Towards Highly Accurate and Stable Face Alignment for High-Resolution Videos

Ying Tai[†*] Yicong Liang[†*] Xiaoming Liu[‡] Lei Duan[§] Jilin Li[†] Chengjie Wang[†] Feiyue Huang[†] Yu Chen[£]

[†]Youtu Lab, Tencent [‡]Michigan State University

[§]Fudan University [£]Nanjing University of Science and Technology

[†]{yingtai, easonliang, jerolinli, jasoncjwang, garyhuang}@tencent.com

[‡]liuxm@cse.msu.edu, [§]15307130193@fudan.edu.cn, [£]chenyu1523@gmail.com

https://github.com/tyshiwo/FHR_alignment

Abstract

In recent years, heatmap regression based models have shown their effectiveness in face alignment and pose estimation. However, Conventional Heatmap Regression (CHR) is not accurate nor stable when dealing with high-resolution facial videos, since it finds the maximum activated location in heatmaps which are generated from rounding coordinates, and thus leads to quantization errors when scaling back to the original high-resolution space. In this paper, we propose a Fractional Heatmap Regression (FHR) for high-resolution video-based face alignment. The proposed FHR can accurately estimate the fractional part according to the 2D Gaussian function by sampling three points in heatmaps. To further stabilize the landmarks among continuous video frames while maintaining the precise at the same time, we propose a novel stabilization loss that contains two terms to address time delay and non-smooth issues, respectively. Experiments on 300W, 300-VW and Talking Face datasets clearly demonstrate that the proposed method is more accurate and stable than the state-of-the-art models.

Introduction

Face alignment aims to estimate a set of facial landmarks given a face image or video sequence. It is a classic computer vision problem that has attributed to many advanced machine learning algorithms Fan et al. (2018); Bulat and Tzimiropoulos (2017); Trigeorgis et al. (2016); Peng et al. (2015, 2016); Kowalski, Naruniec, and Trzcinski (2017); Chen et al. (2017); Liu et al. (2017); Hu et al. (2018). Nowadays, with the rapid development of consumer hardwares (e.g., mobile phones, digital cameras), High-Resolution (HR) video sequences can be easily collected. Estimating facial landmarks on such high-resolution facial data has tremendous applications, e.g., face makeup Chen, Shen, and Jia (2017), editing with special effects Korshunova et al. (2017) in live broadcast videos. However, most existing face alinement methods work on faces with medium image resolutions Chen et al. (2017); Bulat and Tzimiropoulos (2017); Peng et al. (2016); Liu et al. (2017). Therefore, developing face alignment algorithms for *high-resolution videos* is at the core of this paper.

To this end, we propose an accurate and stable algorithm for high-resolution video-based face alignment, named Fractional Heatmap Regression (FHR). It is well known that heatmap regression have shown its effectiveness in landmark estimation tasks Chen et al. (2017); Newell, Yang, and Deng (2016); Chen* et al. (2018). However, Conventional Heatmap Regression (CHR) is not accurate nor stable when dealing with high-resolution facial images, since it finds the maximum activated location in heatmaps which are generated from *rounding* coordinates, and thus leads to quantization errors as the heatmap resolution is much lower than the input image resolution (e.g., 128 vs. 930 shown in Fig. 1) due to the *scaling operation*. To address this problem, we propose a novel transformation between heatmaps and coordinates, which not only preserves the fractional part when generating heatmaps from the coordinates, but also accurately estimate the fractional part according to the 2D Gaussian function by sampling three points in heatmaps.

Using our proposed FHR, we can estimate more accurate landmarks compared to the conventional heatmap regression model, and achieve state-of-the-art performance on popular video benchmarks: Talking Face FGNET (2014) and 300-VW datasets Shen et al. (2017) compared to recent video-based face alignment models Liu et al. (2017); Peng et al.

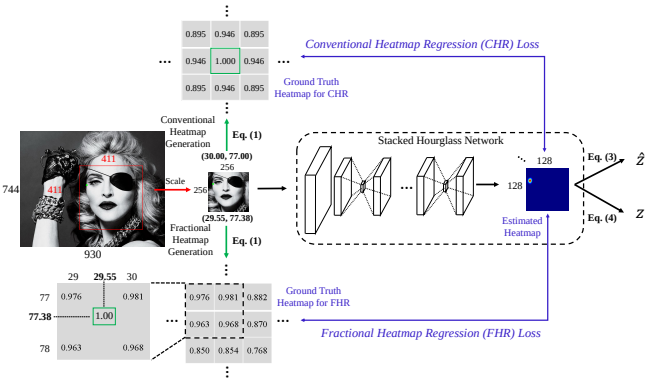


Figure 1: Comparisons between fractional regression heatmap and conventional heatmap regression. Our method differs conventional one in two aspects: 1) the ground truth heatmap for FHR maintains the precision of fractional coordinate, while the conventional one *discards* (e.g., from 29.55 to 30.00, 77.38 to 77.00); and 2) three sampled points on the heatmap analytically computes the fractional peak location of the heatmap (Eq. 4), while the conventional one only finds the maximum activated location (Eq. 3) that loses the fractional part and thus leads to *quantization error*.

∗ indicates equal contributions.

(2016). However, real-world applications such as face makeup in videos often demands extremely high *stability*, since the makeup jumps if the estimated landmarks oscillate between consecutive frames, which negatively impacts the user's experience. To make the sequential estimations as stable as possible, we further develop a novel stabilization algorithm on the landmarks estimated by FHR, which contains two terms, a regularization term ($\mathcal{L}_{reg}$) and a temporal coherence term ($\mathcal{L}_{tm}$), to address two common difficulties: time delay and non-smooth problems, respectively. Specifically, $\mathcal{L}_{reg}$ combines traditional Euclidean loss and a novel loss account for time delay; $\mathcal{L}_{tm}$ generalizes the temporal coherence loss in Cao, Hou, and Zhou (2014) to better handle nonlinear movement of facial landmark.

In summary, the main contributions of this paper are:

- A novel Fractional Heatmap Regression method for high-resolution video based face alignment that leverages 2D Gaussian prior to preserve the fractional part of points.
- A novel stabilization algorithm that addresses time delay and non-smooth problems among continuous video frames is proposed.
- State-of-the-art performance, both in accuracy and stability, on the benchmarks of 300W Sagonas et al. (2013), 300-VW Shen et al. (2017) and Talking Face FGNET (2014) datasets.

