Two-Stream Transformer Networks for Video-Based Face Alignment

Article *in* IEEE Transactions on Pattern Analysis and Machine Intelligence · August 2017

DOI: 10.1109/TPAMI.2017.2734779

Two-Stream Transformer Networks for Video-Based Face Alignment

Hao Liu, Jiwen Lu, *Senior Member, IEEE*, Jianjiang Feng, *Member, IEEE*, and Jie Zhou, *Senior Member, IEEE*

Abstract—In this paper, we propose a two-stream transformer networks (TSTN) approach for video-based face alignment. Unlike conventional image-based face alignment approaches which cannot explicitly model the temporal dependency in videos and motivated by the fact that consistent movements of facial landmarks usually occur across consecutive frames, our TSTN aims to capture the complementary information of both the spatial appearance on still frames and the temporal consistency information across frames. To achieve this, we develop a two-stream architecture, which decomposes the video-based face alignment into spatial and temporal streams accordingly. Specifically, the spatial stream aims to transform the facial image to the landmark positions by preserving the holistic facial shape structure. Accordingly, the temporal stream encodes the video input as active appearance codes, where the temporal consistency information across frames is captured to help shape refinements. Experimental results on the benchmarking video-based face alignment datasets show very competitive performance of our method in comparisons to the state-of-the-arts.

Index Terms—Face alignment, convolutional neural networks, recurrent neural networks, face tracking, biometrics

1 Introduction

Face alignment (*a.k.a.* facial landmark detection) attempts to localize facial landmarks for a given face image, which plays an important role in many facial analysis tasks, such as face verification [15], [38], face recognition [16] and facial attribute analysis [19]. Conventional methods [27], [46], [47], [48], [50] address the face alignment as a cascaded regression problem, which seeks a series of linear feature-to-shape mappings to refine the initial shape to the final shape in a coarse-to-fine manner. However, features employed in these methods are hand-crafted, which requires strong prior knowledge by hand. To address this issue, deep learning [40], [43], [45], [48] has been applied to learn discriminative features directly from image pixels to exploit the complex and nonlinear relationship between face data and facial shapes, which achieves significant improvements for the face alignment performance based on the image-based benchmarking datasets.

Recently, efforts have been devoted to address the problem of video-based face alignment, which aims to localize facial landmarks in unconstrained videos. In contrast to image-based face alignment methods, tracking-by-detection methods [5], [10], [42] have been proposed in video-base face alignment, which employ an incremental learning technique to detect facial landmarks on still frames. However, these methods cannot explicitly capture the temporal dependency relationship on adjacent frames, which is useful for video-based face alignment. To address this challenge, Peng et al. [25], [26] proposed two video-based face alignment methods, which utilize temporal information to flow across frames in a sequential manner. While encouraging performance has been obtained, these methods cannot explicitly exploit the complementary information of the appearance features in the spatial dimension and the consistency information in the temporal dimension accordingly.

In this paper, we propose a two-stream deep learning method for video-based face alignment. Motivated by the fact that the temporal consistency related to facial landmarks in videos is helpful to regarding with the variations of large poses, expressions and occlusions over time, our model aims to exploit the complementary information of the appearance information on still frames and the temporal consistency information across consecutive frames accordingly. To achieve this, we carefully design two-stream transformer networks (TSTN) which specifically consist of the spatial and temporal streams. The spatial stream network learns to transform the facial appearance features to landmark positions by preserving the holistic facial shape structure. Accordingly, the temporal stream network learns to embed the face sequence as the active appearance codes where the consistency information is integrated to refine the landmarks in the temporal dimension. The network parameters of the designed two-stream architecture are optimized by back-propagation in an end-to-end manner. Fig. 1 shows the pipeline of our proposed TSTN. Experimental results show that our method achieves very competitive performance compared with the state-of-the-arts on both controlled and uncontrolled video-based face alignment datasets.

- *The authors are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, and Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing 100084, China. E-mail: h-liu14@mails.tsinghua.edu.cn, {lujiwen, jfeng, jzhou}@tsinghua.edu.cn.*

Manuscript received 1 Mar. 2017; revised 26 June 2017; accepted 27 July 2017. Date of publication 31 July 2017; date of current version 10 Oct. 2018.
(Corresponding author: Jiwen Lu.)
Recommended for acceptance by S. Escalera, X. Baró, I. Guyon, H. J. Escalante, G. Tzimiropoulos, M. Valstar, M. Pantic, J. Coh, and T. Kanade.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2017.2734779

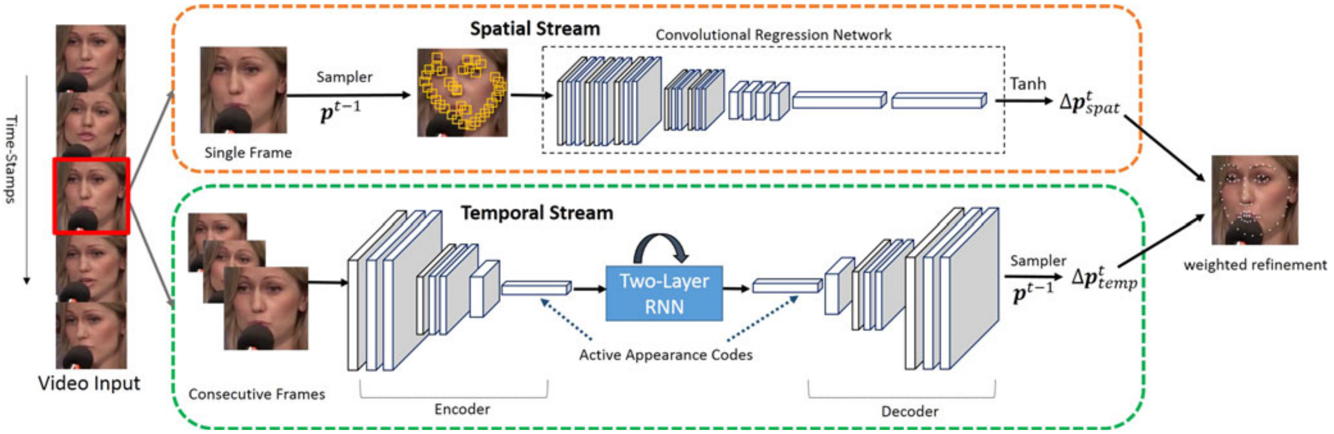


Fig. 1. The pipeline of the proposed two-stream transformer networks (TSTN). The input to our TSTN consists of both color (RGB) channels of still images and a video clip that contains a sequence of face images. The main objectives of our TSTN are two-fold: 1) the RGB channels are exploited as the appearance information in the spatial stream, and 2) the multiple frames (a video clip) are directly encoded as active appearance codes by exploiting the consistency information in the temporal stream. The final positions of facial landmarks are determined by a weighted fusion of both the spatial and temporal streams. The parameters of the designed networks are jointly optimized by back-propagation. During the testing phase, we feed a face sequence as the input to the learned two-stream networks and then predict a series of landmark positions for the face video clip.

