Enhanced Normalized Mean Error loss for Robust Facial Landmark detection

Shenqi Lai
laishenqi@meituan.com
Zhenhua Chai
chaizhenhua@meituan.com
Shengxi Li
lishengxi@meituan.com
Huanhuan Meng
menghuanhuan02@meituan.com
Mengzhao Yang
yangmengzhao@meituan.com
Xiaoming Wei
weixiaoming@meituan.com

Vison and Image Center (VIC) of Meituan
Beijing, China

Abstract

Normalized Mean Error (NME) is one of the most popular evaluation metrics in facial landmark detection benchmark. However, the commonly used loss functions (L1 and L2) are not designed to optimize NME directly, and thus there might be a gap between optimizing the distance losses for regressing the parameters of landmark coordinates and minimizing this metric value. In this paper, we will try to address this issue, and propose a novel loss function named Enhanced Normalized Mean Error (ENME) loss, which will consider both the final metric and the attention mechanism for different NME intervals. In order to evaluate the effectiveness of our proposed loss, we design and train a light-weight regressing model we call Thin Residual Network (TRNet). Extensive experiments are conducted on three popular public datasets such as AFLW, COFW and challenging 300W, and the results show that TRNet when trained with the Enhanced NME loss will exhibit better performance than the state of the art methods.

1 Introduction

Facial landmark detection aims to locate the coordinates of a set of predefined key points on human faces, which is an essential step for some important applications such as face verification [22] and face emotion recognition [5]. In the past few years, several regression based methods have been proposed to deal with this task. For instance, Supervised Descent Method (SDM) [24] is proposed to learn the mapping from feature space to coordinates by minimizing a Non-linear Least Square function. In Ensemble of Regression Trees (ERT) [3], GBDT is explored in order to better estimate the coordinates while keeping a real-time speed during inference. Light-weight LBF [17] used as the regression features can run much faster

(a)SDM

(b)ERT

(c)Our

Figure 1: The performance of SDM, ERT and the proposed method under some challenging cases.

(e.g. in 3000fps) but at the cost of reduced accuracy. All these methods can still achieve good performance even in some complex environment (the top three columns in Figure 1). However, in some unconstrained scenarios the test cases could be even more challenging (e.g. the last two columns in Figure 1 with large pose, partial occlusion, large expression and changing illumination), the performance will degrade dramatically. The reason could be due to the limitation of the used handcrafted features. Thanks to the development of deep learning and establishing of large scale public available databases [12, 15], deep neural networks could be a good solution to this task. Cascade CNN [21] is one of the early works which use deep learning to facial landmark detection. The model is trained in a coarse to fine manner and the results show its superiority over the previous works. The TCDCN [26] and FLD [27] also use deep learning based structure, but they do not treat the facial landmark detection as a stand alone problem. Instead, they train the landmark detection model together with the facial attribute prediction, and they find that the auxiliary attributes will help the convergence of landmark detection especially at the beginning of the training procedure. Tweaked CNNs (TCNN) [23] add an unsupervised clustering after the network output, in this way the problem can be divided according to some semantic attributes (e.g. the head pose or facial attribute). The specific parameters to each pose are further finetuned in each sub problem. DenseU-Net [8] explores the information of adjacent layers from different scales, and the performance can be further improved. Most of the methods mentioned above aim to propose a more robust model and we believe the design of a suitable loss function could also bring some good effects.

Normalized Mean Error (NME) is usually used to measure the performance for facial landmark detection. However, it can be found in Fig. 2 that the two predictions (denoted as red and orange) by two different models will have the same L2 or L1 values while the NME values are still different, which motivates us to introduce the proposed NME loss. Besides, inspired by the recent proposed Wing loss [7], even under the same network structure, the models trained with a well designed loss function can achieve much better performance. The

