Sequential Face Alignment via Person-Specific Modeling in the Wild

Xi Peng
Rutgers University
Piscataway, NJ 08854
xpeng.nb@cs.rutgers.edu

Junzhou Huang
University of Texas at Arlington
Arlington, TX 76019
jzhuang@uta.edu

Dimitris N. Metaxas
Rutgers University
Piscataway, NJ 08854
dnm@cs.rutgers.edu

Abstract

Sequential face alignment, in essence, deals with non-rigid deformation that changes over time. Although numerous methods have been proposed to show impressive success on still images, many of them still suffer from limited performance when it comes to sequential alignment in wild scenarios, e.g., involving large pose/expression variations and partial occlusions. The underlying reason is that they usually perform sequential alignment by independently applying models trained offline in each frame in a tracking-by-detection manner but completely ignoring temporal constraints that become available in sequence. To address this issue, we propose to exploit incremental learning for person-specific alignment. Our approach takes advantage of part-based representation and cascade regression for robust and efficient alignment on each frame. More importantly, it incrementally updates the representation subspace and simultaneously adapts the cascade regressors in parallel using a unified framework. Person-specific modeling is eventually achieved on the fly while the drifting issue is significantly alleviated by erroneous detection using both part and holistic descriptors. Extensive experiments on both controlled and in-the-wild datasets demonstrate the superior performance of our approach compared with the state of the arts in terms of fitting accuracy and efficiency.

1. Introduction

Fitting facial landmarks on sequential images plays a fundamental role in many computer vision tasks, such as face recognition [41, 39, 44], expression analysis [12, 13, 19], and facial unit detection [22, 40, 42]. It is a challenging task since the face undergoes drastic non-rigid deformations caused by extensive pose and expression variations, as well as unconstrained imaging conditions like illuminations changes and partial occlusions.

Despite the long history of research in rigid and non-rigid face tracking [4, 21], current efforts have mostly focused on face alignment on a single image [6, 30, 32, 37, 38, 43, 45, 46, 47, 48]. Generally speaking, they usually accomplish the task by achieving a direct mapping from facial appearance to landmark coordinates. The mapping could be either nonlinear regression [43] or deep neural networks [32]. They have shown great success with impressive results in standard benchmark datasets [29]. However, when it comes to sequential images, many of them still suffer from significant performance degradation especially in real-world scenarios under wild conditions [31]. They usually rely on models trained offline on still images and perform sequential alignment in a tracking-by-detection manner [7, 31, 34]. They lack the capability to capture neither the specifics of tracked subjects nor the imaging continuity in successive frames. To this end, person-specific modeling rather than generic detection is preferred.

One rational way to achieve personalized modeling is to perform joint face alignment [25, 28], which takes the advantage of the shape and appearance consistency in the sequence to minimize fitting errors of all frames at the same time. However, joint alignment is restricted to offline tasks since it usually requires all images are available before image congealing. It also suffers from low-efficiency issue which severely impedes its application on real-time or large-scale tasks [24].

To avoid these limitations, other approaches attempt to incrementally construct person-specific models instead of joint alignment. They either adapt the holistic face representation using incremental subspace learning [33], or update the cascade mapping using online regression [2]. However, how to jointly update both the representation and fitting strategy to achieve more faithful personalized models still remains an open question without investigation. Besides, former approaches often employ holistic face models for person-specific modeling, which has been proved to be inferior to part-based models in challenging conditions [30, 48]. Moreover, some of them attempt to achieve personalized modeling without correction, which may inevitably result in model drifting [24].

