An Efficient Multitask Neural Network for Face Alignment, Head Pose Estimation and Face Tracking

Jiahao Xia, Haimin Zhang, Shiping Wen, Shuo Yang, Min Xu*

School of Electrical and Data Engineering, Faculty of Engineering and IT, University of Technology Sydney

15 Broadway, Ultimo, NSW 2007, Australia

Abstract

While Convolutional Neural Networks (CNNs) have significantly boosted the performance of face related algorithms, maintaining accuracy and efficiency simultaneously in practical use remains challenging. The state-of-the-art methods employ deeper networks for better performance, which makes it less practical for mobile applications because of more parameters and higher computational complexity. Therefore, we propose an efficient multitask neural network, Alignment & Tracking & Pose Network (ATPN) for face alignment, face tracking and head pose estimation. Specifically, to achieve better performance with fewer layers for face alignment, we introduce a shortcut connection between shallow-layer and deep-layer features. We find the shallow-layer features are highly correspond to facial boundaries that can provide the structural information of face and it is crucial for face alignment. Moreover, we generate a cheap heatmap based on the face alignment result and fuse it with features to improve the performance of the other two tasks. Based on the heatmap, the network can utilize both geometric information of landmarks and appearance information for head pose estimation. The heatmap also provides attention clues for face tracking. The face tracking task also saves us the face detection procedure for each frame, which also significantly boost the real-time capability for video-based tasks. We experimentally validate ATPN on four benchmark datasets, WFLW, 300VW, WIDER Face and 300W-LP. The experimental results demonstrate that it achieves better performance with much less parameters and lower computational complexity compared to other light models.

Keywords: Face alignment, Head pose estimation, Face tracking, Multitask neural network

1. Introduction

Face alignment and head pose estimation, which aims at localizing a group of predefined facial landmarks and estimating the Euler angle of head, have a wide range of applications, such as facial expression recognition (Wu et al., 2018b), face morphing (Hassner et al., 2015) and 3D face reconstruction (Dou et al., 2017). Although CNNs have boosted the performance of face-related problems and other prediction tasks significantly (Esen et al., 2017, 2008a,b,c, 2009), the low real-time capability is still the main barrier to practical use since many applications run on the mobile devices (Bakker et al.,

2021). Developing a practical face alignment and head pose estimation method remains challenging.

The facial structural information contained in facial boundaries is crucial to face alignment (Huang et al., 2020), especially in unconstrained scenarios (large pose, extreme lighting condition and occlusion). To explicitly utilize the structural information, many recent studies (Huang et al., 2020; Wu et al., 2018a) utilize several cascaded hourglass modules (Newell et al., 2016) to generate facial boundary heatmaps and the heatmaps are then fused with intermediate features. However, compared to regressing landmarks from the input image directly, these methods are much less efficient because of a large number of bottom-up and top-down convolution layers.

Existing head pose estimation methods only utilize the geometric information of facial landmarks (model-based approach (Baltrusaitis et al., 2018; Martins & Batista, 2008)) or appearance information of the input

---
*Corresponding author

Email addresses: `Jiahao.Xia@student.uts.edu.au` (Jiahao Xia), `Haimin.Zhang@uts.edu.au` (Haimin Zhang), `Shiping.Wen@uts.edu.au` (Shiping Wen), `Shuo.Yang@student.uts.edu.au` (Shuo Yang), `Min.Xu@uts.edu.au` (Min Xu)
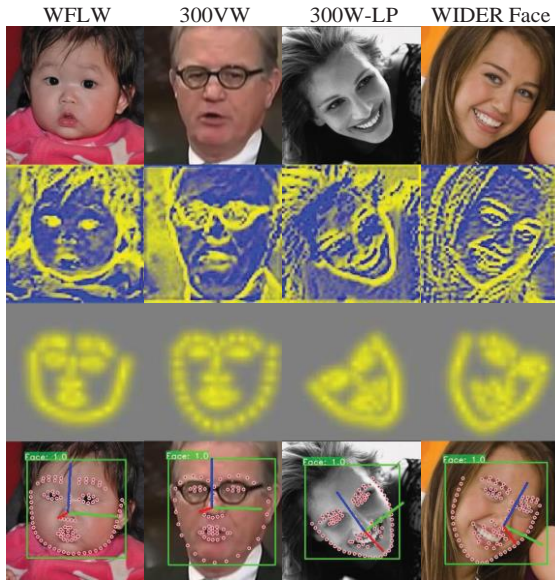
Figure 1: The first row shows the input images from different datasets and the second row illustrates the features maps at low-level. The heatmaps generated by face alignment result, shown in the third row, can provide attention clues for the CNN. The fourth row shows the estimated results of the ATPN (the **red axis** points towards the front of the face, **blue** pointing upward and **green** pointing left side).

image (appearance-based approach (Ahn et al., 2018; Drouard et al., 2017; Ranjan et al., 2019; Ruiz et al., 2018)) to estimate the Euler angle of faces. Model-based approaches fit a predefined 3D face model to the face in image according to facial landmarks. As a result, they ignore the appearance information of image features. Appearance-based approaches only rely on the appearance information of the input image. Nevertheless, the features outputted by a CNN or manifold embedding do not necessarily correspond to pose angle (Drouard et al., 2017), which leads to fragile robustness in unconstrained scenarios.

Moreover, most existing face alignment and head pose estimation algorithms require a face detector before the process. Therefore, they cannot run as fast as the speed in theory. In the video-based processing, face tracking makes face detection unnecessary in each frame, by which face alignment and head pose estimation can be accelerated further.

In this paper, we find out that the shallow-layer features are highly correspond to facial boundaries and they contain the facial structural information(Fig.1 $2th$ column). We give a shortcut to the features in shallow layers so that a light model can also explicitly employ the structural information for accurate face alignment. Then, a cheap heatmap (Fig.1 $3rd$ column) is generated

directly based on face alignment result and fused with the intermediate features. Based on the procedure, the ATPN can utilize both the appearance information of input image and the geometric information of the facial landmarks to achieve accurate and robust head pose estimation. It also provides the attention clues for face tracking task. Moreover, the face tracking task can save the face detection procedure to further accelerated face alignment and head pose estimation in video-based processing.

In a nutshell, our main contributions include:

- Proposing a light architecture that employs the structural information in low-level features for the face alignment. Compared to other light models, it achieves the best performance with the least parameters and lowest computational complexity.

- Providing the geometric information for head pose estimation and attention clues for face tracking by a heatmap generated from the face alignment results.

- Proposing a practical multitask framework for face alignment, head pose estimation and face tracking for video-based processing.

- Conducting extensive experiments and ablation studies on various datasets to prove the effectiveness of ATPN.