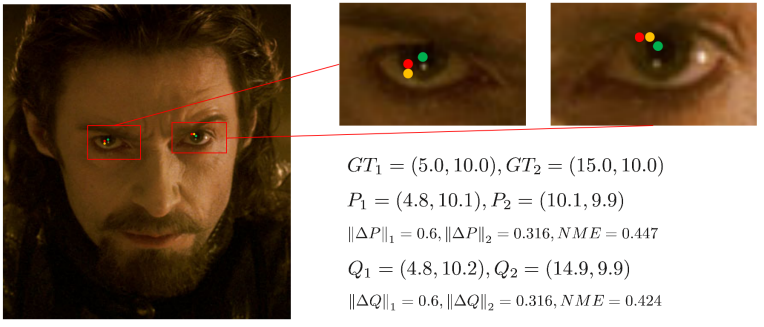

Figure 2: $\|\cdot\|_1$ denotes L_1 norm distance, and $\|\cdot\|_2$ denotes L_2 norm distance. GT(green) is the ground truth. For two different points P(red) and Q(orange), they have same norm distances, but their NME values are very different.

core innovation is that more attention should be paid to the samples with small or medium range NME values during the training process. However the Wing loss has some drawbacks, e.g. there are two hyper parameters needed to tune and even in the same task the best parameters for CNN-6 and CNN-7 in [7] could be totally different, the search for the best parameters in quadratic space could be exhausting. In this paper, we propose a novel loss function named Enhanced NME loss, which has two good properties: 1) it inherits the advantages of Wing loss and has only one hyper parameter to tune which is easier to train; 2) a joint Enhanced version which will consider the NME in the loss function has the potential to achieve even better performance.

The rest part of this paper is organized as follows. Section 2 will introduce the proposed NME loss and the enhanced version. Section 3 is about the design details of Thin Residual Network (TRNet) and the novel Refined Module (RM). The extensive experiments have been conducted and the results will be discussed in Section 4. Finally in Section 5 the conclusion will be drawn.

2 The proposed loss function

2.1 NME Loss

The design of a proper loss function is important to CNN based facial landmark detection. In the past, the L2 loss has been used in deep-neural-network-based facial landmarking systems. Its equation and derivative can be written as:

$$L2(\Delta x, \Delta y) = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(\Delta x_i^2 + \Delta y_i^2)}, \tag{1}$$

$$\frac{\partial L2(\Delta x, \Delta y)}{\partial \Delta x_n} = \frac{1}{N}\frac{\Delta x_n}{\sqrt{\sum_{i=1}^{N}(\Delta x_i^2 + \Delta y_i^2)}}, \tag{2}$$

$$\frac{\partial L2(\Delta x, \Delta y)}{\partial \Delta y_n} = \frac{1}{N}\frac{\Delta y_n}{\sqrt{\sum_{i=1}^{N}(\Delta x_i^2 + \Delta y_i^2)}}, \tag{3}$$

where Δx and Δy are the deviations of the predict coordinates from ground truth values in x-axis and y-axis respectively. It is wellknown that the L2 loss is sensitive to outliers, and we can find from the formulae above that the predictions from different points will affect each other. If there exist outliers, the results will be even worse. Besides, the L2 loss will not reflect the NME directly, which has been discussed in previous section and has been shown in Fig. 2. Similar case will also happen for L1 case.

The original NME metric (Eq. 4) and the proposed NME loss (Eq. 5) with its derivatives (Eq. 6 and 7) are listed as follows:

$$NME = \frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{\Delta x_i^2+\Delta y_i^2}}{d}, \tag{4}$$

$$NME(\Delta x,\Delta y) = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\Delta x_i^2+\Delta y_i^2}, \tag{5}$$

$$\frac{\partial NME(\Delta x,\Delta y)}{\partial \Delta x_n} = \frac{1}{N}\frac{\Delta x_n}{\sqrt{\Delta x_n^2+\Delta y_n^2}}, \tag{6}$$

$$\frac{\partial NME(\Delta x,\Delta y)}{\partial \Delta y_n} = \frac{1}{N}\frac{\Delta y_n}{\sqrt{\Delta x_n^2+\Delta y_n^2}}, \tag{7}$$

where d denotes the normalized term and N is the number of facial landmarks. Δx_i and Δy_i are deviations between the i_{th} predicted landmark and ground truth in x-axis and y-axis.

It's noteworthy that we do not use the normalized term d in our proposal because its definition is not unified for different datasets, e.g. in AFLW [12] the d is the image width while in some datasets like COFW [2] it's inter-pupil distance. As to the proposed NME loss, the main difference from the L2 loss is obvious. We can find from Eq. 6 and 7 that the gradient computing for weight updating is relatively independent for a single landmark, in this way the weight updating for different landmark points will not affect each other and for some hard point it could probably lead to better training results.