Related Work

Heatmap Regression Heatmap regression is one of the most widely used approaches for landmark localization tasks, which estimates a set of heatmaps rather than coordinates. Stacked Hourglass Networks (SHN) are popular architectures in heatmap regression, which have symmetric topology that capture and consolidate information across all scales of the image. Newell et al. Newell, Yang, and Deng (2016) proposed SHN for 2D human pose estimation, which achieved remarkable results even for very challenging datasets Andriluka et al. (2014). With the hourglass structure, Chu et al. Chu et al. (2017) introduced multi-context attention mechanism into Convolutional Neural Networks (CNN). Apart from applications in human pose estimation, there are also several heatmap regression based models for face alignment. Chen et al. Chen et al. (2017) proposed a structure-aware fully convolutional network to implicitly model the priors during training. Bulat et al. Bulat and Tzimiropoulos (2017) built a powerful CNN for face alignment based on the hourglass network and a hierarchical, parallel and multi-scale block.

However, all existing models drops the fractional part of coordinates during the transformation between heatmaps and points, which brings quantization errors to high-resolution facial images. On the contrary, our proposed FHR can accurately estimate the fractional part according to the 2D Gaussian function by sampling three points in heatmaps, and thus achieves more accurate alignment.

Video-based Face Alignment Video-based face alignment estimates facial landmarks in video sequences Liu (2010). Early methods Black and Yacoob (1995); Shen et al. (2017) used incremental learning to predict landmarks on still frames in a tracking-by-detection manner. To address the issue that generic methods are sensitive to initializations, Peng et al. Peng et al. (2015) exploited incremental learning for personalized ensemble alignment, which samples multiple initial shapes to achieve image congealing within one frame. To explicitly model the temporal dependency Oh et al. (2015) of landmarks across frames, the authors Peng et al. (2016) further incorporated a sequence of spatial and temporal recurrences for sequential face alignment in videos. Recently, Liu et al. Liu et al. (2017) proposed a Two-Stream Transformer Networks (TSTN) approach, which captures the complementary information of both the spatial appearance on still frames and the temporal consistency across frames. Different from Peng et al. (2016); Liu et al. (2017) that require temporal landmark labels across frames, our proposed method achieves state-of-the-art accuracy only by making full use of the spatial appearance on still frames, which is able to remedy the problem that *labeled sequential data are very limited*.

Apart from the accuracy of landmarks, stabilization is also a key metric to video-based alignment. Typically, two terms Cao, Hou, and Zhou (2014) are adopted for stabilization, where a regularization term drives the optimized results to be more expressive, and a temporal coherence term drives the results to be more stable and smooth. However, the existing stabilization algorithm is sensitive to time delay and nonlinear movements. Our proposed algorithm takes these into account and thus are overall more robust.

The Proposed Approach

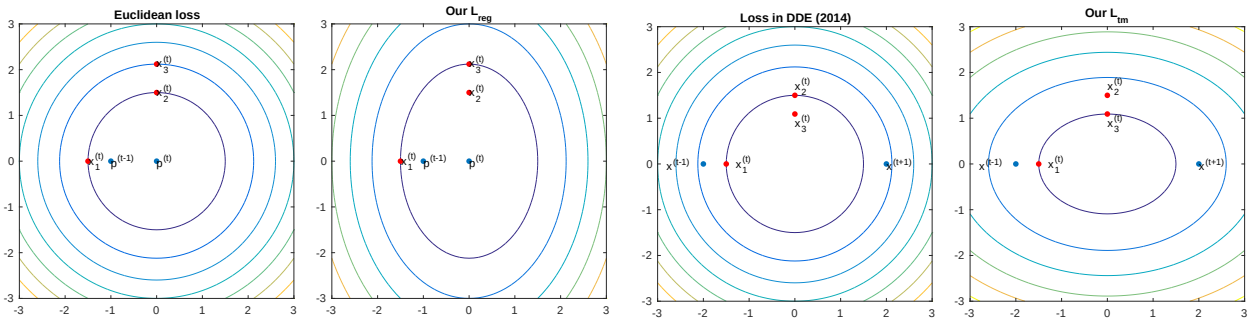
In this section, we introduce the details of the proposed approach based on heatmap regression. A key point of heatmap regression is the transformation between the heatmaps and coordinates. Specifically, before model training, a *pre-process step* is conducted to convert the coordinates to the heatmaps, which are used as the ground truth labels. After estimating the heatmaps, a *post-process step* is conducted to obtain the coordinates from the estimated heatmaps. In this work, we propose a novel transformation between the heatmaps and coordinates that is different from the conventional heatmap regression, which is demonstrated to be simple yet effective.

Fractional Heatmap Regression

As shown in Fig. 1, conventional heatmap regression mainly generates the heatmaps from integral coordinates, despite that the ground truth coordinates are usually with fractions.As a result, it causes quantization errors when scaling back to the original image resolution, since the heatmaps are usually of much lower resolution than the input image. To address this problem, our proposed fractional heatmap regression generates ground truth heatmaps based on the intact ground truth coordinates (see Fig. 1) as follows:

$$\bar{\mathbf{H}}_{(m)}(\mathbf{c}) = \exp\left(-\frac{1}{2\sigma^2}((c_x - c_{(m)x_0})^2 + (c_y - c_{(m)y_0})^2)\right), \tag{1}$$

where $\mathbf{c} = (c_x, c_y) \in \Omega$ represents the coordinate, Ω is the domain of the heatmap $\mathbf{H}$, σ denotes the standard deviation and $(c_{(m)x_0}, c_{(m)y_0})$ is the center of the 2D Gaussian in the mth heatmap.



