Attentive One-Dimensional Heatmap Regression for Facial Landmark Detection and Tracking

Shi Yin[1], Shangfei Wang[*,1,2], Xiaoping Chen[2], Enhong Chen[3] and Cong Liang[1]

davidyin@mail.ustc.edu.cn,{sfwang,xpchen,cheneh}@ustc.edu.cn,lc150303@mail.ustc.edu.cn

[1]Key Lab of Computing and Communication Software of Anhui Province,
School of Computer Science and Technology, University of Science and Technology of China
[2]Anhui Robot Technology Standard Innovation Base, University of Science and Technology of China
[3]Anhui Province Key Lab of Big Data Analysis and Application,
School of Computer Science and Technology, University of Science and Technology of China

ABSTRACT

Although heatmap regression is considered a state-of-the-art method to locate facial landmarks, it suffers from huge spatial complexity and is prone to quantization error. To address this, we propose a novel attentive one-dimensional heatmap regression method for facial landmark localization. First, we predict two groups of 1D heatmaps to represent the marginal distributions of the x and y coordinates. These 1D heatmaps reduce spatial complexity significantly compared to current heatmap regression methods, which use 2D heatmaps to represent the joint distributions of x and y coordinates. With much lower spatial complexity, the proposed method can output high-resolution 1D heatmaps despite limited GPU memory, significantly alleviating the quantization error. Second, a co-attention mechanism is adopted to model the inherent spatial patterns existing in x and y coordinates, and therefore the joint distributions on the x and y axes are also captured. Third, based on the 1D heatmap structures, we propose a facial landmark detector capturing spatial patterns for landmark detection on an image; and a tracker further capturing temporal patterns with a temporal refinement mechanism for landmark tracking. Experimental results on four benchmark databases demonstrate the superiority of our method.

CCS CONCEPTS

• **Computing methodologies → Biometrics**.

KEYWORDS

Facial landmark detection, facial landmark tracking, heatmap regression

*Corresponding author.

1 INTRODUCTION

As a fundamental computer vision task, face alignment [7, 33] consists of two sub-tasks, i.e., facial landmark detection on a static image, and facial landmark tracking in a video with continuous frames. Regression-based methods for face alignment can be divided into two categories, i.e., a coordinate regression approach, or a heatmap regression approach. Coordinate regression approaches directly map facial appearances to the continuous values of landmark coordinates or their displacements. This technique has difficulty handling spatial distributions around ground truth coordinates, especially for the coordinates with large variances. Therefore, as reported by Sun *et al.* [28], it is sub-optimal on spatial generalization performance.

Recently, heatmap regression approaches have been proposed. These approaches assume that a landmark coordinate obeys Gaussian distribution around its ground truth label. These approaches predict the discrete probability distribution that a landmark occurs in each point of the heatmap, and transform points on the heatmaps to the predicted landmark coordinates. Heatmap regression approaches successfully capture the spatial distributions around ground truths. They are thus theoretically capable of better spatial modeling than coordinate regression approaches. However, the quantization process, i.e., using discrete heatmaps to represent continuous landmark coordinates with both integer and fractional parts, may cause quantization error. One way to alleviate such a problem is to improve the heatmap resolution. Unfortunately, current heatmap regression methods suffer from high spatial complexity, as they use 2D heatmaps to represent joint distributions on the x and y axes. Specifically, a landmark produces a heatmap with $L{\times}L$ points, where L is the output resolution on the x and y axes. The total output size of N landmarks is NL^2. The space occupation of 2D heatmaps increases dramatically as L increases. Due to the limited machine memory, L is typically restricted to a value lower than the input face [2, 18]. This compromise may cause severe quantization errors as shown in Fig. 1, limiting the practical performance of the heatmap regression method. Although some methods [19, 28, 30, 37] are proposed to estimate landmark coordinates with fractional parts from a 2D heatmap, they still suffer from the information loss caused by low resolution.
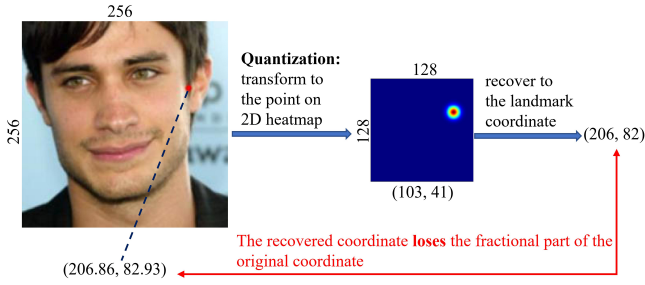
Figure 1: Illustration of the quantization error caused by low-resolution 2D heatmap. The facial image is sampled from the 300W [24] dataset.

To solve the disadvantage of 2D heatmaps, we propose a novel attentive one-dimensional heatmap regression approach that achieves prominent spatial and temporal modeling capability while significantly decreasing the output complexity. The basic idea is to replace the 2D heatmaps by 1D heatmaps that represent marginal distributions on the x and y axes as the output structure. Considering the correlation between x and y coordinates, we capture the x-y joint distribution implicitly by applying a co-attention mechanism on the distributional features of the two axes. Our method significantly decreases the spatial complexity of the output. The total output size of 1D heatmaps on two axes is $2NL$ ($N \times 2L$), much smaller than the 2D heatmaps (NL^2). The small output size allows us to fully boost the resolution despite limited GPU memory, significantly alleviating quantization errors. Based on the proposed heatmap structure, we design a facial landmark detector and a tracker, as shown in Fig. 2a and 2b, respectively. The detector captures spatial patterns in a static image, while the tracker integrates both spatial and temporal patterns from the facial sequence in a video to locate landmarks. For the detector, we design two groups of convolutional neural networks (CNNs) to compress the facial representation tensor to 1D heatmaps on the x and y axes. The co-attention mechanism is adopted between the features extracted by the CNNs. In the tracker, spatial patterns from the current frame are extracted by the proposed detector, temporal patterns among multiple frames are embedded by a novel temporal refinement mechanism integrating features on the current frame and features from past frames.

We conduct facial landmark detection experiments on the 300W dataset and the AFLW dataset, and conduct facial landmark tracking experiments on the 300VW dataset and the TF dataset. Experimental results on these datasets show that the proposed method outperforms state-of-the-art coordinate regression and heatmap regression methods.

The main contributions of our method are three folds. First, we are the first that propose to predict 1D heatmaps on the x and y axes instead of using 2D heatmaps to locate landmarks and successfully alleviate the quantization error with a fully boosted output resolution. Second, we propose a co-attention module to capture the joint coordinate distribution on the two axes. Third, based on the proposed heatmap regression method, we design a facial landmark detector and tracker which achieve state-of-the-art performance.

