Attentional Alignment Network

Lei Yue[1,2]
yuelei@buaa.edu.cn
Xin Miao[3]
xin.miao@mavs.uta.edu
Pengbo Wang[4]
wangpengbo_vincent@sjtu.edu.cn
Baochang Zhang[1]
bczhang@buaa.edu.cn
Xiantong Zhen[1,2]
zhenxt@buaa.edu.cn
Xianbin Cao[1,2]
xbcao@buaa.edu.cn

[1] Beihang University, Beijing, China

[2] The Key Laboratory of Advanced Technologies for Near Space Information Systems
Ministry of Industry and Information Technology of China

[3] University of Texas at Arlington
TX, USA

[4] Shanghai Jiao Tong University
Shanghai, China

Abstract

Face alignment has recently generated great popularity in computer vision due to its widespread applications. The cascaded regression model has dominated and achieved great progress in the last decade, which however suffers from innate shortcomings, e.g., reliance on initialization. In this work, we propose attentional alignment networks (AAN), a novel end-to-end convolutional architecture for direct face alignment without relying on cascaded regression. AAN incorporates the attention mechanism into a convolutional regression network, which generates multiple attention maps for different convolutional layers to capture distinctive features in different granularity; by introducing intermediate supervision to create top-down attention maps, AAN attends to regions around facial landmarks, which enables it to establish more informative and discriminative representation closely related to facial landmarks. Extensive experiments on four commonly-used benchmark datasets demonstrate that the proposed AAN consistently delivers high performance on all datasets, surpassing previous methods by large margins, which shows its great effectiveness for direct face alignment.

1 Introduction

Face alignment, also known as facial landmark detection, is to localize the coordinates of a set of predefined points on the face. Face alignment has recently drawn increasing research interest since it serves as an important prerequisite for face related tasks including face recognition, face animation and face reconstruction. However, face alignment is essentially a challenging task due to the great variations of facial images and huge variabilities of the associated facial shapes.

The cascaded regression model has dominated in face alignment and achieved significant progress on several benchmarks in the past decade. Nevertheless, cascaded regression models suffer from several innate shortcomings. They require a shape initialization and tend to be trapped in local minima when the shape initial is far from the true shape. Moreover, since cascaded models work with local features, e.g., SIFT, they are innately unable to fully capture the holistic facial shape information, which however is of great importance for landmark detection. Recently, convolutional networks have been introduced for face alignment in [23] which simply uses a CNN cascade to regress facial landmark locations, suffering from the same shortcomings of cascaded models. Under a multi-task learning framework, the task constrained deep convolutional network (TCDCN) [34] was developed for face alignment without using iterative cascaded regression; while it requires auxiliary facial attributes, e.g., facial expression, pose orientation and gender, which however would not be always available and therefore limit its applications to different datasets. DSRN [15] constructed a multi-target regression model [35] [36] to disentangle highly nonlinear relationships between images and shapes, and incorporated a linear layer of low-rank learning to improve the performance.

In this work, we propose an attentional alignment network (AAN) to achieve direct face alignment with a fully end-to-end convolutional regression network. Instead of using the iterative cascaded regression, AAN takes the image as input and outputs the coordinates of face shape directly. Regular deep convolutional architectures do not consciously extract detailed features complementary to high-level global features, while desired features located within a small region around the facial landmark. In order to extract features that are relevant to facial landmarks, we introduce the attention mechanism associated with intermediate supervision in AAN to steer feature extraction to focus on regions around landmarks. Specifically, as shown in Fig. 1, we adopt the residual bottleneck block as the basic building unit; spatial attention maps are generated from the output features of each residual block to assign attention scores to the output feature maps, which produce prediction associated with a loss for intermediate supervision. Moreover, to fully capture multi-scale features, we generate multiple attention maps from intermediate convolutional layers to explore distinctive features residing in different granularity. To the best of our knowledge, the top-down spatial attention is for the first time incorporated in convolutional regression networks for end-to-end face alignment though it has achieved great success in various tasks. Consequently, high-resolution attention maps focus on local fine appearance while low-resolution attention maps capture a relatively coarse holistic view of facial shapes.