2 Related Work

Conventional Face Alignment. Conventional image-based face alignment methods [7], [9], [17], [46], [47], [48], [49], [50] aim to learn a series of feature-to-shape regression functions to refine the facial shape progressively, which significantly improve the alignment accuracy. For example, Xiong et al. [46] proposed a supervised decent method (SDM) to learn a sequence of feature-to-shape mapping functions to refine the facial shapes. Zhu et al. [50] developed a coarse-to-fine shape searching (CFSS) approach to gradually shrink the possible facial shape space, which exhibits significant performance for the image-based face alignment. Recently, stereo-based models [6], [13], [14], [33] have also been proposed to explore 3D face alignment on a large, diverse corpora of multi-view face images annotated with 3D spatial information. However, these methods cannot directly model the temporal information on the consecutive frames, which is useful for video-based face alignment. In contrast to these methods, our model learns to exploit the temporal consistency information of landmarks of multiple frames.

Video-based face alignment [32], [34] focuses on detecting facial landmarks in a sequence of face images. Unlike existing image-based face alignment methods, efforts have been devoted to this problem recently [5], [8], [10], [34], [41], [42], which perform the incremental learning method to predict facial landmarks on still frames in a tracking-by-detection manner. However, these methods cannot explicitly model the temporal dependency relationship of landmarks across frames. To exploit this dependency relationship on frames, Peng et al. [25] proposed a sequential face alignment method in videos, which consists of a sequence of spatial and temporal recurrences. While their model has obtained promising performance on the video-based benchmark datasets, their sequential model cannot exploit the complementary information from the spatial and temporal dimensions accordingly. To address this, we propose a two-stream framework in this paper, which learns to capture the complementary information of the appearance information on still images in the spatial stream and the continuous consistency across multiple frames in the temporal stream. The final prediction of facial landmarks is determined by a weighted fusion of both spatial and temporal streams.

Face Alignment by Deep Learning. In recent years, deep learning has been adopted to face alignment [20], [36], [40], [44], [45], [48], which learns discriminative and robust features directly from pixels to model the nonlinear relationship between the face images and facial shapes. For example, Zhang et al. [48] presented a coarse-to-fine auto-encoder networks (CFAN) method to refine the landmark locations iteratively. Zhang et al. [49] developed a multi-task deep learning method dubbed TCDCN to learn feature representation for face alignment with additional expression and gender attributes. However, these methods have been designed to explore spatial and appearance features, which cannot explicitly model the temporal consistency information across frames. In order to include temporal information into the deep neural network, several attempts have been made in video-based visual analysis recently such as the two-stream convolutional networks [11], [21], [35] and the sequential pooling networks [22], [28]. Our motivation of this work is to propose a two-stream video-based face alignment framework by leveraging both the feed-forward and feed-back networks, which specifically refines the positions of facial landmarks by exploiting complementary information of the appearance information in static images and the temporal consistency information across frames accordingly.

3 Two-Stream Transformer Networks

Unlike the static image-based face alignment methods [7], [29], [41], [46], [50] cannot directly model the temporal consistency for facial landmarks across frames, in this work, we propose a two-stream transformer networks approach which decomposes the video input to the spatial and temporal streams. To achieve this, our model learns to transform image sequences to a series of facial shapes via the designed two streams, where the shape-sensitive appearance information in the spatial stream and the consistency information in the temporal stream are exploited simultaneously. Specifically, the spatial stream learns to transform image pixels to the landmark positions by preserving the global shape constraints on still frames. Accordingly, the temporal stream encodes the video input as active appearance codes across frames, where the temporal consistency information for each landmark is exploited to improve the face alignment accuracy in videos.

3.1 Problem Formulation

Suppose we have a training set $\{\mathbf{y}_i\}_{i=1:N}$, where N represents the number of training samples and $\mathbf{y}_i$ denotes the ith sample for the specific video clip. Let $\mathbf{y}_i^{1:T} = \{\mathbf{y}_i^1, \mathbf{y}_i^2, \ldots, \mathbf{y}_i^t, \ldots, \mathbf{y}_i^T\}$ denote a face sequence consisting of T frames, where $\mathbf{y}_i^t = (\mathbf{x}_i^t, \mathbf{p}_i^{t*})$ denotes the frame at time t and $\mathbf{p}_i^{t*}$ is the corresponding ground-truth of the facial landmark positions (T may be different for different person identities). Let $\mathbf{p}_i^t = [p_1, p_2, \ldots, p_l, \ldots, p_L]_i^{t'} \in \mathbb{R}^{2L\times 1}$ denote the coordinates of the global facial shape, the goal of face alignment is to transform face image $\mathbf{x}^t$ (a set of local patches) to facial shape residual $\Delta\mathbf{p}^t = \mathbf{p}^{t*} - \mathbf{p}^{t-1}$ at time t. Hence, the facial shape for the current tth frame is refined based on the shape of previous frame incrementally by ignoring index i for simplicity

$$\mathbf{p}^t = \mathbf{p}^{t-1} + \Delta\mathbf{p}^t. \tag{1}$$

Our basic idea in this work is to transform the face sequence $\{\mathbf{x}^t\}^{t=1:T}$ to a set of facial landmark locations $\{\mathbf{p}^t\}^{t=1:T}$ by exploiting the complementary information of the spatial appearance features in still frames in the spatial dimension and the consistency information of the adjacent frames in the temporal dimension. Moreover, we employ a pair of fusion weights β_1 and β_2 for the spatial stream and temporal stream, respectively. To achieve this, we formulate the following optimization objective function

