The 2nd 106-Point Lightweight Facial Landmark Localization Grand Challenge

Yinglu Liu[1], Peipei Li[1], Xin Tong[1], Hailin Shi[1(✉)], Xiangyu Zhu[2], Zhenan Sun[2], Zhen Xu[3], Huaibo Liu[3], Xuefeng Su[3], Wei Chen[3], Han Huang[4], Duomin Wang[4], Xunqiang Tao[4], Yandong Guo[4], Ziye Tong[5], Shenqi Lai[5], and Zhenhua Chai[5]

[1] JD AI Research, Beijing, China
{liuyinglu1,lipeipei32,tongxin,shihailin}@jd.com
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China
{xiangyu.zhu,znsun}@nlpr.ia.ac.cn
[3] SogouAI, Beijing, China
{xuzhen216234,liuhuaibiao,SuXueFeng,chenweibj8871}@sogou-inc.com
[4] OPPO Research Institute, Beijing, China
huangh92@gmail.com, wangduomin@gmail.com, taoxunqiang@gmail.com, yandong.guo@live.com
[5] Vision Intelligence Center, Meituan, Beijing, China
{tongziye,laishenqi, chaizhenhua}@meituan.com

Abstract. Facial landmark localization has been applied to numerous face related applications, such as face recognition and face image synthesis. It is a very crucial step for achieving high performance in these applications. We host the 2nd 106-point lightweight facial landmark localization grand challenge in conjunction with ICPR 2020. The purpose is to make effort towards benchmarking lightweight facial landmark localization, which enables efficient system deployment. Compared with the 1st grand challenge (https://facial-landmarks-localization-challenge.github.io/), the JD-landmark-v2 dataset contains more than 24,000 images with larger variations in identity, pose, expression and occlusion. Besides, strict limits of model size (≤20M) and computational complexity (≤1G Flops) are employed for computational efficiency. The challenge has attracted attention from academia and industrial practitioners. More than 70 teams participate in the competition, and nine of them involve in the final evaluation. We give a detailed introduction of the competition and the solution by the winners in this paper.

Keywords: Facial landmark localization · Lightweight · Challenge

1 Overview

Facial landmark localization, also known as facial landmark detection, is to locate a set of predefined facial fiducial points on facial images. It has been successfully

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12668, pp. 327–338, 2021.
https://doi.org/10.1007/978-3-030-68793-9_25

applied to many face related applications. For example, facial landmarks are usually employed to implement face alignment for the recognition tasks including face identity recognition, facial expression recognition, facial attribute recognition, *etc.* The landmark-based alignment is crucial for the high recognition performance. Besides, facial landmarks are also taken as features for face manipulation such as face aging, face swapping, face cartoonization and face attribute editing. Furthermore, the methods of facial landmark localization can also be applied on other fields such as pose estimation [10,23]. Recent years, with the popularity of the internet and smart phones, more and more face-related applications are implemented on the mobile device, thus the lightweight models are required to enable efficient system deployment. However, the prior competitions of facial landmark localization (*i.e.,* 300-W [1,24], Menpo [2,34], 300-VW [3,26] and JD-landmark [4,20]) focus only on accuracy, without consideration on the efficiency. To push the frontier of the lightweight facial landmark localization algorithm, we host the 2nd 106-point lightweight **F**acial **L**andmark **L**ocalization **C**hallenge (FLLC[1]) in conjunction with the 25$^{\text{th}}$ International Conference on Pattern Recognition (ICPR2020). We extend the JD-landmark dataset in the 1$^{\text{st}}$ challenge with thousands of in-the-wild facial images. The upgraded JD-landmark-v2 dataset contains more than 24,000 images. Figure 2 shows some examples in this dataset. The challenge has attracted much attention from both academia and industrial practitioners. We will introduce the details of the challenge along with the methods of the winner teams in this paper (Fig. 1).

Fig. 1. Example images of the grand challenge dataset.

2 Related Work

In order to provide a fair comparison between the different methods of automatic facial landmark localization, the Intelligent Behaviour Understanding Group (I·BUG[2]) from Imperial College London held a series of competitions, including

[1] https://fllc-icpr2020.github.io/home/.
[2] https://ibug.doc.ic.ac.uk/home.

2D/3D facial landmark localization in static imagery and 2D/3D facial landmark tracking in videos. The annotated data has been used by the academia and industrial community for training and testing facial landmark localization models. Before presenting FLLC, we outline the previous competitions along with the related datasets.