(a) (b)

Figure 2: (a) Histograms of the Euclidean loss (left) and our $\mathcal{L}_{reg}$ (right). $\mathbf{p}^{(t)}$ and $\mathbf{p}^{(t-1)}$ are ground truths of frame t and $t-1$, respectively. $\mathbf{x}_1^{(t)}$ has the same Euclidean loss as $\mathbf{x}_2^{(t)}$. However, since $\mathbf{x}_1^{(t)}$ lies on the line $\overline{\mathbf{p}^{(t)}\mathbf{p}^{(t-1)}}$, it indicates a loss caused by time delay, which is more likely to happen in the stabilization process. Thus our model prefers to assign it a larger loss (equal to $\mathbf{x}_3^{(t)}$). (b) Histograms of the loss in Cao, Hou, and Zhou (2014) (left) and our $\mathcal{L}_{tm}$ (right). $\mathbf{x}^{(t-1)}$ and $\mathbf{x}^{(t+1)}$ are stabilization outputs of frame $t-1$ and $t+1$, respectively. $\mathbf{x}_1^{(t)}$ has the same loss as $\mathbf{x}_2^{(t)}$ in Cao, Hou, and Zhou (2014). However, since $\mathbf{x}_1^{(t)}$ lies on the line $\overline{\mathbf{x}^{(t-1)}\mathbf{x}^{(t+1)}}$, we argue that movement trajectory $\overrightarrow{\mathbf{x}^{(t-1)}\mathbf{x}_1^{(t)}\mathbf{x}^{(t+1)}}$ is more smooth than trajectory $\overrightarrow{\mathbf{x}^{(t-1)}\mathbf{x}_2^{(t)}\mathbf{x}^{(t+1)}}$. Thus our model assigns it a smaller loss (equal to $\mathbf{x}_3^{(t)}$). Note that the short axis of the ellipse in (a) is $\overline{\mathbf{p}^{(t)}\mathbf{p}^{(t-1)}}$ while the long axis of the ellipse in (b) is $\overline{\mathbf{x}^{(t-1)}\mathbf{x}^{(t+1)}}$, thus the two terms are not in contradiction.

Denoting $\mathbf{I}$ an input image and $\mathcal{F}$ the deep alignment model, we can estimate the recovered heatmaps by

$$\hat{\mathbf{H}} = \mathcal{F}(\mathbf{I}), \tag{2}$$

where $\hat{\mathbf{H}} = [\hat{\mathbf{H}}_{(1)}, ..., \hat{\mathbf{H}}_{(m)}, ..., \hat{\mathbf{H}}_{(M)}]$, and M is the number of landmarks. Given Eq. (1) and $\hat{\mathbf{H}}_{(m)}$, estimating the fractional coordinate $\mathbf{z}$ amounts to solving a binary quadratic equation, which has a closed-form solution as long as we can sample any *three* non-zero points from the heatmap. Specifically, we first obtain $\hat{\mathbf{z}}$ as the location $\mathbf{c}$ with the *maximum likelihood*, as in conventional heatmap regression:

$$\hat{\mathbf{z}}_m = \underset{\mathbf{c}}{\operatorname{argmax}}\, \hat{\mathbf{H}}_{(m)}(\mathbf{c}),\ m = 1, ..., M. \tag{3}$$

Conventional heatmap regression directly takes $\hat{\mathbf{z}}_m$ as the output, which loses the fractional part. In our method, we further sample another two points, e.g., $\hat{\mathbf{z}}_m^1 = (\hat{z}_{(m)x} + 1, \hat{z}_{(m)y}), \hat{\mathbf{z}}_m^2 = (\hat{z}_{(m)x}, \hat{z}_{(m)y} + 1)$ near $\hat{\mathbf{z}}_m = (\hat{z}_{(m)x}, \hat{z}_{(m)y})$ [1]. Let $h_{(m)}^1 = \hat{\mathbf{H}}_{(m)}(\hat{\mathbf{z}}_m^1)$, $h_{(m)}^2 = \hat{\mathbf{H}}_{(m)}(\hat{\mathbf{z}}_m^2)$ and $h_{(m)} = \hat{\mathbf{H}}_{(m)}(\hat{\mathbf{z}}_m)$, we then estimate the *fractional coordinate* $\mathbf{z}_m = (z_{(m)x}, z_{(m)y})$ as follows:

$$\begin{aligned} z_{(m)x} =& \sigma^2(\ln h_{(m)}^1 - \ln h_{(m)}) - \\ & \frac{1}{2}((\hat{z}_{(m)x})^2 - (\hat{z}_{(m)x}^1)^2 + (\hat{z}_{(m)y})^2 - (\hat{z}_{(m)y}^1)^2), \\ z_{(m)y} =& \sigma^2(\ln h_{(m)}^2 - \ln h_{(m)}) - \\ & \frac{1}{2}((\hat{z}_{(m)x})^2 - (\hat{z}_{(m)x}^2)^2 + (\hat{z}_{(m)y})^2 - (\hat{z}_{(m)y}^2)^2). \end{aligned} \tag{4}$$

Detailed derivation can be found in the supplementary material. It should be noted that our fractional heatmap regression is applicable to any heatmap based methods. In this paper, we focus on face alignment, and adopt the stacked hourglass network Newell, Yang, and Deng (2016) as the alignment model $\mathcal{F}$ that minimizes the loss of $\|\bar{\mathbf{H}} - \hat{\mathbf{H}}\|^2$ across the entire training set, where $\bar{\mathbf{H}} = [\bar{\mathbf{H}}_{(1)}, ..., \bar{\mathbf{H}}_{(m)}, ..., \bar{\mathbf{H}}_{(M)}]$.

[1] In case $\hat{\mathbf{z}}_m$ is located at the edge of the heatmap, we would sample the points in the opposite directions.

Stabilization Algorithm

We now introduce our stabilization algorithm for video-based alignment, which takes the alignment results of $\mathcal{F}$ in all the past frames as input, and outputs a more accurate and stable result for the current frame.

Stabilization Model We denote $\mathbf{z}^{(t)}$ as the output of $\mathcal{F}$ at frame t and the stabilization model as $\mathcal{M}_\Theta$, which has parameters Θ to be optimized. $\mathcal{M}_\Theta$ takes $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(t)}$ as input, and outputs the stabilized landmarks of frame t, which is denoted as $\mathbf{x}^{(t)}$. Therefore we have

$$\mathbf{x}^{(t)} = \mathcal{M}_\Theta\left(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(t)}\right). \tag{5}$$

Assume there are V videos in the training set, and the ith video has T_i frames. For frame t in the ith video, we denote its ground truth landmarks as $\mathbf{p}_i^{(t)}$, the output of $\mathcal{F}$ as $\mathbf{z}_i^{(t)}$, and the stabilized output as $\mathbf{x}_i^{(t)}$ ($\mathbf{p}_i^{(t)}, \mathbf{z}_i^{(t)}, \mathbf{x}_i^{(t)} \in \mathbb{R}^{2M\times 1}$). Here, we have $\mathbf{x}_i^{(t)} = \mathcal{M}_\Theta\left(\mathbf{z}_i^{(1)}, \ldots, \mathbf{z}_i^{(t)}\right)$.