In this paper, we further exploit person-specific modeling for sequential face alignment to address aforementioned

issues. We first learn the part-based representation to model the facial shape and appearance respectively, where the fitting parameters are learned through a cascade of nonlinear mappings. The representation is then incrementally updated using very efficient subspace learning, and the cascade mappings are decoupled for an online update in parallel. Personalized modeling is eventually achieved on the fly while the drifting issue is significantly alleviated by the proposed erroneous detection. In summary, our work makes the following **contributions**:

- We propose a novel approach for sequential face alignment. To the best of our knowledge, this is the first time that person-specific modeling is investigated to jointly learn the representation subspace and the fitting model in a unified framework.
- The proposed part-based representation together with the cascade regression guarantees robust alignment in wild conditions. More importantly, the framework is critical to efficiently construct personalized models for real-time or large-scale applications.
- We propose to leverage both local and global descriptors for fitting evaluation. It significantly mitigates the model drifting that is common in former incremental learning based approaches.
- We provide a detailed experimental analysis of each component of our approach, as well as thorough performance comparisons with existing approaches. The results show that our approach has an average of 13.6% fitting accuracy improvement as well as affordable computational cost compared with the state of the arts.

2. Relate Work

Face alignment in a single image has attracted intensive research interest for decades. Numerous methods have been proposed with varying degrees of success. Generally speaking, most of them consist of *a representation model* and *a fitting model*.

The representation model can be either ***holistic*** or ***part*** based on different *facial deformable models* (FDMS) employed. On the one hand, the holistic model, such as *active appearance models* (AAMs) [9] and *morphable models* (MMs) [5], takes the entire face as a whole texture representation. On the other hand, the part-based model, such as *active shape models* (ASMs) [8], *constrained local models* (CLMs) [30] and *tree structure deformable part models* [48], uses a set of local image patches centered at salient landmarks to model the face appearance. These approaches can be further categorized as ***generative*** or ***discriminative*** based on the different fitting model used. The former uses an analysis-by-synthesis framework to minimize the reconstruction residual [15], the later uses either landmark classifier [30, 1] or nonlinear mapping [6, 43] for optimal fitting.

It has been proved that the part-based rather than the holistic representation is more robust to the extensive variations in unconstrained settings. For instance, Saragih *et al.* [30] proposed the *regularized landmark mean-shift* (RLMS) to maximize the joint probability of the reconstructed shape based on a set of response maps extracted around each landmark using expectation maximization. Asthana et al. [1] proposed the *discriminative response map fitting* (DRMF) to learn boosted mappings from the joint response maps to shape parameters. Cao et al. [6] combined a two-level regression to achieve *explicit shape regression* (ESR) by shape-indexed feature selection. Xiong et al. [43] proposed *supervised descent method* (SDM) to learn a sequence of descent directions using nonlinear least squares.

More recently, *deep neural networks* (DNNs) based methods have made significant progress towards systems that work in real-world scenarios [35, 36]. For example, Sun *et al.* [32] proposed to concatenate three-level convolutional neural networks to refine the fitting results from the initial estimation. Zhang *et al.* [45] employed the similar idea of coarse-to-fine framework but using *auto-encoder netowrks* instead of CNNs. Zhang *et al.* [46] showed that learning face alignment together with other correlated tasks, such as identity recognition and pose estimation, can improve the landmark detection accuracy.

The aforementioned methods have shown impressive results in standard benchmark datasets [3, 16, 14, 29]. However, they still suffer from the significant performance degradation in sequential task as they completely rely on static models trained offline. To address this limitation, efforts of constructing person-specific models are made to improve the performance of sequential face alignment.

Some of them achieve person-specific modeling via joint face alignment. A representative example was proposed in [28], which used a clean face subspace trained offline for constrained optimization to minimize fitting errors of all frames at the same time. However, these methods are usually limited to offline tasks due to the image congealing manner as well as intensive computational costs.

Others employ incremental learning to update either the representation or the fitting strategy on the fly. For instance, Sung *et al.* [33] proposed to update *incremental principle component analysis* (IPCA) to adapt the holistic AAMs to achieve personalized representation. Asthana *et al.* [2] further explored SDM in *incremental face alignment* (IFA) by flatting the cascade regressors into decoupled mappings, and simultaneously update each of them independently using incremental least squares functions. However, it can hardly achieve robust and faithful personalized models without jointly adapting the representation and fitting strategy in a unified framework. Besides, blind adaptation without effective correction would result in modeling drifting, and inevitable failure in the end.