2.1 Competitions

300-W Challenge. The first Automatic Facial Landmark Detection in-the-Wild Challenge (300-W Challenge [1,24]) is held in conjunction with ICCV 2013 in Sydney, Australia. It was the first event to benchmark the efforts in the facial landmark localization field. The competition provides 4,350 "in-the-wid" images with around 5,000 faces. All the faces are annotated using a 68-landmark frontal face mark-up scheme as Multi-PIE [5].

300-VW Challenge. In conjunction with ICCV 2015, Zafeiriou *et al.* held the 300 Videos in the Wild (300-VW [3,26]) challenge. The purpose is to develop a comprehensive benchmark for evaluating in-the-wild facial landmark tracking algorithms. The competition collects a large number of long face videos recorded in the wild. Each video has a duration of about 1 min. (at 25–30 FPS). In total, the 300-VW benchmark consists of 114 videos and 218,595 frames. All frames have been annotated with regards to the same 68 points mark-up used in the 300-W competition.

Menpo Challenge. The 300-W and 300-VW challenges have two limitations: 1) lack of faces in extreme poses; 2) limited test images (around 600). To address these issues, the Menpo [2,34] challenge is held in conjunction with CVPR 2017. It consists of 5,658 semi-frontal and 1,906 profile facial images in the training set, and 5,335 frontal and 1,946 profile facial images in the test set. Besides, the 68-point mark-up scheme is used for frontal faces while a 39 points mark-up scheme is adopted for profiles.

3D Menpo Challenge. The I·BUG held the 3D Menpo Challenge [6,33] in conjunction with ICCV 2017 to develop a comprehensive benchmark for evaluating 3D facial landmark localization algorithms in the wild in arbitrary poses. They fitted all the 2D faces provided by the 300-W and Menpo challenges with the state-of-the-art 3D facial morphable models. They also provided 3D facial landmarks for all the videos of 300-VW competition.

106-Point Facial Landmark Localization Challenge. As mentioned above, many efforts have been made for the 68-point facial landmark localization. However, the 68-point landmarks are incompetent to depict the detailed structure of facial components. For example, the lower boundary of eyebrows and the wing of nose are out of the definition in 68-point landmarks, while they are important

in some cases such as face parsing [21]. To overcome this problem, a challenging dataset (named as JD-landmark) is constructed and employed for the competition [4,20] of 106-point facial landmark localization in conjunction with ICME 2019.

2.2 Datasets

Large amount of annotated data are important for training the high performance landmark localization model, especially for the deep learning based methods. We summary the commonly used 2D facial landmark datasets in static images as follows.

LFPW. The Labeled Face Parts in the Wild (LFPW [7,11]) dataset consists of 1,432 face images downloaded from the internet using simple text queries on sites such as google.com, flickr.com, and yahoo.com. Each image was labeled by three MTurk workers with 29 fiducial points.

HELEN. The HELEN [8,19] dataset collected images from the Flickr. It contains 2,330 images of high resolution. Each image is annotated with 194 points. It is also extended to a face parsing benchmark [27].

AFW. The Annotated Face in-the-Wild (AFW [35]) dataset is also built using Flickr images. It includes 205 images with 473 labeled faces. For each image, six landmarks along with the pose angles and a rectangular bounding box are provided.

AFLW. The Annotated Facial Landmarks in the Wild (AFLW [9]) dataset provides a large-scale collection of images gathered from Flickr. It consists of 25,993 faces in 21,997 real-world images, each of them is annotated with up to 21 landmarks.

300-W. The training images of the 300-W dataset [1] consists of the LPFW, AFW, Helen and XM2VTS datasets. Each image is re-annotated using the 68-point markup as the landmark configuration of MultiPIE. Besides, a new dataset (IBUG), which includes 135 images with large variations in expression, illumination conditions and pose, are released as part of 300W dataset. The test set consists of 300 images captured indoors and 300 images captured outdoors.

Menpo. The training set of Menpo dataset [2] consists of 5,658 semi-frontal and 1,906 profile facial images. The test set contains 5,335 frontal and 1,946 profile facial images. The frontal/semi-frontal images employ the same landmark configuration of 300W with 68 points, while the profile facial images are annotated with a 39 profile landmark scheme. All the images are taken from LFW and FDDB datasets.

WFLW. The Wider Facial Landmarks in-the-wild (WFLW [32]) contains 10,000 faces (7,500 for training and 2,500 for testing) with 98 fully manual annotated landmarks. Apart from the landmark annotation, this dataset provides several attribute annotations, *i.e.,* occlusion, pose, make-up, illumination, blur and expression for comprehensive analysis of existing algorithms.

