AnchorFace: An Anchor-based Facial Landmark Detector Across Large Poses

Zixuan Xu[1*], Banghuai Li[2*], Miao Geng[3], Ye Yuan[2], and Gang Yu[2]

[1] Peking University zixuanxu@pku.edu.cn
[2] MEGVII Technology {libanghuai,yuannye,yugang}@megvii.com
[3] Beihang University geng_m@buaa.edu.cn

Abstract. Facial landmark localization aims to detect the predefined points of human faces, and the topic has been rapidly improved with the recent development of neural network based methods. However, it remains a challenging task when dealing with faces in unconstrained scenarios, especially with large pose variations. In this paper, we target the problem of facial landmark localization across large poses and address this task based on a split-and-aggregate strategy. To split the search space, we propose a set of anchor templates as references for regression, which well addresses the large variations of face poses. Based on the prediction of each anchor template, we propose to aggregate the results, which can reduce the landmark uncertainty due to the large poses. Overall, our proposed approach, named AnchorFace, obtains state-of-the-art results with extremely efficient inference speed on four challenging benchmarks, i.e. AFLW, 300W, Menpo, and WFLW dataset. Code will be released for reproduction.

1 Introduction

Facial landmark localization, or face alignment, refers to detect a set of predefined landmarks on the human face. It is a fundamental step for many facial related applications, e.g. face verification/recognition, expression recognition, and facial attribute analysis.

With the recent development of convolutional neural network based methods [29], the performance for facial landmark localization in constrained scenarios has been greatly improved [12,57,42]. However, unconstrained scenarios, for example, faces with large pose, still limit the wide application of the existing landmark algorithms. In this paper, we target to address the problem of facial landmark localization across large poses.

There are two challenges for facial landmark detection across large poses. On one hand, faces with large poses will significantly increase the difficulty for landmark localization due to the **large variations among different poses**. As shown in Fig. 1, directly regressing the point coordinates may not be able to localize every landmark point precisely. On the other hand, there usually

[*] Equal Contribution

exists a large probability of **uncertainty** due to the self-occlusion and noisy annotations. For example, occlusion will usually lead to invisible landmarks, which will increase the uncertainty for the landmark prediction. Besides, the faces with a large pose will also cause difficulty during the data annotation process.

To address the above two challenges, we propose a novel pipeline for facial landmark localization based on an anchor-based design. The new pipeline includes two steps: split and aggregate. An overview of our pipeline can be found in Fig. 2. To deal with the first challenge with large pose variations, we adopt the divide-and-conquer way following an anchor-based design. We propose to use the anchor templates to split the search space, and each anchor will serve as a reference for regression. This can significantly reduce the pose variations for each anchor. To address the second issue with pose uncertainty, we propose to aggregate each anchor result weighted by the predicted confidence.

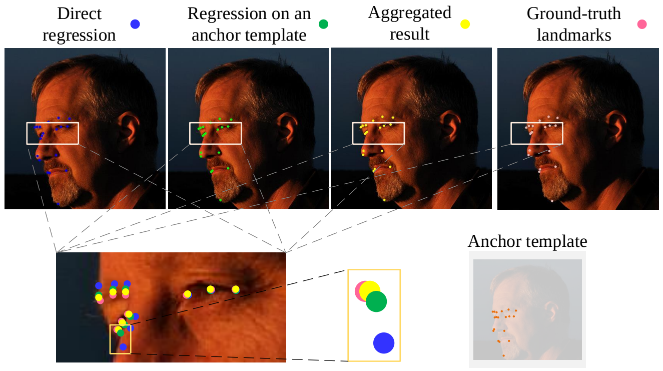

Fig. 1: A comparison between direct regression and anchor-based regression (AnchorFace). Our AnchorFace includes two steps. The first step is to introduce the anchor templates and regress the offsets based on each anchor template (Second Column). The second step is to aggregate the prediction results from multiple anchor templates (Third Column)

In summary, we propose AnchorFace to implement the split-and-aggregate strategy. There are three contributions in our paper.

- We propose a novel pipeline with a split-and-aggregate strategy which can well address the challenges for face alignment across large poses.
- To implement the split-and-aggregate strategy, we introduce the anchor design into the facial landmark problem, which can simplify the search space for each anchor template and meanwhile improve the robustness for landmark uncertainty.

- Our proposed AnchorFace can achieve promising results on four challenging benchmarks with an impressive inference speed of 4050 FPS[†].

2 Related Work

Facial Landmark Localization. In the literature of facial landmark localization, a number of achievements have been developed including the classic ASMs [32], AAMs [17,20,30,39], CLMs [8,9], and Cascaded Regression Models [5,6,7,15,46,59,58]. Nowadays, more and more deep learning-based methods have been applied in this area. These deep learning based methods could be divided into two categories, i.e. coordinate regression methods and heatmap regression methods.

Coordinate regression methods directly map the discriminative features to the target landmark coordinates. The earliest work can be dated to [40]. Sun et al. [40] used a three-level cascade CNN to do facial landmark localization in a coarse-to-fine manner, and achieved promising localization accuracy. MDM [41] was the first to apply a recurrent convolutional network model for facial landmark localization in an end-to-end manner. Zhang et al. [55] utilized a multi-task learning framework to optimize facial landmark localization and correlated facial attributes analysis simultaneously. Recently, Wingloss [16] was proposed as a new loss function for landmark localization, which can obtain robust performance against widely used $L2$ loss.