2.2 Enhanced NME Loss

According to the comprehensive analysis in [7], when trained under the commonly used loss functions (e.g. L1, L2 and smooth L1), CNN based models for facial landmark detection will perform well where the results only contain a small fraction with large NME value. But when referring to the small or medium NME, L1 and smooth L1 will perform much better than L2. This indicates that the training of deep neural networks should pay more attention to the samples with small or medium range errors. The Wing loss is thus proposed and designed as a piecewise function, which acts like log function (with an offset) when NME is small and acts like L1 function when NME is large. In this way, the definition of the Wing loss can be written as:

$$Wing(\Delta p) = \begin{cases} \omega \ln(1+|\Delta p|/\varepsilon), & \text{if } |\Delta p| < \omega \\ |\Delta p| - C, & \text{otherwise} \end{cases}, \tag{8}$$

where the non-negative ω sets the range of the nonlinear part to $(-\omega,\omega)$, ε limits the curvature of the nonlinear part, and $C = \omega - \omega\ln(1+\omega/\varepsilon)$ is a constant.

However, the Wing loss has some shortcomings. First, the Wing loss is a piecewise function, which will make the computation of the forward and backward propagation complicated

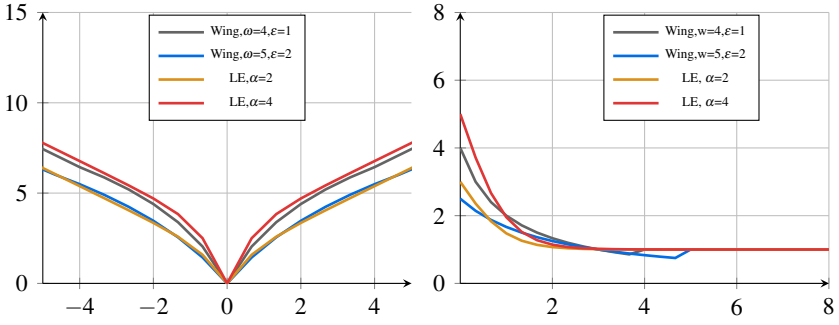

Figure 3: Plots of the Wing loss and LE loss functions with respect to mean errors (left) and their derivative curves (right) under different parameter settings.

and can not be concluded in a unified way. Then, the Wing loss has two hyper parameters and even in the same task the best parameters for CNN-6 and CNN-7 in [7] could be remarkably different, the search for the best parameters in quadratic space will require large computation and time. Therefore, in this paper we propose a novel loss function named Log Enhanced loss (LE loss), which has only one hyper-parameter and at the time maintains the same merits as the Wing loss behaves. The LE loss is defined as:

$$LE(\Delta p) = (\alpha+1) * |\Delta p| - \alpha * ln(\frac{e^{|\Delta p|} + e^{-|\Delta p|}}{2}). \tag{9}$$

To be detailed, LE loss function (Figure. 3 left) is designed as an even function and is composed of two parts. The derivative of the former part is a constant, which is used to make the minimal derivative value close to 1 when the mean error is large. While the latter part is a composite function, its derivative is actually tanh function whose value ranges from -α to α. So when combining them together (Figure. 3 right), with the increase of the mean error, the derivative of LE function will decrease and finally approach to 1, which will strengthens the role of samples with small and medium errors. What's more, the α is used to control the curvature of the function. Besides, in Figure. 3 (left) it can be easily found that the Wing loss with $\omega = 4$, $\varepsilon = 1$ looks close to LE loss with $\alpha = 4$, and the shape of the Wing loss with $\omega = 5$, $\varepsilon = 2$ is similar to LE loss with $\alpha = 2$. The Wing loss with different hyper parameters could be well approximated by LE loss by adjusting the parameter α. Finally, the LE loss and NME loss can be reformulated as a joint version which we named Enhanced NME (ENME) loss. The equation for the joint loss can be written as:

$$ENME(\Delta x, \Delta y) = (\alpha+1) * NME(\Delta x, \Delta y) - \alpha * ln(\frac{e^{NME(\Delta x,\Delta y)} + e^{-NME(\Delta x,\Delta y)}}{2}). \tag{10}$$