JD-landmark. The JD-landmark dataset [4] is an incremental dataset based on 300W, composed of LFPW, AFW, Helen and IBUG, and re-annotated with the 106-point mark-up. The dataset contains 11,393 face images for training. Besides, 2,000/2,000 facial images are collected from the open-source face dataset Megaface as validation/test set. The JD-landmark covers a large variation of pose, illumination and expression.

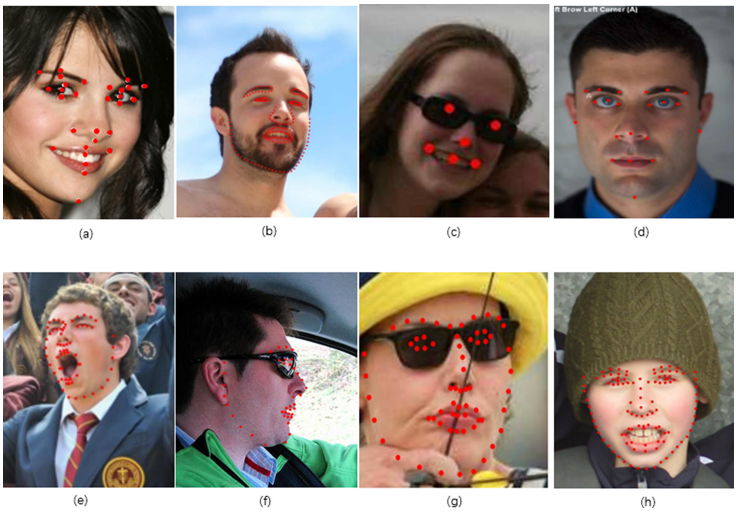

Fig. 2. Example images of the 2D facial landmark localization datasets. (a) LFPW; (b) HELEN; (c) AFW; (d) AFLW; (e) 300-W; (f) Menpo; (g) WFLW; (h) JD-landmark.

3 Introduction of Competition

3.1 Datasets

Compared with the first version of JD-landmark dataset, we expand it by about 9,000 in-the-wild facial images, which are collected from the Megaface dataset [15]. Each sample is annotated with 106-point landmarks. Expect for

the facial images in extreme poses and expressions, many low quality (low resolution) images are added to increase the difficulty of the competition. In total, the second version, *i.e.* JD-landmark-v2 dataset consists of 20,386 images for training, 2,000 images for validation and 2,000 images for testing. Each image is provided with the 106-point landmarks along with the referenced bounding box.

3.2 Evaluation Criteria

The submissions are ranked according to the Area-Under-the-Curve (AUC) from the Cumulative Errors Distribution (CED) curves. Furthermore, the statistics from the CED curves such as the failure rate and average Normalized Mean Error (NME) are also taken into account. The CED curve reflects the proportion of the test images with regard to the NME less than a threshold α. The AUC is the area under the CED curve calculated up to the threshold, then divided by the threshold α. In this competition, we set the value of α to 0.08. Similarly, we regard each image with a NME larger than α as a failure case. NME is computed as:

$$NME = \frac{1}{N} \sum_{k=1}^{N} \frac{\|y_k - \hat{y}_k\|_2}{d} \tag{1}$$

where k refers to the index of landmarks. y and $\hat{y}$ denotes the ground truth and the prediction of landmarks for a given facial image, respectively. In order to alleviate the bias in profile faces caused by the small interocular distance, we employ the square-root of the ground truth bounding box as the normalization factor d, computed as $d = \sqrt{w_{bbox} \times h_{bbox}}$. Here w_{bbox} and h_{bbox} are the width and height of the enclosing rectangle of the ground truth landmarks, respectively. If no face is detected, the NME will be set to infinite.

3.3 Detailed Requirements

The upper bound of computational complexity is set to 1G Flops, and the upper bound of model size is set to 20 MB. For the training/validation/testing images, we provide the bounding boxes obtained by an off-the-shelf face detector. Nevertheless, the participants are allowed to employ their own face detector. Except for the face detectors, any external datasets and models are not allowed. Any test augmentation or multi-model ensemble strategy is not allowed, either.

The 2nd 106-point Lightweight Facial Landmark Localization (FLLC) grand challenge began by July 13, 2020. During the validation phase (from July 27 to October 08), the participants were allowed to evaluate their models on the validation set, and the leaderboard on the validation set was updated every day with respect to the submissions. The test images were released on October 09. To prevent cheating on the test set, each team was given an 24-h window to submit their predicted test results (Fig. 3).

