Combining Local and Global Features for 3D Face Tracking

Pengfei Xiong, Guoqing Li, Yuhang Sun
Megvii (face++) Research
{xiongpengfei, liguoqing, sunyuhang}@megvii.com

Abstract

In this paper, we propose to combine local and global features in a carefully designed convolutional neural network for 3D face alignment. We firstly adopt a part heatmap regression network to predict the landmark points on a local granularity by generating a series of heatmaps for each 3D landmark point. To enhance the ability of local feature representation, we incorporate the designed network with a part attention module, which transfers the convolution opereation into a channelwise attention opereation. Additionally, we take all these heatmaps alongside the input image as the input of another shape regression network in order to model the feature representations from local discrete regions to a global semantically continuous space. Extensive experiments on challenging datasets, AFLW2000-3D, 300VW and the Menpo Benchmark, show the effectiveness of both the global consistency and local description in our model, and the proposed algorithm outperforms state-of-the-art baselines.

1. Introduction

The main idea of face alignment is to regress a function that maps 2D face images to their corresponding landmark points. In the past few decades, extensive studies [39, 40, 31, 11, 26, 41, 33, 43] have been proposed and singnificant improvements have been achieved, especially since a comprehensice benchmark [32] was made public, and deep convolution neural networks [33, 43] were applied in face shape regression. Benefiting from the increasing number of training data and better feature representation capacity, deep learning based approachs are proved to be capable of handling with various conditions.

However, most of these prior works struggle when the faces are confronted with large pose variations and partial occlusions. Intuitively, global regression model captures global spatial representations, while posture variation and occlusions lead to local teature changes. Previous methods usually combine the local and global representations with manually designed frameworks, e.g., global model based on local features [39, 40, 31], or coarse-to-fine framework from global regression to local ones [41, 43, 15, 20, 33]. Both the global consistency and multi-granularity representations are damaged in the model.

Figure 1. 3D face tracking results. The first row depicts the detection results of the first apperance of faces which is normalized with 81 2D landmark. Second row is the same image aligned with 84 3D landmark. Third row is the tracking results of the following frame, and the last row is randomly selected from the corresponding tracking sequences. Our tracking results perform well under various illumination, expression, posture, occlusion and blurriness.

Different with the point definitions of 2D landmark, 3D landmark regression are more susceptible due to self-occlusion of points under an arbitrary poseture. A larger capacity model is needed to learn the local evidence of each points with textureless image. Also a strategy to simultaneously model the local and global representations is essential. Inspired by this, this paper proposes a local-to-global framework to predict the 3D landmark points by concentrating on subsequently local learning and global learning. Our method builds upon the idea of part heatmap regres-

sionfirstly. In order to learn a coherent understanding of local regions, a modifed staked hourglass network is adpoted to generate a series of heatmaps for each 3D landmark point. In the filed of human pose estimation, the architecture [30] combining deep redusial network modules [17, 18] has achieved huge success. Being extended for 3D landmark alignment, it is capable of dealing with large pose variations and extreme expressions well, with sufficient capacity and the strategy of multi-scale learning. Despite these benefits, a part attention module applying a channelwise attention opereation instead of the original convolution opereation is integrated into the network, which effectively increase the diversity of local representations. Although the local heatmap model provides precise description of local points, it lacks of spatial contextual information of different regions. We employ another global regression network to predict the final smooth 3D landmark taking both the input image and all the heatmaps as input. All these networks are trained end-to-end to model the feature representations from local granularity to global consistent space, and erase the noises of independent heatmaps.

Based on the proposed algorithm, we also provide a 3D landmark tracking framework. Experiments on publicly available datasets, together with the results of Menpo challenge competition, confirm the effectiveness of the present method. As shown in Figure 1, our method performs well in arbitrary poses, illumination and occlusions.