Heatmap regression methods generate a probability heatmap for each landmark, respectively. Benefit from FCN [26] and Hourglass [33], heatmap regression methods have been successfully applied to landmark localization problems and have achieved state-of-the-art performance. JMFA [11] achieved high localization accuracy with a stacked hourglass network [33] for multi-view facial landmark localization in the Menpo [52] competition. Yang et al. [49] adopted a supervised face transformation to normalize the faces, then employed an Hourglass network to regress it. Recently, LAB [42] proposed to use additional boundary lines as the geometric structure of a face image to help facial landmark localization.

Faces with Large Pose. Large pose is a challenging task for facial landmark localization, and different strategies have been proposed to address the difficulty. Multi-view framework and 3D model are two popular ways. Multi-view framework uses different landmark configurations for different views. For example, TSPM [61] and CDM [50] employ DPM-like [14] method to align faces with different shape models, and choose the highest possibility model as the final result. However, multi-view methods have to cover each view, making it impractical in the wild. 3D face models have been widely used in recent years, which fit a 3D morphable model (3DMM) [4] by minimizing the difference between face image and model appearance. Lost face information can be recovered to localize the invisible facial landmarks [3,18,19,24,63]. However, 3D face models are limited by their own database and the iterative label generation method.

[†] The computational speed of 4050 FPS is calculated on Nvidia 2080 Ti GPU with batchsize 256. If batchsize is set as 1, the FPS is 320.

Besides, researchers have applied multi-task learning to address the difficulties resulting from pose variations. Other facial analysis tasks, such as pose estimation or facial attributes analysis, can be jointly trained with facial landmark localization [35,47,54]. With joint training, multi-task learning can boost the performance of each subtask. The facial landmark localization task can achieve robust performance. But the multi-task framework is not specially designed for landmark localization, it contains much redundant information and contributes to large models.

In this paper, we propose an anchor-based model for facial landmark localization. Different from [45], which utilized anchor points to predict the positions of a human 3D pose, our approach introduces a split-and-aggregate pipeline for the facial landmark localization. Anchor is utilized as a reference for regression in our approach. Overall, our model requires neither cascaded networks nor large backbones, leading to a great reduction in model parameters and computation complexity, while still achieving comparable or even better accuracy.

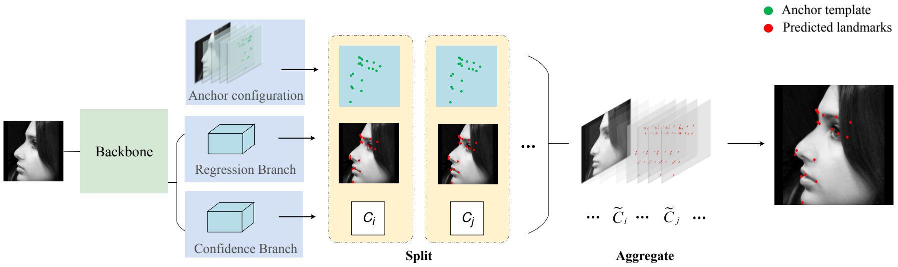

Fig. 2: The pipeline of our proposed AnchorFace landmark detector. AnchorFace is based on a split-and-aggregate strategy, which consists of the backbone and two functional branches: the offset regression branch and the confidence branch. In the split step, we predict the landmark position based on each anchor template. During aggregate step, the predictions of multiple anchor templates are averaged by weighted confidence

3 Proposed Method

In this paper, we propose a new split-and-aggregate strategy for facial landmark detector across large poses. An overview of our pipeline can be found in Fig. 2. To implement the split-and-aggregate strategy, we introduce the anchor-based design, and our approach is named AnchorFace. In the following section, we will discuss the **split** and **aggregate** steps separately, followed by the details on the network training.

3.1 Split Step

Due to the large pose variations among different poses, it is a challenging problem to directly regress the facial landmarks while maintaining high localization precision. In this paper, we propose to utilize the divide-and-conquer way to address the issue from large pose variations. More specifically, we propose to employ the anchor templates as regression references to split the search space. Different from the traditional methods which regress the landmarks with a uniform facial landmark detector, we propose to regress the offsets base on a set of anchor templates.

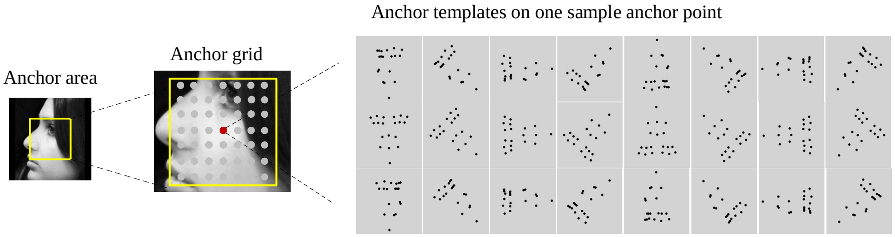

Fig. 3: An illustration of our anchor configuration. Anchor area is a region centered at the image center with a spatial neighborhood. Based on the anchor area, we setup a grid of anchor points, where each anchor point contains a set of anchor templates to model various pose variations

Anchor Configuration. As shown in Fig. 3, there are three hyper-parameters for designing the anchor configuration: anchor area, anchor grid, and anchor templates.

Anchor area is denoted as the region to set the anchors. It is usually centered at the image center with a spatial neighborhood. The reason to define the anchor area is that the input image is cropped to put the face near the image center. Thus, we select a region near the image center, which is called anchor area, to set up the anchors. Based on the anchor area, we sample a set of anchor points in a grid, e.g. a 7 × 7 grid, as shown in Fig. 3. Each anchor point can be considered as the center of a set of anchor templates. The anchor templates are designed to address the challenges from large pose variations. Intuitively, these anchor templates are used to split the search space for regression and can serve as references for offsets prediction. Therefore, the sampling of anchor templates should be able to cover different variations of large poses and reduce the redundancy for the anchor sets.