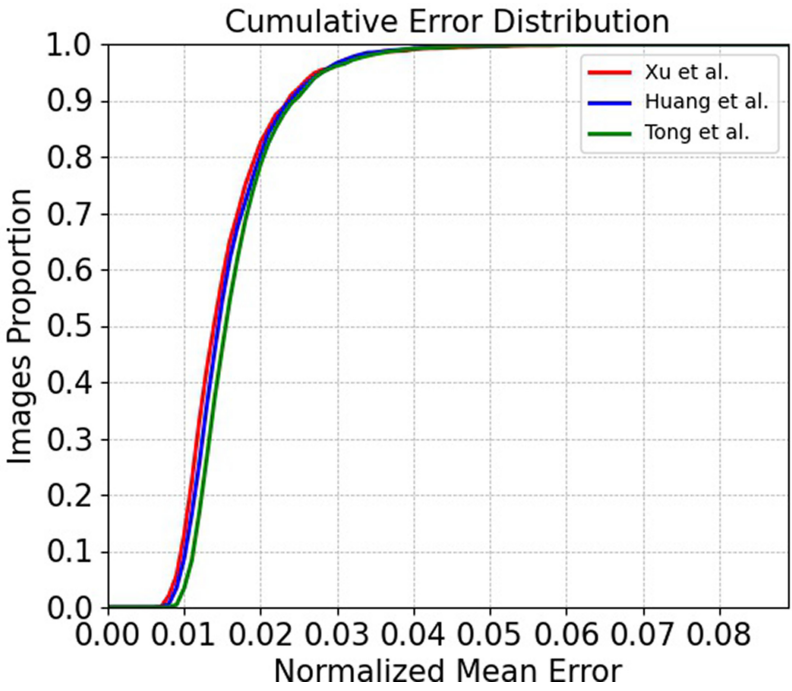

Fig. 3. The CED curve of the top three teams.

4 Summary of Participants

The competition has attracted much attention from both academia and industry. There are more than 70 teams registered in the competition. 15 teams participate in the validation phase, and 9 of them enter in the final test phase.

As shown in Table 1, the champion of the competition is Xu *et al.* from SogouAI. Huang *et al.* from OPPO Research Institute achieve the second place. The third place goes to Tong *et al.* from the Vision Intelligence Center of Meituan. Next, we will give a brief algorithm introduction of the top three winners.

Table 1. Leaderboard of the 2nd 106-point Lightweight Facial Landmark Localization Competition. The top three teams are ranked according to the AUC of the CED curve.

Rank	AUC (%)	Failure rate (%)	NME (%)	Model size (M)	Flops (M)
1	80.52	0.05	1.58	12	887.073
2	79.92	0.05	1.63	17	995.922
3	78.88	0.05	1.71	16	996.190

Xu *et al.* employ the advanced HRNet [14] for facial landmark prediction, which is able to maintain high-resolution representations through the whole process and connect the multi-resolution sub-networks in parallel. In order to reduce computational complexity, the inverted residuals [25] are adopted. The expansion ratio is set to 1 in all inverted residuals. To achieve good results, they increase the number of channels in the branch of HRNet blocks, and utilize group convolution [16] in the last few layers. The settings of network structure is given in Table 2. During the training phase, they apply some forms of data augmentation, including randomly rotating and randomly cropping. Specially, they employ the PDB strategy [12] against pose variations which duplicates large samples many times. Finally, they won the first place with the AUC of 80.52%, NME of 1.58% and Failure rate of 0.05%. The model size is about 12M and the FLOPS is 887.073M.

Table 2. The network structure settings of Xu *et al.*

Operator	Settings
conv2d	kernel_size = 3, stride=2 channel_in = 3, channel_out = 32
conv2d	kernel_size = 3, stride=2 channel_in = 32, channel_out = 64
bottleneck	stride =1, channel_in = 64 channel_out = 64
HRNet_block	number_blocks = 4, 4 number_channels = 24, 48
HRNet_block	number_block = 4, 4, 4 number_channels = 24, 48, 96
HRNet_block	number_block = 4, 4, 4, 4 number_channels = 24, 48, 96, 192
conv2d	kernel_size = 1, stride = 1 channel_in = 360, channel_out = 360
conv2d	kernel_size = 1, stride = 1 channel_in = 360, channel_out = 106

Huang *et al.* propose a multi-level supervision strategy to train the facial landmark localization models. They take ResNet-18 [13] as the backbone and reduce the channel size of the last two residual blocks from 256/512 to 192/256 due to the limits of computational complexity. Instance Normalization [31] is

adopted instead of Batch Normalization, which further improves the details of individual differences without increasing computational overhead. As shown in Fig. 4, apart from the main branch, an additional branch from feature map of the 3rd blocks is introduced for the contour landmarks prediction. Finally, mean aggregation is used for the final output. The AUC, NME and Failure rate are 79.92%, 1.63% and 0.05%, respectively. The model size is 17M while FLOPS is 995.922M. Huang *et al.* gain the second place in the competition.