2. Related Work

2.1. Face Alignment

In the early stage, face alignment algorithms can be split into two categories: the model-based method (Zadeh et al., 2017) and the regression-based method (Cao et al., 2012; Liu et al., 2019; Xiong & De la Torre, 2013; Zhu et al., 2015). With the development of neural network, the performance of face alignment has been improved significantly. Many frameworks, such as DVLN (Wu & Yang, 2017) and UFLD (Zou et al., 2020), achieve higher accuracy and stronger robustness by taking advantages of the stronger expressiveness of CNNs. To improve the performance of face alignment further, (Browatzki & Wallraven, 2020; Dapogny et al., 2019; Dong et al., 2018; Dong et al., 2021; Qian et al., 2019) utilize style transfer or semi-supervised learning to augment training datasets, which addresses the problem of scarce training samples. However, with more large-scale datasets (Liu et al., 2020; Wu et al., 2018a) released, the number of training samples is not the main issue that limits the performance of face alignment. (Wu

et al., 2018a) find the structural information contained in facial boundary is crucial to face alignment. Following their work, (Huang et al., 2020) present a Landmark-Boundary Propagation Module to generate more accurate boundary heatmaps; (Chandran et al., 2020) utilize regional latent heatmap to achieve attention-driven face alignment in very high resolution. Nevertheless, it is difficult for the light model (Shao et al., 2021; Shapira et al., 2021) to utilize the information since generating an accurate boundary heatmap requires a very deep network. Therefore, maintaining both real-time capability and performance is quite challenging to face alignment.

In terms of video-based face alignment, many works try to model the temporary dependency explicitly for better performance. (Simonyan & Zisserman, 2014) propose a two-stream architecture to utilize single frame and the optical flow of multi-frame simultaneously; (Liu et al., 2018) optimize the method further by encoding the video input as active appearance codes in temporal stream. To improve the performance in motion-blurred video, (Sun et al., 2019) deblur the image based on two previous boundary maps and recent frames. However, these methods are also too complex to run on the mobile devices.

2.2. Head Pose Estimation

(Huang et al., 1995) has proven that the pose of a 3-point configuration is uniquely determined up to a reflection by its 2D projection under weak perspective model. Based on the conclusion, many works, such as (Martins & Batista, 2008), fit a predefined 3D face model to images based on the geometric information of facial landmark to estimate Euler angles, which is named as the model-based method. To address the problem that model-based methods degrade dramatically in large pose, (Baltrusaitis et al., 2018) replace the 2D landmarks by the 3D landmarks (Zadeh et al., 2017) that contains more geometric information. However, compared to the appearance information, the geometric information of 3D landmarks is still insufficient for robust head pose estimation.

Different from model-based methods, appearance-based methods regress head pose directly from images. (Drouard et al., 2017) adopt a mixture of linear regressions with partially-latent output to estimate head pose; (Patacchiola & Cangelosi, 2017) utilize CNN to regress head pose and study the influence of different optimizers and CNN structures; (Ruiz et al., 2018) and (Yang et al., 2019) classify head pose into several ranges and then regress a fine-grained result based on the range. Nevertheless, the features outputted by a CNN or man-

ifold embedding do not necessarily correspond to pose angle because of the lack of attention clues.

2.3. Multitask Network

Multitask learning enables CNNs to learn a task together with other related tasks simultaneously using a shared representation. It improves the performance and real-time speed significantly. (Zhang & Zhang, 2014) propose a one-stage multitask CNN to three face-related tasks: face detection, face view prediction and 5 facial landmarks localization. (Zhang et al., 2016) optimize the framework further by a cascaded network; (Ahn et al., 2018) propose a cascaded framework for face detection and head pose estimation; (Mo et al., 2020) present a hardwired accelerator for the framework. Hyperface (Ranjan et al., 2019) utilize both the global and local information on face for four tasks (face detection, landmarks localization, head pose estimation and gender recognition). Nevertheless, multitask framework can share more information for better performance, such as geometric information. Unfortunately, there is no existing work yet.

3. Approach

In this section, we introduce a multitask framework, ATPN, for face alignment, head pose estimation and face tracking. As illustrated on Fig.2(a), the ATPN consists of four parts: the backbone network, alignment branch, pose branch and tracking branch. The input is the Region of Interest (ROI) of face that is acquired by a face detector or the minimum bounding rectangle of the face landmarks in previous frames (25% extension of each boundary). The backbone network firstly learns features from input image with Mobilenet-V3 block (Howard et al., 2019). Then the alignment branch localize facial landmarks by the features at different levels ($28{\times}28$, $14{\times}14$, $7{\times}7$). Based on the predicted facial landmarks, a heatmap is generated directly and fused with low-level features ($28{\times}28$) to provide geometric information and attention clue for other tasks. Finally, the two branch regress the Euler angles and the confidence of face respectively. If the confidence is larger than a certain threshold (0.7 for ATPN), the ROI of next framework is calculated based on the predicted facial landmarks.

3.1. Alignment Branch

The structural information of the facial boundaries is crucial for facial alignment. Different from other works (Huang et al., 2020; Wu et al., 2018a) that utilize
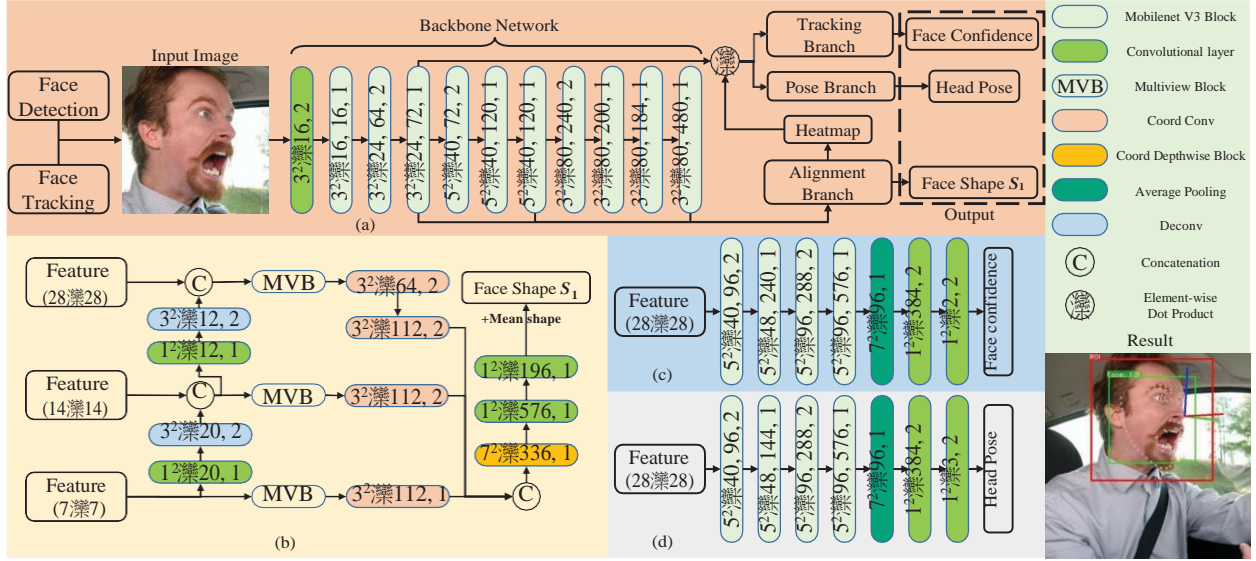
Figure 2: An overview of our multitask framework. The kernels of Mobilenet V3 block are described as kernel size × channels, expansion size, stride. And the kernels in other layers are described as kernel size × channels, stride. (a) The overall framework. The backbone architecture based on Moblienet-V3 module is used to extract image features for facial landmarks localization, head pose estimation and face tracking. (b) The alignment branch is used for facial landmarks localization with multiscale features. (c) The tracking branch output a confidence of face based on the features fused with heatmap. (d) The pose branch regresses head pose from the features fused with heatmap.

