2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)

PLA: Fast and Accurate Face Alignment Network based on Prior Landmarks

Haobin Li
School of Cyberspace Security
Dongguan University of Technology
Dongguan, China
gduflhb@163.com

Ye Ding
School of Cyberspace Security
Dongguan University of Technology
Dongguan, China
* Corresponding author: dingye@dgut.edu.cn

Mingyu Shao
School of Cyberspace Security
Dongguan University of Technology
Dongguan, China
18846849369@163.com

Li Lu
School of Cyberspace Security
Dongguan University of Technology
Dongguan, China
2360890221@qq.com

Juncai Huang
School of Cyberspace Security
Dongguan University of Technology
Dongguan, China
huangjc@dgut.edu.cn

2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE) | 979-8-3503-3597-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/NNICE58320.2023.10105788

***Abstract*—In this paper, we proposed a new method called Prior Landmark Algorithm (PLA) to address the limitations of traditional face landmark prediction techniques. PLA predicts the offset between real and prior landmarks to overcome the issue of solvable domain coverage and to handle different face poses and shapes. A new loss function with weights (PL-Loss) and the strategy of Online Hard Example Mining were also introduced to improve the accuracy of the model. The adoption of prior landmarks and an auxiliary classifier eliminate the need for face normalization during training. The results show that PLA achieved comparable accuracy to state-of-the-art models and had a small model size of only 10 MB with fast processing speed of 162 FPS on average.**

Keywords-component; face alignment, landmark analysis, object detection, prior knowledge

I. Introduction

Face landmarks detection, also known as face alignment, involves identifying key regions of the face, such as eyebrows, eyes, nose, mouth, and facial contours in a given face image. It serves as the foundation for various face-related tasks like face recognition, expression analysis, and 3D face reconstruction. Despite its significance, face landmarks detection remains a challenging task, requiring robustness in different conditions like poses, expressions, lighting, and image quality.

Inspired by traditional object detection algorithms, we introduce prior landmarks to predict the offset between natural landmarks and prior landmarks, leading to improved accuracy and faster inference speed. To resolve ambiguity between samples, we introduce a weighted regression loss function called PL-Loss. Additionally, a classifier is used to categorize face types based on pose and shape, and theprediction results are fused by selecting the top-N results. Our approach addresses the limitations of traditional face landmarks detection methods, making it a promising solution for practical applications. The contributions of this paper include: 1) we propose a novel approach to face landmarks detection that utilizes prior landmarks to improve accuracy and speed; 2) we introduce PL-Loss to resolve ambiguity between faces; 3) we design a classifier to categorize face types based on pose and shape.

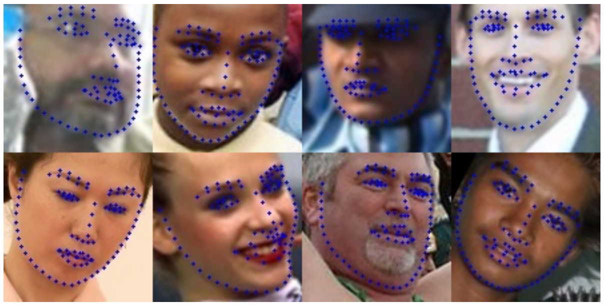


Figure 1. Landmarks inferred by PLA where the dots repre- sent predited landmarks. PLA is capable of handling head deflection and partially occluded images.

II. Related Works

Recently, face landmark detection has been approached as a regression problem, with methods based on convolutional regression networks showing promising results and significantly improving detection performance. Sun et al. [1] was the first to apply a CNN to face landmark detection,