Tong *et al.* take the improved HRNet [28] structure as the backbone, in which the bottleneck block [13] and group convolution [29] are used to replace the standard residual block in the original HRNet. In order to prevent the accuracy loss by the coordinates quantization, they use a mapping function named Dual Soft Argmax (DSA [18]) to map the heatmap response to final coordinates, which overcomes the problem of weight imbalance problem of Soft Argmax (SA [22]). The Normalized Mean Error (NME) loss [17] is taken as the training loss. Besides, inspired by [30], they propose a Similarity-FeatureMap knowledge distillation model. As Figure 5 shows, it guides the training of a student network by keeping the feature maps' similarity of input pairs according to the teacher network. Specifically, similarity matrices are derived from the feature maps and a distillation loss is computed on the matrices produced by the student network and the teacher network. Finally, the submitted model achieves 78.88%, 1.71%, 0.05% of the AUC, NME, and Failure rate, respectively. The model size is about 16M and the FLOPS is 996.190M. Tong *et al.* won the third place.

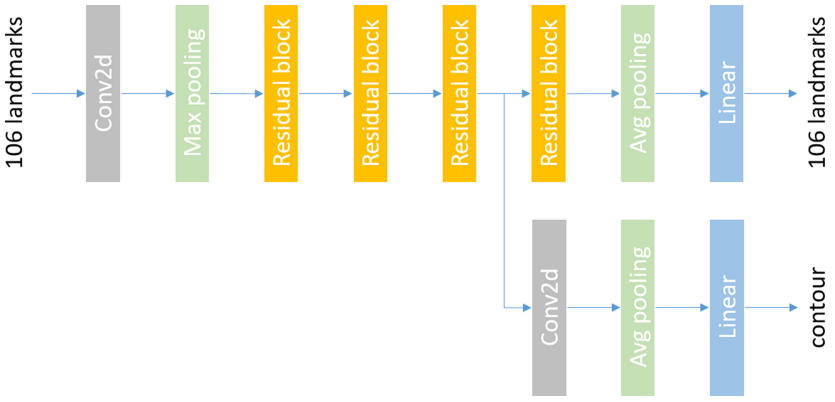

Fig. 4. The network structure of Huang *et al.*

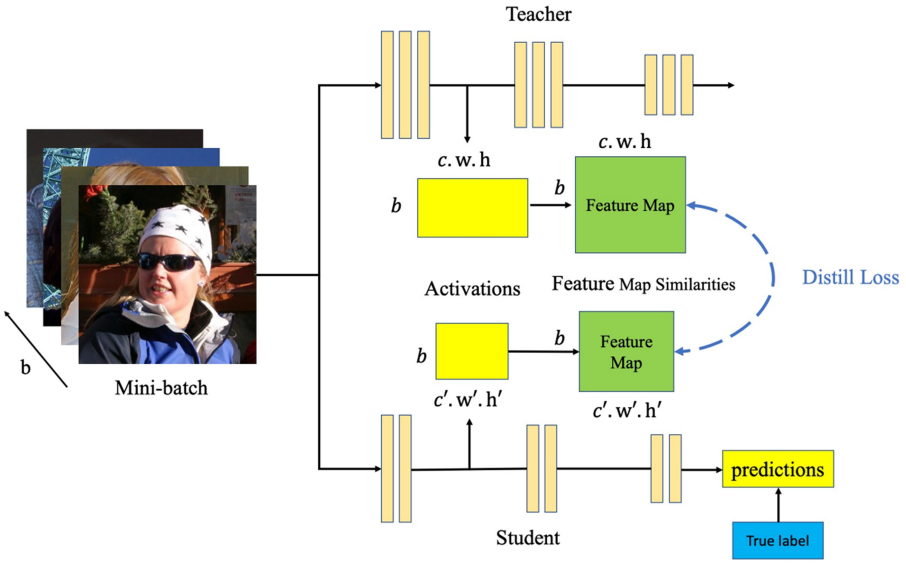

Fig. 5. Similarity-FeatureMap knowledge distillation guides the training of a student network such that input pairs that produce similar (dissimilar) feature maps in the pre-trained teacher network produce similar (dissimilar) feature maps in the student network. Given an input mini-batch of b images, we derive similarity matrices from the feature maps, and compute a distillation loss on the matrices produced by the student and the teacher.