In addition, to further enhance the ability of AAN in handling the highly nonlinear relationship between image representations and facial shapes, we introduce a new nonlinear embedding layer with a cosine activation, which is derived from kernel approximation and therefore inherits the strong nonlinear learning ability of kernels.

In summary, we make contributions in the following three major aspects.

- We propose a new convolutional regression network, called attentional alignment networks, which establishes a fully end-to-end learning architecture for direct face alignment without relying on indirect and iterative cascaded regression.
- We introduce the attention mechanism associated with intermediate supervision. It not only allows to explore the distinctive features of the different scale from intermediate convolutional layers, but also drives the networks to extract features that are most relevant to facial landmarks.
- We design a new nonlinear embedding layer to handle the complex relationship between image representations and facial shapes. It leverages the strong nonlinear learn-

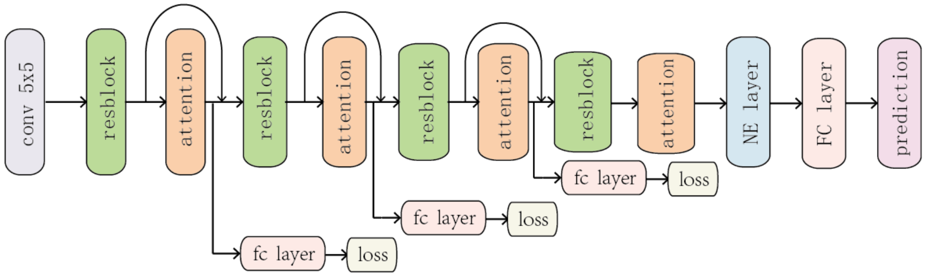

Figure 1: The end-to-end learning architecture of attentional alignment networks (AAN).

ing ability of kernels without forgoing the end-to-end learning architecture of convolutional neural networks.

We have conducted the extensive experimental evaluation on four benchmark datasets including AFLW, 300W, 300VW and CelebA, for face alignment. Our proposed AAN achieves state-of-the-art performance on all tasks and consistently surpasses previous methods, which demonstrates its great effectiveness as an end-to-end learning architecture for direct face alignment.

2 Related Work

Face alignment has witnessed great progress in the last decades, while most of existing methods were developed based on the cascaded regression models [27]. Despite of the great success in many face alignment tasks, cascaded models suffer from innate shortcomings due to its iterative regression with required shape initialization.

Recently, alternative approaches have been developed to avoid iterative regression for face alignment. Zhu *et al*. propose a coarse-to-fine searching method to handle large poses [37], which achieved impressive performance on challenging datasets, e.g., AFLW, by combining multiple handcrafted features, e.g. SIFT, HOG, and BRIEF. Recently, convolutional neural networks (CNNs) have been introduced for face alignment. Sun *et al*. [23] proposed to use a three-stage algorithm with CNNs, which estimated positions of different facial positions at the first stage, and refined the positions in the remaining stage. Face alignment algorithm based on RNN was further developed by Liu *et al*. [11], to take temporal information of the video-based datasets. However, due to their cascaded regression based structure, they still suffered from improper initialization. TCDCN [34] use one-stage multi-task CNNs to estimate facial landmarks and auxiliary attributes at the same time. Compared to the prior multi-stage framework, it can predict the coordinates of the landmarks instead of shape increments, and optimize jointly. However, TCDCN needs to provide extra auxiliary information to guarantee the performance of face alignment in the test stage. In contrast, we directly predict the coordinates of facial landmarks with no need of auxiliary information, which makes our method more generalized and applicable for the wider range of tasks.

Face alignment via 3D face models was recently shown to handle large pose variations [2, 7, 40]. Zhu et al. [40] proposed a solution to a new alignment framework, 3D Dense Face Alignment (3DDFA), for face alignment. In 3DDFA, a dense 3D face model is fitted to the image via a CNN and a method to synthesize many profile training samples is presented to solve the problem of labeling invisible landmarks. To handle faces with large pose variations, the powerful cascaded CNN regressor is combined with a 3D Morphable Model (3DMM) [7]. Face alignment is formulated as a 3DMM fitting problem, where the camera projection matrix and 3D shape parameters are estimated by a cascade of CNN-based regressors. It is interesting to resort to 3D modeling for face alignment, but in so doing, one increases the computational cost.