an additional network or branch to generate boundary heatmaps, we directly employ the structural information in low-level features, as shown in Fig.2(b). Referred to feature pyramid network (Lin et al., 2017), we firstly fuse the low-level features with high-level features by $1 \times$ Conv and Deconv layers. Then, the features at different levels are fed into the multiview block (Bulat & Tzimiropoulos, 2017). As illustrated in Fig.3, it consists of three CNN layers to create three different receptive fields. Besides, the residual connection of the multiview block also preserves the low-level features. The outputs of the multiview block are downsampled directly by CoordConv (R.Liu et al., 2018). On the one hand, it shorten the information path between low-level featuress and output layers. On the other hand, it introduces a coordinate conception into CNN, which is significant to coordinate regression. Finally, we utilize a $7 \times$ Coord Depthwise Block to project the feature map into a vector. Compared to pooling or linear layers, the Coord Depthwise Block can preserve the spatial structure of the input image. Instead of predicting the facial landmarks directly, the alignment branch predicts the residual error between target face shape S_1 and mean shape S_0. The mean shape S_0 provides a good initial face shape for the framework to make the result more stable.
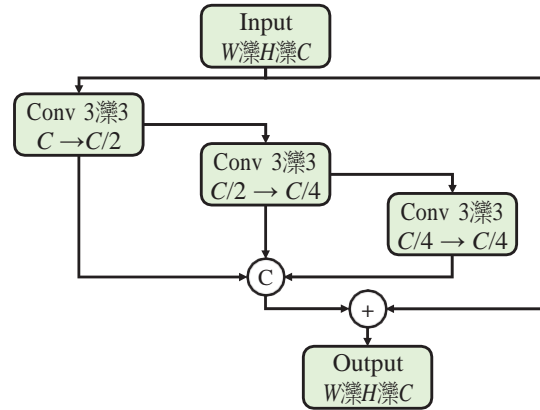
Figure 3: The structure of a multiview block. W, H, C indicate the width, hight and channel of feature respectively

3.2. Pose Branch and Tracking Branch

The tracking branch (Fig.2(c)) and pose branch (Fig.2(d)) mainly consist of MobileNet-V3 Block for better efficiency. In the pose branch, the output layer is activated by Tanh and then multiplied by π to normalize the output P_h into $[-\pi, +\pi]$. Hence, the predicted result can be more stable. In the tracking branch, the output is activated by Softmax to normalize face confidence C_f in $[0, 1]$.

4

3.3. Heatmap

Different from the heatmap in other works (Dapogny et al., 2019; Huang et al., 2020; Wu et al., 2018a), the heatmap in ATPN is generated directly based on the predicted landmarks by a formula. Therefore, the parameters and computational complexity can be reduced significantly. The formula can be written as:

$$H(x, y) = \frac{1}{1 + \min_{(x_i^I, y_i^I) \in \boldsymbol{S}_1} |(x, y) - (x_i, y_i)|},$$

(1)

where $H(x, y)$ is the intensity of point (x, y). (x_i^I, y_i^I) indicates the coordinate for the i-th landmark of $\boldsymbol{S}_1$. To avoid the problem that the CNN completely ignore the feature far from facial landmarks, $H(x, y)$ is set to 0.5 if the value is less than 0.5. Then, heatmap is fused with features by element wise multiplication, as shown below:

$$\boldsymbol{F}_O = \boldsymbol{F}_I \otimes \boldsymbol{H},$$

(2)

where $\boldsymbol{F}_O$ indicates the output features, $\boldsymbol{H}$ and $\boldsymbol{F}_I$ indicates the heatmap and the features learned by backbone network. By fusing the heatmap with the intermediate features, the output features contains both appearance information of the input images and the geometric information of facial landmarks. Moreover, the heatmap also provides the attention clue for ATPN to eliminate the interference of background and improve the performance of face tracking.

3.4. Training Strategy

We train the ATPN with three stages to alleviate the problem that a difficult task is still underfitting while a easy task is overfitting.

In the first stage, we train the backbone network and face alignment branch together. We calculate the point-to-point Euclidean and normalize it with the Inter-ocular distance (Sagonas et al., 2013), which can be written as:

$$L_A = \frac{\|\boldsymbol{S}_1 - \boldsymbol{S}_1^I\|_2}{\boldsymbol{d}_{ION}},$$

(3)

where L_A indicates the normlizaed error of face alignment. $\boldsymbol{d}_{ION}$ is the Inter-ocular distance, $\boldsymbol{S}_1^I$ and $\boldsymbol{S}_1$ are the predicted and annotated landmarks respectively. Then, the loss function is formulated as:

$$\min \frac{1}{N_1} \left(\sum_{i=1}^{N_1} L_{A-i} \right) + w_1 l_{2-A},$$

(4)

where N_1 indicates the number of samples which are used in stage 1. l_{2-A} is the L2-regularization loss of

Figure 4: Column 1 shows the input image and column 2 to column 7 illustrate the low-level feature maps in the ATPN and two single task CNN.

backbone network and face alignment branch, and w_1 is the weight of l_{2-A}.

Then, we freeze the weights of the backbone and train the tracking branch and pose branch in stage 2 and stage 3 respectively. The face tracking error L_T can be written as:

$$L_T = -(y_t \log(p_t) + (1 - y_t)(1 - \log(p_t))),$$

(5)

where p_t is the predicted face confidence and $y_i \in [0, 1]$ is the annotation (0: background, 1: face). The learning object of the second stage can be written as:

$$\min \frac{1}{N_2} \left(\sum_{i=1}^{N_2} L_{T-i} \right) + w_2 l_{2-t},$$

(6)

where N_2 indicates the number of samples used in the second stage. l_{2-t} is the L2-regularization loss of the tracking branch and w_2 is the weight of l_{2-f}. The error of head pose estimation L_p can be written as:

$$L_P = \frac{\|\boldsymbol{P}_h - \boldsymbol{P}_h^I\|}{3},$$

(7)

where $\boldsymbol{P}_h$ and $\boldsymbol{P}_h^I$ are the predicted and annotated head pose respectively. Finally, the learning object of third stage can be written as:

$$\min \frac{1}{N_3} \left(\sum_{i=1}^{N_3} L_{P-i} \right) + w_3 l_{2-p},$$

(8)

where N_3 indicates the number of samples used in the third stage. l_{2-p} is the L2-regularization loss of pose branch and w_3 is the weight of l_{2-p}.

Based on the training strategy, head pose estimation and face tracking can take advantages of a part of knowledge of face alignment. We visualize low-level features in the ATPN and two single task CNNs with same layers, as shown in Fig.4. Compared to single task CNNs, the low-level features in the ATPN focus more on the edge of face rather than background. Therefore, sharing low-level features makes CNNs learn more effective features.