2. Related Work

In the domain of face alignment, shape regression is the most straight-forward way to solve this landmark localization problem. In order to improve the location accuracy, plenty of algorithms are carried out to model both the global and local features. All these methods can be divided into two categories. The first category follows a global regression framework by extracting local features [39, 40, 31, 11, 26]. Majority of such approaches adopt the architecture cascading multiple weak regressors to obtain better and better local feature descriptions. The other category takes the Coarse-to-Fine framework to generate global and locale regression results successively [41, 43, 15, 20, 33, 14, 42, 44, 9]. Although convolutional neural networks produce better feature representation capacity, the performance of single stage network is still not good enough. These two categories discuss the relationship between global and local respectively in terms of feature and texture.

However, 3D face alignment, which aims to estimate 3d landmark from a 2D image, has only a few related studies in 3D shape regression. A relevant but different problem is the 3D face reconstruction [19, 21, 5, 46, 4], where 3DMM [3] is the most effective method to generate dense 3D shape including the face posture and expression. By employing convolution neural networks, [23, 45] propose a similar framework iteratively updating the 3D fitting parameters by combining the cascaded CNN regressor method with 3DMM. The method are used to enhance the performance of 2D landmark regression on large-pose face images. As 3D landmark brings more posture information, [24, 37] also adopt a 3D model as an auxiliary input to solve 2D alignment with large pose variations.

As we all know, what makes deep learning based shape regression successful is massive training datasets. In view of the lack of 3D shape training data, zhu *et al.* [45] sysnthetically generated a series of datasets based on 300W [32], which gradually become the benchmark of 3D shape regression methods. Based on these 3D shape datasets, Adrian Bulat [7, 8] proposes a regression model to verifiy the validity of part heatmap regression in 3D shape regression. He discusses the relationship between 2D and 3D shape regression, and tries to generate a much Larger 3D facial landmark dataset by appliying a 2D guided network which converts 2D landmark annotations to 3D. However, he only experimentally applies a single end-to-end network based on local heatmaps. [6] describes a two-stage network, which is similar with our proposed work. His method predicts the first two axises of 3D landmark with local heatmaps regression at first, then generats the depth of landmark points taking both the heatmaps and the image as input of a global regression network.

Unlike the scarce researchs in 3D shape regression, these exists a great deal of works in human pose estimation. There are many correspondences between these two research topic. [30] designs a hourglass network. It is a symmetric topology consisting of the successive processions of pooling and upsampling along with intermediate supervision, which enables the network to capture features from different scales as well as global contextual information. [12] increases model capacity by applying an self adversarial module. [13] proposes a multi-context attention mechanism onto the modified hourglass network. The attention module mainly consists of a multi-semantics attention obtained by generating attention maps for each stack of the hourglass, and a hierarchical coarse to fine attention scheme to zoom from gloabel into local part regions for more precise localization. This paper further investigates the relationships of global and local features, with all these modifications carried out to encode the local appearance and global representations. [29] discusses this problem from another perspective by designed a network to learn the affine transformation matrix between each local heatmaps. Both of them achieve the state-of-the-art performance on the human pose estimation benchmarks.