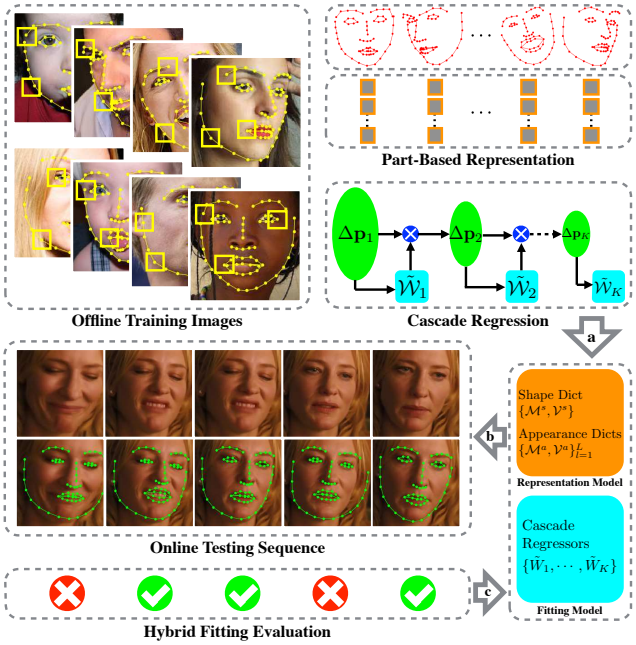

Figure 1: Overview of our approach. The part-based representation and discriminative fitting are: **(a)** trained offline, **(b)** applied on sequential testing images, and **(c)** updated incrementally on the fly.

3. Our Approach

In this paper, we propose a novel approach for sequential face alignment in the wild. We first learn the *part-based representation* to model the facial shape and appearance respectively. The *discriminative fitting* is performed by learning a cascade of regression that maps from the appearance representation to the shape parameters. Then person-specific modeling is achieved by *incremental representation update* and *fitting adaptation in parallel*. Finally, we propose *hybrid fitting evaluation* for erroneous detection to avoid modeling drifting. An overview of our approach is shown in Figure 1.

3.1. Part-Based Representation

We aim to achieve a part-based representation that is compact and easy to update for efficient person-specific modeling. A feasible solution is to learn subspace to model the shape and appearance, respectively.

The *shape representation* is learned using point distributed models [8]. Given a set of training images $\{\mathbf{I}_i\}_{i=1}^{M}$ annotated with L landmarks, we can first perform Procrustes analysis for shape normalization and then apply principle component analysis [20] to obtain the mean shape and eigenvectors $\{\mathcal{M}^s, \mathcal{V}^s\}$. A shape can be represented as:

$$\mathbf{s}(\mathbf{p}) = \mathcal{M}^s + \mathcal{V}^s\mathbf{p}, \tag{1}$$

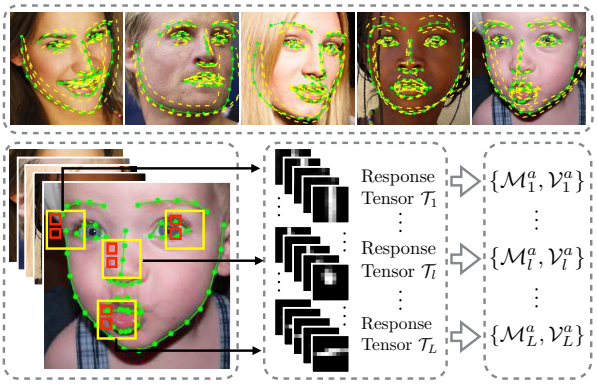

Figure 2: Top: perturbations (yellow dash) are sampled around the ground-truth shape (green dash). Bottom: response maps (yellow box) of the same landmark are arranged as tensor to learn the appearance representation.

where $\mathbf{p}$ is the shape parameters.