$$\begin{aligned}\min_f J = \sum_i^N \sum_t^T \frac{1}{2} \left\| \Delta\mathbf{p}_i^t - \beta_1 f_{\text{spat}}(\mathbf{x}_i^t) - \beta_2 f_{\text{temp}}(\mathbf{x}_i^t) \right\|_2^2, \\ \text{subject to} \quad \beta_1 + \beta_2 = 1,\end{aligned} \tag{2}$$

where $f_{\text{spat}}(\cdot)$ and $f_{\text{temp}}(\cdot)$ denote the image-to-shape transformer functions of the spatial and temporal streams, respectively (typically, f denotes the network parameters), β_1 and β_2 are fusion weights to balance the importance between the predicted residuals of the spatial and temporal streams. $\mathbf{x}_i^t$ denotes the ith face sample for the time step t in the training set.

Our model is trained with the mixed coefficients β_1 and β_2, which performs the weighted fusion of the residuals $f_{\text{spat}}(\cdot)$ and $f_{\text{temp}}(\cdot)$ estimated by both streams. How to learn the spatial and temporal transformers $f_{\text{spat}}(\cdot)$ and $f_{\text{temp}}(\cdot)$ in (2) is the crucial part of our model. In this work, we develop a two-stream transformer networks architecture which incorporates the spatial stream transformer $f_{\text{spat}}(\cdot)$ to transform face image to landmark locations and the temporal stream transformer $f_{\text{temp}}(\cdot)$ to exploit temporal consistency information over time under the unified deep learning architecture. Next, we detail the two-stream architecture of the spatial and temporal networks, respectively.

3.2 Spatial Stream Network

In the spatial stream network, the motivation of this part is to localize facial landmarks directly from still face images. Hence, the spatial network learns to transform face image or local raw patches to facial shape residuals and then to refine the current facial shape based on the previous shape. To achieve this, we design a sampling transformer, dubbed sampler in this paper, which aims to extract the shape-index raw patches [46] directly from face images based on the initial shape and is plugged to the designed spatial stream. Moreover, we develop a convolutional regression network to predict facial shape residuals by utilizing these sampled shape-index features under the deep convolutional architecture. Fig. 1 demonstrates the network design of the spatial stream that consists of the sampler and convolutional regression net modules.

Sampler. The goal of the sampler layer is to sample local patches surrounding with the initial shape $\mathbf{p}$ as the input of the regression network. The main advantage of the proposed sampler module lies on that the shape-index patches succeed in preserving the holistic facial shape constraints during shape updating iterations [7], [46], [48]. Suppose we have these cropped patches denoted by $\mathbf{x}^t(\mathbf{p}+d)$, where d is the patch size that was specified to 26 in our experiments. To perform an end-to-end optimization procedure, we also provide the derivatives of the shape with respect to the loss, which is computed for each landmark p (certain point from the shape $\mathbf{p}$) as follows (ignoring the time step t)

$$\frac{\partial J}{\partial p} = \frac{\partial \mathbf{x}}{\partial p} \frac{\partial J}{\partial \mathbf{x}}, \tag{3}$$

$$\frac{\partial \mathbf{x}}{\partial p} = \nabla(\mathbf{x}(p+d)), \tag{4}$$

where d is the size of sampled shape-index patches, $\mathbf{x}(p)$ denotes the pixel value located at the landmark p and ∇ denotes the gradient-image w.r.t the cropped image patch, respectively. Since the derivatives of the shape-image are not strictly differentiable for 2D images, the value is approximated by the gradient of the image. Specifically, $\nabla(\mathbf{x}(p+d))$ is calculated by the Sobel operator [12] in size of $d \times d$ which is convolved on the image patches. The final result is summed up by performing gradients of total landmarks.

Convolutional Regression Net. The convolutional regression network consists of a sequence of convolutional layer, pooling layer, nonlinear ReLU [18] layer and inner product (fully connected) layer, which learns to predict the facial shape residual based on the extracted shape-index patches from pixels. Given a facial image $\mathbf{x}^t$ at the tth time step, we feed shape-index patches to the network and compute the feature representation $\mathbf{a}$ as follows:

$$\mathbf{a}^t = \text{Pool}\left(\text{ReLU}(\mathbf{W}^S \otimes \mathbf{x}^t(\mathbf{p}+d) + \mathbf{b}^S)\right), \tag{5}$$

where $\mathbf{W}^S$ and $\mathbf{b}^S$ represent the filter weights and bias of the spatial network, respectively, $\mathbf{x}^t(\mathbf{p}+d)$ denotes cropped local patches indexed by the initial shape, $\text{Pool}(\cdot)$ denotes the max-pooling operation, $\text{ReLU}(\cdot)$ denotes the nonlinear *ReLU* [18] function and $\otimes$ denotes the convolution operation of the designed convolutional regression network. Having obtained the immediate feature $\mathbf{a}$, we append two layers of fully connected neural networks to transform the feature $\mathbf{a}^t$ to facial shape residual $\Delta\mathbf{p}^t$ in the spatial stream. At the top of the raw prediction layer, we employ a nonlinear