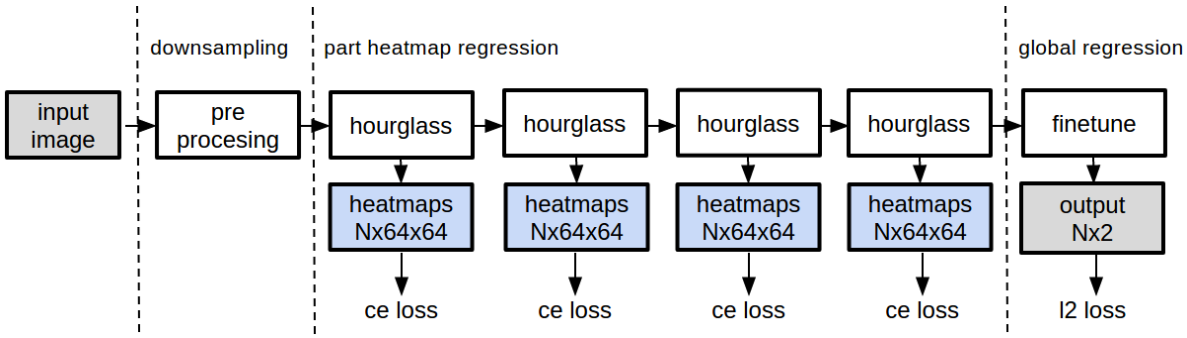

Figure 2. The proposed shape regression network. The preprocessing module downsamples the input image size from 256 to 64. Then four staged hourglass networks are employed to extract local landmark heatmaps, followed by a fintune module to generate the final smooth results. Also the details of the preprocessing and finetune modules are depicted here, where $c\_k\_s\_p$ notes the channel num, kernel size, stride and pad of each convolution opereationsuccessively. $celoss$ and $l2loss$ are cross entropy loss and L2 loss.

3. The Proposed Network

The proposed shape regression network is made up of two main steps. The first part foucuses on estimating the raw landmark locations by generating a series of regression heatmaps, one for each landmark, in which, a local attention mechanism is carried out from global to local heatmaps to explore global context information combined with local evidence. These two modules are described in section 3.1 and 3.2 separately. The second part, presented in section 3.3, is aggregated to produce the final smooth 3D prediction. It takes both the point heatmaps and the original input image as input. These two networks are trained end-to-end to become the core of 3D landmark tracking system. The overall architecture of the proposed network is illustrated in Figure 2.

3.1. Heatmap Regression Network

The architecture of the heatmap regression network is based on the houreglass model proposed in [30]. It introduces an efficient way to capture and consolidate features at different scales and resolutions with the skip route. The structure of hourglass is a symmetric topology, consisting of the successive steps of pooling and upsampling. With downsampling operation, it increases receptive field to allow smaller spatial filters to compare features across the entire space of the image. With upsampling operation, it acts on local information to identify each semantic landmark in a human face. Residual modules [17] is employed between two downsampling operations.

As employed, the network was adapted by: (1) changing the convolution opereation after hourglass module with a residual module, (2) replaceing the original nearest neighbour upsampling by learnable deconvolution layers, (3) replacing the residual module [17] to the Identity version in [18], (4) removing the dropout module, and (5) adopting the pixel wise sigmoid cross entropy loss function instead

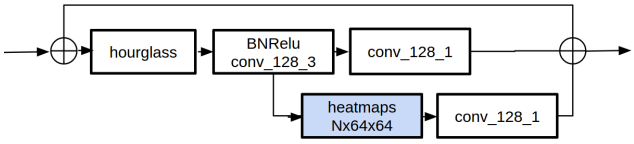

Figure 3. Single heatmap regression module. Two convolution operation and an attention module are adopted to generate the heatmaps and the input of following regression module.

of L2 loss function. All these changes result in a slignt boost to the final performance. The first three amendments enrich the information received by the output of each building block, and makes the whole framework more robust to scale change. Dropout has been proved by experiment that it leads to performance degradation without enhancing generalization quality. And cross entropy loss is more appropriate then L2 loss in the gaussian distribution of heatmaps. Based on the modified hourglass module, Figure 3 depicts the structure of single heatmap regression module.

Empirically, we decode each landmark as a heatmap using 2D Gaussian with radius=5 pixels centered at pixel location of that landmark. Each heatmap has a resolution of 64×64 to reduce the GPU memory consumption in training. However, the input RGB image is aligned with the given landmarks and cropped with size 256×256 to locate the face at the centre of the image. Then the full network starts with 3x3 convolution layer with stride 2, followed by a residual module and a max pooling to bring the resolution down from 256 to 64. Two subsequent residual modules are employed before the hourglass, and four stages hourglass networks are stacked in our algorithm.