Attention mechanism has recently been introduced into deep learning to augment its performance, which shows great effectiveness in a broad range of visual tasks, e.g., image captioning [28] [30], image question answering [29] , object detection [18], pedestrian analysis [12] and image classification [26]. In this work, we explore the soft attention mechanism, which is computationally efficient and can be updated by back-propagation. To the best of our knowledge, our attentional alignment network, is the first to introduce attention mechanism for face alignment. More importantly, instead of using a single attention map, we generate multiple attention maps in different resolutions, which explore both local and global features, achieving more comprehensive representation for improved face alignment.

3 Attentional Alignment Networks

In this section, we provide the architectural overview of the proposed attentional alignment networks in Sec. 3.1, describe the attention mechanism in Sec. 3.2 and introduce the newly designed nonlinear embedding layer in Sec. 3.3.

3.1 Architecture Overview

Face alignment is to find the mapping relationship between the input image I and the face shape S which is represented by the coordinates of landmarks. Our attentional alignment network (AAN) directly predicts the coordinates of face shape from images in an end-to-end learning architecture, distinguishing from cascaded regression methods which update the face shape increments iteratively. Our AAN is composed of four stacking building blocks as illustrated in Fig. 1. In each building block, one residual bottleneck block is followed by one attention block, and the residual bottleneck block is the same as [6] . Attention block is an 1×1 convolutional layer with BN and relu activation function, and generates an attention map using the output features of the residual block. The attention map is normalized by a sigmoid layer into the range of $[0,1]$. If the network is more interested in a specific location, the corresponding attention score will be higher. The attention map would assign attention scores to the output feature of the residual block through elementary-wise multiplication. Then this output feature and the original one would be passed into the next building block together. To be more precise, given the input $\mathbf{x}$, the output of building block F is :

$$F(\mathbf{x}) = (1 + M(\mathbf{x})) * H(\mathbf{x}) \tag{1}$$

where $H(\mathbf{x})$ is the output of the residual block, $M(\mathbf{x})$ denotes the attention map in the range of $[0,1]$. $*$ means elementary-wise multiplication. This residual connection is benefitted from the effectiveness of residual learning. connection way as Equation2

$+$

$$F(\mathbf{x}) = H(\mathbf{x}) * M(\mathbf{x}) \tag{2}$$

Feeding the output $F(\mathbf{x})$ as the input to the next building block directly leads to sharp plunge in feature values due to the repeated dot product with $M(\mathbf{x})$.

To leverage the effectiveness of residual learning which offers a powerful mechanism in solving back propagation degradation problem, showing impressive performance in various tasks, we build our attentional network based on the residual learning module, that is, $F(\mathbf{x})$ in (2) is replaced by

We would like to highlight that deploying the residual block can effectively alleviate the value degradation problem due to the use of multiple attention maps.

3.2 Top-Down Spatial Attention

Face images usually exhibit greatly varied appearances from which it is highly desired to extract features that are directly and closely related to facial landmarks. However, ordinary convolutional networks for facial feature extraction are not necessarily aware of the locations of facial landmarks. In this work, we incorporate the attention mechanism into the convolutional networks and design top-down spatial attention by introducing the intermediate supervision, which steers the network to automatically focus on regions directly related to facial landmarks. Moreover, facial attribute locations are mainly determined by two factors, that is, local features related to facial parts, e.g., eyes, and the holistic view of each part of the whole face. We therefore propose to generate multi-scale attention maps from intermediate convolutional layers of different resolutions, which enables it to capture complementary information for more informative facial representation.