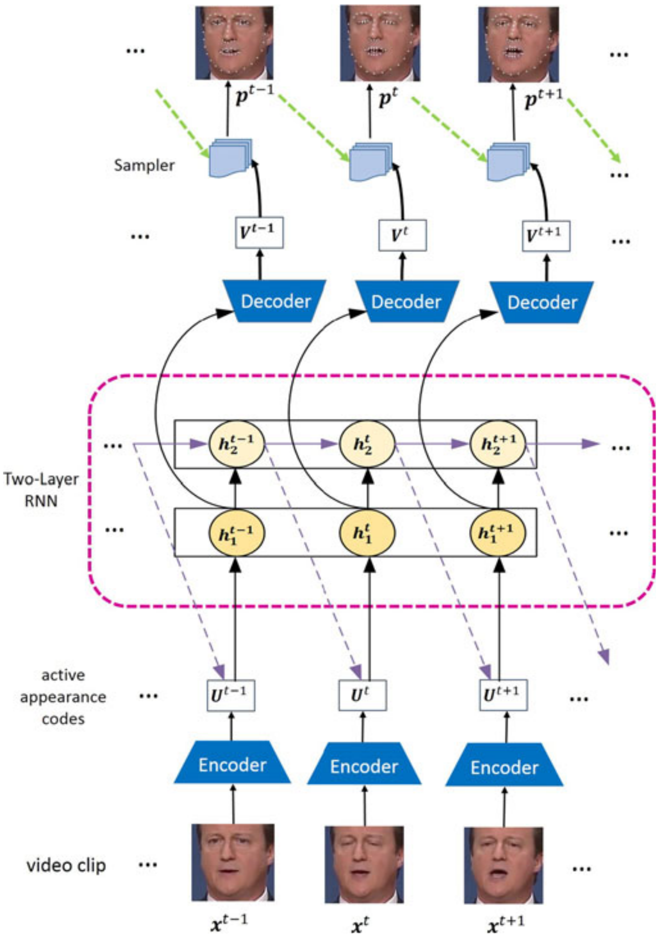


Fig. 2. The work flow of the temporal stream. The proposed temporal stream network consists of an encoder-decoder module to encode the spatial information as active appearance codes $\mathbf{U}$, and a two-layer RNN module to memorize the temporal information to flow across frames. Having obtained the decoded feature maps $\mathbf{V}$, we sample a set of patches and transform these spatial-temporal patches to facial landmark locations by a convolutional neural network.

hyperbolic-tangent (*tanh*) nonlinear function, which efficiently constrains the predicted shape residual within the range of $[-1, 1]$ [4], [37]. To further improve the alignment accuracy, we extend our model as a cascaded framework by sharing two cascaded executions, where the inputs of the current stage is fed with the outcomes of the previous stage and this enables a coarse-to-fine process for face alignment.

Overall, the positions of facial landmarks for still frames are predicted based on the proposed spatial stream, which contributes a reliable shape refinement in the spatial dimension. However, the spatial stream cannot explicitly model the temporal consistency for a video clip, so that the dependency relationship of frames cannot be utilized to disambiguate the challenging cases when the face data encounters large variations of facial aspect ratios, expressions and partial occlusions over time.

3.3 Temporal Stream Network

The basic idea of our temporal stream network is to discover shape-sensitive and spatial-temporal features for facial dynamics across the temporal dimension. To realize the temporal refinement mechanism, we carefully design a temporal stream network, which is equipped with an encoder-decoder module to compress the face sequence as the *active appearance codes* that encode the whole face changes in the temporal dimension, and a two-layer recurrent neural networks (RNN) module to memorize and flow the temporal information across consecutive frames by capturing the consistency information over time. Fig. 2 shows the network design of the proposed temporal stream network.

Encoder-Decoder Module. The encoder-decoder module plays an important role in the temporal stream network, which attempts to encode the image pixels across the timestamps as active appearance codes. Generally, the encoder-decoder module is equipped with an encoder network and an decoder network. To be specific, the encoder network aims to learn the spatial-temporal features, which embeds the local context details to refine each landmark. The decoder network remaps the learned codes to the same size of the origin face input, which preserves the spatial structural information for robust shape estimation.

The input of the encoder requires a sequence of facial images of $\mathbf{x}^t \in \mathbb{R}^{W\times H\times C}$ with width W, height H and channel C (RGB). The encoder architecture consists of a series of convolution, pooling and inner product layers and the encoded feature maps are computed as: $\mathbf{U}^t \in \mathbb{R}^{W'\times H'\times C'}$ with new width W', height H' and channel C' by $\mathbf{U}^t = \mathrm{Encoder}(\mathbf{x}^t)$, where $\mathrm{Encoder}(\cdot)$ denotes the parameters of the encoder network. By the encoder network, face frames are encoded as a set of active appearance codes, where the shape-informative features are captured across time, which is useful for video-based face alignment. Symmetrically, the decoder network performs a sequence of inner product, uppooling and deconvolution layers [23] to rescale the learned codes $\mathbf{U}^t$ from the encoder net to a multi-channel response maps $\mathbf{V}^t \in \mathbb{R}^{W\times H\times C''}$, the size of which is equivalent dimensions to the image input, where C'' denotes the channel dimension of the decoded feature maps. The decoded maps $\mathbf{V}^t$ are computed as: $\mathbf{V}^t = \mathrm{Decoder}(\mathbf{U}^t)$, where $\mathrm{Decoder}(\cdot)$ denotes decoder parameters.

Having obtained these learned feature maps $\mathbf{V}^t$ at time t, we take these maps as the input to a tiny convolutional network and obtain the final shape estimation by adding the facial shape predicted by that of the previous time step $t-1$.

Recurrent Module. As mentioned previously, the temporal stream learns to exploit the dependency relationship across frames by allowing the temporal information flow over time. Conventional feed-forward deep architectures [18], [24] cannot feasibly explore the temporal information from the full video input space. To achieve the contextual dependencies for facial landmark across frames, we develop a stacked two-layer recurrent neural network architecture. More intuitively, the first layer learns to capture the holistic spatial-temporal appearance features to be decoded for shape estimation, and the second layer memorizes the temporal information across frames.

As described in Fig. 2, the input of the recurrent module contains a set of feature maps $\mathbf{U}^t$ which are learned by the encoder for the observation at time t. For the first RNN layer, we formulate the hidden state $\mathbf{h}_1^t$ for the time step t as follows:

$$\mathbf{h}_1^t = \sigma(\mathbf{W}_1^I * \mathbf{U}^t + \mathbf{b}_1), \tag{6}$$

where $\mathbf{W}_1^I$ denotes the weights of the input-to-hidden fully connected layers and $\mathbf{b}_1$ denotes the bias. The input of the second RNN layer is the output of the first RNN layer, making this module a stacked RNN network. For the second RNN layer, the hidden stage $\mathbf{h}_2^t$ of the second RNN layer is formulated as follows:

$$\mathbf{h}_2^t = \sigma(\mathbf{W}_2^I * \mathbf{h}_1^t + \mathbf{W}_2^R * \mathbf{h}_2^{t-1} + \mathbf{b}_2), \tag{7}$$

where $\mathbf{W}_2^I$ denotes the weights of the input-to-hidden fully connected layers, $\mathbf{W}_2^R$ denotes the hidden-to-hidden connections across adjacent time steps and $\mathbf{b}_1$ denotes the bias. Each output of the second RNN layer is also concatenated with the encoded active appearance codes $\mathbf{U}^t$ as the spatial-temporal features, where the complementary information of the spatial and temporal streams is exploited accordingly for landmark localization. The hidden connections of the RNN layers share parameters across time steps, where the information on the RNN module's state at the previous time step can be memorized to flow across frames.