Next, we present the specific form of $\mathcal{M}_\Theta$ as well as its parameters Θ. Our model follows a Bayesian framework. Specifically, we model the prior distribution of $\mathbf{x}_i^{(t)}$ given $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(t-1)}$ as a K-component Gaussian mixture:

$$Pr\left(\mathbf{x}^{(t)}|\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(t-1)}\right) = \sum_{k=1}^{K} \mathcal{N}\left(\mathbf{x}^{(t)}; \mu_k^{(t)}, \Sigma_k^{(t)}\right), \tag{6}$$

where $Pr(\cdot)$ indicates the density function, $\mathcal{N}(\cdot; \mu, \Sigma)$ indicates a normal distribution with mean μ and covariance Σ. We then model the likelihood of $\mathbf{z}^{(t)}$ given $\mathbf{x}^{(t)}$ as Gaussian

$$Pr\left(\mathbf{z}^{(t)}|\mathbf{x}^{(t)}\right) = \mathcal{N}\left(\mathbf{z}^{(t)}; \mathbf{x}^{(t)}, \Sigma_{noise}\right), \tag{7}$$

and use the Bayesian rule to obtain the most probable value of $\mathbf{x}^{(t)}$:

$$\begin{aligned}\mathbf{x}^{(t)} &= \underset{\mathbf{x}}{\operatorname{argmax}}\, Pr\left(\mathbf{x}|\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(t)}\right)\\ &= \underset{\mathbf{x}}{\operatorname{argmax}}\, \frac{Pr\left(\mathbf{x},\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(t)}\right)}{Pr\left(\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(t)}\right)}\\ &= \underset{\mathbf{x}}{\operatorname{argmax}}\, \frac{Pr\left(\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(t-1)}\right) Pr\left(\mathbf{x}|\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(t-1)}\right) Pr\left(\mathbf{z}^{(t)}|\mathbf{x}\right)}{Pr\left(\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(t)}\right)}\\ &= \underset{\mathbf{x}}{\operatorname{argmax}}\, Pr\left(\mathbf{x}|\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(t-1)}\right) Pr\left(\mathbf{z}^{(t)}|\mathbf{x}\right).\end{aligned} \tag{8}$$

Combining (6),(7) and (8), we can obtain the closed-form solution of (5). In practice, we fix $K = 2$ since it already achieves satisfactory results and larger K may cause both efficiency and overfitting issues. Moreover, to reflect the fact that $\mathbf{x}^{(t)}$ has a decreasing correlation with $\mathbf{x}^{(t-\tau)}$ when τ increases, we assume that

$$\begin{aligned}\mu_k^{(t)} &= \frac{\sum_{\tau=1}^{t-1}\gamma^\tau \mathbf{x}^{(t-\tau)}}{\sum_{\tau=1}^{t-1}\gamma^\tau},\\ \Sigma_k^{(t)} &= \beta_k\Sigma_k + \\ &\quad (1-\beta_k)\frac{\sum_{\tau=1}^{t-1}\gamma^\tau\left(\mathbf{x}^{(t-\tau)}-\mu_k^{(t)}\right)^T\left(\mathbf{x}^{(t-\tau)}-\mu_k^{(t)}\right)}{\sum_{\tau=1}^{t-1}\gamma^\tau}.\end{aligned} \tag{9}$$

where $\gamma, \{\beta_k\}_{k=1}^K \in [0,1]$ along with $2M \times 2M$ positive semi-definite matrices $\{\Sigma_k\}_{k=1}^K$ and Σ_{noise} are unknown model parameters. In practice Eq. (9) can be calculated recursively, whose computational complexity remains constant when t increases.

To further reduce the number of parameters, we calculate the covariance matrix $\mathbf{S}$ of all $\mathbf{p}_i^{(t)} - \mathbf{p}_i^{(t-1)}$ in the training set, denote the matrix of eigenvectors of $\mathbf{S}$ as $\mathbf{V}$, and finally assume

$$\begin{aligned}\Sigma_{noise} &= \mathbf{V}^T\Gamma_{noise}\mathbf{V},\\ \Sigma_k &= \mathbf{V}^T\Gamma_k\mathbf{V}, k = 1,\ldots,K,\end{aligned} \tag{10}$$

where Γ_{noise} and $\{\Gamma_k\}_{k=1}^K$ are diagonal matrices. In summary, we have $\Theta = \left[\gamma, \{\beta_k\}_{k=1}^K, \Gamma_{noise}, \{\Gamma_k\}_{k=1}^K\right]$, which will be optimized with the loss function in the next subsection.

Loss Function Design We now introduce and optimize a novel loss function so as to estimate the above stabilization model parameter Θ. Throughout this section, we denote all $\mathbf{x}_i^{(t)}$ as $\mathbf{x}_i^{(t)}(\Theta)$ to emphasize that the stabilized landmarks are functions of model parameter Θ. Our loss function is defined as follows:

$$\mathcal{L}(\Theta) = \mathcal{L}_{reg}(\Theta) + \lambda_1\mathcal{L}_{tm}(\Theta). \tag{11}$$

This loss has two terms. The first term, $\mathcal{L}_{reg}(\Theta)$ is called the regularization loss, which regularizes the stabilized output to be close to the ground truth, and is defined as follows:

$$\mathcal{L}_{reg}(\Theta) = \frac{\sum_{i=1}^V\sum_{t=1}^{T_i}\|\mathbf{x}_i^{(t)}(\Theta)-\mathbf{p}_i^{(t)}\|_2^2}{\sum_{i=1}^V T_i} + \lambda_2\frac{\sum_{i=1}^V\sum_{t=2}^{T_i}\left(\left(\mathbf{p}_i^{(t)}-\mathbf{p}_i^{(t-1)}\right)^+\left(\mathbf{x}_i^{(t)}(\Theta)-\mathbf{p}_i^{(t)}\right)\right)^2}{\sum_{i=1}^V(T_i-1)}, \tag{12}$$

where $\mathbf{x}^+$ indicates the Moore-Penrose general inverse of vector/matrix $\mathbf{x}$. We can see that the first term of (12) is the average Euclidean distance of the ground truth and the model output. The second term aims to fit every $\mathbf{x}_i^{(t)}(\Theta)$ in terms of $\alpha\mathbf{p}_i^{(t-1)} + (1-\alpha)\mathbf{p}_i^{(t)}$, where α is the coefficient, and estimate the expectation of α^2 (detailed derivation can be found in the supplementary material). If this expectation is large, it means that 1) $\mathbf{x}_i^{(t)}(\Theta)$ is more similar to $\mathbf{p}_i^{(t-1)}$ than $\mathbf{p}_i^{(t)}$, and 2) the model output has a significant time delay, which is undesirable. Since our stabilization model uses the alignment results of the past frames, how to avoid time delay is a critical task. Therefore we emphasize the time delay loss as an individual term in $\mathcal{L}_{reg}$ (see Fig. 2(a)).