3.2. Local Attention Module

Since the visual attention model is computationally efficient and is effective in understanding images, it has

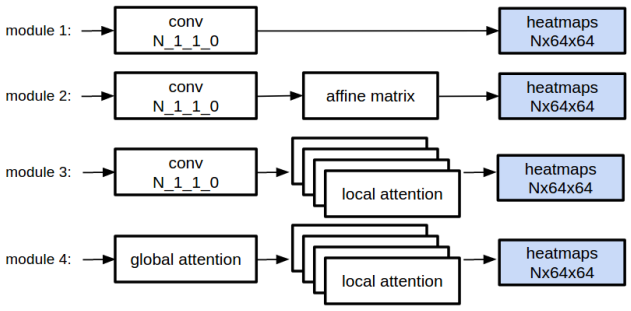

Figure 4. Differnet heatmap modules. the first row is the traditional network, The second row is the method

achieved great success in image classification [38, 16, 2, 36], saliency detection [27], human pose estimation [13]. The traditional soft attention mechanism can be defined as follows:

$$y = f(W * x + b) \tag{1}$$

where $*$ denotes convolution, W and b denote the convolution filters and bias. f is the nonlinear activation function, where sigmoid is usually applied to normalize the attention map. y is the final attention map that summarizes information of all channels in x, and projects into $C \times H \times W$ space. In [36], a squeeze operation is employed to squeeze the y into $C \times 1 \times 1$, which greatly improve the performance of image classification. However, we find out that 1×1 convolution operation performs better by generating $1 \times H \times W$ attention map.

$$h = y \odot x \tag{2}$$

Then, a channel-wise matrix product operation $\odot$ on the input feature x and the attention map y is applied to generate the refined feature map h, which has the same size with x, and refines the feature x.

In our paper, we adopte a local attention scheme to generate reweighted heatmaps. As shown in Figure 4 (module 4), an attemtion module is applied firstly to generate the global heatmap.

$$h_{global} = y_{global} \odot x \tag{3}$$

Then, the global heatmap h_{global} is used as an input of N local attention modules to generate a series of local heatmaps. one module for each landmark.

$$h_i = y_i \odot h_{global} \tag{4}$$

All these reweighted features are concatenated as the final landmark heatmaps. Different from the traditional $N \times H \times W$ convolution operation, the proposed attention scheme strengthens the feature response of each point, and increases the local neighboring spatial correlations.

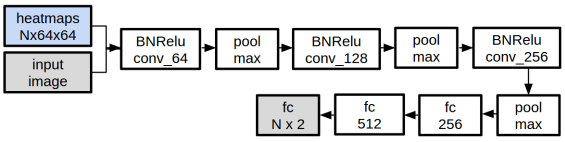

Figure 5. Finetune regression Module.

3.3. Finetune Regression Module

The finetune network is depicted in Figure 5. The input of this module is the staged heatmaps alongside the input RGB image. A simplified VGG-like[34] module is carried out here followed by two full-connection layers to produce $N \times 2$ output. $L2$ loss is applied here to minimize the normalized euclidean distance between the ground truth landmarks and the prediction results. To simultaneously model the global and local features, both the heatmap regression network and the finetune regression network are trained end-to-end.

3.4. 3D Landmark Tracking Framework

Based on the proposed shape regression network, we carry out a 3D landmark tracking system to evaluate the test videos provided by Menpo challenge competitions, as depicted in Figure 6. 'face++ API' [1] is employed to detect the face inside a given image. When applied on a video, it proceeds on each frame to generate all candidate face rects and their corresponding 81 2D points above a appropriate threshold. We manually check the detected faces and find 0 false detection in the test videos.

Two models are trained for different initializations. Model 1 is learned from the cropped training set by aligning 2D landmarks (81 points) to the meanshape of given 3D landmarks (84 points). In model 2, images are directlly aligned to the 3D landmark meanshape, and randomly altered by slight similarity transformation (rotating, translating and scaling) before feeding into the network to deal with the face texture variation between two adjacent frames.

4. Experiments

Given that this work focuses on 3D face landmark tracking, we test our proposed method on static images and landmark tracking on videos separately. Six publicly available datasets are used. Also the performance of our system on Menpo challenge competition [] is present.