The *appearance representation* is learned from local response maps around landmarks. More specifically, the local response map of the l^{th} landmark in image $\mathbf{I}_i$ is:

$$\mathcal{A}_l(\mathbf{p}; \mathbf{I}_i) = \frac{1}{1 + \mathbf{exp}(a_l\Phi(\mathbf{s}(\mathbf{p}); \mathbf{I}_i) + b_l)}, \tag{2}$$

where $\{a_l, b_l\}_{l=1}^{L}$ are patch experts [30] learned using SVM by cross-validation. $\Phi(\cdot)$ is the feature vector with a possible choice of SIFT, HOG, LBP, etc.

As illustrated in Figure 2, to simulate the appearance variation and obtain more robust fittings, we sample perturbations $\{\Delta\mathbf{p}_{ij}\}$ around the ground truth $\mathbf{p}_i^*$ and arrange response maps as a tensor $\{\mathcal{A}_l(\mathbf{p}_i^* + \Delta\mathbf{p}_{ij}; \mathbf{I}_i)\}_{i=1,j=1}^{M,N}$, where i and j count images and perturbations, respectively.

It have been proved that the response maps extracted from different images lie in a low-dimensional manifold embedded in the high-dimensional feature space [2]. Therefore, similar to the shape representation, we can apply principle component analysis on the tensor to obtain a set of mean appearance and eigenvectors $\{\mathcal{M}_l^a, \mathcal{V}_l^a\}_{l=1}^{L}$ for facial parts representation. The whole face is then modeled as $\left[\mathbf{c}_1^T; \cdots; \mathbf{c}_L^T\right]^T$, where the appearance parameters $\mathbf{c}_l$ is calculated by fast projection:

$$\mathbf{c}_l = (\mathcal{V}_l^a)^{-1}(\mathcal{A}_l(\mathbf{p}; \mathbf{I}_i) - \mathcal{M}_l^a). \tag{3}$$

Now we can model an instance face using shape parameters $\mathbf{p}$ and appearance parameters $\left[\mathbf{c}_1^T; \cdots; \mathbf{c}_L^T\right]^T$. Different from former approaches [43, 47] that directly concatenate feature vectors for the high-dimensional representation, our part-based representation is highly compact and efficient. Besides, it is robust to variations even for unseen images considering the generative nature of parametric models. All these merits facilitate the incremental representation adaptation which will be explained soon in Section 3.3.

3.2. Discriminative Fitting

The goal is to learn a cascade of non-linear mappings from the part-based appearance representation $\mathbf{x}(\mathbf{p};\mathbf{I})$ to the shape parameter update $\Delta\mathbf{p}$. We refine the shape parameter $\mathbf{p}$ from an initial guess $\mathbf{p}^0$ to the ground-truth $\mathbf{p}^*$ step by step:

$$\mathbf{p}^{k+1} = \mathbf{p}^k + \mathbf{x}(\mathbf{p}^k, \mathbf{I})\mathbf{r}^k + \mathbf{b}^k, \tag{4}$$

where $\{\mathbf{r}^k, \mathbf{b}^k\}$ are regressors at step k. Let $\Delta\mathbf{p}_{ij}^k = \mathbf{p}_i^\star - \mathbf{p}_{ij}^k$, the regressors can be obtained by solving least square:

$$\arg\min_{\mathbf{r}^k,\mathbf{b}^k} \sum_{i=1}^{M}\sum_{j=1}^{N} ||\Delta\mathbf{p}_{ij}^k - \mathbf{x}(\mathbf{p}_{ij}^k;\mathbf{I}_i)\mathbf{r}^k - \mathbf{b}^k||^2, \tag{5}$$

with very efficient closed-form solution:

$$\tilde{\mathbf{r}}^k = \left[\tilde{\mathbf{x}}^T\tilde{\mathbf{x}} + \lambda\mathbf{I}\right]^{-1}\tilde{\mathbf{x}}^T\Delta\mathbf{p}^k, \tag{6}$$

where $\tilde{\mathbf{r}}^k = \left[{\mathbf{r}^k}^T{\mathbf{b}^k}^T\right]^T$, $\tilde{\mathbf{x}} = \left[\mathbf{x}(\mathbf{p}^k;\mathbf{I}_i)^T\mathbf{1}\right]^T$, and $\lambda\mathbf{I}$ is used for *Ridge Regression*.