Metric	Method	Testset	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME(%)↓	LAB† (Wu et al., 2018a)	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN† (Dong et al., 2018)	5.22	10.39	5.71	5.19	5.49	6.83	5.8
	2 Hourglass† (Newell et al., 2016)	5.19	9.03	5.53	5.08	4.97	6.45	5.93
	DeCaFA† (Dapogny et al., 2019)	4.62	8.11	4.65	4.41	4.63	5.74	5.38
	PropNet† (Huang et al., 2020)	4.05	6.92	3.87	4.07	3.76	4.58	4.36
	MuSiCa98‡ (Shapira et al., 2021)	7.90	15.80	8.52	7.49	8.56	10.04	8.92
	3FabRec‡ (Browatzki & Wallraven, 2020)	5.62	10.23	6.09	5.55	5.68	5.92	6.38
	$G\&LSR_\omega$‡ (Shao et al., 2021)	5.26	-	-	-	-	-	-
	Res18+AVS‡ (Qian et al., 2019)	5.25	9.10	5.83	4.93	5.47	6.26	5.86
	ATPN	5.13	8.97	5.49	4.95	4.94	6.30	5.78
$FR_{0.1}$(%)↓	LAB†	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN†	6.32	27.91	7.01	4.87	6.31	11.28	6.60
	2 Hourglass†	6.04	25.46	5.41	5.59	5.34	12.5	8.40
	DeCaFA†	4.84	21.4	3.73	3.22	6.15	9.26	6.61
	PropNet†	2.96	12.58	2.55	2.44	1.46	5.16	3.75
	MuSiCa98‡	-	-	-	-	-	-	-
	3FabRec‡	8.28	34.35	8.28	6.73	10.19	15.08	9.44
	$G\&LSR_\omega$‡	5.72	-	-	-	-	-	-
	Res18+AVS‡	7.44	32.52	8.3	4.3	8.25	12.77	9.06
	ATPN	6.27	26.99	6.05	4.72	7.28	12.22	7.89
$AUC_{0.1}$↑	LAB†	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	SAN†	0.536	0.236	0.462	0.555	0.552	0.456	0.493
	2 Hourglass†	0.531	0.337	0.509	0.540	0.543	0.475	0.488
	DeCaFA†	0.563	0.292	0.546	0.579	0.575	0.485	0.494
	PropNet†	0.615	0.382	0.628	0.616	0.638	0.572	0.583
	MuSiCa98‡	-	-	-	-	-	-	-
	3FabRec‡	0.484	0.192	0.448	0.496	0.473	0.398	0.434
	$G\&LSR_\omega$‡	0.493	-	-	-	-	-	-
	Res18+AVS‡	0.503	0.229	0.453	0.525	0.484	0.431	0.453
	ATPN	0.557	0.337	0.528	0.568	0.565	0.495	0.516

Table 1: Performance comparison of the ATPN and the state-of-the-art methods on WFLW and its subsets (the methods in the table are ranked by the NME on testset). Key: [**Best**, **Second Best**, ↓=the lower the better, ↑=the larger the better, †=the method is based on structural information or semi-supervised learning, ‡=the method is for mobile devices]

4. Experiments

4.1. Datasets

- **WFLW** (Wu et al., 2018a) contains 10,000 faces (7,500 for training and 2,500 for testing) with 98 landmarks. Besides, each face also has attributes annotation in terms of large pose, expression, illumination, make-up, occlusion and blur.

- **300W-LP** (Zhu et al., 2016) is a synthetic dataset. It fits every sample of 300W (Sagonas et al., 2013) to 3D models by 3DDFA (Zhu et al., 2016) and rotates 3D model to generate 122450 samples with head pose label.

- **WIDER Face** (Yang et al., 2016) is a face detection benchmark dataset, including 32,203 images and 393,703 labeled faces with a high degree of
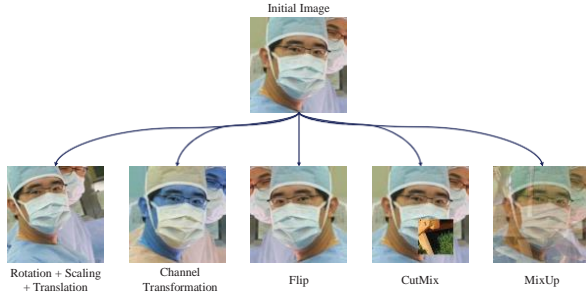
Figure 5: Data augmentations method used in our method.

variability in scale, pose and occlusion.

- **300VW** (Shen et al., 2015) is a video face alignment dataset, containing 50 videos for training and 64 videos for testing. The testing set is divided into three subsets (category A, B and C).

4.2. Implementation Detail

We utilize the samples with facial landmark annotation for the first stage training. The learning rate is set to 0.001 and reduced by a factor of 0.03 after every 4 epochs. In the second stage, tracking branch is trained with positive and negative samples created with WIDER Face. We only utilize the face area whose margin is more than 60 for positive samples. For each positive sample generation, 10 negative samples whose IoU is smaller than 0.3 are generated. The learning rate is set to 0.001 and decayed by a factor 0.04 after every 2 epochs. In the third stage, we only use the samples of 300-LP without 3DDFA (Zhu et al., 2016) data augmentation (3148 for training, 689 for testing) for training. It can keep domain consistency. The method used for data augmentation is the same as that used in the first stage. The learning rate setting is the same as the second stages.

Moreover, each sample randomly translates, rotates, flips, transforms channels or mixes up with the background image created by WIDER to augment the training data, as shown in Fig.5. we utilize Adam (D.P. Kingma & J.Ba, 2014) optimizer with a β_1 of 0.9 and a β_2 of 0.999 in the three stages. The batch size is set to 128.

4.3. Evaluation Metrics

4.3.1. Face Alignment

We evaluate the proposed method with Normalized Mean Error (NME). In WFLW, we use the inter-ocular distance (Sagonas et al., 2013) as the normalization factor. In 300VW, we also use the inter-pupil distance (Wu et al., 2018a) as the normalization factor to compare ATPN with other methods. Besides, we also report the Failure Rate (FR) for a maximum error of 0.1 and Area Under Curve (AUC) derived from Cumulative Errors Distribution (CED) curve by plotting the curve from 0 to 0.1.

4.3.2. Head Pose Estimation

Following the works of (Huang et al., 2020; Wu et al., 2018a), we divide the test set into two subsets (common subset and challenging subset) and report the Mean Average Error (MAE), as given below:

$$MAE = \frac{1}{N_{test}} \sum_{l=1}^{N^{test}} \frac{|\alpha_l - \alpha_l{}^l| + {}^1\beta_l - \beta_l{}^{l1} + |\gamma_l - \gamma_l{}^l|}{3},$$
(9)

where (α, β, γ) and $(\alpha^l, \beta^l, \gamma^l)$ are respectively the predicted and annotated roll, pitch, yaw angle of head. N_{test} is the total number of testing samples. Besides, we also report the FR in for a maximum error of $10°$ and CED curve.

4.3.3. Face Tracking

We evaluate ATPN with the NME and tracking failure rate. The threshold of failure rate is set to 0.7. Besides, we also report the Precision-Recall (PR) curve and Average Precision (AP) on the test set of WIDER Face.