3 Thin Residual Network

The network architectures play an important role to the facial landmark detection task. In [7], when replacing the backbone CNN-6/7 network with ResNet50 [9], the performance has been further improved by around 10%. Unfortunately, large capacity networks will bring some disadvantages, e.g. making the model size large and slowing down the application

which could be unacceptable in some realtime cases. Based on the discussion above, we propose to use a compressed version of residual network. First, three convolutional layers are used to extract feature maps, and only convolutional kernel sized in 3x3 is used. Then, four residual blocks are followed and each of them is with a max pooling layer, which can be used to enhance the nonlinearity and can probably deal with the challenging cases (e.g. image rotation, changing illumination, occlusion and deformation). After that we use global average pooling layer to further reduce the parameter volume and finally a fully connected layer is used for computing the regression. In order to control the model size, we strictly control the number of the convolutional kernels used in both feature extraction layer and each residual unit. So the final network structure will look much thinner than the original ResNet, this is the reason we name it Thin Residual Network (TRNet). The details can be found in Table 1.

Numbers	Type	Filters	Size	Stride	Output
3	Conv	32	3*3	1	64*64
1	Max Pool	32	2*2	2	32*32
2	Conv	48	3*3	1	-
	Conv	48	3*3	1	-
	Residual	-	-	-	32*32
1	Max Pool	48	2*2	2	16*16
2	Conv	64	3*3	1	-
	Conv	64	3*3	1	-
	Residual	-	-	-	16*16
1	Max Pool	64	2*2	2	8*8
2	Conv	80	3*3	1	-
	Conv	80	3*3	1	-
	Residual	-	-	-	8*8
1	Max Pool	80	2*2	2	4*4
2	Conv	96	3*3	1	-
	Conv	96	3*3	1	-
	Residual	-	-	-	4*4
1	Avg Pool	-	Global	4	1*1
	Connected	-	96	-	96

Table 1: TRNet-20

Some existing methods (e.g. [7]) make use of the global structure information of facial landmarks, which can still predict the correct outputs even when there are occlusions on the faces and even some of the landmarks become invisible. However, when the diversity of the training set is relatively limited, the trained model will have some drawbacks. For instance, if there is one model which can correctly deal with cases for sunglasses occlusion or scarf occlusion separately, but when handling a test case with both occlusions the model could probably fail to predict. Besides the data augmentation, two-stage cascade structure is a usual way to solve this issue, but as we mentioned before the cascade structure will make the model size almost twice (e.g. in [7] CNN-7 in the second stage is even larger than the CNN-6 in the first stage) which will limit its use especially in some mobile cases.

Another way to improve the robustness of the model is the divide-and-conquer approach. In [20], the Multi Center Learning (MCL) [20] is proposed to decompose the whole face shape into several patches, and then predicts the location of facial landmarks from each

local patch independently. In this way, the prediction will not be affected by the global correlation and still keep the structure information in each local area. In this paper, we will take advantage of both global and local information. At first, one backbone network will be trained as usual and the global prediction will be obtained. Then, we will use the local refined module. To be detailed, we will keep all the parameters fixed in the global model, and retrain the parameters in the full connected layer in a local way. It can be found that the facial contour does not contain obvious corner so the landmarks located on these area will probably have large NME values, and interestingly the local refined module will further improve the performance. The detailed results will be shown in next section.

4 Experiments

4.1 Datasets

Three popular datasets were used to evaluate the performance of our proposed method. AFLW [12] is one of the largest datasets, which contains 20,000 images for training and 4,386 images for testing. For each image 19 facial landmarks have been annotated manually. While COFW [2] is a relatively smaller dataset, there are 1345 images for training and 507 images for testing respectively. All the images are captured in realistic conditions, and there exists some faces with heavy occlusions due to the use of accessories such as sunglasses and hats and interactions with objects (e.g. food). Hence it will require the algorithm be capable of dealing with the occlusion cases. For each image there are 29 manually annotated facial landmarks. 300W [15] could be the more complex one, which is made up of several sub datasets including LFPW[1], HELEN[13], AFW[29] and IBUG[19]. According to the protocol, the full set of AFW and the training subsets of LFPW and HELEN are used as the training set which contains 3148 samples in all, and the test set includes the test subsets of LFPW and HELEN as well as 135 extra face images from IBUG. The final size of the test set is 689. What is more important, different from the former two datasets, the annotation of 300W [15] has been conducted in a semi-automatic way and 68 landmarks are annotated in all, which include the landmarks located on the facial contour.