2 RELATED WORK

2.1 Coordinate Regression Methods

Coordinate regression approaches [3, 4, 14, 22, 29, 34, 35, 38] directly predict coordinate values or their increments by a mapping function. As early works, Xiong *et al.* [35] proposed a Supervised Descent Method (SDM) method, which maps Scale-Invariant Feature Transform (SIFT) features to landmark displacements between the current output and the ground truth by minimizing a nonlinear least square objective function. Cao *et al.* [4] and Burgos-Artizzu *et al.* [3] learned fern regressors to predict landmark increments. Ren *et al.* [22] proposed to learn local binary features for every facial landmark by random forests.

Recently, deep learning based methods are proposed for face alignment. Sun *et al.* [29] proposed a CNN based method to predict facial landmarks in a cascaded way. Zhang *et al.* [38] proposed Tasks-Constrained Deep Convolutional Network (TCDCN), a multi-task learning method that learns to predict landmark coordinates as well as other facial attributes, including expression, gender, etc. Liu *et al.* [14] combined a CNN network with a RNN based encoder-decoder network to learn spatial and temporal patterns of landmarks in adjacent frames.

To capture dependencies among landmark labels, some works adopted probabilistic graphical models, such as a Dynamic Bayesian Network (DBN) [13], or a Restricted Boltzmann Machine (RBM) [34], as shape constraints. Yin *et al.* [36] utilized adversarial learning to explore the inherent dependencies among the movement of facial landmarks.

Despite these progresses, it is still challenging for a coordinate regression method to capture spatial distributions around ground truth coordinates, especially when the coordinate values vary in a wide range. This weakness leads to a sub-optimal spatial generalization performance of these methods.

2.2 Heatmap Regression Methods

Heatmap regression methods [2, 5, 8, 18, 20, 28, 30, 32] capture spatial distributions around ground truths by likelihood heatmaps of landmarks. Newell *et al.* [18] proposed a stacked hourglass network to generate heatmaps for 2D human pose estimation. Bulat and Tzimiropoulos [2] enhanced the stacked hourglass network with hierarchical, parallel and multi-scale residual blocks. Xi *et al.* [20] proposed a spatial and temporal recurrent learning method for landmark detection and tracking. Based on the heatmap technique, Chu *et al.* [8] proposed a multi-context attention mechanism to focus on informative feature regions. Wu *et al.* [32] proposed to estimate the heatmap of facial boundary as auxiliary features to locate landmarks. Chen *et al.* [6] designed an adversarial learning method to learn structural patterns among landmarks. Chen *et al.* [5] proposed a Conditional Random Field (CRF) method to embed geometric relationships among landmarks based on their heatmaps. Liu *et al.* [15] proposed a heatmap correction unit which uses global shape constraints to refine heatmaps.

The heatmap structure achieves good theoretical performance for spatial generalization. However, it suffers from huge spatial complexity. Under limited space, the heatmap resolution is typically compressed to a value smaller than the input face. Consequently, fractional parts of landmark coordinates are totally dropped, leading
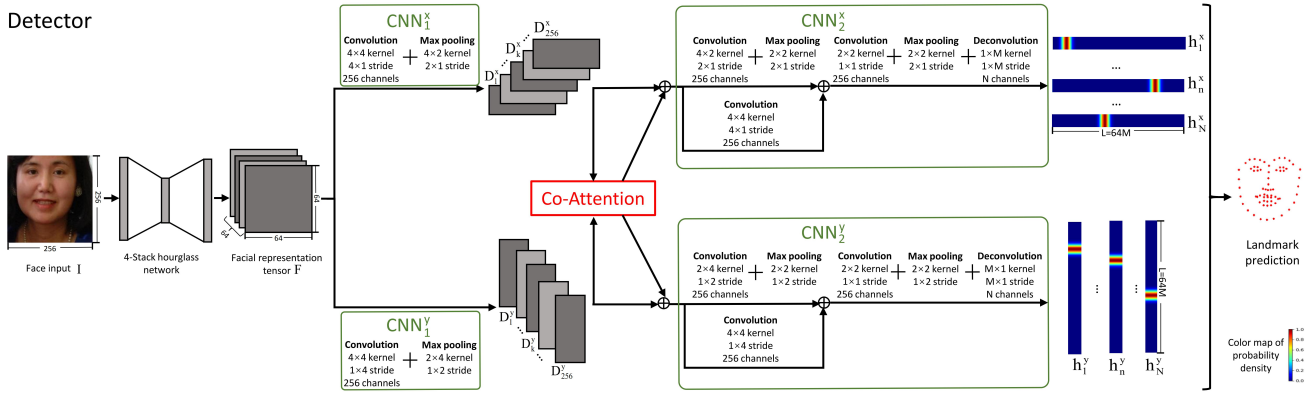
(a) 1D heatmaps on the x and y axes are generated by CNN networks. The co-attention mechanism is applied between the output features of CNN_1^x and CNN_1^y to capture joint distribution on the two axes.
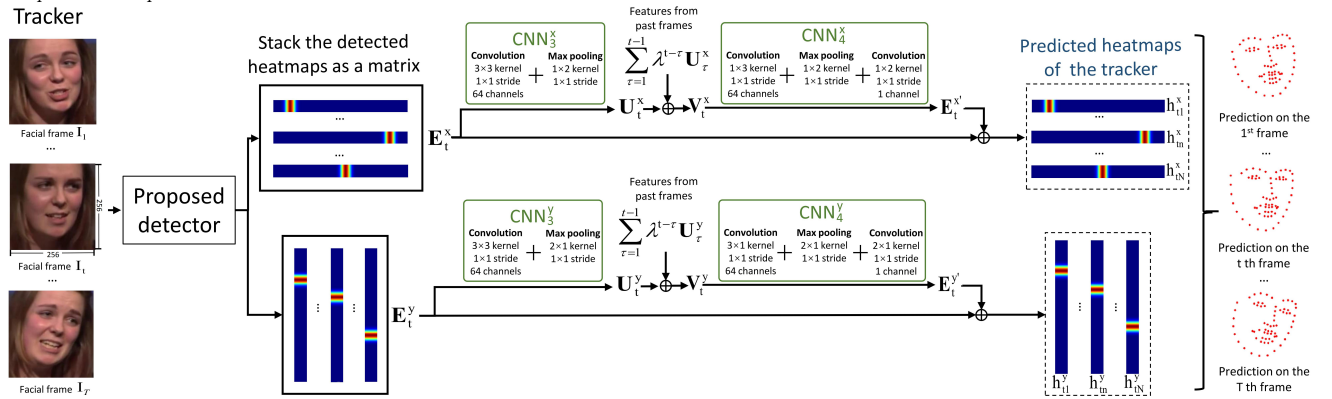
(b) Based on the spatial patterns captured by the detector, the tracker further captures temporal patterns by fusing features on the current frame with features from past frames. Temporal patterns are used to refine the detected heatmaps.