To implement the anchor templates, we present two potential ways. The first one is to hand-design the anchors based on prior knowledge. The second one integrates the proposals generated by the data distribution.

An overview of our hand-designed anchors can be found in Fig. 3. For each anchor point, we explore the 3D pose spaces (yaw, roll, pitch) and design the pose-level anchor set as follows. As unconstrained large-pose faces have large variations on the yaw direction, we first select N_{yaw} base anchors ($N_{yaw} = 3$ in our paper representing the anchors for the left, frontal, right faces). To generate the N_{yaw} base anchors, we utilize a heuristic approach to divide the training faces into three buckets and compute the average face landmarks for each bucket to obtain the anchor proposal. More specifically, we use the ratio of two eyes' width for bucket assignment. We define an indicator to estimate the yaw angle of each training face:

$$r = \frac{|p_{l_1} - p_{l_2}|_2}{|p_{r_1} - p_{r_2}|_2} - \frac{|p_{r_1} - p_{r_2}|_2}{|p_{l_1} - p_{l_2}|_2}, \tag{1}$$

where $p_{l_1}, p_{l_2}, p_{r_1}, p_{r_2}$ are the coordinates of left eye inner corner, left eye outer corner, right eye inner corner, and right eye outer corner respectively. With a threshold γ, we put the faces into the left or right bucket, when $r > \gamma$ or $r < -\gamma$. The other faces will be kept into the frontal bucket. We set $\gamma = 6$ in our experiments, as shown in Fig. 4.

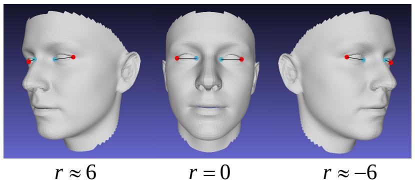

Fig. 4: An illustration of the metric r for classifying the faces into three buckets along the yaw direction

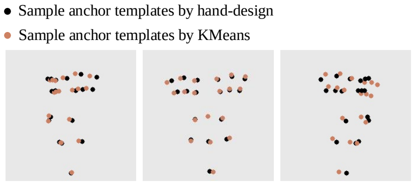

Fig. 5: A comparison of three base anchors generated by hand-design approach and KMeans clustering based on AFLW [22] dataset

Based on the N_{yaw} base anchors, to cover the roll variations, we rotate each anchor on the roll dimension. For example, we can get twenty-four templates by rotating the basic three anchors each 45° from 0° to 360°. Optionally, we can involve the pitch variations by directly projecting (rotating) along the pitch dimension. However, based on our experimental results, the anchors designed along the pitch view cannot further improve the performance but compromise the computational speed. Thus, in our final design, only anchors along the yaw and roll dimensions are utilized as shown in Fig. 3.

An alternative solution for the anchor design is based on the data distribution among the training faces. We first perform KMeans clustering, and we can generate a set of base anchors. One example is shown in Fig. 5. We can see that the clustered anchors among all the training faces obtain similar anchors along the yaw direction, as discussed in hand-designed anchors. Following the similar steps for the hand-designed anchors, we can rotate the generated prototypes along the roll and pitch direction to generate more anchors.

Regression and Confidence Branch. Based on the anchor proposals, we design a new head structure which involves two branches: regression branch and confidence branch. Regression branch aims to regress the landmark coordinate offsets based on each anchor. Confidence branch assigns each anchor with a confidence score. Among all the anchor templates, those anchors which are close to the pose of the ground-truth face should be given higher confidence.

As shown in Fig. 2, both the confidence branch and the regression branch are built upon the output feature map of the backbone network. While we set $h \cdot w$ anchors in the image, the output of the confidence and regression branch are $C_{con} \cdot h \cdot w$ and $C_{reg} \cdot h \cdot w$ respectively, where C_{con} and C_{reg} are denoted as the output channel number of the confidence branch and the regression branch respectively. Here $C_{con} = K$ and $C_{reg} = K \cdot 2L$, where K, L refer to the number of anchor templates on each anchor point and the number of facial landmarks respectively.

Table 1: The definition of Symbols

Symbol	Definition
A	A set of Anchor points on the spatial anchor grid
a	One anchor point $a \in A$
T	A set of anchor templates as in Fig. 3
$T(a,t)$	Anchor template $t \in T$ centering at anchor point a
$T_j(a,t)$	Landmark j on anchor template $T(a,t)$
$O(a,t)$	Output from the regression branch based on $T(a,t)$
$\overline{O}(a,t)$	Ground-truth (GT) offsets based on $T(a,t)$
$C(a,t)$	Output from the confidence branch based on $T(a,t)$
$\overline{C}(a,t)$	Confidence GT label based on $T(a,t)$

3.2 Aggregate Step

Large-pose faces will increase the uncertainty for the landmark prediction. To address this problem, we propose to aggregate the predictions from different anchor templates. More specifically, we first set a threshold C_{th} to pick up the reliable anchor predictions. The anchor predictions with low confidence scores are regarded as outliers and will be discarded. The remaining anchor predictions will be averaged by the weighted confidence for each prediction. As a result, the position of landmark j can be obtained as the weighted average of the outputs of all anchor faces as:

$$\widetilde{S}_j = \frac{\sum_{a \in A, t \in T} \widetilde{C}(a,t) \cdot (O_j(a,t) + T_j(a,t))}{\sum_{a \in A, t \in T} \widetilde{C}(a,t)}, \tag{2}$$

where

$$\widetilde{C}(a,t) = \begin{cases} 0, & C(a,t) < C_{th} \\ C(a,t), & others \end{cases} \tag{3}$$

The definition of the symbols can be found from Table 1, and the threshold C_{th} is set to 0.6 in our experiments.