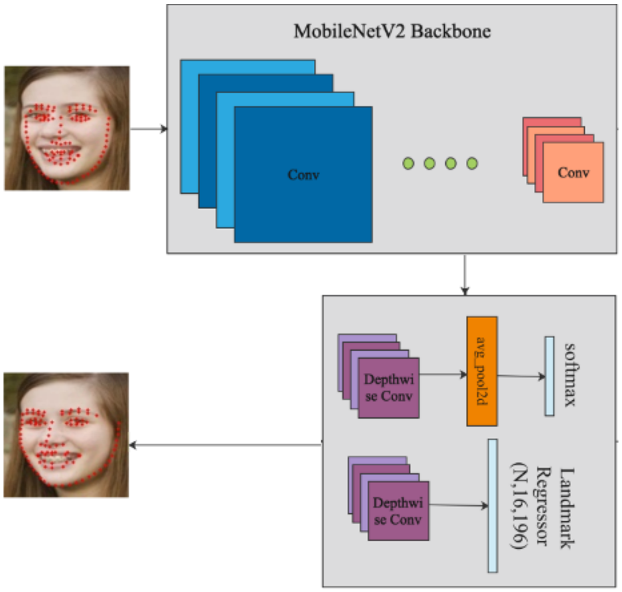


Figure 2. The network structure of PLA

presenting a coarse-to-fine method, named DCNN, which incrementally estimates the positions of facial landmarks. Subsequently, CNNs [2–4] were integrated as a regression module in facial landmark detection. This unification allowed for the learning of relevant and specific facial regions, leading to improved robustness across various alignment tasks.

Combining face landmark detection with other tasks has also been explored. Zhang et al. introduced the Multi-Task Cascaded Convolutional Network (MTCNN) [5] to tackle face detection and alignment problems simultaneously. Wu et al. [6] presented the boundary-aware LAB face alignment algorithm, using boundary information as geometric structure to better-fit landmarks within the face boundary and reduce landmark drift. Zheng et al. [7] propose a pre-training method, FaRL, that employs image-text contrastive learning and masked image modelling to learn facial representations that are more generalizable, resulting in improved performance on downstream tasks such as face alignment.

All the above methods predict face landmarks directly, while PLA was inspired by object detection algorithms and predicts the offset between a prior anchor and ground truth, offering accuracy and speed suitable for real-world applications.

III. THE NETWORK STRUCTURE OF PLA.

Prior Landmark Algorithm (PLA) is a one-stage, regression convolutional neural network (CNN) incorporating a classifier. As shown in Fig. 2, PLA is composed of the following modules:

1) The backbone network was inspired by MobileNet-V2 [8] and was designed with some blocks such as depth-wise separable convolution, linear bottlenecks, and inverse residuals to maintain high performance while minimizing computational complexity.

2) The classification module employs convolutional and pooling layers. Traditional classification networks use fully connected layers, which can account for up to 80% of the total parameters. To avoid an excessive number of model parameters, we use the global average pooling (GAP) to merge the learned features, thereby reducing model complexity. Our results surprisingly demonstrate that using GAP instead of fully connected layers enhances classification accuracy.

3) The regression module utilizes a single convolutional layer. While detection networks utilize multi-layer feature fusion and scale-aware feature selection to improve performance on complex tasks, these techniques are unnecessary for facial landmark detection, which is performed on images of a fixed size containing a single face. Thus, a simpler network architecture with a single convolutional layer maintains accuracy and efficiency for this task.

PLA leverages prior information to improve network convergence and simplify data preprocessing. Inspired by YOLOv2 [9], we also use K-Means [10] to cluster landmarks in the training set and generate *k*(depends on the dataset) landmark anchors assuming that the dataset contains *k* reference faces with different poses, scales, etc. The network predicts the offset between the ground truth landmarks and these prior landmark-anchors and simultaneously identifies the type of landmark-anchor. Incorporating prior landmarks in our approach eliminates the need for image normalization and simplifies data preprocessing.

The loss of PLA is composed of the cross entry loss for classifier and the prior landmarks loss for regression.