The second term in (11), $\mathcal{L}_{tm}(\Theta)$ is called the smooth loss which favors the stabilized output to be smooth, and is defined as follows:

$$\mathcal{L}_{tm}(\Theta) = \min_{\boldsymbol{q}}\frac{\sum_{i=1}^V\sum_{t=2}^{T_i-1}\left(\|\mathbf{x}_i^{(t)}(\Theta)-q_i^{(t)}\mathbf{x}_i^{(t-1)}(\Theta)-(1-q_i^{(t)})\mathbf{x}_i^{(t+1)}(\Theta)\|_2^2+\lambda_3\|q_i^{(t)}-.5\|_2^2\right)}{\sum_{i=1}^V(T_i-2)}, \tag{13}$$

where $q_i^{(t)} \in \mathbb{R}$ and $\boldsymbol{q} = (q_1^{(2)},\ldots,q_V^{(T_V-1)})$. $\mathcal{L}_{tm}(\Theta)$ can be seen as a trade-off between two stability losses, controlled by λ_3. When $\lambda_3 = 0$, it is equivalent to

$$\min_{\boldsymbol{q}}\frac{\sum_{i=1}^V\sum_{t=2}^{T_i-1}\|\mathbf{x}_i^{(t)}(\Theta)-q_i^{(t)}\mathbf{x}_i^{(t-1)}(\Theta)-(1-q_i^{(t)})\mathbf{x}_i^{(t+1)}(\Theta)\|_2^2}{\sum_{i=1}^V(T_i-2)}, \tag{14}$$

which is the average distance from $\mathbf{x}_i^{(t)}$ to the line $\overline{\mathbf{x}_i^{(t-1)}\mathbf{x}_i^{(t+1)}}$.

On the other hand, when $\lambda_3 \to \infty$, $\mathcal{L}_{tm}(\Theta)$ is equivalent to

$$\frac{\sum_{i=1}^V\sum_{t=2}^{T_i-1}\|\mathbf{x}_i^{(t)}(\Theta)-\frac{1}{2}\left(\mathbf{x}_i^{(t-1)}(\Theta)+\mathbf{x}_i^{(t+1)}(\Theta)\right)\|_2^2}{\sum_{i=1}^V(T_i-2)}, \tag{15}$$

which is the average distance from $\mathbf{x}_i^{(t)}$ to the midpoint of the line $\overline{\mathbf{x}_i^{(t-1)}\mathbf{x}_i^{(t+1)}}$. The trade-off smooth loss $\mathcal{L}_{tm}(\Theta)$ will cause the loss contour of $\mathbf{x}_i^{(t)}$ to have an ellipse with long axis $\overline{\mathbf{x}_i^{(t-1)}\mathbf{x}_i^{(t+1)}}$ (see Fig. 2(b)), which we argue is a more reasonable indicator of the smoothness of the movement trajectory $\overrightarrow{\mathbf{x}_i^{(t-1)}\mathbf{x}_i^{(t)}\mathbf{x}_i^{(t+1)}}$.

With the stabilization model and the loss function, we can estimate the model parameters using the standard optimization method Lagarias et al. (1998). The proposed stabilization algorithm is trained by various videos, which learns the statistics of different kinds of movements and thus is more robust than the traditional stabilization model Cao, Hou, and Zhou (2014). The optimization process converges within 12 hours on a conventional laptop.

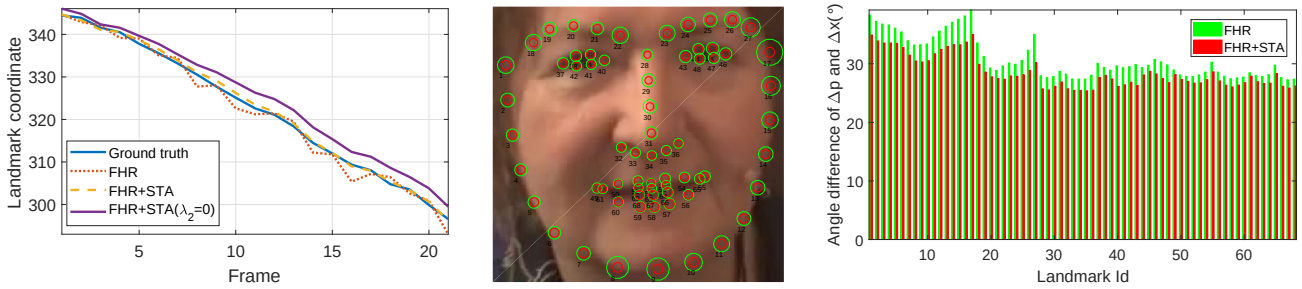
Experiments

Experimental Setup

Datasets We conduct extensive experiments on both image and video-based alignment datasets, including 300W Sagonas et al. (2013), 300-VW Shen et al. (2017) and Talking Face (TF) FGNET (2014). To test on 600 images of 300W private set, we follow Chen et al. (2017) to use use $3,148$ training images from LFPW, HELEN and AFW datasets. To

Table 1: Comparisons of NRMSE, AUC and failure rate (at 8.00% NRMSE) on 300W test set.

Methods	SDM (2013)	CFAN (2014)	CFSS (2015)	MDM (2016)	DAN (2017)	GAN (2017)	CHR (2016)	FHR
NRMSE	5.83	5.78	5.74	5.05	4.30	3.96	4.07	**3.80**
AUC	36.3	34.8	36.6	45.3	47.0	53.6	50.9	**55.9**
Failure	13.0	14.0	12.3	6.80	2.67	2.50	2.33	**1.33**



(a) (b) (c)

Figure 3: (a) The effect of stabilization loss for time delay; (b) The magnitute of the stability metric for FHR (green) and FHR+STA (red). (c) The orientation of the stability metric for FHR (green) and FHR+STA (red).