Figure 2: The proposed detector (a) and tracker (b).

to serious quantization errors. To estimate the landmark coordinates with fractional parts based on 2D discrete heatmaps, Sun *et al.* [28] proposed to integrate all point locations weighted by their probabilities in the 2D heatmap as the predicted coordinate. Nibali *et al.* [19] proposed a differentiable layer at the end of CNN networks to convert heatmaps as continuous coordinates. Tai *et al.* [30] proposed Fractional Heatmap Regression (FHR), which uses three heatmap points to predict the fractional parts of landmark coordinates according to a Gaussian function. Zhang *et al.* [37] proposed to estimate the landmark coordinates by a Taylor-expansion based distribution approximation. However, for these 2D heatmap-based methods, when the heatmap resolution is low, the heatmap points are sparse and some detailed information on the spatial distribution is lost, making it very hard for coordinate estimation.

To address the disadvantages of 2D heatmap regression, we propose a new regression method based on 1D heatmaps which represent the marginal distribution on each axis. We capture joint distributions between the x and y axes by a co-attention mechanism, instead of using 2D heatmaps. The proposed regression method is much more space-efficient and the output resolution can be fully boosted despite limited space. Therefore, the quantization error is significantly alleviated.

3 ANALYSIS ON THE QUANTIZATION ERROR OF 2D HEATMAP

Conventional heatmap regression methods, denoted as $f(\cdot)$, predict discrete 2D joint distributions (heatmaps) for N predefined landmarks from a facial image $\mathbf{I} \in \mathbb{R}^{F \times F \times 3}$, as shown in Equation (1):

$$\mathbf{H}_n = f(\mathbf{I}, n; \theta_f), 1 \le n \le N \quad (1)$$

where θ_f is the parameters of $f(\cdot)$ and $\mathbf{H}_n \in \mathbb{R}^{L \times L}$ denotes the heatmap of the n ($1 \le n \le N$) th landmark with resolution L. The ground truth heatmap, represented as $\mathbf{H}_n^*$, is considered a discrete Gaussian distribution centered on the ground truth position, i.e., (x_n^*, y_n^*). The probability density on an arbitrary heatmap point (x, y) follows Equation (2):

$$\mathbf{H}_n^*(x, y) \propto exp(-\frac{1}{2\sigma^2}((x - x_n^*)^2 + (y - y_n^*)^2)) \quad (2)$$

where σ is the variance of the Gaussian distribution.

The heatmap structure models the spatial distributions of landmarks to obtain spatial generalization. However, using a discrete heatmap to represent continuous landmark coordinates with both integer and fractional parts may cause quantization error. This is

because the rounding-down operation $\lfloor \ \rfloor$ is applied to convert the continuous coordinate (p, q) to the discrete heatmap point (x, y), as shown in Equation (3):

$$x = \lfloor \frac{p \cdot L}{F} \rfloor, \quad y = \lfloor \frac{q \cdot L}{F} \rfloor \tag{3}$$

The rounding-down operation drops the fractional part of its input. Therefore, we could only recover an approximate value of (p, q) from (x, y), as shown in Equation (4):

$$p' = x \cdot \frac{F}{L} = \frac{\lfloor p \cdot (\frac{L}{F}) \rfloor}{(\frac{L}{F})}, \quad q' = y \cdot \frac{F}{L} = \frac{\lfloor q \cdot (\frac{L}{F}) \rfloor}{(\frac{L}{F})} \tag{4}$$

where (p', q') is the coordinate recovered from the heatmap. Quantization error E is defined as the Euclidean distance between (p, q) and (p', q'), as shown in Equation (5):

$$E = \sqrt{(p - p')^2 + (q - q')^2} \tag{5}$$

From Equation (3) and Equation (4), we find that the higher the value of $\frac{L}{F}$, the closer p comes to p' and q comes to q'. In other words, the quantization error E is decreasing. For example, suppose (p, q) is $(142.84, 188.72)$. When $\frac{L}{F} = 0.5, E = 1.11$. When $\frac{L}{F}$ increases to 3.0, E decreases to 0.18. Therefore, with a given F, one way to reduce E is to improve L.

Unfortunately, the 2D heatmap structure is very spatially complex. For N landmarks, a total of NL^2 heatmap points are generated. Due to the limited machine memory, current heatmap regression methods usually set L as a value smaller than F, resulting in severe quantization errors.

4 METHODOLOGY

The key to reducing quantization errors is to boost the output resolution L despite limited machine memory. For that purpose, we propose a new method with good capability of spatial and temporal modeling while significantly decreasing output complexity compared to 2D heatmap regression methods. Instead of predicting joint distributions on the x and y axes explicitly by 2D heatmaps that occupy huge space, we propose to model joint distributions implicitly and just predict 1D heatmaps that represent marginal distributions.

Based on such an idea, we propose a new landmark detector, as depicted in Fig. 2a. The detector $g_{de}(\cdot)$ with parameters θ_{de} is formalized as Equation (6), where $\mathbf{h}_n^x$ and $\mathbf{h}_n^y$ are the predicted 1D heatmaps on the x and y axes, respectively, for the n th landmark.

$$\mathbf{h}_n^x = g_{de}(\mathbf{I}, n, x; \theta_{de}), \quad \mathbf{h}_n^y = g_{de}(\mathbf{I}, n, y; \theta_{de}), \quad 1 \le n \le N \tag{6}$$

Coordinate prediction for the n th landmark $(\hat{p}_n, \hat{q}_n)$ on a facial image is obtained from the maximum points of $\mathbf{h}_n^x$ and $\mathbf{h}_n^y$, as shown in Equation (7):

$$\hat{p}_n = \arg\max(\mathbf{h}_n^x) \cdot \frac{F}{L}, \quad \hat{q}_n = \arg\max(\mathbf{h}_n^y) \cdot \frac{F}{L} \tag{7}$$

We also extend the detector as a tracker, as depicted in Fig. 2b. The tracker $g_{tr}(\cdot)$ with parameters θ_{tr} predicts landmark positions in a facial video, denoted as $\mathbf{I}_{1:T} = (\mathbf{I}_1, ..., \mathbf{I}_t, ..., \mathbf{I}_T)$, where $\mathbf{I}_t$ is the t $(1 \le t \le T)$ th frame of the video. For the t th frame, the tracker captures not only spatial patterns on $\mathbf{I}_t$, but also temporal patterns

inherent in the sequence from $\mathbf{I}_1$ to $\mathbf{I}_t$. The tracker is formalized as Equation (8).

$$\mathbf{h}_{tn}^x = g_{tr}(\mathbf{I}_{1:t}, n, x; \theta_{tr}),$$
$$\mathbf{h}_{tn}^y = g_{tr}(\mathbf{I}_{1:t}, n, y; \theta_{tr}), 1 \le n \le N, 1 \le t \le T \tag{8}$$

where $\mathbf{h}_{tn}^x$ and $\mathbf{h}_{tn}^y$ are the output heatmaps on the t th frame.