A. Cross Entry Loss for Classifier

The class label of the classifier is determined by assigning a value of 1 to the nearest landmark-anchor and 0 to the others based on the Euclidean distance between the priors and reals. The classifier uses cross-entropy loss as its loss function, which is calculated as follows:

$$L_{cls} = -\log(p_t) \quad (1)$$

B. Prior Landmarks Loss for Regression

Object detection tasks are typically treated as regression tasks and often use the L2 loss function. However, for the regressor in PLA, we use the wing loss to account for the impact of minor errors. The function is given as:

$$\text{wing}(x) = \begin{cases} w\ln(1 + |x|/\epsilon) & \text{if } |x| < w \\ |x| - C & \text{otherwise} \end{cases} \quad (2)$$

Given *k* groups of prior and predicted landmarks, the real correspond to only one prior, leading to high error rates when considering only the closest prior. This hinders the convergence of network regression, necessitating careful consideration of loss weights. Therefore, we propose Prior Landmarks Loss(PL-loss) to account for blurry samples at class boundaries and is designed to mitigate losses.

First, calculate the error caused by the distance. The *k* groups losses generated by the regressor are respectively marked as l_i, the ground truth landmarks are marked as *gt*, and the *k* prior landmarks are marked as pt_i. The Euclidean distance between *gt* and pt_i is given as:

$$d_i = \text{dist}\,(gt, pt_i) = \| gt - pt_i \|_2, i \in (0,1,\dots,k) \qquad (3)$$

where d_i can also be calculated according to Eq. (2), $gt = (gt^1, gt^1, \dots, gt^m)$, $pt_i = (pt_i^1, pt_i^2, \dots, pt_i^m)$, m represents the number of landmarks in the dataset. And we normalize d_i as:

$$\text{norm}_{\,i} = \frac{d_i}{\sum_{i=0}^{15} d_i} \qquad (4)$$

Calculate the weights according to the result obtained by Eq. (4) and normalize it:

$$wl_i = \frac{1}{\text{normd}_i^{\theta} + \epsilon} \qquad (5)$$

ε means infinitesimal, and θ is used to control the expansion of the weight. Generally, θ is set to 2.

We try to pay more attention to the nearest prior:

$$I = \text{argmin}_{i \in (0,\dots,15)}\,(\text{normd}_{\,i}) \qquad (6)$$

α is set to the expansion parameter. Generally, we set $\alpha = 2$, the weight of each loss is finally given as:

$$wf_i = \begin{cases} \alpha * wl_i, & i = I \\ wl_i, & \text{otherwise} \end{cases} \qquad (7)$$

$$w_i = \frac{wf_i}{\sum_{i=0}^{15} wf_i} \qquad (8)$$

Following the equations below, weight each loss of the landmarks, $e_i = (e_i^1, e_i^2, \dots, e_i^m)$ means the error from the i-th prior-landmarks, the weight of each landmark l_i is given as:

$$\mu_i^j = m * \frac{e_i^j}{\sum_{j=1}^{m} e_i^j}, i = 0,1,\dots,15; j = 1,2,\dots,m \qquad (9)$$

$$l_i = \sum_{j=1}^{m} \mu_i^j * e_i^j \qquad (10)$$

where the external is still multiplied by a factor m to ensure that the dimension of l_i remains unchanged. Moreover, in order to pay more attention to landmarks that are difficult to converge, OHEM [11] is used to optimize. Each landmark's loss is sorted, 30 landmarks with the largest loss are selected to expand their weight. The expansion factor is set to 2.

The final regression loss function is given as:

$$L_{reg} = \sum_{i=0}^{15} w_i * \sum_{j=1}^{m} \mu_i^j * e_i^j \qquad (11)$$

Our study diverges from previous research by not calculating regression error as the distance between predicted and actual landmark coordinates. Following Faster-RCNN [12] and YOLO v2 [9], we utilize prior landmark positions and adopt a more sophisticated approach to bias calculation, which means we just predict the offset instead of predicting landmarks directly. (x_{pt}, y_{pt}) refers to a prior landmark, (x^g, y^g) refers to the real landmark, while (x, y) refers to the predicted landmark, and (t_x, t_y) refers to the predicted offset, while (t_x^g, t_y^g) refers to the real offset, and d_{iod} is set to the inter-ocular distance of the prior landmarks. Then regression of ground truth offset can be calculated as:

$$\begin{gathered} t_x = (x - x_{pt})/d_{\text{iod}}\,, t_y = (y - y_{pt})/d_{\text{iod}} \\ t_x^g = (x^g - x_{pt})/d_{\text{iod}}\,, t_y^g = (y^g - y_{pt})/d_{\text{iod}} \end{gathered} \qquad (12)$$

C. Total Loss

Based on Eq. (1) and Eq. (11), the total loss of PLA is:

$$L_{total} = \frac{1}{N} * L_{cls} + \frac{\lambda}{N} * L_{reg} \qquad (13)$$

where N is equal to the numbers of priors we set, λ = 2.