Table 2: RMSE comparisons on 5 face scales on 300W.

Face scale	Ave. inter-ocular dis.	CHR/FHR (RMSE ↓)
Very small	61.45	2.66/**2.42** (0.24 ↓)
Small	91.50	3.71/**3.35** (0.36 ↓)
Medium	127.68	5.06/**4.55** (0.51 ↓)
Large	179.55	7.59/**7.11** (0.48 ↓)
Very large	296.28	11.67/**10.67** (**1.00** ↓)

test on 300-VW, we follow Shen et al. (2017) to use 50 videos for training and the rest 64 videos for testing. Specifically, the 64 videos are divided to three categories: well-lit (Scenario1), mild unconstrained (Scenario2) and challenging (Scenario3) according to the difficulties. To test on $5,000$ frames of TF dataset, we follow Liu et al. (2017) to use the model trained by the training set of 300-VW dataset.

Training Setting Training faces are cropped using the detection bounding boxes, and scaled to 256×256 pixels. Following Chen et al. (2017), we augment the data (e.g., scaling, rotation) for more robustness to different face boxes. We use the stacked hourglass network Newell, Yang, and Deng (2016); Chen et al. (2017) as the alignment model. The network starts with a 7×7 convolutional layer with stride 2 to reduce the resolution to 128×128, followed by stacking 4 hourglass modules. For evaluation, we adopt the standard Normalized Root Mean Squared Error (NRMSE), Area-under-the-Curve (AUC) and the failure rate (at 8.00% NRMSE) to measure the accuracy, and use the consistency between the movement of landmarks and ground truth as the metric to measure the stability. We train the network with the Torch7 toolbox Collobert, Kavukcuoglu, and Farabet (2011), using the RMSprop algorithm with an initial learning rate of 2.5×10^{-4}, a mini-batch size of 6 and $\sigma = 3$. Training a fractional heatmap based hourglass model on 300W takes $\sim$7 hours on a P100 GPU.

During the stabilization training, we set $\lambda_1 = \lambda_3 = 1$ and $\lambda_2 = 10$ to make all terms in the stabilization loss (11) on the same order of magnitude. We estimate the average variance ρ of $\mathbf{z}_i^{(t)} - \mathbf{p}_i^{(t)}$ across all training videos and all landmarks, and empirically set the initial value of Γ_{noise} as $\rho\mathbb{I}$. Also, we initialize Γ_1 as a zero matrix $\mathbb{O}_{2M\times 2M}$, Γ_2 as $10\rho\mathbb{I}$, and $\gamma = \beta_1 = \beta_2 = 0.5$.

Ablation Study

Fractional vs. Conventional Heatmap Regression We first compare our fractional heatmap regression with the conventional version Newell, Yang, and Deng (2016) and other state-of-the-art models Xiong and De la Torre (2013); Zhang et al. (2014); Zhu et al. (2015); Trigeorgis et al. (2016); Kowalski, Naruniec, and Trzcinski (2017); Chen et al. (2017) on 300W test set. The training set includes $3,148$ images, and the test set contains 600 images. Note that the stacked hourglass networks for our fractional method and the conventional one are the *same*. The only difference is the transformation between the heatmaps and the coordinates, where our method preserves the fractional part. As shown in Tab. 1, our method significantly outperforms the conventional version with an improvement of 0.27% on NRMSE, and is also better than the state-of-the-art model Chen et al. (2017).

What's more, the standard NRMSE cannot fully reflect the advantage of our FHR compared to CHR, since it eliminates the scaling effects by dividing the inter-ocular distance. To further demonstrate the effects of FHR, we calculate RMSE *without* normalization. Specifically, we collect the inter-ocular distances of all 600 images and evenly divide the distances to five groups w.r.t. face scales. Tab. 2 shows that with larger scales, the gap between FHR and CHR is bigger. Especially in the largest scale, FHR achieves 1 pixel promotion for each landmark on average.

Stabilization Loss for Time Delay Here we demonstrate the effectiveness of our proposed time delay term (i.e., the

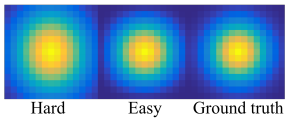


Figure 4: Averaged heatmap distributions.

Table 3: NRMSE/stability comparisons on 300-VW test set using 68 landmarks.

Methods	FHR	FHR+STA	FHR+STA($\lambda_2 = 0$)	FHR+STA($\lambda_3 = \infty$)	FHR+STA($\lambda_2 = 0, \lambda_3 = \infty$)
Scenario1	5.07/2.79	4.42/1.67	5.55/**1.64**	**4.40**/1.78	4.49/1.68
Scenario2	4.34/1.85	4.18/**1.15**	4.74/1.16	**4.16**/1.19	4.33/1.17
Scenario3	7.36/4.48	5.98/2.74	7.58/**2.57**	**5.96**/2.86	6.74/2.76

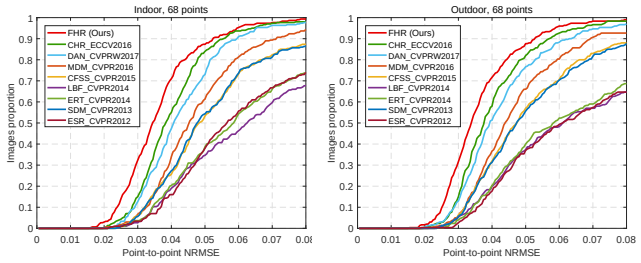


Figure 5: CED curves on 300W.

Table 4: NRMSE/stability comparisons with REDnet on 300-VW test set using 7 landmarks.

Methods	REDnet (2016)	FHR	FHR+STA
Scenario1	8.03/10.3	4.44/5.49	**3.93/3.04**
Scenario2	10.1/9.64	3.96/3.55	**3.82/3.44**
Scenario3	16.5/15.9	5.45/4.72	**4.91/4.45**

right part of Eq. 12). As in Fig. 3(a), compared to the fractional heatmap regression's output, the stabilized output is not only more smooth, but also closer to ground truth landmarks. Besides, when the time delay term is removed ($\lambda_2 = 0$), the stabilized outputs behave lagged behind the ground truth while this phenomena is largely suppressed by using our proposed loss (Eq. 12).