The detector and the tracker output $\mathbf{h}_n^x$ and $\mathbf{h}_n^y$ (or $\mathbf{h}_{tn}^x$ and $\mathbf{h}_{tn}^y$) $\in \mathbb{R}^L$ for N landmarks, and the total output size for a facial image or a facial frame is $N \times 2L = 2NL$, much smaller than that of the 2D heatmaps (NL^2). In other words, the light-weight structure of 1D heatmap allows us to boost its resolution L to a large value without heavy space occupation, and therefore the quantization error is significantly alleviated.

4.1 Detector

4.1.1 Generating 1D Heatmaps Using CNNs. First, a facial representation $\mathbf{F}$ is learned from the input face through a stacked hourglass network [2], as shown in the left part of Fig. 2a. Then, 1D heatmaps on two axes are generated by two groups of CNNs, respectively, as depicted in the green border boxes of Fig. 2a. The first group of CNNs, composed of CNN_1^x and CNN_2^x, converts $\mathbf{F}$ to 1D heatmaps on the x axis, i.e., $\mathbf{h}_1^x, ..., \mathbf{h}_n^x, ..., \mathbf{h}_N^x$, by compressing features along the y axis with a striding operation. At the end of CNN_2^x, a deconvolution module is adopted to generate heatmaps and the heatmap resolution is proportional to the kernel and stride size (M) of deconvolution. The second group of CNNs, composed of CNN_1^y and CNN_2^y, generates heatmaps on the y axis by compressing features along the x axis.

4.1.2 Capturing Joint Distribution on the x and y Axes by Co-Attention. Co-attention [16] is a category of attention methods which capture the correlation between pairwise features. As shown in the red border box of Fig. 2a, we design a co-attention module to capture the x-y joint distributions of landmark coordinates, . First, we encode correlations between features representing distributions on the x and y axes as affinity matrices. Second, features are converted by the affinity matrices and then fused together to embed the joint distribution into their representations.

The co-attention mechanism is adopted between the output features of CNN_1^x and CNN_1^y, denoted as $\mathbf{D}^x$ and $\mathbf{D}^y$, respectively. Both $\mathbf{D}^x$ and $\mathbf{D}^y$ have multiple channels, $\mathbf{D}_k^x$ and $\mathbf{D}_k^y$ represent their k th channel. $\mathbf{D}_k^x$ is a feature representing the distribution on the x axis, while $\mathbf{D}_k^y$ represents the distribution on the y axis. The shape of $\mathbf{D}_k^x$ is the same as $(\mathbf{D}_k^y)^\mathrm{T}$. Two affinity matrices, i.e., $\mathbf{W}_k^{xy}$ and $\mathbf{W}_k^{yx}$, are adopted to encode the correlation between $\mathbf{D}_k^x$ and $\mathbf{D}_k^y$, as shown in Equation (9):

$$\mathbf{W}_k^{xy} = softmax(\frac{(\mathbf{D}_k^y)^\mathrm{T} \cdot \mathbf{P} \cdot (\mathbf{D}_k^x)^\mathrm{T}}{\sqrt{d}}),$$
$$\mathbf{W}_k^{yx} = softmax(\frac{\mathbf{D}_k^x \cdot \mathbf{Q} \cdot \mathbf{D}_k^y}{\sqrt{d}}) \tag{9}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are parameter matrices trained with the whole network. The $softmax(\cdot)$ operation is applied to each row vector of its input matrix, and d is the column number of $\mathbf{P}$ and $\mathbf{Q}$. Following Vaswani *et al.* [31], $\frac{1}{\sqrt{d}}$ is used as a normalization factor to keep

$softmax(\cdot)$ from the region with an extremely small gradient. Based on the affinity matrices, $\mathbf{D}_k^x$ and $\mathbf{D}_k^y$ are fused to capture the joint distribution on the x and y axes, as shown in Equation (10):

$$\mathbf{D}_k^{x'} = \mathbf{D}_k^x + \gamma \mathbf{W}_k^{yx} \cdot (\mathbf{D}_k^y)^{\mathrm{T}}, \quad \mathbf{D}_k^{y'} = \mathbf{D}_k^y + \gamma (\mathbf{W}_k^{xy} \cdot \mathbf{D}_k^x)^{\mathrm{T}} \quad (10)$$

where γ is the weight of the attentive feature. Next, $\mathbf{D}_k^{x'}$ is fed into CNN_2^x as an input channel, and $\mathbf{D}_k^{y'}$ is fed into CNN_2^y.

4.1.3 Loss Function. $g_{de}(\cdot)$ is trained by supervised regression. The error between the prediction and ground truth is minimized as shown in Equation (11). L_{de} is optimized by an Adam optimizer with a learning rate of 1e-4.

$$
\begin{aligned}
\min_{\theta_{de}} L_{de} &= \sum_{n=1}^{N} (||\mathbf{h}_n^x - \mathbf{h}_n^{x*}||_2^2 + ||\mathbf{h}_n^y - \mathbf{h}_n^{y*}||_2^2) \\
&= \sum_{n=1}^{N} (||g_{de}(\mathbf{I}, n, x; \theta_{de}) - \mathbf{h}_n^{x*}||_2^2 + ||g_{de}(\mathbf{I}, n, y; \theta_{de}) - \mathbf{h}_n^{y*}||_2^2)
\end{aligned}
$$
(11)

where $\mathbf{h}_n^{x*}$ and $\mathbf{h}_n^{y*}$ are the ground truths. They are the marginal distributions of $\mathbf{H}_n^*(x, y)$, as shown in Equation (12):

$$\mathbf{h}_n^{x*} = \sum_y \mathbf{H}_n^*(x, y), \quad \mathbf{h}_n^{y*} = \sum_x \mathbf{H}_n^*(x, y) \quad (12)$$

4.2 Tracker

4.2.1 Integrating Spatial and Temporal Patterns. First, spatial patterns on the current frame are encoded by the proposed detector, as shown in the left part of Fig. 2b. The detected heatmaps on the x and y axes are stacked as matrices, denoted as $\mathbf{E}_t^x$ and $\mathbf{E}_t^y$, respectively, for the t th frame.