IV. Experiments

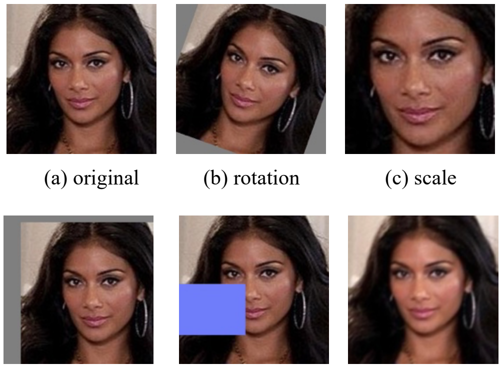
We evaluate PLA on two well-known facial landmarks detection datasets: WFLW [6] and 300W [13]. WFLW comprises 10,000 images, with 7,500 used for training and 2,500 for testing. Each image features 98 manually annotated facial landmarks, as well as rich attribute annotations such as occlusion, pose, makeup, lighting, blur, expression, etc. which is available for evaluating the model's robustness easily. 300W comprises 3,837 training images and 600 testing images from multiple sources, each of them annotated 68 facial landmarks.

In this paper, the performance of our model is assessed using the normalized mean error (NME), which is given as:

$$\text{NME} = \frac{\sum_{i=1}^{N} \| \mathrm{x}_{(i)}^{g} - \mathrm{x}_{(i)} \|_2}{N \times d} \qquad (14)$$

where d is the normalized distance. NME is a commonly used metric in face alignment datasets, with the inter-pupil and inter-ocular distances expressed to provide a more comprehensive evaluation.

We obtained face images based on a given bounding box and used a batch size of 32. The network is implemented by PyTorch,



(a) original (b) rotation (c) scale

(d) shift (e) occlusion (f) blur

Figure 3. Data augmentation

TABLE I. COMPARISON ON # OF PRIOR LANDMARKS

Dataset	*WFLW*			*300W*		
# of Prior Landmarks	4	9	16	4	9	16
Inter-ocular NME %	6.59	6.29	5.59	4.59	4.54	4.67

TABLE II. COMPARISON ON DISTANCE OF PRIOR LANDMARKS

di	7	8	13	12	17	23
norm di	0.088	0.100	0.162	0.150	0.213	0.288
wli	130.612	100.00	37.870	44.444	22.145	12.098
wfi	261.224	100.00	37.870	44.444	22.145	12.098
wi	0.547	0.209	0.079	0.093	0.046	0.025

TABLE III. EXAMINATION OF WEIGHTING EQUATION

Top No.	0	1	2	3	4	5
li	10	12	17	14	25	37
wi	0.547	0.209	0.079	0.093	0.046	0.025
wili	5.467	2.418	1.343	1.302	1.158	0.936

and the Adam optimization with a fixed learning rate of 0.001 is utilized. The maximum number of iterations is set to 100K, and the experiments are run on an RTX-1080Ti GPU. To improve the model's generalization ability and increase the diversity of the dataset, as shown in Fig. 3, data augmentation such as random rotation, scale, shift, and blur are applied.

Additionally, all images are resized to 256*256 to maintain consistency.

A. Comparison on # of Prior Landmarks

K-Means was employed to cluster the landmarks in the WFLW and 300W training sets, resulting in distinct groups of prior landmarks with varying poses and shapes. The evaluation results for models trained on different numbers of prior landmarks are presented in Table I. WFLW obtained the best performance using 16 priors, while 300W is with more minor variations in head pose and expression and performs best when the number of priors is set to 9.