3.3 Network Training

In this subsection, we will discuss the ground-truth setting for the regression and confidence branch as well as the related losses. For the regression branch, the target is to regress the offsets against each of the predefined anchor. The regression loss L_{reg} can be defined as:

$$L_{reg} = \sum_{a \in A, t \in T} C(a,t) \sum_{j} |O_j(a,t) - \overline{O}_j(a,t)|, \tag{4}$$

where $O_j(a,t)$ and $\overline{O}_j(a,t)$ refer to the prediction offsets and the ground-truth offsets for jth landmark. $C(a,t)$ is denoted as the confidence weight for the anchor template $T(a,t)$. The detailed symbol definitions can be found in Table 1.

For the confidence branch, we set the targeted confidence output $\overline{C}(a,t)$ as the L2 distance between the anchor pose $\boldsymbol{v_1}$ and the ground-truth pose $\boldsymbol{v_2}$ as $||\boldsymbol{v_1} - \boldsymbol{v_2}||_2$, where $\boldsymbol{v_1}, \boldsymbol{v_2}$ refer to flatten landmark coordinates. To normalize the pose difference, we perform a tanh operation as:

$$\overline{C} = \tanh(\frac{||\boldsymbol{v_1} - \boldsymbol{v_2}||_2}{\beta \cdot 2L}), \tag{5}$$

where β is a hyperparameter and L refers to the count of facial landmarks. The confidence loss is then defined as:

$$\begin{aligned} L_{con} = \sum_{a \in A, t \in T} & (-C(a,t) \cdot \log \overline{C}(a,t) \\ & - (1 - C(a,t)) \cdot \log(1 - \overline{C}(a,t))). \end{aligned} \tag{6}$$

The network is jointly supervised by the two loss functions above with end-to-end training. The final training loss is then defined as:

$$L_{total} = L_{reg} + \lambda \cdot L_{con} \tag{7}$$

where λ is a hyperparameter in our method, and it is insensitive to the localization accuracy in our experiments.

4 Experiment

4.1 Experiment settings

Datasets. The experiments are evaluated on four challenging datasets, i.e. AFLW, 300W, Menpo, and WFLW.

AFLW [22] dataset: AFLW contains 24386 in-the-wild faces with a large head pose up to 120° for yaw direction. We follow the standard setting [59,58], which ignores two landmarks of ears and evaluates the remaining 19 landmarks. AFLW is split into two sets: (i) AFLW-Full: 20000 and 4386 images are used for training and testing, respectively; (ii) AFLW-Frontal: 1314 images are selected from 4386 testing images for evaluation on frontal faces.

300W [38] dataset: 300W is a collection of LFPW [2], AFW [62], HELEN [23], XM2VTS [31], IBUG [37], which have 68 landmark annotations. Following the common settings [27,59], we utilize all the training samples from LFPW, HELEN and the full set of AFW as the training set, with a total of 3148 training images. 554 testing images from LFPW and HELEN are used as the common testing subset; 135 images from IBUG are used as the challenging testing subset. These two subsets constitute the full testing set.

Menpo [10,52] dataset: The Menpo challenge dataset consists of two subsets: semi-frontal and profile face image datasets. The former is annotated with the standard 68 point landmarks following the principle of [37], and the latter has symmetric 39 landmarks for the left and right profile faces.

WFLW [42] dataset: WFLW is a recently proposed facial landmark dataset based on WIDER Face. There are 10000 faces (7500 for training and 2500 for testing) with 98 annotated landmarks. Faces in WFLW are collected under unconstrained conditions, including large variations in pose. Several subsets are extracted from the full testing set for further analysis, including large pose (326 images), expression (314 images), illumination (698 images), make-up (206 images), occlusion (736 images) and blur (773 images).

Evaluation metric. We adopt the normalized mean error (NME) for evaluation. The normalized mean error is defined as the average Euclidean distance between the predicted facial landmark locations $O_{i,j}$ and their corresponding ground-truth facial landmark annotations $\overline{O}_{i,j}$:

$$NME = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{L}\sum_{j=1}^{L}|O_{i,j}-\overline{O}_{i,j}|_2}{d} \tag{8}$$

where N is the number of images in the testing set, L is the number of landmarks, and d is the normalization factor. On AFLW dataset, we follow [59] to use face size as the normalization factor. On Menpo dataset, we use the distance between left-top corner and right-bottom corner as the normalization factor. On 300W and WFLW dataset, we follow MDM [41] and [37] to use the "inter-ocular" normalization factor, i.e. the distance between the outer eye corners.

In addition, on WFLW dataset, two further statistics i.e. the area-under-the-curve (AUC) [48] and the failure rate (which is defined as the proportion of failed detected faces) are measured for furthter analysis. Especially, any normalized error above 0.1 is considered as a failure [42].

Implementation details. In our method, the original images are cropped and resized to a fixed resolution, i.e. 224 × 224, according to the provided bounding boxes. Anchor templates are generated based on KMeans clustering following 3.1, while anchor area and anchor grid are set as 56 × 56 and 7 × 7 respectively. Random rotation and translation are applied for data augmentation. We apply the Adam optimizer with the weight decay of 1×10^{-5} and train the network for 50 epochs in total. The learning rate is set to 1×10^{-3} and divided by ten at 20-th, 30-th, 40-th epoch. $\beta = 0.05$ and $\lambda = 0.5$ are applied to all models across four benchmarks. ShuffleNet-V2 [29] is utilized as the backbone for our algorithm.

4.2 Comparison with the state-of-the-art methods

For a fair comparison, we only compare the methods following the standard settings, as discussed in Section 4.1. Therefore, those methods which are trained from external datasets or combined multiple datasets are not compared.