Intermediate Supervision. We introduce the intermediate supervision into the network to create top-down spatial attention by injecting an intermediate branch with a shortcut connection with landmarks. We use the intermediate features weighted by attention maps to conduct early prediction of facial landmarks upon which a loss can be applied. Back-propagation proceeds from not only the final output, but also the local predictions from each attention module. Making early prediction force attention maps represent the local appearance in a fine-grained way and have a high-level understanding of the image. To be more specific, we apply the same ground truth to all the intermediate prediction and sum up into the overall loss in the optimization process. The intermediate supervision works seamlessly with back-propagation and allows errors to flow directly to the intermediate layers to adjust network parameters. As a result, the network is able to refine features at both local appearance and global contexts. In contrast to saliency-based bottom-up attention, the top-down attention drives the network to extract features around the locations of facial landmarks, while filtering out irrelevant features through suppression. In the test stage, we remove the intermediate branch and take the final output as the result directly.

Multi-Scale Attention. It is widely acknowledged that in CNNs, features in different depths contain different levels of semantics, that is, features in deeper layers encode high-order semantic information while those in shallower layers represents the local appearance. To fully capture the information for face alignment, we generate multiple attention modules to consolidate features of various semantics. Therefore, the attention maps from shallow layers would pay greater attention to local regions, e.g., eyes, nose, mouth and contour, while attention maps from deeper layer can capture the holistic view of the face. Compared to single attention design, which is limited to capture only one certain level of semantic features, our design of multi-scale attention enables the alignment network to have a comprehensive and coherent understanding of both whole facial shapes and local regions.

3.3 Nonlinear Embedding Layer

Although features from the convolutional layers reach a relatively high semantic level, it would still not be strong enough to handle the highly nonlinear relationship between images and facial shapes by simply feeding the features into a linear fully connected layer. We therefore design a new nonlinear embedding layer by leveraging the strength of kernels.

We derive the nonlinear embedding layer by kernel approximation based on Bochner's theorem. It is well known that a shift-invariant kernel, e.g., radius basis function (RBF), can be approximated by finding explicit feature maps, that is,

$$\phi(\mathbf{x}_i) = \sqrt{\frac{2}{d}}[\cos(\omega_i^\top \mathbf{x}_i + b_i)]_{1:d} \tag{3}$$

where cos is the elementary-wise cosine function, $\phi(\mathbf{x})$ is called the random Fourier feature [16] which has been successfully used for kernel approximation. We propose learning features in data-driven way rather than using random sampling. The nonlinear embedding layer takes the form of

$$\phi = \cos(W\mathbf{x} + \mathbf{b}) \tag{4}$$

where $\mathbf{x} \in \mathbf{R}^D$ is the feature vector produced by the last residual block, $W \in \mathbf{R}^{d \times D}$ is the wight matrix of the layer, and b is the bias. (4) turns out to be fully connected layer with cosine activations, which can be injected into convolutional networks for end-to-end training. Following the nonlinear embedding layer, a linear fully connected layer is deployed to directly predict the coordinates of landmarks.

4 Experiments and Results

We conduct extensive experiments on four benchmark datasets, i.e., AFLW, 300W, CelebA, and 300VW including both images and videos. Experiments demonstrate that the proposed AAN delivers high performance on all datasets, largely exceeding previous methods.

4.1 Datasets

We provide the description of the datasets used in this work, associated with experimental settings to benchmark with previous methods. Facial images are cropped according to the bounding boxes provided with datasets. We do not use any data augmentation though the performance could be further improved with more augmented data for training.

AFLW [9] contains around 25,000 face images collected in the wild. It is regarded as one of most challenging datasets for face alignment due to huge face yaw angles between ±90°, extreme expressions, large variety in face appearance (e.g. gender, occlusion and ethnicity). The images were originally labeled with 21 landmarks and a bounding box. We discard the ear landmarks to keep the experiment settings consistent with cascaded compositional learning [38]. Following common setting, we use the subset with 20,000 images for training and 4,386 images for testing, respectively.

300W [19] [20] consists of several databases with 68 landmarks, including AFW [39], HELEN [10], LFPW [1], XM2VTS [14] and IBUG [21]. It is widely used to evaluate near-frontal face alignment. Following the widely-used settings, we test on a common set and a challenging set, respectively.