B. Comparison on Distance of Prior Landmarks

To verify the rationality of the weight setting in the loss function of regression, we take 6 samples at different distances to examine the weight calculation according to Table II. Generally, $\alpha = 2$, $\gamma = 2$. For samples with two very close distances, the closest prior landmark is assigned a relatively large weight of 0.547, while the second closest distance is also considered to some extent. This approach effectively balances the importance of the different distances in calculating the overall weight. As given in Table III, the greater the loss between real landmarks and the farther away priors, the more the weighting effect is countered, which supports the validity of Eq. 11.

C. Model Analysis

Each real landmark may correspond to more than one prior landmark. As shown in Table IV, PLA_top3 considers the average of the top-3 predictions with highest classification score as the final result to ensure the stability of the landmark position. As given in Table V, ablation study of different number of prior landmarks setting was also operated. The NME evaluation metric uses the inter-ocular normalization factor for the WFLW, while the inter-pupil and inter-ocular normalization factors are used for the 300W.

TABLE IV. COMPARISON OF LATEST MODELS ON WFLW

Method	*NME (%), Inter-ocular*						
Testset	*test*	*pose*	*Expres-sion*	*Illumi-nation*	*make-up*	*Occlus-ion*	*blur*
ESR [14]	11.13	25.88	11.47	10.49	11.05	13.75	12.20
SDM [15]	10.29	24.10	11.45	9.32	9.38	13.03	11.28
CFSS [16]	9.07	21.36	10.09	8.30	8.74	11.76	9.96
DVLN [17]	6.08	11.54	6.78	5.73	5.98	7.33	6.88
LAB [6]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
PLA top1	5.62	10.34	6.04	5.45	5.65	6.76	6.29
PLA top3	5.59	10.3	5.98	5.43	5.59	6.74	6.27

TABLE V. COMPARISON OF LATEST MODELS ON 300W

Method	*Common Subset*	*Challenging Subset*	*Full Subset*
Inter-pupil Normalization			
RCPR [18]	6.18	17.26	8.35
CFAN [19]	5.50	16.78	7.69
ESR [14]	5.28	17.00	7.58
SDM [15]	5.60	15.40	7.52
3DDFA [20]	6.15	10.59	7.01
PLA Prior16 top3	5.60	10.46	6.55
PLA Prior4 top3	5.50	10.61	6.46
PLA Prior9 top3	5.42	10.43	6.37
Inter-ocular Normalization			
PIFA-CNN [21]	5.43	9.88	6.30
RDR [22]	5.03	8.95	5.80
PCD-CNN [23]	3.67	7.62	4.44
PLA Prior16 top3	4.02	7.46	4.67
PLA Prior4 top3	3.96	7.35	4.59
PLA Prior9 top3	3.91	7.22	4.54

TABLE VI. COMPARISON ON MODEL SIZE & PROCESSING SPEED

Model	*SAN [24]*	*LAB [6]*	*SDM [15]*	*PFLD 1X[25]*	*PLA*
Size (Mb)	798	50.7	10.1	12.5	10.4
Speed(FPS)	3	6	62.5	163	162

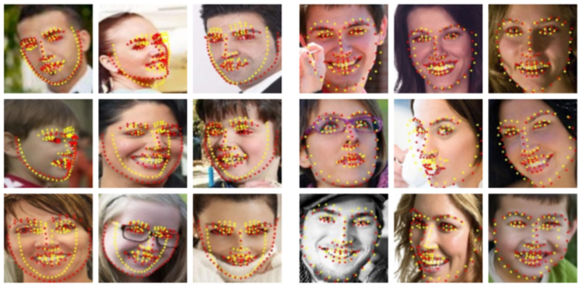


Figure 4. Inference results of PLA.

We evaluate the inference performance of the PLA algorithm on the CPU and compare it to other models and the comparing result is given in Table VI. PLA enjoys a smaller model size and faster processing speed, which shows it is a promising candidate for various applications.

Fig. 4 shows the performances on the test sets from WFLW [6] and 300W [13] (red dots are predicted landmarks, and yellow dots are prior landmarks for classification top-1 output), PLA can handle some head-deflection and partially occluded images.

