict 3D landmarks with gradually reduced variance. As shown in Fig. 4, we first utilise the 2D multi-view Hourglass model to supervise the removal of the spatial transformation (i.e. translation, scale and rotation). We apply Procrustes analysis between the 2D alignment results and the mean 2D shapes to remove the difference of rigid face transformation. Then, we employ another Hourglass network to predict the 3D facial landmarks. The 3D alignment network only focuses on the non-rigid face transformation, thus the variance of the regression target is decreased. During training, instead of implementing back-propagation on the affine transformation step, we simply stop the gradient from 3D network to 2D network in our implementation.

IV. Experiments

A. Experiment Setting

1) Training Data: **Menpo Benchmark dataset** [41] consists of 5658 semi-frontal and 1906 profile face images, which are selected from FDDB [17] and ALFW [21]. These annotated face images are collected from completely unconstrained conditions, which exhibit large variations in pose, expression, illumination, etc. In this dataset, semi-frontal faces and profile faces are annotated with 68 and 39 2D landmarks respectively. In order to evaluate our method in different 3D datasets, we use two 3D annotation schemes (68-point markup and 84-point markup) for training. In [9], 3D annotations with 68 landmarks are generated by stacking Hourglass model based on the ground-truth 2D landmarks. In [40], 3D annotations with 84 landmarks are generated by applying the state-of-the-art 3DMM fitting algorithm of [5] driven by the ground-truth 2D landmarks.

2) Testing data: Evaluations of 3D face alignment and tracking are performed in two *in-the-wild* databases.

AFLW2000-3D dataset [45] contains 2,000 static face images captured in the wild with large pose variations, severe occlusions and extreme illuminations, with each annotated with 68 3D landmarks. To evaluate the alignment performance under different poses, AFLW2000-3D is divided into three subsets: 1306 samples from 0° to 30°, 462 samples from 30° to 60°, and 232 samples in [60°, 90°].

Menpo-3D Tracking dataset [40] consists of 35 videos (∼35k frames in total), captured in the wild with large pose variations. Each face from these videos is annotated with 84 3D landmarks.

3) Training Setting: The training of the proposed method follows a similar design as the Hourglass Model in [29]. According to the centre and size of bounding box provided by the face detector [43], each face region is cropped and scaled to 256×256. To improve the robustness of our method, we increase the number of training examples by randomly perturbing the ground truth image with a different combination of rotation (+/- 45 degrees), scaling (0.75 - 1.25), and translation (+/- 20 pixels).

The full network starts with a 7×7 convolutional layer with stride 2, followed by a residual module and a round of max pooling to bring the resolution down from 256 to 64, as it could reduce GPU memory usage while preserving alignment accuracy. The network is trained using Tensorflow with an initial learning rate of 10^{-4}, batch size of 8, and 100k learning steps. We drop the learning rate to 10^{-5} after 20 epochs. The Mean Squared Error (MSE) loss is applied to compare the predicted heatmaps to the ground-truth heatmaps. Each training step takes 1.02s on one NVIDIA GTX Titan X (Pascal). During testing, face regions are cropped and resized to 256×256, and it takes 20.76ms to generate the response maps. By contrast, the baseline method, two stack Hourglass model [29], takes 24.42ms to generate the response maps.

B. 3D Face Alignment on Images

We firstly present ablation experiment results on the AFLW2000-3D dataset [45]. The alignment accuracy is evaluated by the Normalised Mean Error (NME), which is the average of landmark errors normalised by the bounding box size [45], [8]. We use the two-stack Hourglass model [29] as the baseline, upon which we incrementally add inception-resnet blocks, joint multi-view 2D supervision, and spatial transformation, and evaluate their performances. We observe from Table IV-B that the overall NME decreases as more modules are incorporated, and eventually drops to 3.08 from 3.78 (equals to a performance gain of 18.5%). The same NME changes clearly hold across different pose ranges, which suggests that our method could improve alignment accuracy for all poses.

TABLE I
3D alignment results on the AFLW2000-3D dataset.