Stabilization Loss for Smooth Next, we evaluate the impact of each term in our loss function (11) on NRMSE and stability. We use the consistency between the cross-time movement of the landmarks and the ground truth as an indicator of stability. Specifically, we calculate $\Delta\mathbf{x}_i^{(t)} = \mathbf{x}_i^{(t)} - \mathbf{x}_i^{(t-1)}$ and $\Delta\mathbf{p}_i^{(t)} = \mathbf{p}_i^{(t)} - \mathbf{p}_i^{(t-1)}$ for every video i in the test set, and calculate the average NRMSE between $\Delta\mathbf{x}_i^{(t)}$ and $\Delta\mathbf{p}_i^{(t)}$. Assuming that the ground truth is stable, a lower value indicates higher stability.

The comparison result is shown in Tab. 3. It can be seen that dropping the time delay term ($\lambda_2 = 0$) causes a higher NRMSE, and changing the smooth loss to the one in Cao, Hou, and Zhou (2014) ($\lambda_3 = \infty$) causes a higher stability loss. Our proposed method achieves a good balance between accuracy and stability.

Comparisons with State of the Arts

We now compare our methods FHR and FHR+STA (i.e., the stabilized version) with state-of-the-art models Peng et al. (2016); Liu et al. (2017); Kowalski, Naruniec, and Trzcinski (2017); Zhang et al. (2017) on two video datasets: 300-VW and Talking Face. The comparison adopts two popular settings (i.e., 7 and 68 landmarks) used in prior works.

Comparison with 7 Landmarks First, we evaluate our method on the 300-VW Shen et al. (2017) dataset and compare with REDnet Peng et al. (2016), using the code released by the authors. Tab. 4 shows the results on the three test sets. Our proposed method achieves much better performance than REDnet in all cases. Especially in the hardest Scenario3, our method achieves a large improvement of $\sim$11%, which shows the robustness of our method to large facial variations.

Then, we evaluate our method on the Talking Face FGNET (2014) dataset compared with state-of-the-art models, such as CFAN Zhang et al. (2014), CFSS Zhu et al. (2015), I-FA Asthana et al. (2014), REDnet Peng et al. (2016) and TSTN Liu et al. (2017). Although the annotations of the TF dataset has the same landmark number as 300-VW dataset, the definitions of landmarks are different. Therefore, following the setting in Liu et al. (2017); Peng et al. (2016), we use 7 landmarks for fair comparisons. The results are shown in Tab. 5, in which the performance of Zhang et al. (2014); Zhu et al. (2015); Asthana et al. (2014) are directly cited from Peng et al. (2016); Liu et al. (2017). Since the images in TF set are collected in controlled environment and with small facial variations, all of the methods achieve relatively small errors, and our proposed method is still the best.

Comparison with 68 Landmarks Next, we evaluate our method on the 300-VW Shen et al. (2017) dataset under the setting with 68 landmarks. The comparison methods include TSCN Simonyan and Zisserman (2014), CFSS Zhu et al. (2015), TCDCN Zhang et al. (2017) and TSTN Liu et al. (2017). We cite the results of Simonyan and Zisserman (2014); Zhu et al. (2015); Zhang et al. (2017) from Liu et al. (2017), and list the performance in Tab. 6. As we can see, our proposed FHR achieves the best NRMSEs in all scenarios, and our stabilized version FHR+STA further improves the performance, especially in Scenario3.

We then illustrate the Cumulative Errors Distribution (CED) curves of FHR, CHR and some SOTA methods on 300W in Fig. 5, where the gap between FHR and CHR is competitive to those gaps in prior top-tier works (e.g., CFSS&MDM). Fig. 5 also shows that our FHR contributes more to those relatively easy samples, which makes sense since the insight of FHR is to find a more precise location near a coarse but correct coordinate, whose heatmap output may accurately model the distribution of Ground Truth (GT). To demonstrate this, we collect some predicted heatmaps from those hard and easy samples, and show their averaged heatmap distributions in Fig. 4 by fixing the centers at the same position respectively. The easy samples' heatmaps better resemble the Gaussian distribution in GT where FHR can improve the most, while hard samples resemble less and thus FHR contributes little.

Comparison on Stabilization We further compare stabilization between our method and REDnet Peng et al. (2016)

Table 5: NRMSE comparison with state-of-the-art methods on Talking Face dataset using 7 landmarks.

Methods	CFAN (2014)	CFSS (2015)	IFA (2014)	REDnet (2016)	TSTN (2017)	CHR (2016)	FHR	FHR+STA
NRMSE	3.52	2.36	3.45	3.32	2.13	2.28	**2.06**	2.14

Table 6: NRMSE comparison with state-of-the-art methods on 300-VW test set using 68 landmarks.

Methods	TSCN (2016)	CFSS (2015)	TCDCN (2016)	TSTN (2017)	CHR (2016)	FHR	FHR+STA
Scenario1	12.5	7.68	7.66	5.36	5.44	5.07	**4.42**
Scenario2	7.25	6.42	6.77	4.51	4.71	4.34	**4.18**
Scenario3	13.10	13.70	15.00	12.80	7.92	7.36	**5.98**

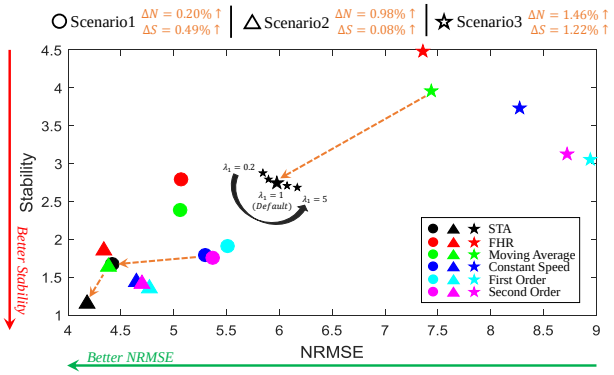


Figure 6: NRMSE/Stability comparisons with four baselines on 300-VW. *Dashed lines* indicate the difference between our STA method and the *closest* competitor, where ΔN, ΔS represent NRMSE, stability improvements, respectively. *Black arrow and pentagrams* further illustrate the parameter sensitivity of λ_1, varying among [0.2, 0.5, 1, 2, 5]), on Scenario3, which indicates λ_1 moderately adjusts our stabilization model between accuracy and stability.