The proposed temporal stream including the encoder-decoder and recurrent modules incorporates the consistency over time during the temporal refinement process. Even regarding with the temporal occlusion, the shape-informative details for non-occluded parts in previous frames are memorized to flow across the succeeding frames for occluded parts. Overall, by incorporating the spatial stream, our model achieves the complementary information of appearance features with global shape-preserving prior in still frames and the temporal consistency across consecutive frames simultaneously, which performs the robust face alignment in videos.

Optimization. In order to train the proposed two-stream network, we employ our objective loss function (2) at the raw predictions of both the spatial and temporal streams. Since each component in the TSTN is differentiable (or approximated for sampler layer), errors can be back-propagated to all network layers and parameters of TSTN, making the parameters in $f_{\text{spat}}(\cdot)$ and $f_{\text{temp}}(\cdot)$ trainable with the stochastic gradient descent method. In terms of the RNN module in the temporal stream network, the parameters $\mathbf{W}_1^I$, $\mathbf{b}_1$, $\mathbf{W}_2^I$, $\mathbf{W}_2^R$, $\mathbf{b}_2$ of the temporal stream $f_{\text{temp}}(\cdot)$ are obtained by RMSProp [39].

3.4 Discussions

Differences with TSCN [35]. The spatial stream in TSCN [35] learns features directly from the whole frame without any shape-sensitive priors, which ignores the shape-informative details to completely depict facial shape structure. Differently, our spatial stream learns to crop shape-index local patches based on the initial shape, and then leverages convolutional neural network to predict the facial shape by global shape regression. As a result, both local and global shape-sensitive information are preserved simultaneously during the shape refining process. Moreover, the temporal stream in TSCN [35] needs pre-processing optical flow which is hand-crafted and may lead to local optima due to the two-stage manner. In contrast to TSCN [35], the employed RNN architecture automatically exploits the shape-sensitive features directly from the image pixels of previous frames, and then making inference for the succeeding frames.

Differences with REDN [25]. Our model leverages a two-stream deep learning method to jointly optimize the complementary residuals of both spatial and temporal streams. Hence, the landmark positions are refined simultaneously in both spatial and temporal dimensions. However, REDN [25] learned the spatial and temporal recurrences sequentially, which makes their framework be sensitive to the accuracy of previous shape prediction more heavily than ours.

4 Experiments

We conducted experiments on the video-based face alignment datasets including 300-VW [34] and TF [1] to evaluate the effectiveness of the proposed approach.

Datasets. 300-VW [34]: The 300 Videos in the Wild (300-VW) is a large dataset for video-based face alignment, which consists of 114 videos in various conditions. Each video lasts around 1 minute (25-30 images per second). By following the settings in [34], we utilized 50 sequences for training and used the remaining 64 sequences for testing in our experiments. Considering the difficulty of different video sequences, the testing set is divided for validation, from easy to hard, into three categories: well-lit, mild unconstrained and challenging. We utilized 300-VW [34] training set to train our temporal stream network and finetuned the pretrained spatial stream network beforehand.

TF [1]: The Talking Face (TF) video dataset consists of 5,000 frames of a person within a conversation. Due to lack of data variance, we only evaluated our model on its testing set, where our model was trained on the training samples from the 300-VW [34] datasets without using any additional training set.

Evaluation Protocols. In our experiments, we employed the standard normalized root mean squared error (RMSE) and cumulative error distribution (CED) curves for the evaluation protocols. The normalized RMSE [31] was employed for averaged error comparisons. In the experiments, we performed our method on the testing video clips for each identity for evaluation, and then averaged the RMSE errors for final performance. We also leveraged the CED curves [46], [50] of RMSE errors to quantitatively evaluate the performance in comparisons to the state-of-the-arts.

4.1 Evaluation on 300-VW

Comparisons with State-of-the-Arts. We compared our TSTN with both conventional methods which utilize hand-crafted features and the deep learning-based methods which learn features directly from pixels. The methods with hand-crafted features include SDM [46], ESR [7], CFSS [50], PIEFA [26] and iCCR [32]. The deep learning methods include TCDCN [49] and REDN [25]. Here SDM [46], ESR [7], CFSS [50], TSCN [35] and TCDCN [49] were utilized to predict landmarks on still images, while PIEFA [26], REDN [25], TSCN [35] and iCCR [32] were designed for video-based face alignment. For SDM [46], CFSS [50] and TCDCN [49], we conducted experiments with their released codes in a tracking-by-detection protocol, where 68 landmarks were employed for evaluation. Since the source codes are not available for PIEFA [26] and REDN [25], we directly cropped the results reported in original papers. Note that

TABLE 1
Averaged Error Comparisons of Our Proposed TSTN with the State-of-the-Art Face Alignment Approaches Including Both Conventional Hand-Crafted Approaches (SDM [46], CFSS [50], PIEFA [26] and iCCR [32]) and Deep-Learning Based Approaches (TCDCN [49], TSCN [35] and REDN [25]) on 300-VW Dataset [34]