It is worthy to mention that in order to have a fair comparison with the methods above we strictly follow the commonly used protocol for each dataset. We adopt the widely used Normalised Mean Error (NME) as the measurement, and the mean error larger than 10% will be treated as a failure. For the AFLW dataset [12], the AFLW-Full protocol [28] is used. The face bounding box of each test sample is given and the shape is square. Finally, the width of the face bounding box is used as the normalisation term. For COFW dataset, the situation will have some differences, e.g. inter-ocular distance is used as the the normalization term. For the 300W dataset, rather than using the outer eye corner distance as the normalization term, we uses the inter-pupil distance which is exactly the same as [7, 18].

4.2 Ablation study

In order to show the effectiveness of the proposed method, we conduct extensive experiments on different combinations of backbone networks and loss functions. We firstly train the existing CNN-6 network [7] with the commonly used L1, L2 and the proposed NME loss separately. In Tab. 2, we can find the model trained with NME loss will surpass the model using traditional losses by a large margin. Then we still use the same network structure and

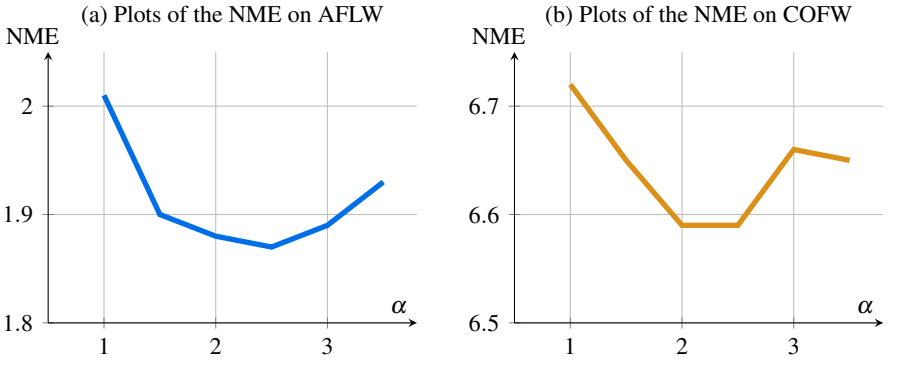

Figure 4: The performance of CNN-6 with LE for different α on AFLW and COFW.

test the attention based losses. It is easy to find in Fig. 4 that the best hyper parameter α in LE is relatively stable for different datasets, and the performance is comparable to the model with Wing loss. After that we train the model with the Enhanced NME, the performance can be further improved and the result show that our proposed method outperform the Wing loss based method. Finally, we replace the CNN-6 with our proposed TRNet and refined module (RM). Even though the used model size is less than one third of the CNN-6, our method can still achieve better performance. The reason could be that our method is benefited from the increase of the network depth.

Method	AFLW	COFW
CNN-6 + L2	2.41	7.03
CNN-6 + L1	2.00	6.70
CNN-6 + NME	1.88	6.60
CNN-6 + Wing	1.88	6.59
CNN-6 + LE	1.87	6.59
CNN-6 + ENME	1.83	6.49
TRNet-20 + ENME	1.60	5.48
TRNet-20 + ENME + RM	1.57	5.39

Table 2: The performance of methods with different networks and losses.

Method	COFW
ESR [3]	11.2
TCDCN [26]	8.05
MCNet [20]	6.08
DU-Net [8]	5.55
TRNet-20 + ENME + RM	5.39

Table 3: A comparison of the proposed method with state-of-the-art approaches on the COFW dataset in terms of the NME.

4.3 Comparison with the state of the art methods

In order to show the effectiveness of the proposed method, we have compared our proposed method with the state of the art methods on AFLW, COFW and the challenging 300W. The experimental results show that our proposed method is comparable with most of the state of the art methods and only performs a little worse than the hourglass based method RF-CHN especially in some challenging cases. The details can be found in Tab. 3, Tab. 4 and Tab. 5. Besides, our model size is only 4 MB and suitable for some mobile applications. We have test our model on cellphone released three year before whose CPU is Snapdragon 820, and 120 fps can be achieved in average.