4.1. Evaluations on benchmarks

This section evaluates the present two stage shape regression algorithm on static face images, where the ground truth landmarks are applied for face alignment to generate a aligned face image as the input of networks, and the prediction 3D landmarks are generated for evaluation. To compare

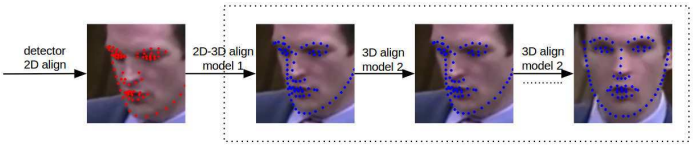

Figure 6. The proposed 3D face tracking framework.

Method	NME
Resnet18 [17]	0.0371
Hourglass [30]	0.0338
+ residual unit modified 3.1 (1,2,3)	0.0332
+ cross entropy loss 3.1 (5)	0.0307

Table 1. NME comparisons of modifications in heatmap regression and fc regression on AFLW2000-3D.

Method	NME
+ 3.2 module 1	0.0307
+ 3.2 module 2	0.0364
+ 3.2 module 3	0.0300
+ 3.2 module 4	0.0296

Table 2. NME comparisons of different types of heatmap generation on AFLW2000-3D.

with the state-of-the-art methods, two well-konwn benchmark datasets (300W-LP and AFLW2000-3D) are used for training and evaluations. Images in both of them provide 68 points landmark annotations.

300W-LP: 300W-LP(300W across Large Poses) [45] contains 61225 samples (1786 from IBUG, 5207 from AFW, and 16556 from LFPW) sysnthetical generated from 300-W [32]. Zhu *et al.* [45] carries out a 3D Dense Face Alignment(3DDFA) algorithm to fit the image into a dense 3D face model to render the faces into larger poses ranging from -90^o to 90^o. The dataset contains large-pose variations, with various expressions and illumination conditions. The 3D landmark annotations 300W-LP-3D is taking as the training set to bring into correspondence with Zhu's original experiment.

AFLW2000-3D: AFLW2000-3D is the most commonly used testset for 3D face alignment in-the-wild. It is also specifically contructed by Zhu *et al.* from the first 2000 AFLW smaples [28]. AFLW2000-3D is suitable for evaluating face alignment performance across large poses.

Following most previous works, we evaluate the alignment accuracy using the standard Normalized Mean Error. As noted in [45], error metric normalized by the inter-pupil distance leads to serious bias for which the interocular distance varies a lot under different face posture, hence the relatively stable face size is adopted as normalization factor. In particular, the Normalized Mean Error is defined as:

$$NME = \frac{1}{N}\sum_{n=1}^{N}\frac{\|s_g - s_p\|_2}{\sqrt{w * h}} \tag{5}$$

where S_g and S_p denote the ground truth landmarks and the corresponding prediction separately, w and b are the bounding box shape calculated from the ground truth 3D landmarks.

Evaluations on heatmap module We separately evaluate all of our proposed modules. Firstly, we try to inverstigate the capacity of heatmap regression and systematically evaluate the performance of each modification. Resnet18 is employed as the baseline of shape regression model. Table 1 reports the NME-based comparisons of all 68 3D points on AFLW2000-3D. As it can be observed, the performance of heatmap regression is far better then shape regression with NME decreased from 0.0371 tp 0.0338. All the modifications described above result in performance improvements step by step, except the finetune module leads to a slight decrease. Cross entropy loss is the most effective one. Based on these modules, NME of AFLW2000-3D reduced about 13% from the originial hourglass based heatmap regression algorithm.

Evaluations on attention module Attention scheme is one of the most important module to explore local representations from global context information. We apply four different types of heatmap generation methods based on the modified heatmap regression network to prove the effectiveness of the present attention module 4 . Module 1 is the original 1×1 convolutional network. In Module 2, sereval affine transformation matrixs are learned between the local heatmap of adjacent points to adjust heatmaps based on points correlations. Difference between Module 3 and Module 4 are whether to apply attention on global heatmaps. As shown in 2, more attention leads to better performance. While applyed on both the global heatmap and the local heatmap, attention scheme enhances the ability of local feature representation accompanied by the global context information. However, module 2 results in significant deterioration, that is because transformation between adjacent points leads to the unreasonable distribution and brought noises to point locations.