Methods	Model Description	Category 1	Category 2	Category 3	Challset [25]	-pts	Year
SDM [46]	Cascaded Linear Regression	7.41	6.18	13.04	7.44		2013
TSCN [35][1]	Two-Stream Action Network	11.61	11.59	17.67	-		2014
TSCN [35][1],[2]	Two-Stream Action Network	12.54	7.25	13.13	-		2014
CFSS [50]	Coarse-to-Fine Shape Searching	7.68	6.42	13.67	5.92	68	2015
PIEFA [26]	Personalized Ensemble Learning	-	-	-	6.37		2015
REDN [25]	Recurrent Auto-Encoder Net	-	-	-	6.25		2016
TCDCN [49]	Multi-Task Deep CNN	7.66	6.77	14.98	7.27		2016
TSTN	Two-Stream Transformer Net	**5.36**	**4.51**	**12.84**	**5.59**		-
CCR [32]*	Cascaded Continuous Regression	7.26	5.89	15.74	-		2016
iCCR [32]*	Cascaded Continuous Regression	*6.71*	**4.00**	*12.75*	-	66	2016
TSTN	Two-Stream Transformer Net	**5.21**	*4.23*	**10.11**	-		-

For fair comparisons, we used the indices of frames that removed from evaluation according to the 300VW organizers [34].
[1] *To make TSCN [35] adaptive to face alignment, we deployed the mean square loss layers [7], [46] at the top of both steams instead of the softmax [18] loss.*
[2] *We revised architecture of the spatial stream in [35] by plugging the proposed sampler module, which explicitly exploits the global shape-preserving prior [7].*
**iCCR integrates with an ensemble learning updates rules based on the cascaded continuous regression (CCR) formulation. We compared the results of our TSTN with CCR and iCCR by the released codes (http://www.cs.nott.ac.uk/~psxes1/) on 300-VW [34], where 66 landmarks were employed for evaluation.*

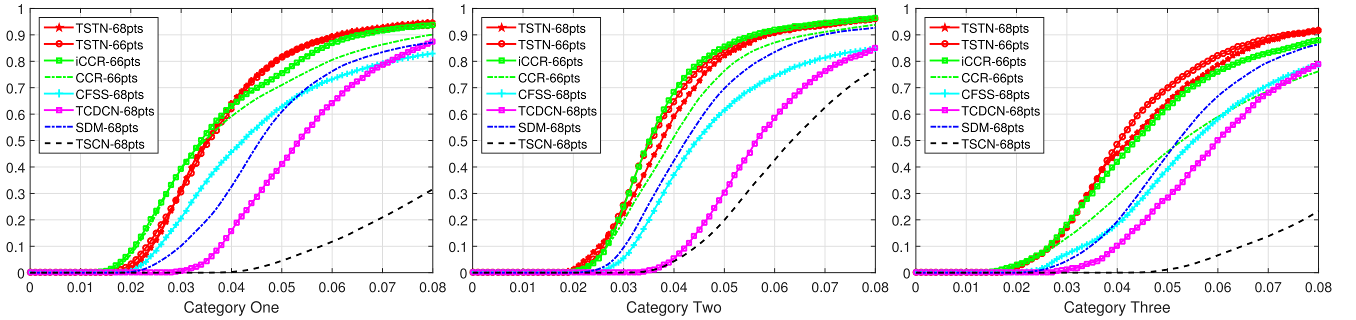


Fig. 3. CED curves of our TSTN compared to the state-of-the-arts on three categories in 300-VW [34] separately. In contrast to the state-of-the-art methods, our TSTN achieves comparable results in category two and superior performance in category one and the most difficult category three.

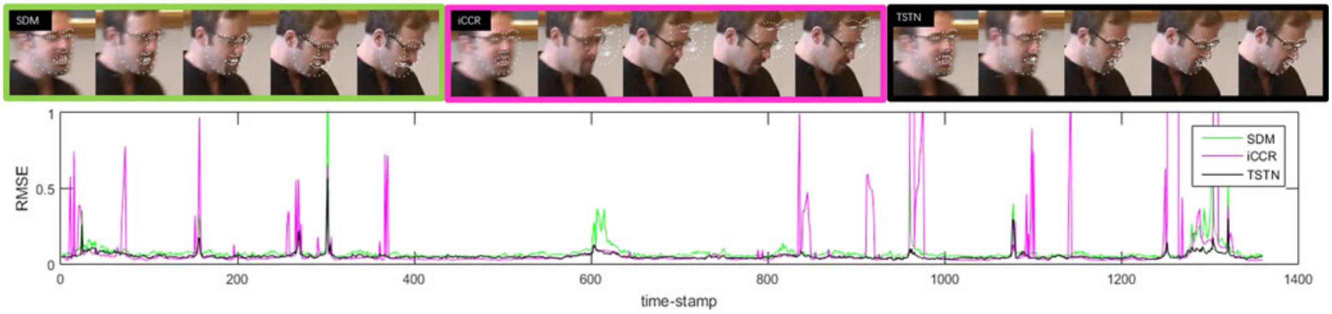


Fig. 4. Resulting examples of our TSTN on the 557th video clip in 300-VW [34] Category Three, where the selected tracked subject undergoes severe poses over time. The bottom subfigure shows that our TSTN exhibits robustness to difficult cases like large variations of facial aspect ratios.

the CED curves of CCR and iCCR of only 66 points were reported in [32], we re-trained our TSTN by the common 66 points for fair comparisons.

Table 1 tabulates the averaged error comparisons and Fig. 3 shows the CED curves of TSTN compared with the state-of-the-art face alignment methods. From these results, we observe that our model exhibits superior performance compared with the state-of-the-arts on Category One and the most challenging Category Three. These achievements benefit from the exploited complementary information that simultaneously encodes both the appearance features on still images and temporal consistency across frames. Compared with CCR and iCCR [32], our model achieves promising results especially in the most challenging Category Three. This is because our proposed recurrent module automatically memorizes the shape-informative details of previous frames to flow across frames in an end-to-end manner. Based on these historical cues, our model accomplishes making inferences for the succeeding frames (e.g., occluded parts) to reinforce the robustness to the variations of the large poses, expressions, occlusions, etc. Besides, our model largely outperforms TSCN [35], because the extracted optical flows in TSCN [35] are hand-crafted which may fall into local optima. Moreover, the spatial stream in TSCN [35] loses shape-informative details without considering the shape-sensitive priors during shape refinement process. Lastly, we

TABLE 2
Comparisons of CED Values Where The RMSEs Are Less Than 0.05 and 0.1 with Different Specifications of $\{\beta_1, \beta_2\}$ on the 300-VW Fullset [34]