4.4. Comparison with State-of-The-Art Methods
4.4.1. WFLW

Table 1 compares the ATPN to the methods based on structural information and the methods for mobile devices respectively. Table 2 illustrates number of parameters and computational complexity of the methods. Despite the PropNet achieves best performance of 4.05% in NME, the computational complexity is much more than the methods for mobile devices (more than $150\times$). By taking advantages of the structural information at low-level features, ATPN achieves best performance in NME among the methods for mobile devices. Compared to $G\&LSR_\omega$, we observe an improvement of 2.47% in NME with comparable computational complexity. Although LAB adopts GAN to generate a boundary heatmap, ATPN still outperforms it with only 0.3% computational complexity and 8.9% parameters. Therefore, ATPN is much more efficient than the state-of-the-art methods.

4.4.2. 300VW

We pretrain the ATPN on WFLW and then retrain it with the training set of 300VW. We carry out two different experiments on three subsets of 300VW. In the

Method	Params(M)↓	FLOPS(G)↓
PropNet (Huang et al., 2020)	36.3	42.83
LAB(Wu et al., 2018a)	12.3	18.85
DeCaFA (Dapogny et al., 2019)	≈10	≈30
2 Hourglass (Newell et al., 2016)	6.30	8.00
MuSiCa98‡ (Shapira et al., 2021)	≈3	≈0.25
$G\&LSR_\omega$ (Shao et al., 2021)	1.83	0.06
ATPN (Backbone + Alignment Branch)	1.1	0.06

Table 2: Parameters and computational complexity for ATPN and other the state-of-the-art methods (the methods in the table are ranked by the number of parameters). Key: [Best, Second Best,↓=the lower the better]

Method	Cat A	Cat B	Cat C
inter-ocular distance normalization (%)↓			
TSCN (Simonyan & Zisserman, 2014)	12.54	7.25	13.13
CFSS (Zhu et al., 2015)	7.68	6.42	13.67
SDM (Xiong & De la Torre, 2013)	7.41	6.08	14.03
TSTN (Liu et al., 2018)	5.21	4.23	10.11
ADC (Chandran et al., 2020)	4.17	3.89	7.28
DeCaFA (Dapogny et al., 2019)	3.82	3.63	6.67
FAB (Sun et al., 2019)	3.56	3.88	5.02
ATPN (Tracking)	3.52	3.64	4.99
ATPN (Detection)	3.49	3.57	4.89
inter-pupil distance normalization (%)↓			
4S-HG (Newell et al., 2016)	6.54	5.65	8.13
4S-HG+UFLD (Zou et al., 2020)	6.09	5.34	7.76
ATPN (Tracking)	4.83	5.04	7.75
ATPN (Detection)	4.79	4.95	7.61

Table 3: NME for the ATPN in tracking and detection mode compared with previous methods on Category A, Category B and Category C of 300VW (the methods in the table are ranked by the NME on Category A). Key: [Best, Second Best, ↓=the lower the better]

	Cat A	Cat B	Cat C
Sequences	146	39	188
Detection frame	233	39	210
Tracking frame	61902	32766	26128
Failure frame	87	0	22
Failure rate	0.14%	0.00%	0.08%

Table 4: Tracking details on 300VW. **Note**: 300VW fails to label some frames in several videos, resulting in a video being divided into dozens of sequences.

Method	Params(M)↓	FLOPS(G)↓
MTCNN (Zhang et al., 2016)	0.47	0.7∼1.4
Tracking Branch	0.19	0.017

Table 5: Parameters and computational complexity for tracking branch and MTCNN (the methods in the table are ranked by the number of parameters). Key: [Best, ↓=the lower the better].

frist one, we evaluate ATPN with detection mode. In this mode, the input image is cropped based on the ground truth. In the second experiment, the input image is cropped based on the face alignment result in the last frame for tracking. The comparison result are illustrated in Table 3. The most samples in 300VW are without occlusion, which enables the low-level features to produce more complete facial boundaries. Therefore, the improvement of ATPN is more significantly compared to other state-of-the-art methods. For example, although DeCaFA achieves excellent performance on WFLW, ATPN still achieves an impressive improvement of 8.64%, 1.66% and 26.69% in NME respectively on the three subsets

Moreover, the tracking failure rates are only 0.14%, 0.00% and 0.08% on the three subsets, which means the tracking branch is with satisfactory robustness. As a result, the performance of tracking mode only degrades a little compared to detection mode. Besides, in Fig.6, we also demonstrate the PR curves of the ATPN and Hyperface (Ranjan et al., 2019) on the validation set of WIDER Face. The average precision of ATPN reaches at 98.82% that outperforms Hyperface a lot. Some predicted results in the validation set can be viewed in Fig.7.

The Table 5 shows the parameters and computational complexity of the tracking branch and a commonly used face detection framework (MTCNN). The MTCNN is with 10×higher computational complexity than ATPN. Hence, the real-time capability of ATPN will degrades dramatically if the face detection framework is employed in each frame. By taking advantages of tracking branch, the processing time can be accelerated more than 50 times.

Method	Common Subset					Challenging Subset				
	Yaw↓	Pitch↓	Roll↓	MAE↓	FR$_{10°}$↓	Yaw↓	Pitch↓	Roll↓	MAE↓	FR$_{10°}$↓
ESR (Cao et al., 2012)	8.54°	7.47°	2.61°	6.21°	19.31%	24.52°	12.70°	10.64°	15.95°	50.62%
Annotated landmarks	7.32°	6.99°	1.88°	5.40°	16.37%	11.64°	9.96°	5.40°	9.00°	32.84%
Hyperface (Ranjan et al., 2019)	3.46°	3.87°	2.92°	3.42°	3.37%	6.30°	5.57°	7.34°	6.40°	19.76%
Openface2.0 (Baltrusaitis et al., 2018)	2.69°	3.56°	**1.10°**	2.45°	**1.72%**	**3.23°**	**3.56°**	**1.93°**	**2.91°**	**1.51%**
FSA-Net (Yang et al., 2019)	**1.85°**	**2.84°**	1.12°	**1.94°**	**2.35%**	4.51°	5.29°	**2.27°**	4.12°	19.26%
ATPN	**1.31°**	**2.38°**	**0.97°**	**1.55°**	**0.12%**	**2.62°**	**3.97°**	**1.93°**	**2.84°**	**1.49%**

Table 6: Mean Average Error (MAE) in degrees and FR$_{10°}$ on the common and challenging subset (the methods in the table are ranked by the MAE on the common subset). Key: [**Best**, **Second Best**, ↓=the lower the better, *=5-landmarks model-based method (eyes corner, mouth corner, nose tip and chin)]

Method	low-level Features	Multiview Block	CoordConv	NME(%)↓							FR$_{0.1}$(%)↓
				Testset	Pose	Expression	Illumination	Makeup	Occlusion	Blur	Testset
Baseline				5.61	9.90	6.11	5.36	5.47	6.77	6.38	8.68
Model1	✓			5.40	9.22	5.92	5.30	5.30	6.43	6.00	7.16
Model2	✓	✓		5.33	9.15	5.75	5.16	5.33	6.36	5.95	7.52
Model3	✓	✓	✓	**5.13**	**8.97**	**5.49**	**4.95**	**4.94**	**6.30**	**5.78**	**6.27**

Table 7: The contribution of different modules (the methods in the table are ranked by the NME in the test set). Key:[**Best**, ↓=the lower the better]
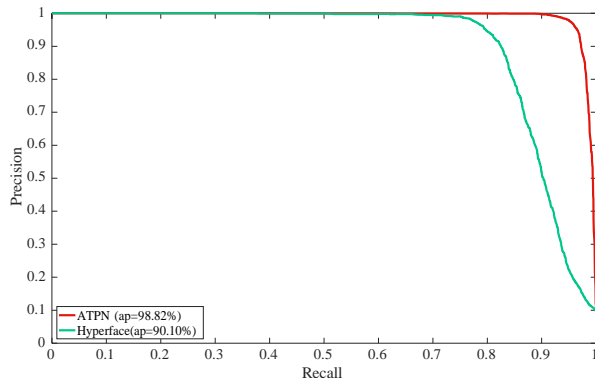

Figure 6: Evaluation on the validation set of WIDER FACE. The number following the method indicates the average accuracy of face tracking.