Evaluations on finetune module Also we evaluate the performance of fintune shape regression network. Evaluated on AFLW2000-3D, the NME has a slight drop on the performance from 0.0296 to 0.0304. As there exists exaggerated expressions, occlusions, and posture, too smooth landmarks results in lower accuray. However, NME of the videoset is remarkable descrease.

Comparisons with other methods The proposed results are compared with other state-of-the-art performances in 3. 3DDFA [45] is applied as a baseline on AFLW2000-3D. we recomplement another heatmap regression method [7]

Method	NME
PCPR [10]	0.0780
ESR [11]	0.0799
SDM [39]	0.0612
3SDFA [45]	0.0542
3DDFA+SDM [45]	0.0494
binary version [7]	0.0326
real-valued version [7]	0.0331
proposed	0.0296

Table 3. NME comparisons of out proposed modules and other state-of-the-art methods on AFLW2000-3D. The results for RCPR, ESR and SDM are taken from [45].

in real-valued version. However, it exists a slight drop on the performance. While both of these two methods are implemented with similar architecture, there is only small performance differences as listed in Table **??**. It also proves the capacity of local heatmap regression network.

4.2. Results of Menpo Benchmark Competion

Furthermore, we evaluate the proposed tracking framework to analysis the performance on videos. Two datasets are provided by Menpo organizer. In which, 300VW [25] includes 50 high resolution video sequences with moderate expression, head pose, and illumination changes. 3D facial landmark annotations are provided with the semi-automatic annotation process [35]. We randomly divide the dataset into training and testset. All frames of 12 persons are selected as testset, and the remaining as training dataset. However, the given dataset only contains sparse frames selected from the raw videos instead of a complete video. To simulate the evaluation standard of Menpo Benchmark Competion, we employ a linear interpolation method to generate continous video frames from the image testset to build a video testset.

The other dataset is about static images. Images from AFLW [28], FDDB[22], 300W, 300W-Test [32] are collected and fitted with 3D facial morphable model [4] to generate 84 3D point annotations. The parameters of the model have been carefully selected and all fittings have been visually inspected. Also the final landmarks have been manually corrected. This dataset is combined with the selected 300VW training set as the final training dataset with about 30000 video frames and 12000 images in-the-wild.

We focus on discuss the two-stage network performance on the videoset. NME results are listed in 4. After finetuning, NME is reduce from 0.054 to 0.043. Different from the static imageset, two stage strategy results in performance promotion. This problem may be caused by overfitting. In our experiments, we find all these models are sensitive to the landmark initialization, while all the initialization of static images are ground truth. Under a more complicated

Method	NME
attention based hourglass 3.1,3.2	0.054
+ finetune module 3.3	0.043

Table 4. NME comparisons of two stage results of videoset.

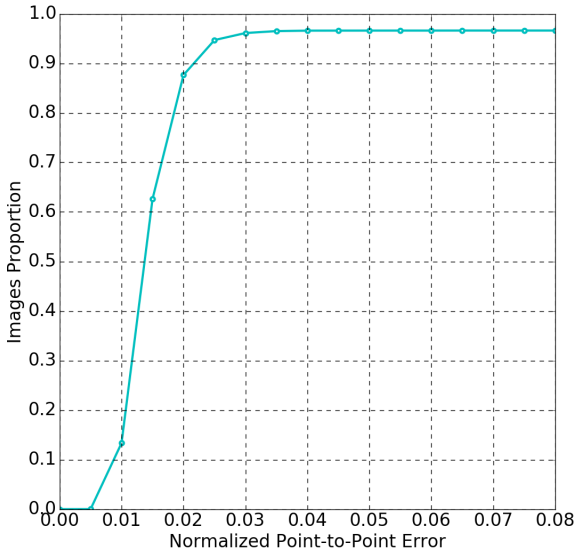

Figure 7. Evaluation on the test sets of Menpo challenge competition benchmark.

situation, two stage network can fully explore both the part details and global context information to some extent. However, it still leads to a huge err increase compared with the evaluation results of static images, although two aligned strategies are employed in the framework. Further research is required.