Second, the detected heatmaps are refined by temporal patterns. This is beneficial because the facial appearance on the current frame may not be reliable due to some "in the wild" disturbances, such as occlusions or uneven illuminations. Integrating temporal patterns from previous frames may help locate landmarks when spatial features on the current frame are unreliable. As depicted in the green border boxes of Fig. 2b, CNN_3^x and CNN_3^y encode $\mathbf{E}_t^x$ and $\mathbf{E}_t^y$ as features, denoted as $\mathbf{U}_t^x$ and $\mathbf{U}_t^y$, respectively. In the feature space, $\mathbf{U}_t^x$ and $\mathbf{U}_t^y$ are fused with features from the past frames, as shown in Equation (13):

$$\mathbf{V}_t^x = \mathbf{U}_t^x + \sum_{\tau=1}^{t-1} \lambda^{t-\tau} \mathbf{U}_\tau^x, \quad \mathbf{V}_t^y = \mathbf{U}_t^y + \sum_{\tau=1}^{t-1} \lambda^{t-\tau} \mathbf{U}_\tau^y \quad (13)$$

where $\mathbf{V}_t^x$ and $\mathbf{V}_t^y$ are features embedded with temporal patterns, $\lambda \in [0, 1]$ is a hyper-parameter to attenuate the weights of frames far from the current. To generate heatmap refinements, CNN_4^x decodes $\mathbf{V}_t^x$ to $\mathbf{E}_t^{x'}$ and CNN_4^y decodes $\mathbf{V}_t^y$ to $\mathbf{E}_t^{y'}$. $\mathbf{E}_t^{x'}$ and $\mathbf{E}_t^{y'}$ are used to refine $\mathbf{E}_t^x$ and $\mathbf{E}_t^y$ respectively by adding together with them as the tracking results.

4.2.2 Loss Function. Similar to the detector, the tracker g_{tr} is also trained by supervised regression. The training loss is shown in

Equation (14):

$$
\begin{aligned}
\min_{\theta_{tr}} L_{tr} &= \sum_{t=1}^{T} \sum_{n=1}^{N} (||\mathbf{h}_{tn}^x - \mathbf{h}_{tn}^{x*}||_2^2 + ||\mathbf{h}_{tn}^y - \mathbf{h}_{tn}^{y*}||_2^2) \\
&= \sum_{t=1}^{T} \sum_{n=1}^{N} (||g_{tr}(\mathbf{I}_{1:t}, n, x; \theta_{tr}) - \mathbf{h}_{tn}^{x*}||_2^2 \\
&\quad + ||g_{tr}(\mathbf{I}_{1:t}, n, y; \theta_{tr}) - \mathbf{h}_{tn}^{y*}||_2^2)
\end{aligned}
$$
(14)

5 EXPERIMENTS

5.1 Experimental Conditions

Facial landmark detection experiments are conducted on the 300W [24] and the AFLW dataset, both of them are image datasets. Facial landmark tracking experiments are conducted on the 300VW [25] and the Talking Face (TF) [12] dataset, they are video datasets.

The 300W dataset contains 68 pre-defined landmarks. For experiments on the 300W dataset, the proposed method is trained on its training set with 3, 148 images, and evaluated on its public testing set, composed of a common subset with 554 images and a challenging subset with 135 images. For saving space, in the following part of the paper, their names are simplified as 300W *com* and *cha*, respectively. The full public testing set is simplified as 300W *full*.

The AFLW dataset has 21 pre-defined landmarks. For experiments on the AFLW dataset, we just use 19 landmarks as previous work did [10]. The proposed method is trained on the training set with 20000 images, and evaluated on the testing set with 4386 images.

The 300VW dataset contains 68 pre-defined landmarks. It has a training set with 50 videos, and a testing set consisting of 60 videos from three difficulty levels, i.e., well-lit (scenario 1), mild unconstrained (scenario 2) and challenging (scenario 3). Their names are simplified as 300VW S1, S2, and S3, respectively. For experiments on the 300VW dataset, following Yin *et al.* [36], our method is trained on the union of training sets from the 300VW and 300W dataset, and evaluated on the 300VW testing set. Since the 300W dataset only contains images with no temporal information, for the tracking task, we only use it to train the detector inside the tracker.

Since the TF dataset only contains one video with 5000 frames, the method is trained on the 300VW dataset and evaluated on the TF dataset. Due to the different landmark definitions between the TF and the 300VW dataset, we follow Liu *et al.* [14] to apply the seven common landmarks for testing.

The detector and tracker are evaluated by the accuracy of their predictions, which is quantified by the *Normalized Root Mean Squared Error* (NRMSE) between the predicted landmark coordinates and the ground truths. A lower NRMSE corresponds to a better accuracy. Following previous works [10, 24, 36], on the 300W, the 300VW, and the TF datasets, the error is normalized by the interocular distance of a face. On the AFLW dataset, it is normalized by the face size.

All experiments are conducted by Tensorflow 1.9.0 on a NVIDIA TESLA V100 GPU with 32GiB memory. A face is firstly cropped from the bounding box and scaled to 256×256 pixels, then fed into the detector or the tracker. A training batch contains 10 facial images or frames. The resolution of the 1D heatmap is set to three

L	64	128	256	64	128	256	384	512	640	768
L/F	0.25	0.5	1.0	0.25	0.5	1.0	1.5	2.0	2.5	3.0
Method	2D heatmap-based detector			The proposed 1D heatmap-based detector						
NRMSE 300W *com*	3.53	3.26	OOM	3.45	3.22	3.10	3.01	2.97	2.93	**2.91**
NRMSE 300W *cha*	6.16	5.62	OOM	6.22	5.68	5.45	5.40	5.39	5.35	**5.31**
NRMSE 300W *full*	4.04	3.72	OOM	3.99	3.70	3.56	3.48	3.44	3.40	**3.38**
Method	2D heatmap-based tracker			The proposed 1D heatmap-based tracker						
NRMSE 300VW S1	4.53	4.27	OOM	3.61	3.47	3.37	3.29	3.20	3.12	**3.06**
NRMSE 300VW S2	4.60	4.34	OOM	3.90	3.71	3.54	3.43	3.33	3.24	**3.17**
NRMSE 300VW S3	5.97	5.72	OOM	4.75	4.52	4.39	4.32	4.24	4.18	**4.12**

Table 1: NRMSE (%) of the 2D heatmap-based detector, the proposed detector, the 2D heatmap-based tracker and the proposed tracker with different output resolutions (L).The input face resolution (F) is fixed as 256. OOM is the abbreviation of "Out of Memory".

times the face size. For the experiment on each dataset, the training set is splitted as 10 folds to conduct cross validation and select optimal values for γ in Equation (10) and λ in Equation (13). γ and λ are searched from {0.1, 0.2, ..., 1.0}. The optimal γ found on the 300W training set, the AFLW training set, and the 300VW training set are 0.4, 0.5 and 0.4, respectively. The optimal λ found on the 300VW training set is 0.3.