Method	AFLW
DSRN [16]	1.86
SAN [6]	1.85
AAN [25]	1.73
FEC [10]	1.70
Wingloss [7]	1.65
TRNet-20 + ENME + RM	1.57

Table 4: A comparison of the proposed method with state-of-the-art approaches on the AFLW dataset in terms of the NME.

Method	Com.	Challenge	Full
ESR [3]	5.28	17.00	7.58
MCNet [20]	-	8.87	-
LDFFA [11]	4.10	9.99	5.26
CVAE [14]	4.96	8.87	5.22
AAN [25]	4.38	9.44	5.39
DSRN [16]	4.12	9.68	5.21
RF-CHN [4]	4.18	7.39	4.81
TRNet-20 + ENME + RM	3.86	7.81	4.63

Table 5: A comparison of the proposed approach with the state-of-the-art approaches on the 300W dataset in terms of the NME.

5 Conclusion

In this paper, we analyze the mismatch problem between the commonly used L2 loss and the NME metric and introduce a novel loss function named Enhanced NME (ENME) loss. We have shown that the proposed loss function has some appealing properties which contains only one hyper parameter and can be directly used to optimize the NME metric. In order to show the effectiveness of the proposed method, a light-weight network structure named Thin Residual Network has been designed, the size of which is almost ten percent of the traditional CNN-6. Besides, a novel one stage refined module is used, which will only increase very limited parameters while the existing cascade structure will make the model size twice. Our method is comparable with the state of the art methods in facial landmark detection, and we believe ENME loss will also be useful in other regression related tasks (e.g. generic object detection and tracking), which will be our future work.

References

[1] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

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

[3] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[4] Weiliang Chen, Qiang Zhou, and Roland Hu. Face alignment by combining residual features in cascaded hourglass network. In *ICIP*, 2018.

[5] Ciprian Adrian Corneanu, Marc Oliu Simon, Jeffrey F. Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE TPAMI*, 38(8): 1548–1568, 2016.

[6] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, 2018.

[7] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *CVPR*, 2018.

[8] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[10] Zhenliang He, Meina Kan, Jie Zhang, Xilin Chen, and Shiguang Shan. A fully end-to-end cascaded CNN for facial landmark detection. In *FG*, 2017.

[11] Ankit Jalan, Siva Chaitanya Mynepalli, Viswanath Veera, and Shankar M. Venkatesan. Low dimensional deep features for facial landmark alignment. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 2612–2616, 2017.

[12] 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 *IEEE ICCV Workshops*, 2011.

[13] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir D. Bourdev, and Thomas S. Huang. Interactive facial feature localization. In *ECCV*, 2012.

[14] Shuying Liu, Yipeng Huang, Jiani Hu, and Weihong Deng. Learning local responses of facial landmarks with conditional variational auto-encoder for face alignment. In *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*, pages 947–952, 2017.

[15] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard G. Medioni. Do we really need to collect millions of faces for effective face recognition? In *ECCV*, 2016.

[16] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *CVPR*, 2018.

[17] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 FPS via regressing local binary features. In *CVPR*, 2014.

[18] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment via regressing local binary features. *IEEE TIP*, 25(3):1233–1245, 2016.

[19] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Elsevier Image and Vision Computing*, 47:3–18, 2016.

[20] Zhiwen Shao, Hengliang Zhu, Yangyang Hao, Min Wang, and Lizhuang Ma. Learning a multi-center convolutional network for unconstrained face alignment. In *ICME*, 2017.

[21] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[22] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[23] Yue Wu, Tal Hassner, KangGeon Kim, Gérard G. Medioni, and Prem Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE TPAMI*, 40(12): 3067–3074, 2018.

[24] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[25] Lei Yue, Xin Miao, Pengbo Wang, Baochang Zhang, Xiantong Zhen, and Xianbin Cao. Attentional alignment networks. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 2018.

[26] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[27] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE TPAMI*, 38(5):918–930, 2016.

[28] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *CVPR*, 2016.

[29] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012.