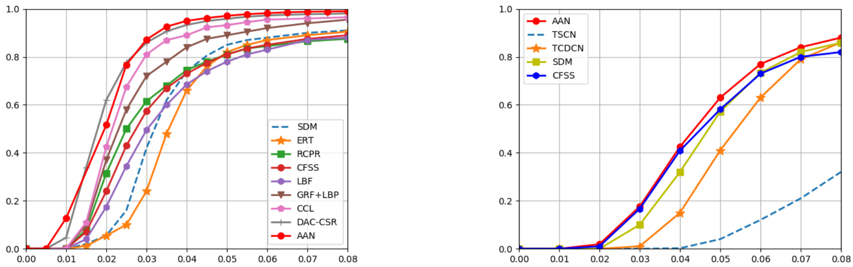

Figure 2: Comparison in terms of CED on AFLW (left) and 300VW-Category 3 (right).

CelebA [13] is one of the largest dataset including 202,599 images with large pose variations and background clutter. Each image contains five landmarks localizing eyes, nose and mouth corners. To facilitate comparison, we follow the original work [13], and take 182,632 images for training and 19,926 for testing.

300VW [21] is a video-based face alignment dataset and provides 114 videos in total. Faces in frames are annotated with 68 landmarks. In [21], a subset of 50 videos is selected to train the model , and the remaining are divided into three categories for testing.

4.2 Implementation Details

We design the network with four residual blocks and four pooling layers to extract features. Following each residual block, an attention block is attached to produce the attention map upon the output features of the residual block. The features from the last attention block are fed to the nonlinear embedding layer followed by a linear fully connected layer to produce the prediction of landmarks for direct face alignment.

In the training stage, we employ the weight decay and batch normalization. The weight decay for parameter is set to 0.001. We train the attention alignment network with stochastic optimization algorithm Adam, and set the mini-batch size 64. The initial learning rate is 1e-3, we decrease the learning rate to 1e-4 and 1e-5 after $25,000$ and $40,000$ epochs. The number d of notes in nonlinear embedding layer is set to 256 by cross validation.

We use the normalized mean error (NME) as the evaluation metric defined as follows:

$$\mathrm{NME} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(\hat{x}_i - x_i)^2 - (\hat{y}_i - y_i)^2}/d, \tag{5}$$

where (x_i, y_i) is the ground truth and $(\hat{x}_i, \hat{y}_i)$ is the prediction coordinates of facial landmarks. N denotes the total number of landmarks on the face and d is the normalization distance.

Method	CDM [31]	PCPR [3]	ERT [8]	SDM [27]	LBF [17]	POCR [25]	CFSS [37]	CLL [38]	CSR [5]	AAN
Error	5.43	3.73	4.35	4.05	4.25	5.32	3.92	2.72	2.27	**1.73**

Table 1: Performance comparison on AFLW (NME error in %)

Method	Common Subset	Challenging Subset	Full Test set
RCPR [3]	6.18	17.26	8.35
SDM [27]	5.57	15.40	7.52
ESR [4]	5.28	17.00	7.58
GN-DPM [25]	5.78	-	-
ERT [8]	-	-	6.40
CFAN [32]	5.50	16.78	7.69
LBF [17]	4.95	11.98	6.32
CFSS [37]	4.73	9.98	5.76
MDM [24]	4.83	10.14	5.88
AAN	**4.38**	**9.44**	**5.39**

Table 2: Performance comparison on 300W (NME error in %)

d is used for normalization, and specifically, we use the inter-ocular distance to normalize the mean error for near-frontal view datasets such as celebA. For the large-pose face dataset, e.g. AFLW, we take the bounding box size as the normalization distance. For brevity, % is omitted in all tables. The evaluation results are also shown in the form of the cumulative error distribution (CED) curve for comprehensive comparison.

4.3 Results

Our AAN achieves state-of-the-art results across all four datasets and outperforms previous methods by large margins in most cases. In which follows, we explain the results accompanied with discussions on each dataset.