AFLW dataset: We first evaluate our algorithm on the AFLW dataset. The performance comparisons are given in Table 2. It can be observed that, on this large dataset, our network outperforms the other approaches. As mentioned in Section 4.1, AFLW contains lots of faces with large poses. Note that our method has a significant improvement on AFLW-Full set against AFLW-Frontal, which means that we achieve more robust localization performance on faces in unconstrained scenarios, including large pose. This essentially validates the superiority of our approach.

Table 2: Normalized mean error (%) on AFLW dataset

Methods	AFLW-Full	AFLW-Frontal
LBF [36]	4.24	2.74
CFSS [59]	3.92	2.69
CCL [60]	2.72	2.17
TSR [27]	2.17	-
SAN [12]	1.91	1.85
Wing [16]	1.65	-
SA [25]	1.62	-
ODN [57]	1.63	1.38
AnchorFace	**1.56**	**1.38**

Table 3: Normalized mean error (%) on 300W dataset

Methods	Common	Challenge	Full
Two-Stage [28]	4.36	7.42	4.96
RDR [44]	5.03	8.95	5.80
Pose-Invariant [19]	5.43	9.88	6.30
SBR [13]	3.28	7.58	4.10
PCD-CNN [21]	3.67	7.62	4.44
LAB [42]	**2.98**	**5.19**	**3.49**
SAN [12]	3.34	6.60	3.98
ODN [57]	3.56	6.67	4.17
AnchorFace	3.12	6.19	3.72

300W dataset: We compare our approach against several state-of-the-art methods on 300W Fullset. The results are shown in Table 3. Since there are fewer large pose variations across the whole dataset and the cropped faces normally center near the image center point, 300W dataset is not very challenging compared with the other three benchmarks. However, our algorithm still can achieve promising localization performance with an efficient speed at 4050fps with batch size 256 and 320 fps with batch size 1. Compared with LAB [42], which is slightly better than our method, our approach is much faster (320 vs 17 fps).

Menpo dataset: Menpo dataset has two subsets: semi-frontal and profile. Follow the standard settings on Menpo dataset, we conduct the experiments on each subset and evaluate the testing set separately. The experiment results are reported in Table 4 with the normalized mean error. Our method achieves state-of-the-art performance. Especially on the profile subset, our method outperforms state-of-the-art methods with a large margin, which validates the effectiveness of our proposed approach across large poses.

Table 4: Normalized mean error (%) on Menpo dataset

Methods	CLNF [1]	CFAN [53]	CFSS [59]	TCDCN [56]	3DDFA [63]	CE-CLM [51]	**AnchorFace**
Frontal	2.66	2.87	2.32	3.32	4.51	2.23	**1.82**
Profile	6.68	25.33	9.99	9.82	6.02	5.39	**2.43**

Table 5: Evaluation on WFLW dataset

Metric	Methods	Flops	Testset	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME(%)	CFSS [59]	-	9.07	21.36	10.09	8.30	8.74	11.76	9.96
	DVLN [43]	-	6.08	11.54	6.78	5.73	5.98	7.33	6.88
	LAB [42]	10.6G	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN [12]	11.3G	5.22	10.39	5.71	5.19	5.49	6.83	5.80
	Wing [16]	3.8G	5.11	8.75	5.36	4.93	5.41	6.37	5.81
	AVS [34]	1.8G	5.25	9.10	5.83	4.93	5.47	6.26	5.86
	AnchorFace	**227M**	5.26	9.35	5.56	5.17	5.63	6.47	6.05
Failure Fate(%)	CFSS [59]	-	29.40	84.36	33.44	26.22	27.67	41.85	35.32
	DVLN [43]	-	10.84	46.93	11.15	7.31	11.65	16.30	13.71
	LAB [42]	-	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN [12]	-	6.32	27.91	7.01	4.87	6.31	11.28	6.60
	Wing [16]	-	6.00	22.70	4.78	4.30	7.77	12.50	7.76
	AVS [34]	-	7.44	32.52	8.60	4.30	8.25	12.77	9.06
	AnchorFace	-	7.00	27.91	5.09	5.73	9.70	13.04	8.27
AUC	CFSS [59]	-	0.3659	0.0632	0.3157	0.3854	0.3691	0.2688	0.3037
	DVLN [43]	-	0.4551	0.1474	0.3889	0.4743	0.4494	0.3794	0.3973
	LAB [42]	-	0.5323	0.2345	0.4951	0.5433	0.5394	0.4490	0.4630
	SAN [12]	-	0.5355	0.2355	0.4620	0.5552	0.5222	0.4560	0.4932
	Wing [16]	-	0.5504	0.3100	0.4959	0.5408	0.5582	0.4885	0.4918
	AVS [34]	-	0.5034	0.2294	0.4534	0.5252	0.4849	0.4318	0.4532
	AnchorFace	-	0.5380	0.2555	0.4961	0.5451	0.5423	0.4540	0.4746

WFLW dataset: A comparison of the performance from our proposed approach as well as state-of-the-art methods on WFLW dataset is shown in Table 5. As indicated in Table 5, benefit from the split-and-aggregate pipeline, our proposed method achieves comparable localization accuracy with much lower computational complexity. The best method just outperforms our model by 0.1% in NME, while its computational cost is ten times larger than ours. It is clear that AnchorFace can achieve the best trade-off between the speed and accuracy.

4.3 Computational cost

Based on the efficient ShuffleNet-V2 backbone, the total FLOPs of our network is 227M. Due to the light design of the network, our approach can run as fast as 4050 fps with batchsize 256 and 320 fps with batchsize 1 on an NVIDIA GeForce RTX 2080Ti GPU. Comparisons with some state-of-the-art methods are shown in Table 5. AnchorFace can not only achieve promising results on the challenging benchmarks but also provide extremely efficient inference speed.