Method	[0°,30°]	[30°,60°]	[60°,90°]	Overall
Baseline	2.68	3.34	5.32	3.78
++ Inception-Resnet	2.41	3.15	4.73	3.43
++ 2D Multi-view	2.38	3.08	4.35	3.27
++ Spatial Transform	2.36	2.80	4.08	3.08
RCPR [10]	4.26	5.96	13.18	7.80
ESR [11]	4.60	6.70	12.67	7.99
SDM [36]	3.67	4.94	9.76	6.12
3DDFA [45]	3.78	4.54	7.93	5.42
3DDFA+SDM [45]	3.43	4.24	7.17	4.94
Bulat et al. [8]	2.47	3.01	4.31	3.26

We compare the proposed CMHM with several 3D face alignment methods, particularly, including two recently proposed 3DDFA [45] and Binarised CNN [8] that exhibit state-of-the-art performance. Our method outperforms 3DDFA by a large margin, decreasing the NME by 37.65%. Compared to the method proposed by Bulat et al. [8], we improve

Fig. 5. Example results of the proposed method on the AFLW2000-3D dataset.

the performance by 5.52%. In Fig. 5, we give some exemplary alignment results, which demonstrate very clear and unambiguous responses even under extreme poses, exaggerate expressions or occlusions. This might well explain the superior performance of our model over other state-of-the-art methods.

C. 3D Face Alignment on Videos

We utilise the test set of the Menpo-3D tracking challenge [40] to evaluate 3D face tracking on videos. We perform a frame-by-frame tracking on the video, specifically, we always initialise the next frame by the previous facial bounding box unless there is a fitting failure, in which case, a face detector [43] would be called to initialise. The fitting failure is judged by the third face region classifier of the face detector in [43]. We follow the same protocol of the Menpo-3D challenge, compare our method with its participants: Xiong et al. [35], Zadeh et al. [39], and Crispell et al. [13].

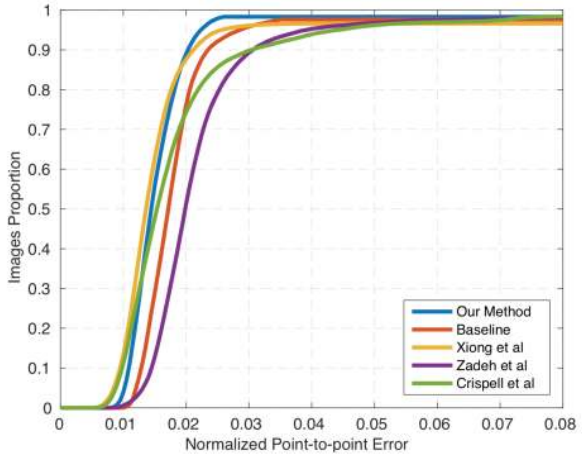


Fig. 6. CED curves on the Menpo-3D tracking test set.

Fig. 6 reports the Cumulative Error Distribution (CED) curves, and Table IV-C reports the Area Under the Curve (AUC) and Failure Rate (FR). The RMS point-to-point error is normalised by the diagonal length of ground truth bounding box [40]. We could observe from the Table IV-C

TABLE II
3D ALIGNMENT RESULTS ON THE MENPO-3D TRACKING TEST SET.

Method	AUC	FR (%)
Our method	0.7977	1.68
Baseline	0.7605	2.35
Xiong et al.	0.7935	3.38
Zadeh et al.	0.7187	1.83
Crispell et al.	0.7617	1.61

Fig. 7. Example results of our method on the Menpo-3D tracking test set.

that CMHM obtains a clear improvement (3.72% in AUC) over the baseline two-stack Hourglass model [29], and it also achieves the best performance (AUC=0.7977, FR=1.68%), which is slightly better than the challenge winner [35], considering that they combined the local heatmap regression and global shape regression. We believe such good performance comes from the robustness of our response maps under large pose variations. This could be proved in Fig. 7, where we select some frames from the Menpo-3D tracking test set and plot their corresponding response maps as well as 3D alignment results. We could easily see that the responses remain clear and evident across different poses.

V. CONCLUSION

In this paper, we proposed a Cascade Multi-view Hourglass Model (CMHM) for 3D face alignment, in which the first Hourglass model is used to jointly predict semi-frontal and profile 2D facial landmarks, after removing similarity transformations, another Hourglass model is used to estimate an accurate 3D facial shape. To improve the capacity without increasing computational complexity of Hourglass model, original residual bottleneck blocks are replaced by a multi-scale and parallel inception-resnet blocks. Extensive experiments on two challenging 3D face alignment datasets, AFLW2000-3D and Menpo-3D, show the accuracy and robustness of the proposed method under continuous pose changes.

VI. ACKNOWLEDGEMENT

J. Deng is supported by the President's Scholarship of Imperial College London. The work of S. Cheng is funded by the EPSRC project EP/N007743/1 (FACER2VM). We thank the NVIDIA Corporation for the GPU donations.