The results on AFLW are reported in Table 1. Our AAN achieves highest performance among all compared methods with an error of 1.73, which is impressive considering the great challenges of AFLW. The intuitive results of our AAN are shown in the top row of Fig 3, from which we can see that AAN can produces high accurate prediction on faces with illumination, occlusion and huge orientation.

The results on 300W are summarized in Table 2. Our AAN also achieves the best results and outperforms previous methods consistently. The intuitive results of our AAN are also illustrated in the middle row of Fig. 3. Our AAN can well handle challenging cases with great head orientation and large occlusions.

The results on CelebA are compared in Table 3. Again, our AAN achieves the best performance with an error of 2.99, largely surpassing representative methods, e.g., RCPR and CFSS. It can also be observed from the second last row of Fig. 3 that our AAN can achieve accurate prediction on facial images with great appearance variations.

The results on 300VW are listed in Table 4. Our AAN achieves the best performance on test sets of Category 1 and 3. It is worth mentioning that Category 3 is regarded as the most challenging subset, while our method achieves the best results with a mean error of

Method	RCPR [3]	SDM [27]	CFSS [37]	**AAN**
Error	4.12	4.35	3.95	**2.99**

Table 3: Performance comparison on CelebA (NME error in %)

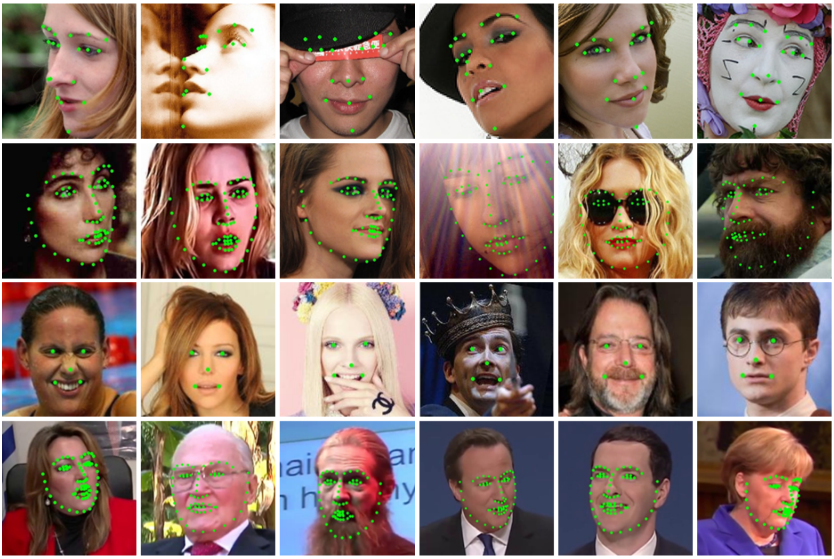

Figure 3: Intuitive illustration of the prediction results.

Method	Category 1	Category 2	Category 3
SDM [27]	7.41	6.18	13.04
TSCN [22]	12.54	7.25	13.13
CFSS [37]	7.68	6.42	13.67
TCDCN [34]	7.66	6.77	14.98
TSTN [11]	5.36	**4.51**	12.84
AAN	**5.03**	4.82	**7.98**

Table 4: Performance comparison on 300VW (NME error in %)

7.98, dramatically surpassing previous methods with large margins, which again indicates its great effectiveness for direct face alignment.

For more comprehensive comparison, we have plotted the cumulative error distribution (CED) curves of the results in Fig. 2. We use AFLW and 300VW as representatives due to the space limit. Our AAN is clearly above all compared methods, which again shows its performance advantages for face alignment. Moreover, we have also conducted an ablation study in Table 5 by removing (w/o) the attention mechanism and the nonlinear embedding layer. The results have shown that both largely improve the overall performance, which verifies their effectiveness for face alignment.