4.4 Model analysis

Our proposed AnchorFace introduces a novel split-and-aggregate strategy based on anchor design to address the face alignment across large poses. In this section, we perform further analysis of its mechanism.

Anchor design. Anchor templates serve as regression references to split the search space in our proposed approach. In comparison with directly regressing target landmark coordinates in whole 2D space, regress offsets based on anchor templates can simplify the search space and boost the robustness of localization accuracy. We conduct several experiments on AFLW dataset and make statistics across yaw dimension which is shown in Fig. 6. It is quite clear that AnchorFace significantly outperforms the baseline with lower NME and smaller variances in each subinterval especially for large pose, which can well verify our assumptions.

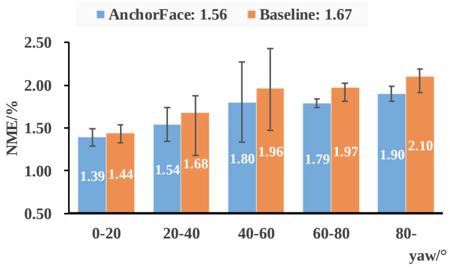

Fig. 6: A comparison of baseline and AnchorFace across yaw dimension

Split-and-aggregate strategy. In our proposed algorithm, we follow the divide-and-conquer way to address the challenges for face alignment across large poses. To verify its effectiveness, we adopt *Pearson correlation coefficient* to measure the correlation between **confidence scores** $C(a,t)$ and **prediction errors** $|O(a,t) - \overline{O}(a,t)|_2$:

$$P = \frac{1}{N}\sum_{i=1}^{N}[\boldsymbol{r}^i_{a,t}(|O(a,t) - \overline{O}(a,t)|_2, C(a,t))] \tag{9}$$

Where $\boldsymbol{r}$ represents the calculation function of *Pearson correlation coefficient*. We conduct experiments on AFLW dataset and get P = -0.82, which means a strong negative correlation between them. In other words, anchor template with larger confidence score can achieve more accurate predicted landmarks. It can help filter prediction outliers and aggregate remaining predictions to mitigate the uncertainty of the localization result on a single anchor face. Due to the confidence score is defined as mathematical modeling of the distance between the anchor pose and the ground-truth pose, we can come to another conclusion that *closer* anchors tend to achieve more accurate localization, which also directly proves our search space split strategy based on anchor. Comparison details can be found in Section 4.5 and intuitive samples are also shown in Fig. 7.

4.5 Ablation study

In this section, we perform the ablation study for our proposed algorithm on the AFLW dataset, which is a challenging benchmark with large pose variations.

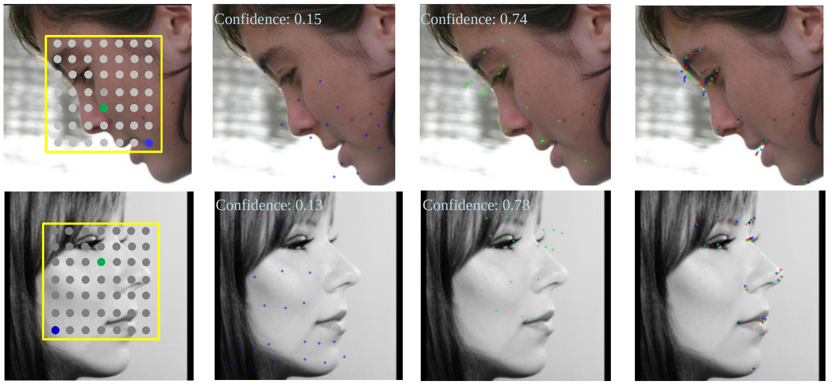

Fig. 7: Algorithm analysis based on different anchors. First column shows anchor grid settings, the green and blue anchors are two selected samples. Second and third column shows two anchor templates with prediction scores, which are randomly selected from the previous two anchors. The last column shows the final prediction results while the groud-truth is in red

More specifically, we divide the test set into four subsets according to the yaw dimension, i.e. Light ($0^\circ \sim 30^\circ$), Medium ($30^\circ \sim 60^\circ$), Large ($60^\circ \sim 90^\circ$), and Heavy ($90^\circ \sim$). Normalized mean error is utilized to evaluate the performance of our algorithm. Without explicitly specified, we use anchor templates as KMeans-24 (KMeans clustering to generate 24 anchor templates), anchor area as 56×56, anchor grid as $7{\times}7$, and the aggregating strategy is weighted average for ablation.

Comparison with the regression baseline. Table 6 compares the performance of our proposed approach with the baseline of direct regression on AFLW dataset. “Baseline” directly maps the discriminative features to the target landmark coordinates with ShuffleNet-V2 backbone. A fully connected layer with length $2L$ is used as the output of the baseline network. As shown in Table 6, our proposed anchor-based method significantly outperforms the baseline by a large margin across yaw variations.The improvements are attributed to two reasons. First, the anchor design can significantly reduce the search space and simplify the regression problem. Second, the aggregating of different anchors can further improve model robustness.

Comparison of various split configurations. Due to the challenges from the large-pose faces, we propose a set of anchor templates as references for regression to split search space. Split strategy consists of three hyper-parameters: anchor templates, anchor area, and anchor grid, as shown in Fig. 3.