REFERENCES

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pages 3444–3451. IEEE, 2013.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *CVPR*, pages 1859–1866, 2014.

[3] Akshay Asthana, Stefanos Zafeiriou, Georgios Tzimiropoulos, Shiyang Cheng, and Maja Pantic. From pixels to response maps: Discriminative image filtering for face alignment in the wild. *TPAMI*, 37(6):1312–1320, 2015.

[4] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *TPAMI*, 25(9):1063–1074, 2003.

[5] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3d face morphable models" in-the-wild". *CVPR*, 2017.

[6] Adrian Bulat and Georgios Tzimiropoulos. Convolutional aggregation of local evidence for large pose face alignment. In *BMVC*, 2016.

[7] Adrian Bulat and Georgios Tzimiropoulos. Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *ECCV*, pages 616–624. Springer, 2016.

[8] Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. *ICCV*, 2017.

[9] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). *ICCV*, 2017.

[10] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pages 1513–1520, 2013.

[11] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, pages 2887–2894. IEEE, 2012.

[12] Shiyang Cheng, Akshay Asthana, Stefanos Zafeiriou, Jie Shen, and Maja Pantic. Real-time generic face tracking in the wild with cuda. In *ACM MMSys*, pages 148–151, 2014.

[13] Daniel Crispell and Maxim Bazik. Pix2face: Direct 3d face model estimation. In *ICCV Workshop*, 2017.

[14] Jiankang Deng, Qingshan Liu, Jing Yang, and Dacheng Tao. M3 csr: Multi-view, multi-scale and multi-component cascade shape regression. *IVC*, 47:19–26, 2016.

[15] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *arXiv:1708.06023*, 2017.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

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

[18] Deng Jiankang, Yubao Sun, Qingshan Liu, and Hanqing Lu. Low rank driven robust facial landmark regression. *Neurocomputing*, 151:196–206, 2015.

[19] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3d face alignment. In *ICCV*, pages 3694–3702, 2015.

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

[21] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshops*, pages 2144–2151. IEEE, 2011.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[23] Hanjiang Lai, Shengtao Xiao, Yan Pan, Zhen Cui, Jiashi Feng, Chunyan Xu, Jian Yin, and Shuicheng Yan. Deep recurrent regression for facial landmark detection. *CSVT*, 2016.

[24] Zhujin Liang, Shengyong Ding, and Liang Lin. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *arXiv:1507.03409*, 2015.

[25] Qingshan Liu, Jiankang Deng, and Dacheng Tao. Dual sparse constrained cascade regression for robust face alignment. *TIP*, 25(2):700–712, 2016.

[26] Qingshan Liu, Jiankang Deng, Jing Yang, Guangcan Liu, and Dacheng Tao. Adaptive cascade regression model for robust face alignment. *TIP*, 26(2):797–807, 2017.

[27] Qingshan Liu, Jing Yang, Jiankang Deng, and Kaihua Zhang. Robust facial landmark tracking via cascade regression. *PR*, 66:53–62, 2017.

[28] Yu Liu, Duc Minh Nguyen, Nikos Deligiannis, Wenrui Ding, and Adrian Munteanu. Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sensing*, 9(6):522, 2017.

[29] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[30] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. *arXiv:1611.00851*, 2016.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[32] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, pages 3476–3483, 2013.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[34] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, pages 57–72. Springer, 2016.

[35] Pengfei Xiong, G Li, and Y Sun. 3d face tracking via two stage hierarchically attentive shape regression network. In *ICCV Workshop*, 2017.

[36] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[37] Jing Yang, Jiankang Deng, Kaihua Zhang, and Qingshan Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *ICCV Workshops*, pages 41–49, 2015.

[38] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshop*, volume 3, page 6, 2017.

[39] Amir Zadeh and Yun Chuan Lim. Convolutional experts constrained local model for 3d facial landmark detection. In *ICCV Workshop*, 2017.

[40] Stefanos Zafeiriou, Grigorios Chrysos, Anastasios Roussos, Evangelos Ververas, Jiankang Deng, and George Trigeorgis. The 3d menpo facial landmark tracking challenge. In *CVPR Workshop*, 2017.

[41] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *CVPR Workshop*, 2017.

[42] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, pages 1–16. Springer, 2014.

[43] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23(10):1499–1503, 2016.

[44] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *ICCV Workshops*, pages 386–391, 2013.

[45] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, pages 146–155, 2016.