5 Conclusion

In this paper, we first summarize the prior facial landmark localization challenges and the commonly used 2D facial landmark datasets in recent years. Then we introduce the detailed information of the 2nd 106-point lightweight facial landmark localization grand challenge. We construct and release a new facial landmark dataset, named JD-landmark-v2. Compared with the previous challenges, our work pays attention on the lightweight facial landmark localization model, which is important for the efficient system deployment. Finally, there are more than 70 teams participate in the competition and 9 teams involve in the final evaluation. We introduce the methods together with the performance of top three teams in this paper. We hope this work could push the frontier of the lightweight facial landmark localization algorithm.

Acknowledgment. This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0103800.

References

1. https://ibug.doc.ic.ac.uk/resources/300-W/

2. https://ibug.doc.ic.ac.uk/resources/2nd-facial-landmark-tracking-competition-menpo-ben/
3. https://ibug.doc.ic.ac.uk/resources/300-VW/
4. https://facial-landmarks-localization-challenge.github.io/
5. http://www.cs.cmu.edu/afs/cs/project/PIE/MultiPie/Multi-Pie/Home.html
6. https://ibug.doc.ic.ac.uk/resources/1st-3d-face-tracking-wild-competition/
7. https://neerajkumar.org/databases/lfpw/
8. http://www.ifp.illinois.edu/~vuongle2/helen/
9. https://www.tugraz.at/institute/icg/research/team-bischof/lrs/downloads/aflw/
10. Bao, Q., Liu, W., Hong, J., Duan, L., Mei, T.: Pose-native network architecture search for multi-person human pose estimation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 592–600 (2020)
11. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2930–2940 (2013)
12. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245 (2018)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
14. Huang, J., Zhu, Z., Huang, G.: Multi-stage HRNet: multiple stage high-resolution network for human pose estimation. arXiv preprint arXiv:1910.05901 (2019)
15. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4873–4882 (2016)
16. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
17. Lai, S., Chai, Z., Li, S., Meng, H., Yang, M., Wei, X.: Enhanced normalized mean error loss for robust facial landmark detection. In: British Machine Vision Conference, p. 111 (2019)
18. Lai, S., Chai, Z., Wei, X.: Improved hourglass structure for high performance facial landmark detection. In: Proceedings of IEEE International Conference on Multimedia & Expo Workshops, pp. 669–672. IEEE (2019)
19. 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.) ECCV 2012. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_49
20. Liu, Y., et al.: Grand challenge of 106-point facial landmark localization. In: IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 613–616. IEEE (2019)
21. Liu, Y., Shi, H., Shen, H., Si, Y., Wang, X., Mei, T.: A new dataset and boundary-attention semantic segmentation for face parsing. In: Association for the Advancement of Artificial Intelligence, pp. 11637–11644 (2020)
22. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372 (2018)
23. Ruan, W., Liu, W., Bao, Q., Chen, J., Cheng, Y., Mei, T.: POINet: pose-guided ovonic insight network for multi-person pose tracking. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 284–292 (2019)

24. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397–403 (2013)
25. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
26. Shen, J., Zafeiriou, S., Chrysos, G.G., Kossaifi, J., Tzimiropoulos, G., Pantic, M.: The first facial landmark tracking in-the-wild challenge: benchmark and results. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 50–58 (2015)
27. Smith, B.M., Zhang, L., Brandt, J., Lin, Z., Yang, J.: Exemplar-based face parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3484–3491 (2013)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)
29. Ting, Z., Guo-Jun, Q., Bin, X., Jingdong, W.: Interleaved group convolutions for deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
30. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1365–1374 (2019)
31. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: the missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
32. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
33. Zafeiriou, S., Chrysos, G.G., Roussos, A., Ververas, E., Deng, J., Trigeorgis, G.: The 3D Menpo facial landmark tracking challenge. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 2503–2511 (2017)
34. Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., Shen, J.: The Menpo facial landmark localisation challenge: a step towards the solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 170–179 (2017)
35. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879–2886. IEEE (2012)