4.4.3. 300W-LP

We compare the pose branch to other state-of-the-arts methods on both the common and challenging subset, as shown in Table 6. Besides, Fig.8 also illustrates the CED curves of the ATPN and other methods on the full set. It is difficult for 5-landmarks model-based method to achieve accurate head pose estimation even if it utilizes the annotated landmarks. Hyperface and FSA-Net utilize the appearance features of the input image. The performance is improved significantly because the ap-pearance features contains more information compared to 5 landmarks. Nevertheless, it degrades significantly for the images with extreme conditions (challenging subset) because of the fragile robustness. OpenPose2.0 adopt 3D geometric information to estimate head pose and it can maintain excellent performance on the challenging subset. ATPN estimates head pose based on both geometric and appearance features, which significantly reduces the MAE by 36.73%, from $2.45°$ to $1.55°$ on the common subset compared to Openface2.0. Besides, the geometric features enables ATPN to maintain comparable performance.

4.5. Ablation Study

4.5.1. Alignment Branch

The alignment branch consists of several pivotal modules: the structural information at low-level, multiview block and CoordConv. We compare a baseline with several models that consist of these modules. The baseline is set as a model that regresses facial landmarks directly from the last layer of the backbone. The evaluation results on WFLW are shown in Table 7.

We observe 3.74% and 17.51% improvement respectively in NME and FR$_{0.1}$ by introducing the structural information into CNN. It illustrates that the structural information of low-level features is essential to face
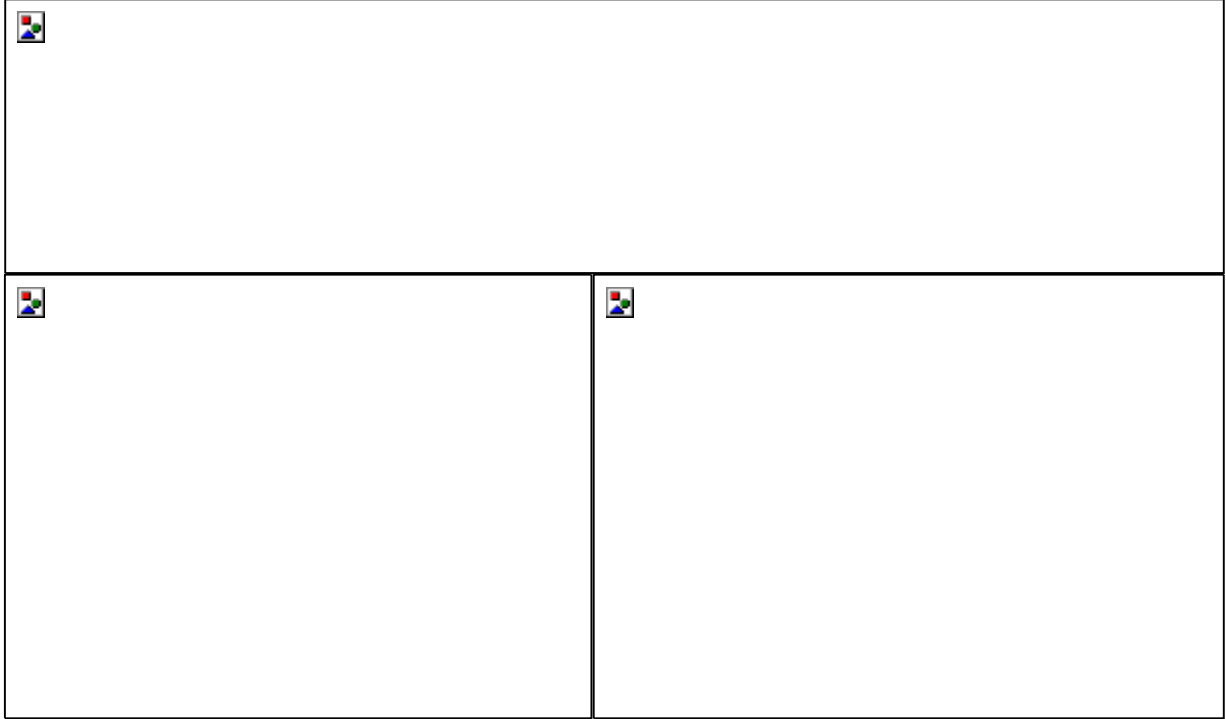
Figure 7: Sample results on WIDER Face validation set (the number above the bounding box indicates the confidence of face, the **red axis** points towards the front of the face, **blue** pointing upward and **green** pointing left side).

Method	Common Subset					Challenging Subset				
	Yaw↓	Pitch↓	Roll↓	MAE↓	FR$_{10°}$↓	Yaw↓	Pitch↓	Roll↓	MAE↓	FR$_{10°}$↓
Baseline	1.47°	2.53°	1.03°	1.68°	0.12%	2.93°	3.98°	2.44°	3.12°	4.70%
Feature sharing	1.42°	2.60°	1.11°	1.71°	0.24%	2.96°	4.17°	2.38°	3.17°	3.46%
Feature sharing+Heatmap	**1.31°**	**2.38°**	**0.97°**	**1.55°**	**0.12%**	**2.62°**	**3.97°**	**1.93°**	**2.84°**	**1.49%**

Table 8: The effects of feature sharing and heatmap on face tracking on common subset and challenging subset (the methods in the table are ranked by the MAE). Key: [**Best**, ↓=the lower the better]

Method	Baseline	Feature Sharing	Feature Sharing + Heatmap
Average Precision ↑	98.57%	98.72%	**98.82%**

Table 9: The influence of feature sharing and heatmap on the face tracking on WIDER Face. Key: [**Best**, ↑=the larger the better]

alignment. Then, CoordConv also improves the performance of face alignment significantly by providing spatial information for ATPN. Without adding much computational complexity, NME and FR are reduced by 3.57% and 16.62% respectively. Finally, the improvement brought by multiview block is not as significant as others, although it obtains the largest number of param-

eters. It suggests that increasing parameters in network cannot improve the performance effectively.