In the following parts, we make quantitative evaluation on the proposed detector and tracker under different parameter settings, and compare them with related works. All hyper-parameters are assigned with their optimal values except for the parameter we want to further study on.

5.2 Experimental Results and Analysis under Different Heatmap Resolutions

Table 1 shows the NRMSE performance of the proposed detector and tracker with different heatmap resolutions. Due to the space limitation of the paper, Table 1 only displays the results on the 300W and the 300VW datasets, which have the most pre-defined landmarks. To compare with 2D heatmaps, we also implement a 2D heatmap-based detector and tracker and display their performance in Table 1. The 2D heatmap-based detector is the same as FAN [2], which takes the output of the stacked hourglass network as 2D heatmaps. Based on the 2D heatmaps predicted by the detector, the temporal recurrent learning method proposed by Xi *et al.* [20] is adopted as the compared tracker.

From Table 1, we have the following observations. First, when L increases to 256, the 2D heatmap-based methods crash by memory overload, which means the memory requirement is larger than the capacity (32 GiB) of the GPU. The huge output complexity of 2D heatmaps restricts L to a low value, i.e., 128, which is lower than the face size and causes great quantization errors. Second, compared to the 2D heatmap regression methods, the proposed method can achieve a much larger output resolution under limited GPU memory because the low spatial complexity ($2NL$) of 1D heatmap does not take significant memory overhead. Third, as L increases, NRMSE decreases. This is because with the increase of L, the quantization error is reduced and more detailed spatial and temporal patterns are captured. For example, as L increases from 64 to 768, the NRMSE of our method decreases by 15.65%, 14.63%, and 15.28% on the 300W
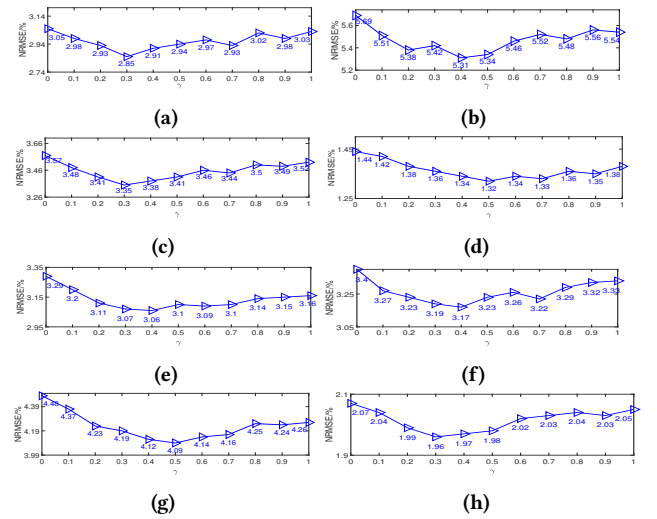

Figure 3: NRMSE (%) performance on (a) 300W *com*, (b) 300W *cha*, (c) 300W *full*, (d) AFLW, (e) 300VW S1, (f) 300VW S2, (g) 300VW S3 and (h) TF with different γ.

com, *cha*, and *full*. The light-weight 1D heatmap allows us to fully boost its resolution despite limited space, and the quantization error is well alleviated. With a much higher output resolution, the proposed detector and tracker outperform the 2D heatmap-based detector and tracker significantly on accuracy. Fourth, the proposed 1D heatmap-based tracker outperforms the 2D heatmap-based tracker significantly even if the output resolution is the same. The reason may be that, it is hard for the tracker to process the huge structure of the temporal sequence of 2D heatmaps. The proposed tracker solves the problem by only processing the light-weight 1D heatmaps, and therefore the performance is much better.

5.3 Ablation Study for the Co-Attention Module

The proposed co-attention module shares distributional features between the x and y axes to implicitly capture joint distributions. According to Equation (10), the weight of the attentive feature is

Dataset	300VW S1	300VW S2	300VW S3	TF
NRMSE	3.31/**3.06**	3.45/**3.17**	4.42/**4.12**	2.02/**1.97**

Table 2: NRMSE (%) of the proposed tracker without/with temporal refinement.

controlled by the parameter γ. When γ is 0.0, the co-attention module is discarded and the method only models marginal distributions separately on the two axes. As γ increases, the weight of the attentive feature in the fused representations $\mathbf{D}_k^{x'}$ and $\mathbf{D}_k^{y'}$ grows. We display NRMSE performance with different γ values in Fig. 3.

From Fig. 3, we find that there is a significant boost on detecting and tracking accuracy when γ increases from 0.0 to 0.4. Specifically, NRMSE decreases by 4.59%, 6.68% and 5.32% on 300W *com*, *cha* and *full*, respectively. It also decreases by 6.94% on the AFLW dataset; and by 6.99%, 6.76% and 8.04% on the three scenarios of the 300VW dataset; and by 4.83% on the TF dataset. That demonstrates the effectiveness of the co-attention module.

5.4 Ablation Study for the Temporal Refinement Mechanism of the Tracker

The tracker refines the heatmaps predicted by the detector by integrating temporal patterns from past frames, as shown in Equation (13). The weight of features from past frames is determined by the parameter λ. We make ablation study for the temporal refinement mechanism by comparing the results of two experimental settings. For the first setting, temporal refinement is discarded by assigning λ as 0.0. For the second setting, temporal refinement is kept by assigning λ as 0.3, the optimal value found by cross validation. Results of the two settings are shown in Table 2 and splitted by a slash. From Table 2, we find that the temporal refinement mechanism is beneficial to tracking accuracy. When λ is 0.3, NRMSE decreases by 7.55%, 8.12%, 6.79% and 2.48% on the three scenarios of the 300VW dataset and the TF datasets, respectively. This comparison demonstrates the effectiveness of the temporal refinement mechanism.

5.5 Comparison with State-of-the-art Methods

The proposed method is compared to other state-of-the-art landmark localization methods. From these approaches, coordinate regression methods include SDM [35], TSCN [26], IFA [1], CFSS [40], TCDCN [38], TSTN [14], DSRN [17], ODN [39], STA [30], Sun *et al.*'s work [27] and GAN [36]. Heatmap regression methods include Newell *et al.*'s work [18], SAN [9], LAB [32], CNN-CRF [5], LaplaceKL [23], Sun *et al.*'s work [28], DSNT [19] , DARK [37], FHR [30], GHCU [15] and Chen *et al.*'s work [6]. From these methods, CFSS, TCDCN, DSRN, ODN, Newell *et al.*'s work, SAN, LaplaceKL, Sun *et al.*'s work [28], DSNT, DARK, FHR, GHCU, Chen *et al.*'s work, LAB and CNN-CRF are detection methods. TSCN, TSTN, STA, Sun *et al.*'s work [27] and GAN are tracking methods. SDM, IFA contain both a detection method and a tracking method. Some newly proposed semi-supervised learning or unsupervised learning methods [10, 11, 21] for landmark detection are trained under different conditions with our method so are not included in our comparison.