The results of the same tracking system are submitted to 1st 3D Face Tracking in-the-wild Competition. The tracking results has been evaluated independently by the organizer using their own ground truth and own evaluation metric. The returned results are illustrate in Figure 7.

5. Conclusion

In this paper, we have presented a novel two-stage 3D landmark regression network which shows strong robustness and high accuracy for face images in-the-wild. An attention-based heatmap regression network followed by another shape regression network is developed to discuss the relationship representations between global and local features in shape regression. We hope this will be helpful for other regression methods. Also attention mechanism is proved to be helpful in several tasks. Future work will concentrate on exploring more expressive feature representations to further improve the accuracy and robustness of the proposed model, and accelerating it for real time face landmark tracking.

References

[1] https://www.faceplusplus.com.cn/face-landmark-sdk/.

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention, 2015. ICLR.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces, 1999. siggraph.

[4] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models "in-the-wild", 2017. CVPR.

[5] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniahy, and D. Dunaway. A 3d morphable model learnt from 10,000 faces, 2016. CVPR.

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

[7] A. Bulat and G. Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources, 2017. arXiv:1703.00862.

[8] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks), 2017. arXiv:1703.07332.

[9] A. Bulat and Y. Tzimiropoulos. Convolutional aggregationof local evidence for large pose face alignment, 2016. British Machine Vision Conference.

[10] X. P. Burgos-Artizzu, P. Perona, and P. Dollr. Robust face landmark estimation under occlusion, 2013. ICCV.

[11] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression, 2014. IJCV.

[12] C.-J. Chou, J.-T. Chien, and H.-T. Chen. Self adversarial training for human pose estimation, 2017. CoRR abs/1707.02439.

[13] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation, 2017. arXiv:1702.07432.

[14] J. Deng, Q. Liu, J. Yang, and D. Tao. M3csr: Multi-view, multi-scale and multi-component cascade shape regression, 2016. Image and Vision Computing.

[15] H. Fan and E. Zhou. Approaching human level facial landmark localization by deep learning, 2016. Image and Vision Computing.

[16] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks, 2015. CVPR.

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

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

[19] G. Hua, F. Yana, J. Kittlera, . H. C. William Christmasa, Z. Fenga, and P. Hubera. Efficient 3d morphable face modelfitting, 2017. Pattern Recognition.

[20] Z. Huang, E. Zhou, and Z. Cao. Coarse-to-fine face alignment with multi-scale local patch regression, 2015. arXiv:1511.04901.

[21] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Ratsch. Fitting 3d morphable face models using local features, 2015. International Conference on Image Processing.

[22] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings, 2010.

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

[24] A. Jourabloo and X. Liu. Pose invariant face alignment via cnn-based dense 3d model fitting, 2016. CVPR.

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

[26] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees, 2014. CVPR.

[27] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection, 2016. CVPR.

[28] M. Kstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, 2011. ICCV Workshop.

[29] A. Newell, Z. Huang, and J. Deng. Associative embedding: End-to-end learning for joint detection and grouping, 2016. CVPR.

[30] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation, 2016. arXiv:1603.06937.

[31] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features, 2014. CVPR.

[32] C. Sagonas, E. Antonakosa, G. Tzimiropoulosb, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results, 2016. Image and Vision Computing (IMAVIS).

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

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2015. CVPR.

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

[36] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification, 2017. CVPR.

[37] S. Xiao, J. Li, Y. Chen, Z. Wang, A. Kassim, J. Feng, and S. Yan. 3d-assisted coarse-to-fine extreme-pose facial landmark detection, 2017. CVPR Workshop.

[38] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification, 2015. CVPR.

[39] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment, 2013. CVPR.

[40] X. Xiong and F. D. la Torre. Global supervised descent method, 2015. CVPR.

[41] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for accurate face alignment, 2013. ICCV Workshops.

[42] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment, 2014. ECCV.

[43] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade, 2013. ICCV Workshop.

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

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

[46] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3d morphable model fitting, 2015. ICCV Workshop.