Additionally, it would be interesting to look into how attention mechanism works in our attentional alignment network. We pint the multiple attention maps of different convolutional layers in Fig. 4 in which the top row shows the original face images. It is easy to find attention maps generally highlight the regions that are closely related to facial landmarks. Specially, attention maps from shallow layers (upper rows) focus on local regions while those from deeper layers (lower rows) tend to capture the holistic view of facial shapes. The intuitive illustration demonstrates that the introduced multi-scale attention helps to learn

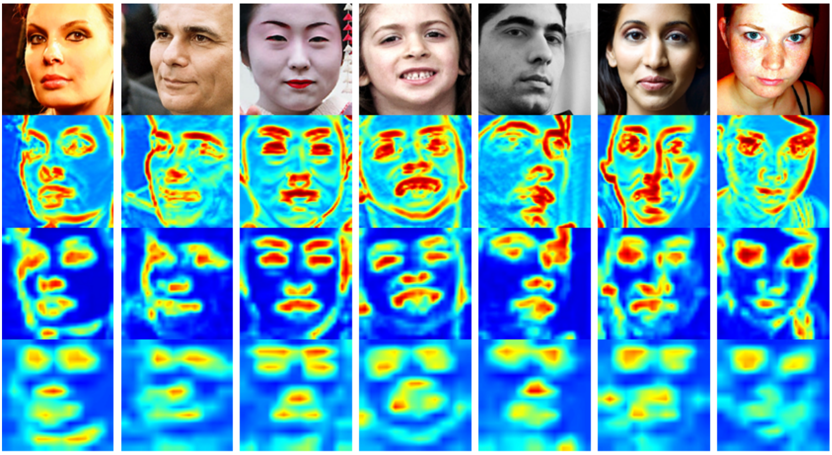

Figure 4: Multiple attention maps from shallow (higher rows) and deep (lower rows) layers.

	AAN	w/o attention	w/o NE layer	w/o both
AFLW	1.73	2.06	1.78	2.10
300VW-Category 3	7.98	9.38	8.15	9.60

Table 5: The effectiveness of the attention mechanism and the nonlinear embedding layer.

more distinctive features, which enables more accurate face alginment.

5 Conclusions

We have presented an attentional alignment network (AAN) for direct face alignment without relying on cascaded regression. We introduce attention mechanism associated with intermediate supervision to face alignment and design a new nonlinear embedding layer derived from kernel approximation to handle nonlinear image-shape relationships. Our AAN achieves a new end-to-end learning architecture that combines the strengths of neural networks and kernels for face alignment. Experiments on four benchmark datasets have demonstrated the effectiveness of our AAN for direct face alignment.

6 Acknowledgments

This paper was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1200100, in part by the National Natural Science Foundation of China under Grant 91538204 and Grant 61425014, in part by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China under Grant 61521091.

References

[1] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, volume 1, page 8, 2017.

[3] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[4] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[5] Zhen-Hua Feng, Josef Kittler, William Christmas, Patrik Huber, and Xiao-Jun Wu. Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. *arXiv preprint arXiv:1611.05396*, 2016.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4188–4196, 2016.

[8] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[9] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151. IEEE, 2011.

[10] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012.

[11] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Two-stream transformer networks for video-based face alignment. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[12] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *arXiv preprint arXiv:1709.09930*, 2017.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[14] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966, 1999.

[15] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[17] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[18] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[19] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 397–403. IEEE, 2013.

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

[21] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 50–58, 2015.

[22] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[23] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

[24] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[25] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[26] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *arXiv preprint arXiv:1704.06904*, 2017.

[27] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013.

[28] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[29] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[30] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016.

[31] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.

[32] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[33] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[34] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930, 2016.

[35] Xiantong Zhen, Mengyang Yu, Xiaofei He, and Shuo Li. Multi-target regression via robust low-rank learning. *IEEE transactions on pattern analysis and machine Intelligence*, 40(2):497–504, 2018.

[36] Xiantong Zhen, Mengyang Yu, Feng Zheng, Ilanit Ben Nachum, Mousumi Bhaduri, David Laidley, and Shuo Li. Multitarget sparse latent regression. *IEEE transactions on neural networks and learning systems*, 29(5):1575–1586, 2018.

[37] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

[38] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained face alignment via cascaded compositional learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3409–3417, 2016.

[39] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.

[40] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016.