V. CONCLUSION

We present a novel facial landmark detection model, named PLA, that leverages prior landmarks. The backbone of PLA is inspired by MobileNet-V2, leading to a smaller model size and faster inference. To enhance training, a classifier is incorporated, and a weighted regression loss function is employed to promote convergence. Our experimental results demonstrate that PLA, a single-stage regression network, surpasses other models in terms of accuracy, model size, and processing speed. Hence, PLA is a viable solution for practical applications.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under grant no. 61976051, U19A2067, and U1811463.

REFERENCES

[1] Yi Sun, Xiaogang Wang, and Xiaoou Tang, “Deep con- volutional network cascade for facial point detection,” in CVPR, 2013, pp. 3476–3483

[2] Arnaud Dapogny, Kevin Bailly, and Matthieu Cord, “Decafa: Deep convolutional cascade for face alignment in the wild,” in Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision, 2019, pp. 6893– 6901.

[3] Yang Zhao, Yifan Liu, Chunhua Shen, Yongsheng Gao, and Shengwu Xiong, “Mobilefan: Transferring deep hidden representation for face alignment,” Pattern Recognition, vol. 100, pp. 107114, 2020.

[4] Jun Wan, Jing Li, Zhihui Lai, Bo Du, and Lefei Zhang, “Robust face alignment by cascaded regression and deocclusion,” Neural Networks, vol. 123, pp. 261–272, 2020.

[5] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE signal processing letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[6] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou, “Look at boundary: A boundaryaware face alignment algorithm,” in *CVPR*, 2018, pp. 2129–2138.

[7] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen, “General facial representa- tion learning in a visual-linguistic manner,” in *Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18697–18709.

[8] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *CVPR*, 2018, pp. 4510–4520.

[9] Joseph Redmon and Ali Farhadi, “Yolo9000: better, faster, stronger,” in *CVPR*, 2017, pp. 7263–7271.

[10] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek, “The global k-means clustering algorithm,” *Pattern recognition*, vol. 36, no. 2, pp. 451–461, 2003.

[11] Abhinav Shrivastava, Abhinav Gupta, and Ross Gir- shick, “Training region-based object detectors with on- line hard example mining,” in *CVPR*, 2016, pp. 761– 769.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *NIPS*, vol. 28, 2015.

[13] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, “300 faces in-the-wild challenge: The first facial landmark localization challenge,” in *ICCV workshops*, 2013, pp. 397–403.

[14] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun, “Face alignment by explicit shape regression,” *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.

[15] Xuehan Xiong and Fernando De la Torre, “Supervised descent method and its applications to face alignment,” in *CVPR*, 2013, pp. 532–539.

[16] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang, “Face alignment by coarse-to-fine shape searching,” in *CVPR*, 2015, pp. 4998–5006.

[17] Wenyan Wu and Shuo Yang, “Leveraging intra and inter-dataset variations for robust face alignment,” in *CVPR workshops*, 2017, pp. 150–159.

[18] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dolla ́r, “Robust face landmark estimation under occlusion,” in *ICCV*, 2013, pp. 1513–1520.

[19] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen, “Coarse-to-fine auto-encoder networks (cfan) for realtime face alignment,” in *ECCV*. Springer, 2014, pp. 1– 16.

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

[21] Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren, “Pose-invariant face alignment with a single cnn,” in *ICCV*, 2017, pp. 3200–3209.

[22] Shengtao Xiao, Jiashi Feng, Luoqi Liu, Xuecheng Nie, Wei Wang, Shuicheng Yan, and Ashraf Kassim, “Recur- rent 3d-2d dual learning for large-pose facial landmark detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1633–1642.

[23] Amit Kumar and Rama Chellappa, “Disentangling 3d pose in a dendritic cnn for unconstrained 2d face alignment,” in *CVPR*, 2018, pp. 430–439.

[24] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang, “Style aggregated network for facial landmark detection,” in *CVPR*, 2018, pp. 379–388.

[25] Wu, Yuxiang, et al. "Foxnet: a multi-face alignment method." in IEEE International Conference on Image Processing , 2019, pp. 1322-1326.