4.5.2. Feature Sharing and Heatmap

Feature sharing and heatmap also play an important role in the tracking and pose branch. To further investigate whether the heatmap can introduce the geometric information into the network, we carry out an additional experiment on background images and some results are shown in Fig.9. Despite the input images do not include any face, the predicted head pose of the model with heatmap is still highly correspond to the predicted facial landmarks, while the results of the model without heatmap are not. It illustrates that when the appearance features are not reliable, the model with heatmap can utilize the geometric information for head pose estima-
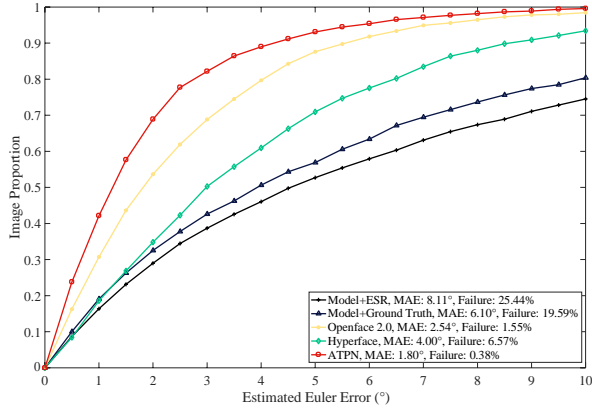
Figure 8: CED for the full set (Yaw, Pitch, Roll). MAE and Failure Rate are also reported.

Figure 9: Estimation results of the model with/without heatmap on the background images (the **red axis** points towards the front of the face, **blue** pointing upward and **green** pointing left side).

tion.

To study the influence of feature sharing and heatmap on two tasks, we set the baseline as a single task CNN with the same layers. In feature sharing model, the low-level layers are shared with face alignment and only high-level layers are trained.

The results of head pose estimation on 300W-LP are shown in Table 8. Compared to the baseline, the feature sharing model exhibits the comparable performance on the both common and challenging subset with much less parameters. With the geometric information of the heatmap, the MAE is significantly improved by 10.41% and 9.36% respectively on the common and challenging subset. Therefore, sharing features can boost the real-time capability and the geometric information of the heatmap is crucial to head pose estimation.

In terms of face tracking, the experiment is carried out on WIDER Face. The influence of feature sharing and heatmap is shown in Table 9. Although the trainable parameters in features sharing model is much less than the baseline, the average precision is still improved from 98.57% to 98.72%. By introducing the attention clue into network with heatmap, the performance is further boosted to 98.82%. It indicates that the features learned

by face alignment task and the attention clue of heatmap can improve the performance of face tracking.

5. CONCLUSION

In this paper, we present the Alignment & Tracking & Pose Network (ATPN), an efficient multitask network for facial landmark localization, face tracking and head pose estimation. Different from other state-of-the-art works that generates the structural information by an additional network or branch, we directly utilize the structural information in the low-level features. By taking advantages of the structural information, the computational complexity is only $1/300$ of the other methods that utilize the structural information. Moreover, it also achieves the best performance with the least parameters and FLOPS compared to other methods for mobile devices. On the video dataset, the improvement of real-time capability is more significant because the tracking branch eliminates the face detection process. Different from other multitask frameworks that only share features, ATPN also generates a heatmap by face alignment result to provide the geometric information for head pose estimation and the attention clues for face tracking. The experiment results illustrate that the geometric information of heatmap can help network maintain robustness on the extreme conditions. And the attention clues provided by heatmap can also improve the accuracy of face tracking.

In our future work, we will further explore the features sharing mechanism and attention mechanism learned by semi-supervised learning. Based on the mechanisms, we will develop a lighter model to boost both the performance and real-time capability of the state-of-the-art methods and apply it on the mobile devices.

11

Appendix A

All abbreviations used in this paper are described in Table 10.

Abbreviation	Description
CNN	Convolutional Neural Network
ATPN	Alignment & Tracking & Pose Network
ROI	Region of Interest
NME	Normalized Mean Error
FR	Failure Rate
AUC	Area Under Curve
CED	Cumulative Errors Distribution
MAE	Mean Average Error
PR	Precision-Recall
AP	Average Precision

Table 10: Lookup table for abbreviations in the paper.

Acknowledgement

References

Ahn, B., Choi, D.-G., Park, J., & Kweon, I. S. (2018). Real-time head pose estimation using multi-task deep neural network. *Robotics and Autonomous Systems*, *103*, 1–12. doi:https://doi.org/10.1016/j.robot.2018.01.005.

Bakker, B., Zabłocki, B., Baker, A., Riethmeister, V., Marx, B., Iyer, G., Anund, A., & Ahlström, C. (2021). A multi-stage, multi-feature machine learning approach to detect driver sleepiness in naturalistic road driving conditions. *IEEE Transactions on Intelligent Transportation Systems*, (pp. 1–10). doi:10.1109/TITS.2021.3090272.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In *AFGR* (pp. 59–66). doi:10.1109/FG.2018.00019.

Browatzki, B., & Wallraven, C. (2020). 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR* (pp. 6109–6119). doi:10.1109/CVPR42600.2020.00615.

Bulat, A., & Tzimiropoulos, G. (2017). Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *ICCV* (pp. 3726–3734). doi:10.1109/ICCV.2017.400.

Cao, X., Wei, Y., Wen, F., & Sun, J. (2012). Face alignment by explicit shape regression. In *CVPR* (pp. 2887–2894). doi:10.1109/CVPR.2012.6248015.

Chandran, P., Bradley, D., Gross, M., & Beeler, T. (2020). Attention-driven cropping for very high resolution facial landmark detection. In *2020 IEEE/CVPR* (pp. 5860–5869). doi:10.1109/CVPR42600.2020.00590.

Dapogny, A., Cord, M., & Bailly, K. (2019). Decafa: Deep convolutional cascade for face alignment in the wild. In *ICCV* (pp. 6892–6900). doi:10.1109/ICCV.2019.00699.

Dong, X., Yan, Y., Ouyang, W., & Yang, Y. (2018). Style aggregated network for facial landmark detection. In *CVPR* (pp. 379–388). doi:10.1109/CVPR.2018.00047.

Dong, X., Yang, Y., Wei, S.-E., Weng, X., Sheikh, Y., & Yu, S.-I. (2021). Supervision by registration and triangulation for landmark detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*, 3681–3694. doi:10.1109/TPAMI.2020.2983935.

Dou, P., Shah, S. K., & Kakadiaris, I. A. (2017). End-to-end 3d face reconstruction with deep neural networks. In *CVPR* (pp. 1503–1512). doi:10.1109/CVPR.2017.164.

D.P. Kingma, & J.Ba (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, .

Drouard, V., Horaud, R., Deleforge, A., Ba, S., & Evangelidis, G. (2017). Robust head-pose estimation based on partially-latent mixture of linear regressions. *IEEE Transactions on Image Processing*, *26*, 1428–1440. doi:10.1109/TIP.2017.2654165.

Esen, H., Esen, M., & Ozsolak, O. (2017). Modelling and experimental performance analysis of solar-assisted ground source heat pump system. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*, 1–17. doi:10.1080/0952813X.2015.1056242.

Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008a). Artificial neural networks and adaptive neuro-fuzzy assessments for ground-coupled heat pump system. *Energy and Buildings*, *40*, 1074–1083. doi:https://doi.org/10.1016/j.enbuild.2007.10.002.

Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008b). Forecasting of a ground-coupled heat pump performance using neural networks with statistical data weighting pre-processing. *International Journal of Thermal Sciences*, *47*, 431–441. doi:https://doi.org/10.1016/j.ijthermalsci.2007.03.004.

Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008c). Performance prediction of a ground-coupled heat pump system using artificial neural networks. *Expert Systems with Applications*, *35*, 1940–1948. doi:https://doi.org/10.1016/j.eswa.2007.08.081.

Esen, H., Ozgen, F., Esen, M., & Sengur, A. (2009). Artificial neural network and wavelet neural network approaches for modelling of a solar air heater. *Expert Systems with Applications*, *36*, 11240–11248. doi:https://doi.org/10.1016/j.eswa.2009.02.073.

Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). Effective face frontalization in unconstrained images. In *CVPR* (pp. 4295–4304). doi:10.1109/CVPR.2015.7299058.

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). Searching for mobilenetv3. In *ICCV* (pp. 1314–1324). doi:10.1109/ICCV.2019.00140.

Huang, T. S., Bruckstein, A. M., Holt, R. J., & Netravali, A. N. (1995). Uniqueness of 3d pose under weak perspective: a geometrical proof. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *17*, 1220–1221. doi:10.1109/34.476515.

Huang, X., Deng, W., Shen, H., Zhang, X., & Ye, J. (2020). Propagationnet: Propagate points to curve to learn structure information. In *CVPR* (pp. 7263–7272). doi:10.1109/CVPR42600.2020.00729.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *CVPR* (pp. 936–944). doi:10.1109/CVPR.2017.106.

Liu, H., Lu, J., Feng, J., & Zhou, J. (2018). Two-stream transformer networks for video-based face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*, 2546–2554. doi:10.1109/TPAMI.2017.2734779.

Liu, L., Wang, Q., Zhu, W., Mo, H., Wang, T., Yin, S., Shi, Y., & Wei, S. (2019). A face alignment accelerator based on opti-

mized coarse-to-fine shape searching. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*, 2467–2481. doi:`10.1109/TCSVT.2018.2867499`.

Liu, Y., Shi, H., Shen, H., Si, Y., Wang, X., & Mei, T. (2020). A new dataset and boundary-attention semantic segmentation for face parsing. In *AAAI 2020* (pp. 11637–11644). URL: `https://aaai.org/ojs/index.php/AAAI/article/view/6832`.

Martins, P., & Batista, J. (2008). Accurate single view model-based head pose estimation. In *AFGR* (pp. 1–6). doi:`10.1109/AFGR.2008.4813369`.

Mo, H., Liu, L., Zhu, W., Li, Q., Liu, H., Yin, S., & Wei, S. (2020). A multi-task hardwired accelerator for face detection and alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*, 4284–4298. doi:`10.1109/TCSVT.2019.2955463`.

Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision – ECCV 2016* (pp. 483–499). Cham: Springer International Publishing.

Patacchiola, M., & Cangelosi, A. (2017). Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods. *Pattern Recognition*, *71*, 132 – 143. doi:`https://doi.org/10.1016/j.patcog.2017.06.009`.

Qian, S., Sun, K., Wu, W., Qian, C., & Jia, J. (2019). Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *ICCV* (pp. 10152–10162). doi:`10.1109/ICCV.2019.01025`.

Ranjan, R., Patel, V. M., & Chellappa, R. (2019). 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*, 121–135. doi:`10.1109/TPAMI.2017.2781233`.

R.Liu, L.Joel, P.Molino, F.P.Such, E.Frank, A.Sergeev, & J.Yosinski (2018). An intriguing failing of convolutional neural networks and the coordconv solution. In *NIPS* NIPS'18 (p. 9628–9639). Red Hook, NY, USA: Curran Associates Inc.

Ruiz, N., Chong, E., & Rehg, J. M. (2018). Fine-grained head pose estimation without keypoints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 2155–215509). doi:`10.1109/CVPRW.2018.00281`.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW* (pp. 397–403). doi:`10.1109/ICCVW.2013.59`.

Shao, X., Xing, J., Lyu, J., Zhou, X., Shi, Y., & Maybank, S. J. (2021). Robust face alignment via deep progressive reinitialization and adaptive error-driven learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (pp. 1–1). doi:`10.1109/TPAMI.2021.3073593`.

Shapira, G., Levy, N., Goldin, I., & Jevnisek, R. J. (2021). Knowing when to quit: Selective cascaded regression with patch attention for real-time face alignment. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 2372–2380).

Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., & Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCVW* (pp. 1003–1011). doi:`10.1109/ICCVW.2015.132`.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *NIPS* NIPS'14 (p. 568–576). Cambridge, MA, USA: MIT Press.

Sun, K., Wu, W., Liu, T., Yang, S., Wang, Q., Zhou, Q., Ye, Z., & Qian, C. (2019). Fab: A robust facial landmark detection framework for motion-blurred videos. In *ICCV* (pp. 5461–5470). doi:`10.1109/ICCV.2019.00556`.

Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., & Zhou, Q. (2018a).

Look at boundary: A boundary-aware face alignment algorithm. In *CVPR* (pp. 2129–2138). doi:`10.1109/CVPR.2018.00227`.

Wu, W., & Yang, S. (2017). Leveraging intra and inter-dataset variations for robust face alignment. In *CVPRW* (pp. 2096–2105). doi:`10.1109/CVPRW.2017.261`.

Wu, W., Yin, Y., Wang, Y., Wang, X., & Xu, D. (2018b). Facial expression recognition for different pose faces based on special landmark detection. In *ICPR* (pp. 1524–1529). doi:`10.1109/ICPR.2018.8545725`.

Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *CVPR* (pp. 532–539). doi:`10.1109/CVPR.2013.75`.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). Wider face: A face detection benchmark. In *CVPR* (pp. 5525–5533). doi:`10.1109/CVPR.2016.596`.

Yang, T.-Y., Chen, Y.-T., Lin, Y.-Y., & Chuang, Y.-Y. (2019). Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1087–1096). doi:`10.1109/CVPR.2019.00118`.

Zadeh, A., Baltrusˇaitis, T., & Morency, L. (2017). Convolutional experts constrained local model for facial landmark detection. In *CVPRW* (pp. 2051–2059). doi:`10.1109/CVPRW.2017.256`.

Zhang, C., & Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. In *WACV* (pp. 1036–1041). doi:`10.1109/WACV.2014.6835990`.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*, 1499–1503. doi:`10.1109/LSP.2016.2603342`.

Zhu, S., Li, C., Loy, C. C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *CVPR* (pp. 4998–5006). doi:`10.1109/CVPR.2015.7299134`.

Zhu, X., Lei, Z., Liu, X., Shi, H., & Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *CVPR* (pp. 146–155). doi:`10.1109/CVPR.2016.23`.

Zou, X., Xiao, P., Wang, J., Yan, L., Zhong, S., & Wu, Y. (2020). Towards unconstrained facial landmark detection robust to diverse cropping manners. *IEEE Transactions on Circuits and Systems for Video Technology*, (pp. 1–1). doi:`10.1109/TCSVT.2020.3006236`.