Weighted Fusion	$\text{CED}_{e<0.05}$	$\text{CED}_{e<0.1}$
TSTN ($\beta_1 = 0, \beta_2 = 0$)	38.24%	72.91%
TSTN ($\beta_1 = 1, \beta_2 = 0$)	65.93%	82.35%
TSTN ($\beta_1 = 0.8, \beta_2 = 0.2$)	73.89%	87.92%
TSTN ($\beta_1 = 0, \beta_2 = 1$)	70.29%	92.13%
TSTN ($\beta = 0.2, \beta_2 = 0.8$)	78.89%	92.19%
TSTN ($\beta_1 = \beta_2 = 0.5$)	**80.33**%	**95.87**%

We leveraged the fitting results of previous frame as the prediction of current frame, leading to the baseline method without refinements.

illustrated some resulting face examples of 300-VW [34] Category Three in Fig. 4 along with the RMSE errors across time steps. Based on these illustrations, we observe our TSTN demonstrates the robustness to the difficult cases compared with SDM [46] and iCCR [32], even if facial landmarks undergo large variances of severe facial aspect ratios.

Influence of Different Weighted Fusion Strategies. We investigated the contributions of different weights for the fusion strategies. To achieve this, we created several alternative baselines to study the importance of different weights for β_1 and β_2 in our objective function, where the sum of β_1 and β_2 is 1. Specifically, we first specified the parameters $\beta_1 = 1$ for the spatial stream and $\beta_2 = 1$ for the temporal stream of our model. Then we conducted the experimental comparisons by setting $\beta_1 = \{0.8, 0.2, 0.5\}$ and $\beta_2 = \{0.2, 0.8, 0.5\}$. Table 2 tabulates the CED values where the RMSE errors are less than 0.05 and 0.1 of our model and other alternative variations on the 300-VW dataset [34]. According to these results, we see that the equivalent weighted fusion strategy of both the spatial and temporal streams obtains the higher performance than other baseline methods, which shows that the complementary information of both appearance information in the spatial dimension and consistency information in the temporal dimension contribute to the alignment performance.

Analysis of Network Decisions. To justify the decisions of our proposed temporal stream, we created three baselines as follows (with spatial stream fixed): 1) TSTN-1: directly feeding cropping patches to regression net across frames; 2) TSTN-2: discarding $\mathbf{h}_2$, making $\mathbf{h}_1$ recurrent and directly feeding $\mathbf{h}_1$ to decoder; 3) TSTN-3: directly feeding $\mathbf{V}$ to regression net; 4) TSTN-4: our proposed model. We conducted experiments of our TSTN compared with these baselines on the 300-VW fullset and Table 3 tabulates the results. From these results, we see that our temporal stream with the proposed modules including encode-decoder, two-layer RNN and sampler together boost the performance.

Computational Time. Our model was built based on the accelerated deep learning toolbox TensorFlow [2]. In terms of the training procedure, we introduce an efficient strategy for fast convergence. Specifically, we first learned the network parameters of the spatial stream by using all images in 300-W [30]. Then we trained the temporal stream and simultaneously fine-tuned the pre-trained spatial stream. The proposed training scheme was roughly 10× faster than training both streams from scratch. The whole training procedure requires 15 hours with a GPU of single NVIDIA GTX 1080 Ti graphic computation card. We also tested our method on the core-i7 CPU@3.6 GHZ platform. Our model runs nearly at 30 frames per second on CPU (without the face detection part), which satisfies the real-time requirements in practice.

TABLE 3
Comparisons of CED Values Where The RMSEs Are Less Than 0.05 and 0.1 with Different Network Decisions on the 300-VW Fullset [34]

Method	TSTN-1	TSTN-2	TSTN-3	TSTN-4
$\text{CED}_{e\leq 0.05}$	60.15%	67.83%	66.26%	$\mathbf{70.29\%}$
$\text{CED}_{e\leq 0.10}$	79.33%	89.76%	88.91%	$\mathbf{92.13\%}$

4.2 Evaluation on TF

We evaluated our method on the TF dataset [1] compared with the state-of-the-arts such as ESR [7], SDM [46], CFAN [48], DCNC [37], CFSS [50], IFA [3] and REDN [25]. Since the mark-up annotations of the TF dataset is partially different with those employed in the 300-VW dataset [34], we generated 7-landmark annotations including the eye corners, nose tip and mouth corners to localize the facial landmarks in our evaluation for fair comparisons. It is noticed that we only utilized 7 landmarks predicted by the pre-trained models including CFSS [50] and CFAN [48]. For other methods, the results were cropped from the original paper [25]. Table 4 shows the averaged errors of our method with the state-of-the-art methods, where 7 landmarks were employed for evaluation. According to these results, we see that our method outperforms the existing video-based face alignment methods including PIEFA [26] and REDN [25], because our TSTN is helpful to video-based face alignment by exploiting the complementary information of the spatial appearance and the temporal consistency information.

5 Conclusions and Future Work

In this paper, we have developed a two-stream transformer networks architecture for video-based face alignment. Specifically, our model learns to exploit the complementary information of appearance features on still frames in the spatial stream and consistency information across frames in the temporal stream accordingly. The performance of our method on the benchmarking datasets verifies the effectiveness of our approach. How to apply learnable weights for

TABLE 4
Averaged Error Comparisons of Our Model with the State-of-the-Arts on the TF Dataset [1], Where 7 Landmarks Were Employed for Evaluation

Methods	ESR [7]	SDM [46]	CFAN [48]	DCNC [37]	CFSS [50]	IFA [3]	REDN [25]	TSTN
RMSE	3.81	4.01	3.52	3.67	2.36	3.45	3.32	**2.13**

The results were directly cropped from the recent work [25]. Our approach achieves very competitive performance compared with the state-of-the-arts.

the fusion of both streams and it is desirable to squeeze our networks to boost the efficiency performance, which are interesting future directions of this work.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB1001004, in part by the National Natural Science Foundation of China under Grant 61672306, Grant 61572271, Grant 61527808, Grant 61373074, and Grant 61373090, in part by the National 1000 Young Talents Plan Program, in part by the National Basic Research Program of China under Grant 2014CB349304, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, in part by the Ministry of Education of China under Grant 20120002110033, and in part by the Tsinghua University Initiative Scientific Research Program.