Table 3 lists the NRSME performance of the proposed detector and the compared methods on the image datasets, i.e., the 300W public testing set and the AFLW dataset. Table 4 lists the NRSME performance of the proposed tracker and the compared methods on the video datasets, i.e., the 300VW and the TF dataset, respectively. Performances of the compared methods are directly copied from literature, except for that of Sun *et al.*'s work [28], DSNT, DARK, FHR in Table 3 because we could not find their published results on the respective datasets. We just re-implement them under the same experimental condition with our method. Their open source codes[1] are used to facilitate re-implementation. Other methods lacking published results or evaluated under different metrics or normalization standards are just left blank.

From Table 3, we find that our method achieves state-of-the-art detecting accuracy on the image datasets. We outperform Sun *et al.*'s work [28], DSNT, DARK and FHR, which also try to capture the fractional parts of coordinates and reduce quantization errors. These methods still maintain a 2D heatmap structure whose resolution is limited by the memory capacity, and therefore suffering from the information loss caused by the quantization process. Our method significantly boosts the output resolution with a lightweight output structure, i.e., the 1D heatmap, which captures more detailed distributional information and significantly alleviates the quantization error. We also combine our method with DARK by adapting DARK to 1D Gaussian distribution to estimate ground truth positions from the predicted 1D heatmaps. As shown in Table 3, the result of Ours+DARK outperforms the original DARK. This result further demonstrates the superiority of the proposed 1D heatmap structure.

Table 4 shows that the proposed tracker achieves state-of-the-art tracking accuracy on the video datasets. Among the compared methods, GAN, a combination of coordinate regression and adversarial learning, achieves the best performance after our method. Our method outperforms the NRMSE of GAN by 12.57%, 13.62% and 7.00% on the 300VW scenarios 1, 2, 3 and by 2.96% on the TF dataset, respectively. For fine-grained analysis, we gather all testing samples in the three scenarios of 300VW and group their landmarks into five facial areas, i.e., eyes, contour, nose, eyebrows and mouth. We display the NRMSE performance on each area in Fig. 4. We also calculate the variances of x and y coordinates for each landmark and depict their average value (s) within each area in Fig. 5. From Fig. 4, our method decreases NRMSE on every facial area compared to GAN, especially in contour by 13.67%. From Fig. 5, we find that the contour area has the highset coordinate variances on both the x and y axes. Although GAN uses adversarial learning for spatio-temporal modeling and promotes the robustness of landmark localization in the challenging facial area, it is still sub-optimal on handling coordinates with large variances due to the intrinsic weakness of the coordinate regression technique. We significantly alleviate this problem with the proposed heatmap regression method which is more adept at capturing spatial and temporal distributions.

---

[1]https://github.com/JimmySuen/integral-human-pose
https://github.com/anibali/dsntnn
https://ilovepose.github.io/coco/
https://github.com/tyshiwo/FHR_alignment

Method	SAN	CNN-CRF	DSRN	LAB	LaplaceKL	ODN	[28]	DSNT	DARK	FHR	[6]	[18]	Ours	Ours+DARK
300W com	3.34	3.33	4.12	2.98	3.19	3.56	3.22	3.17	2.98	3.04	-	3.30	2.91	**2.68**
300W cha	6.60	6.29	9.68	5.19	6.87	6.67	5.62	5.54	5.48	6.21	-	5.69	5.31	**4.92**
300W full	3.98	3.91	5.21	3.49	3.91	4.17	3.69	3.63	3.47	3.66	-	3.77	3.38	**3.12**
AFLW	1.91	-	1.86	1.25	1.97	1.63	1.80	1.74	1.43	1.58	1.39	1.95	1.32	**1.17**

Table 3: NRSME (%) of the proposed detector and the compared methods on the image datasets.

Method	SDM	TSCN	CFAN	CFSS	IFA	TCDCN	TSTN	GHCU	[27]	FHR	FHR+STA	GAN	Ours
300VW S1	7.41	12.54	-	7.68	-	7.66	5.36	3.85	3.56	4.82	4.21	3.50	**3.06**
300VW S2	6.18	7.25	-	6.42	-	6.77	4.51	3.46	3.88	4.23	4.02	3.67	**3.17**
300VW S3	13.04	13.13	-	13.67	-	14.98	12.84	7.51	5.02	7.09	5.64	4.43	**4.12**
TF	4.01	-	3.52	2.36	3.45	-	2.13	-	-	2.07	2.10	2.03	**1.97**

Table 4: NRSME (%) of the proposed tracker and the compared methods on the 300VW and the TF datasets
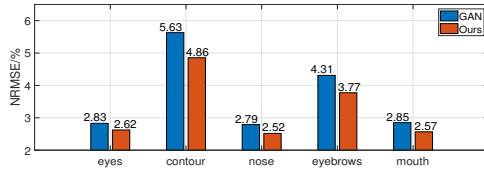

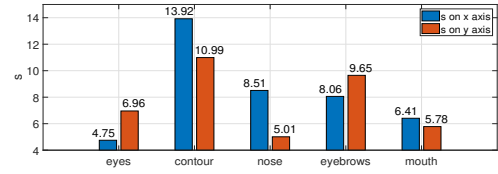Figure 4: NRMSE (%) performance on each facial area.

Figure 5: Average value (s) of coordinate variances within each facial area.
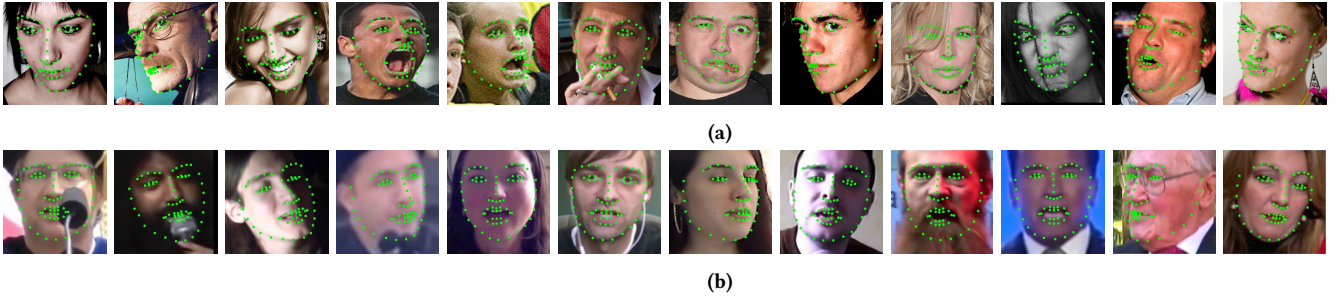

(a)

(b)

Figure 6: Visualization of (a) detection results on the 300W testing set and (b) tracking results on the 300VW testing set.