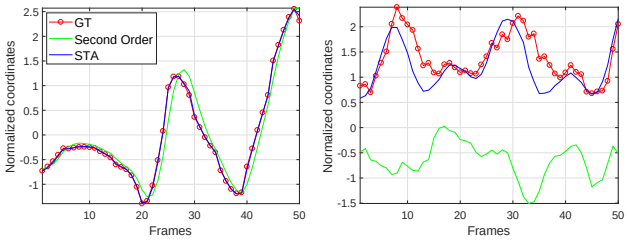


Figure 7: Stability comparisons with second order stabilization method for time delay (left) and complex movements (right) issues.

on 300-VW Shen et al. (2017) with 7 landmarks. As shown in Tab. 4, our FHR is much more stable than REDnet according to the metric mentioned in Ablation Study. To visualize stabilization improvement, we compute $E(||(\Delta\mathbf{x} - \Delta\mathbf{p})||_2^2)$ for each landmark estimated by FHR and our proposed FHR+STA, and plot them in Fig. 3(b). Fig. 3(c) plots the difference of the stability orientation $E(|<\Delta\mathbf{x} - <\Delta\mathbf{p}|)$ of two methods, where $<$ indicates the orientation of a vector. We provide videos in supplementary material, which can effectively demonstrate the stability and superiority of our method. In some continuous frames, our stabilized landmarks are *even more stable than the ground truth annotations*.

In addition, Fig. 6 shows that our stabilization model (i.e., STA) significantly outperforms other four baselines in *both of NRMSE and stability*, where all 5 methods take the same $\mathbf{z}$ from FHR as the input. Especially in the most challenging set Scenario3 where lots of complex movements exist, our method is much better than the *closest* competitor (i.e., moving average), with the improvements of NRMSE and Stability to be 1.46% and 1.22%, respectively. The reasons include: 1) moving average filter and first order smoother may cause serious time delay problem; 2) although second order and constant speed methods can handle time delay, it cannot handle multiple movement types (e.g., blinking and turning head). In contrast, our algorithm can effectively address time delay issue and multiple movement types through the Gaussian mixture setting, and hence is more *precise and stable*. Fig. 7 shows the stability comparisons with the second order stabilization method, which is chosen as a good baseline considering its stability performance. As we can see, our method significantly outperforms the second order method when handling complex movements, and also shows better ability for time delay issue which is very close to the GT.

Time Complexity Note that our fractional heatmap regression *does not* impose any additional complexity burden during training compared to the conventional heatmap regression. For inference, our method provides a closed-form solution to estimate the coordinates from the heatmaps as in Eq. (4), whose runtime is negligible. Besides, after the parameter Θ of our stabilization model is learnt, our stabilization algorithm costs $\sim$6s to process the entire 64 test videos of 300-VW, which costs 5×10^{-5}s per image, and can also be ignored.

Conclusions

In this paper, a novel Fractional Heatmap Regression (FHR) is proposed for high-resolution video-based face alignment. The main contribution in FHR is that we leverage 2D Gaussian generation prior to accurately estimate the fraction part of coordinates, which is ignored in conventional heatmap regression based methods. To further stabilize the landmarks among video frames, we propose a novel stabilization model that addresses the time-delay and non-smooth issues. Extensive experiments on popular benchmarks demonstrate our proposed method is more accurate and stable than the state of the arts. Except for the facial landmark estimation task, the proposed FHR has the potential to be plugged into any existing heatmap based system (e.g., human pose estimation task) and boost the accuracy.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Asthana, A.; Zafeiriou, S.; Cheng, S.; and Pantic, M. 2014. Incremental face alignment in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Black, M. J., and Yacoob, Y. 1995. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Bulat, A., and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Cao, C.; Hou, Q.; and Zhou, K. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. In *ACM Transactions on Graphics (SIGGRAPH)*.

Chen, Y.; Shen, C.; Wei, X.-S.; Liu, L.; and Yang, J. 2017. Adversarial learning of structure-aware fully convolutional networks for landmark localization. *arXiv:1711.00253*.

Chen*, Y.; Tai*, Y.; Liu, X.; Shen, C.; and Yang, J. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Y.; Shen, X.; and Jia, J. 2017. Makeup-go: Blind reversion of portrait edit. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Chu, X.; Ouyang, W.; Li, H.; and Wang, X. 2017. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Collobert, R.; Kavukcuoglu, K.; and Farabet, C. 2011. Torch7: A matlab-like environment for machine learning. In *NIPS Workshop*.

Fan, X.; Liu, R.; Kang, H.; Feng, Y.; and Luo, Z. 2018. Self-reinforced cascaded regression for face alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

FGNET. 2014. Talking face video. *Technique Report, online*.

Hu, T.; Qi, H.; Xu, J.; and Huang, Q. 2018. Facial landmarks detection by self-iterative regression based landmarks-attention network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Korshunova, I.; Shi, W.; Dambre, J.; and Theis, L. 2017. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Kowalski, M.; Naruniec, J.; and Trzcinski, T. 2017. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Lagarias, J. C.; Reeds, J. A.; Wright, M. H.; and Wright, P. E. 1998. Convergence properties of the neldermead simplex method in low dimensions. *SIAM J. OPTIM.*

Liu, H.; Lu, J.; Feng, J.; and Zhou, J. 2017. Two-stream transformer networks for video-based face alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, X. 2010. Video-based face model fitting using adaptive active appearance model. *Image and Vision Computing* 28(7):1162–1172.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Oh, J.; Guo, X.; Lee, H.; Lewis, R.; and Singh, S. 2015. Action-conditional video prediction using deep networks in atari games. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Peng, X.; Zhang, S.; Yang, Y.; and Metaxas, D. N. 2015. Piefa: personalized incremental and ensemble face alignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Peng, X.; Feris, R. S.; Wang, X.; and Metaxas, D. N. 2016. A recurrent encoder-decoder network for sequential face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; and Pantic, M. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Shen, J.; Zafeiriou, S.; Chrysos, G. G.; Kossaifi, J.; Tzimiropoulos, G.; and Pantic., M. 2017. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*.

Trigeorgis, G.; Snape, P.; Nicolaou, M. A.; Antonakos, E.; and Zafeiriou, S. 2016. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiong, X., and De la Torre, F. 2013. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, J.; Shan, S.; Kan, M.; and Chen, X. 2014. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhang, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2017. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(5):918–930.

Zhu, S.; Li, C.; Loy, C. C.; and Tang., X. 2015. Face alignment by coarse-tofine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.