REFERENCES

[1] FGNET: Talking Face Video. 2014, [Online]. Avilable: http://www-prima.inrialpes.fr/FGnet/data/01-TalkingFace/talking_face.html

[2] M. Abadi, et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software Available: tensorflow.org.

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1859–1866.

[4] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proc. IEEE Conf. Comput. Vis.*, 2015, pp. 2830–2838.

[5] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion," in *Proc. IEEE Conf. Comput. Vis.*, 1995, pp. 374–381.

[6] A. Bulat and G. Tzimiropoulos, "Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAW) challenge," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 616–624.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2887–2894.

[8] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2015, pp. 954–962.

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[10] D. DeCarlo and D. N. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *Int. J. Comput. Vis.*, vol. 38, no. 2, pp. 99–127, 2000.

[11] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1933–1941.

[12] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[13] C. Gou, Y. Wu, F. Wang, and Q. Ji, "Shape augmented regression for 3D face alignment," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 604–615.

[14] C. M. Grewe and S. Zachow, "Fully automated and highly accurate dense correspondence for facial surfaces," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 552–568.

[15] J. Hu, J. Lu, and Y. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1875–1882.

[16] Z. Huang, X. Zhao, S. Shan, R. Wang, and X. Chen, "Coupling alignments with recognition for still-to-video face recognition," in *Proc. IEEE Conf. Comput. Vis.*, 2013, pp. 3296–3303.

[17] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1867–1874.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[19] N. Kumar, P. N. Belhumeur, and S. K. Nayar, "FaceTracer: A search engine for large collections of images with faces," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 340–353.

[20] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," pp. 1894–1903, 2016.

[22] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1325–1334.

[23] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. British Mach. Vis. Conf.*, 2015, pp. 41.1–41.12.

[25] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 38–56.

[26] X. Peng, S. Zhang, Y. Yang, and D. N. Metaxas, "PIEFA: Personalized incremental and ensemble face alignment," in *Proc. IEEE Conf. Comput. Vis.*, 2015, pp. 3880–3888.

[27] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris, "3D facial landmark detection under large yaw and expression variations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1552–1564, Jul. 2013.

[28] L. Pigou, A. van den Oord, S. Dieleman, M. V. Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *arxiv: 1506.01911*, pp. 1-10, 2015.

[29] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1685–1692.

[30] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, 2016.

[31] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2013, pp. 397–403.

[32] E. Sánchez-Lozano, B. Martínez, G. Tzimiropoulos, and M. F. Valstar, "Cascaded continuous regression for real-time incremental face tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 645–661.

[33] Z. Sánta and Z. Kato, "3D face alignment without correspondences," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 521–535.

[34] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2015, pp. 1003–1011.

[35] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[36] P. Sun, J. K. Min, and G. Xiong, "Globally tuned cascade pose regression via back propagation with application in 2D face pose estimation and heart segmentation in 3D CT images," *arXiv:1507.07508*, vol. abs/1503.08843, 2015.

[37] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3476–3483.

[38] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1701–1708.

[39] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[40] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4177–4187.

[41] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3659–3667.

[42] X. Wang, M. Yang, S. Zhu, and Y. Lin, "Regionlets for generic object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2071–2084, Oct. 2015.

[43] Y. Wu and Q. Ji, "Discriminative deep face shape model for facial point detection," *Int. J. Comput. Vis.*, vol. 113, no. 1, pp. 37–53, 2015.

[44] Y. Wu, Z. Wang, and Q. Ji, "A hierarchical probabilistic model for facial feature detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1781–1788.

[45] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. A. Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 57–72.

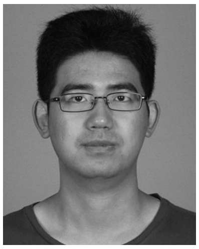
[46] X. Xiong and F. D. la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[47] X. Yu, J. Huang, S. Zhang, and D. N. Metaxas, "Face landmark fitting via optimized part mixtures and cascaded deformable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 11, pp. 2212–2226, Nov. 2016.

[48] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.

[49] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.

[50] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4998–5006.



Hao Liu received the BS degree in software engineering from Sichuan University, China, in 2011 and the Engineering Master degree in computer technology from University of Chinese Academy of Sciences, China, in 2014. He is currently pursuing the PhD degree at the department of automation, Tsinghua University. His research interests include face alignment, facial age estimation and deep learning.

Jiwen Lu received the BEng degree in mechanical engineering and the MEng degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the PhD degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. From 2011 to 2015, he was a Research Scientist with the Advanced Digital Sciences Center, Singapore. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and machine learning. He has authored or co-authored over 130 scientific papers in these areas, where 33 were the IEEE Transactions papers. He is the Workshop Chair/Special Session Chair/Area Chair for over ten international conferences. He was a recipient of the National 1000 Young Talents Plan Program in 2015. He serves as an Associate Editor of the Pattern Recognition Letters, the Neurocomputing, and the IEEE ACCESS, a Managing Guest Editor of Pattern Recognition and Image and Vision Computing, a Guest Editor of Computer Vision and Image Understanding and Neurocomputing, and an Elected Member of the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society.

Jianjiang Feng is an associate professor in the Department of Automation at Tsinghua University, Beijing. He received the BS and PhD degrees from the School of Telecommunication Engineering, Beijing University of Posts and Telecommunications, China, in 2000 and 2007. From 2008 to 2009, he was a Post Doctoral researcher in the PRIP lab at Michigan State University. He is an Associate Editor of Image and Vision Computing. His research interests include fingerprint recognition and computer vision.

Jie Zhou received the BS and MS degrees from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 1995. From 1995 to 1997, he served as a Post-Doctoral Fellow with the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a Full Professor with the Department of Automation, Tsinghua University. In recent years, he has authored over 100 papers in peer-reviewed journals and conferences. Among them, over 40 papers have been published in top journals and conferences, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and CVPR. His current research interests include computer vision, pattern recognition, and image processing. He received the National Outstanding Youth Foundation of China Award. He is an Associate Editor of the International Journal of Robotics and Automation and two other journals.