6 VISUALIZATION OF THE DETECTION AND TRACKING RESULTS

We visualize the detection and tracking results of our method in Fig. 6, respectively. Testing images are sampled from the 300W and the 300VW testing sets. From these testing samples, we find that the proposed method can perform well under challenging conditions, such as occlusions, low resolution, non-frontal head poses, dramatic facial expressions and uneven illuminations.

7 CONCLUSION

To address the issues caused by the huge spatial complexity of 2D heatmaps, we propose a new method that predicts 1D heatmaps as marginal distributions on each axis to detect facial landmarks. Instead of modeling joint distribution explicitly that occupies much memory, we design a co-attention mechanism to share features between the x and y axes and implicitly capture joint distributions. The light-weight 1D heatmap structure enables us to boost the output resolution to a large value despite limited GPU space, making more accurate predictions. Based on such an idea, we propose a novel landmark detector and a tracker. The detector captures spatial patterns in an image, while the tracker further captures temporal patterns in a video by a temporal refinement mechanism. Experiments on the 300W, AFLW, 300VW and TF datasets demonstrate that the proposed detector and tracker outperform state-of-the-art methods.

REFERENCES

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. 2014. Incremental Face Alignment in the Wild. In *CVPR*. 1859–1866.

[2] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*. 1021–1030.

[3] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. 2013. Robust Face Landmark Estimation under Occlusion. In *ICCV*. 1513–1520.

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

[5] Lisha Chen, Hui Su, and Qiang Ji. 2019. Deep Structured Prediction for Facial Landmark Detection. In *NeurIPS*.

[6] Y Chen, C Shen, H Chen, XS Wei, L Liu, and J Yang. 2019. Adversarial Learning of Structure-Aware Fully Convolutional Networks for Landmark Localization. *TPAMI* (2019).

[7] Grigorios G Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, and Stefanos Zafeiriou. 2018. A comprehensive performance evaluation of deformable face tracking "In-the-Wild". *IJCV* 126, 2-4 (2018), 198–232.

[8] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang. 2017. Multi-context Attention for Human Pose Estimation. In *CVPR*. 5669–5678.

[9] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. 2018. Style Aggregated Network for Facial Landmark Detection. In *CVPR*. 379–388.

[10] Xuanyi Dong and Yi Yang. 2019. Teacher Supervises Students How to Learn From Partially Labeled Images for Facial Landmark Detection. In *ICCV*.

[11] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. 2018. Supervision-by-Registration: An Unsupervised Approach to Improve the Precision of Facial Landmark Detectors. In *CVPR*. 360–368.

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

[13] Y. Li, S. Wang, Y. Zhao, and Q. Ji. 2013. Simultaneous Facial Feature Tracking and Facial Expression Recognition. *IEEE Transactions on Image Processing* 22, 7 (2013), 2559–2573.

[14] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2018. Two-stream transformer networks for video-based face alignment. *TPAMI* 40, 11 (2018), 2546–2554.

[15] Zhiwei Liu, Xiangyu Zhu, Guosheng Hu, Haiyun Guo, Ming Tang, Zhen Lei, Neil M. Robertson, and Jinqiao Wang. 2019. Semantic Alignment: Finding Semantically Consistent Ground-Truth for Facial Landmark Detection. In *CVPR*.

[16] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NeurIPS*. 289–297.

[17] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. 2018. Direct Shape Regression Networks for End-to-End Face Alignment. In *CVPR*. 5040–5049.

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *ECCV*. 483–499.

[19] Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. 2018. Numerical Coordinate Regression with Convolutional Neural Networks. *CoRR* abs/1801.07372 (2018).

[20] Xi Peng, Rogerio S Feris, Xiaoyu Wang, and Dimitris N Metaxas. 2016. A recurrent encoder-decoder network for sequential face alignment. In *ECCV*. 38–56.

[21] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. 2019. Aggregation via Separation: Boosting Facial Landmark Detector With Semi-Supervised Style Translation. In *ICCV*.

[22] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. 2014. Face Alignment at 3000 FPS via Regressing Local Binary Features. In *CVPR*. 1685–1692.

[23] Joseph P. Robinson, Yuncheng Li, Ning Zhang, Yun Fu, and Sergey Tulyakov. 2019. Laplace Landmark Localization. In *ICCV*.

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

[25] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. 2015. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *ICCV Workshops*. 50–58.

[26] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*. 568–576.

[27] Keqiang Sun, Wayne Wu, Tinghao Liu, Shuo Yang, Quan Wang, Qiang Zhou, Zuochang Ye, and Chen Qian. 2019. FAB: A Robust Facial Landmark Detection Framework for Motion-Blurred Videos. In *ICCV*.

[28] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. 2018. Integral Human Pose Regression. In *ECCV*. 536–553.

[29] Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2013. Deep Convolutional Network Cascade for Facial Point Detection. In *CVPR*. 3476–3483.

[30] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. 2019. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*. 8893–8900.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.

[32] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. 2018. Look at Boundary: A Boundary-Aware Face Alignment Algorithm. In *CVPR*. 2129–2138.

[33] Yue Wu and Qiang Ji. 2019. Facial Landmark Detection: A Literature Survey. *IJCV* 127, 2 (2019), 115–142.

[34] Yue Wu, Ziheng Wang, and Qiang Ji. 2014. A Hierarchical Probabilistic Model for Facial Feature Detection. In *CVPR*. 1781–1788.

[35] Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *CVPR*. 532–539.

[36] Shi Yin, Shangfei Wang, Guozhu Peng, Xiaoping Chen, and Bowen Pan. 2019. Capturing Spatial and Temporal Patterns for Facial Landmark Tracking through Adversarial Learning. In *IJCAI*. 1010–1017.

[37] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. 2020. Distribution-Aware Coordinate Representation for Human Pose Estimation. *CVPR* (2020).

[38] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. Learning deep representation for face alignment with auxiliary attributes. *TPAMI* 38, 5 (2016), 918–930.

[39] Meilu Zhu, Daming Shi, Mingjie Zheng, and Muhammad Sadiq. 2019. Robust Facial Landmark Detection via Occlusion-Adaptive Deep Networks. In *CVPR*.

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