Anchor template plays a voting role in our method as a reference for regression. As mentioned in Section 3.1, we get three basic template faces from the training dataset by hand design or KMeans clustering. Then we do some transformations to get more templates, corresponding to the pose variations in yaw, roll, and pitch dimension. By comparing KMeans-24 against HandDesign-24 in Table 7, KMeans is better than hand-design approach based on the same anchor number (24). The potential reason is that KMeans utilizes more data features to generate the base anchors, which should be more general compared with hand-designed based anchors. Besides, as shown in Table 7, 24 may be a good option for the number of anchor templates compared with 3 or 48 in our algorithm.

Table 6: A comparison of direct regression and anchor-based regression

Method	Full	Light	Medium	Large	Heavy
Baseline	1.67	1.46	1.92	1.99	2.13
AnchorFace	**1.56**	**1.40**	**1.74**	**1.80**	**1.96**

Table 7: Comparisons of different template settings

Template	Full	Light	Medium	Large	Heavy
Kmeans-3	1.60	1.43	1.80	1.83	2.10
Kmeans-24	**1.56**	**1.40**	**1.74**	**1.80**	**1.96**
Kmeans-48	1.58	1.41	1.79	1.82	2.06
HandDesign-24	1.58	1.42	1.76	1.84	2.00

Table 8: Comparisons of different anchor area settings

Anchor area	Full	Light	Medium	Large	Heavy
112 × 112	1.58	1.40	1.79	1.83	2.06
56 × 56	**1.56**	**1.40**	**1.74**	**1.80**	**1.96**
28 × 28	1.58	1.42	1.79	1.82	2.04
14 × 14	1.58	1.41	1.81	1.83	2.01

Table 9: Comparisons of different anchor grid settings

Anchor grid	Full	Light	Medium	Large	Heavy
3 × 3	1.58	1.40	1.78	1.82	2.17
5 × 5	1.57	1.40	1.77	1.82	2.05
7 × 7	**1.56**	**1.40**	**1.74**	**1.80**	**1.96**
13 × 13	1.56	1.40	1.75	1.81	2.07

Table 10: Comparisons of different aggregation strategies

Aggregate	Full	Light	Medium	Large	Heavy
Argmax	1.58	1.42	1.76	1.83	2.00
Weighted	**1.56**	**1.40**	**1.74**	**1.80**	**1.96**
Mean	1.61	1.43	1.80	1.89	2.22

Anchor area is the area where we set anchors in the image for the spatial domain. As the input image is cropped and resized to 224 × 224 and the face is around the image center, we set anchor points at a center area with size 14 × 14, 28 × 28, 56 × 56, 128 × 128, respectively. As shown in Table 8, 56 × 56 around the image center would be a good choice for putting anchors in the spatial domain.

Anchor grid defines how many anchors we set in the anchor area. For example, 7 × 7 means we sample 49 spatial points in a 7 × 7 grid from the anchor area to generate the anchor templates. As shown in Table 9, it is a good choice to set at 7 × 7.

Comparison of various aggregate strategies. To mitigate the uncertainty of the localization result on a single anchor face, we aggregate the predictions from different anchor templates. We introduce three aggregate strategies: Mean, Argmax, confidence weighted voting (Weighted). As shown in Table 10, aggregating the predictions with weighted confidences can obtain superior results compared with the argmax choice without aggregating. Besides, the confidence generated by the confidence branch is important if we compare the strategy of "Weighted" and "Mean".

5 Conclusions

In this paper, a novel split-and-aggregate strategy is proposed for large-pose faces. By introducing an anchor-based design, our proposed approach can simplify the regression problem by splitting the search space. Moreover, aggregating the prediction results contributes to reducing uncertainty and improving the localization performance. As validated on four challenging benchmarks, our proposed AnchorFace obtains state-of-the-art results with extremely fast inference speed.

References

1. Baltrušaitis, T., Robinson, P., Morency, L.P.: Continuous conditional neural fields for structured regression. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 593–608. Springer International Publishing, Cham (2014)
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). vol. 35, pp. 2930–2940 (December 2013)
3. Bhagavatula, C., Zhu, C., Luu, K., Savvides, M.: Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
4. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(9), 1063–1074 (Sep 2003). https://doi.org/10.1109/TPAMI.2003.1227983
5. Burgos-Artizzu, X.P., Perona, P., Dollar, P.: Robust face landmark estimation under occlusion. In: The IEEE International Conference on Computer Vision (ICCV) (December 2013)
6. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. International Journal of Computer Vision **107**(2), 177–190 (Apr 2014). https://doi.org/10.1007/s11263-013-0667-3, `https://doi.org/10.1007/s11263-013-0667-3`
7. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 109–122. Springer International Publishing, Cham (2014)
8. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition **41**(10), 3054 – 3067 (2008). https://doi.org/https://doi.org/10.1016/j.patcog.2008.01.024, `http://www.sciencedirect.com/science/article/pii/S0031320308000630`
9. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: BMVC (2006)

10. Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., Zafeiriou, S.: The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. International Journal of Computer Vision **127**(6), 599–624 (Jun 2019). https://doi.org/10.1007/s11263-018-1134-y, `https://doi.org/10.1007/s11263-018-1134-y`
11. Deng, J., Trigeorgis, G., Zhou, Y., Zafeiriou, S.: Joint multi-view face alignment in the wild. IEEE Transactions on Image Processing **28**, 3636–3648 (2017)
12. Dong, X., Yan, Y., Ouyang, W., Yang, Y.: Style aggregated network for facial landmark detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 379–388 (June 2018). https://doi.org/10.1109/CVPR.2018.00047
13. Dong, X., Yu, S.I., Weng, X., Wei, S.E., Yang, Y., Sheikh, Y.: Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
14. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9), 1627–1645 (Sep 2010). https://doi.org/10.1109/TPAMI.2009.167
15. Feng, Z., Kittler, J., Christmas, W., Huber, P., Wu, X.: Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3681–3690 (July 2017)
16. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks. arXiv e-prints arXiv:1711.06753 (Nov 2017)
17. Ikeuchi, K., Hebert, M., Delingette, H.: A spherical representation for recognition of free-form surfaces. IEEE Transactions on Pattern Analysis & Machine Intelligence **23**(07), 681–690 (jul 1995). https://doi.org/10.1109/34.391410
18. Jourabloo, A., Liu, X.: Large-pose face alignment via cnn-based dense 3d model fitting. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
19. Jourabloo, A., Ye, M., Liu, X., Ren, L.: Pose-invariant face alignment with a single cnn. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
20. Kahraman, F., Gokmen, M., Darkner, S., Larsen, R.: An active illumination and appearance (aia) model for face alignment. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–7 (June 2007). https://doi.org/10.1109/CVPR.2007.383399
21. Kumar, A., Chellappa, R.: Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
22. Kstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). pp. 2144–2151 (Nov 2011). https://doi.org/10.1109/ICCVW.2011.6130513
23. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 679–692. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
24. Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense face alignment. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2017)

25. Liu, Z., Zhu, X., Hu, G., Guo, H., Tang, M., Lei, Z., Robertson, N.M., Wang, J.: Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. ArXiv **abs/1903.10661** (2019)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
27. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3691–3700 (July 2017). https://doi.org/10.1109/CVPR.2017.393
28. Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
29. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: The European Conference on Computer Vision (ECCV) (September 2018)
30. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60**(2), 135–164 (Nov 2004). https://doi.org/10.1023/B:VISI.0000029666.37597.d3, `https://doi.org/10.1023/B:VISI.0000029666.37597.d3`
31. Messer, K., Matas, J., Kittler, J., Jonsson, K.: Xm2vtsdb: The extended m2vts database. In: In Second International Conference on Audio and Video-based Biometric Person Authentication. pp. 72–77 (1999)
32. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) Computer Vision – ECCV 2008. pp. 504–513. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
33. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 483–499. Springer International Publishing, Cham (2016)
34. Qian, S., Sun, K., Wu, W., Qian, C., Jia, J.: Aggregation via Separation: Boosting Facial Landmark Detector With Semi-Supervised Style Translation p. 11
35. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(1), 121–135 (Jan 2019). https://doi.org/10.1109/TPAMI.2017.2781233
36. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment via regressing local binary features. IEEE Transactions on Image Processing **25**(3), 1233–1245 (March 2016). https://doi.org/10.1109/TIP.2016.2518867
37. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: 2013 IEEE International Conference on Computer Vision Workshops. pp. 397–403 (Dec 2013). https://doi.org/10.1109/ICCVW.2013.59
38. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. 2013 IEEE International Conference on Computer Vision Workshops pp. 397–403 (2013)
39. Saragih, J., Goecke, R.: A nonlinear discriminative approach to aam fitting. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8 (Oct 2007). https://doi.org/10.1109/ICCV.2007.4409106

40. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2013)
41. Trigeorgis, G., Snape, P., Nicolaou, M.A., Antonakos, E., Zafeiriou, S.: Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4177–4187 (June 2016). https://doi.org/10.1109/CVPR.2016.453
42. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: CVPR (2018)
43. Wu, W., Yang, S.: Leveraging intra and inter-dataset variations for robust face alignment. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
44. Xiao, S., Feng, J., Liu, L., Nie, X., Wang, W., Yan, S., Kassim, A.: Recurrent 3d-2d dual learning for large-pose facial landmark detection. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
45. Xiong, F., Zhang, B., Xiao, Y., Cao, Z., Yu, T., Zhou Tianyi, J., Yuan, J.: A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In: Proceedings of the IEEE Conference on International Conference on Computer Vision (ICCV) (2019)
46. Xiong, X., De la Torre, F.: Global supervised descent method. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
47. Xu, X., Kakadiaris, I.A.: Joint head pose estimation and face alignment framework using global and local cnn features. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017). pp. 642–649 (May 2017). https://doi.org/10.1109/FG.2017.81
48. Yang, H., Jia, X., Loy, C.C., Robinson, P.: An empirical study of recent face alignment methods. CoRR **abs/1511.05049** (2015), `http://arxiv.org/abs/1511.05049`
49. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
50. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: 2013 IEEE International Conference on Computer Vision. pp. 1944–1951 (Dec 2013). https://doi.org/10.1109/ICCV.2013.244
51. Zadeh, A., Chong Lim, Y., Baltrusaitis, T., Morency, L.P.: Convolutional experts constrained local model for 3d facial landmark detection. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2017)
52. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The menpo facial landmark localisation challenge: A step towards the solution. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 2116–2125 (2017)
53. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 1–16. Springer International Publishing, Cham (2014)
54. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (Oct 2016). https://doi.org/10.1109/LSP.2016.2603342

55. Zhang, Z., Luo, P., Change Loy, C., Tang, X.: Learning Deep Representation for Face Alignment with Auxiliary Attributes. arXiv e-prints arXiv:1408.3967 (Aug 2014)
56. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 94–108. Springer International Publishing, Cham (2014)
57. Zhu, M., Shi, D., Zheng, M., Sadiq, M.: Robust facial landmark detection via occlusion-adaptive deep networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
58. Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3409–3417 (June 2016). https://doi.org/10.1109/CVPR.2016.371
59. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4998–5006 (2015)
60. Zhu, S., Li, C., Loy, C.C., Tang, X.: Unconstrained face alignment via cascaded compositional learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
61. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2879–2886 (June 2012). https://doi.org/10.1109/CVPR.2012.6248014
62. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. pp. 2879–2886 (06 2012). https://doi.org/10.1109/CVPR.2012.6248014
63. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)