Note that former approaches [1, 6] also employ boosted regression for discriminative fitting. However, it is difficult to extend the boosting framework for personalized modeling as updating a large number of week regressors would be extremely time-consuming. In contrast, our approach is easy to train, fast in test, and can be effectively online updated in parallel. We leave the details in Section 3.4.

3.3. Incremental Representation Update

Given the offline trained shape representation $\{\mathcal{M}^s, \mathcal{V}^s\}$ and part-based appearance representation $\{\mathcal{M}_l^a, \mathcal{V}_l^a\}_{l=1}^L$, we propose to incrementally update both the shape and appearance subspace for personalized representation in a unified framework.

Suppose the offline model is trained on m offline data $T_A = \left[\mathbf{O}_1, \cdots, \mathbf{O}_m\right]$ with mean $\mathcal{M}_A$ and eigenvectors $\mathcal{V}_A$. Given n new online observations $T_B = \left[\mathbf{O}_1, \cdots, \mathbf{O}_n\right]$ with mean $\mathcal{M}_B$, our task is equivalent to efficiently compute the SVD of the concatenation $\left[T_A\ T_B\right] = U'\Sigma'V'^T$.

It is infeasible to directly calculate the SVD, since the entire offline training data need to be stored and computed online, which inevitably results in extensive computational cost. Instead, we follow the motivation of the *sequential Karhumem-Loeve* (SKL) algorithm [17, 27] to rewrite the concatenation as:

$$\left[U\ E\right]\begin{bmatrix}\Sigma & U^T\hat{T}_B \\ \mathbf{0} & E(\hat{T}_B - UU^T\hat{T}_B)\end{bmatrix}\begin{bmatrix}V^T & \mathbf{0} \\ \mathbf{0} & I\end{bmatrix}, \tag{7}$$

where $\hat{T}_B = \left[T_B\ \sqrt{\frac{mn}{m+n}}(\mathcal{V}_B - \mathcal{V}_A)\right]$, $E = orth(\hat{T}_B - UU^T\hat{T}_B)$. Then we only need to perform SVD on the middle term instead of the entire concatenation

$$T_C = \tilde{U}\tilde{\Sigma}\tilde{V}^T,\ T_C = \begin{bmatrix}\Sigma & U^T\hat{T}_B \\ \mathbf{0} & E(\hat{T}_B - UU^T\hat{T}_B)\end{bmatrix}. \tag{8}$$

By inserting T_C back to Equation 7, we have

$$\left[T_A\ T_B\right] = \left(\left[U \quad E\right]\tilde{U}\right)\tilde{\Sigma}\left(\tilde{V}^T\begin{bmatrix}V^T & \mathbf{0} \\ \mathbf{0} & I\end{bmatrix}\right), \tag{9}$$

and we can update mean and eigenvectors instantly

$$\begin{aligned}\mathcal{M}_{AB} &= \frac{m}{m+n}\mathcal{M}_A + \frac{n}{m+n}\mathcal{M}_B, \\ U' &= \left[U\ E\right]\tilde{U},\ \Sigma' = \tilde{\Sigma}.\end{aligned} \tag{10}$$

Let d denotes the length of observation and give the fact that $m \gg n$, compared with the naive approach, the proposed subspace learning can significantly reduce space complexity from $O(d(m+n))$ to $O(dn)$ and cut down the computational complexity from $O(d(m+n)^2)$ to $O(dn^2)$. This approach fits the proposed part-based representation very well, which is critical to incrementally construct the personalized models.

3.4. Fitting Adaptation in Parallel

Once the representation model is updated, the fitting model needs to be adapted instantly to catch up the online changes. In cascade regression, computing $\tilde{\mathbf{r}}^k$ depends on $\tilde{\mathbf{r}}^{k-1}$. Directly adapting the cascade of regressors in a sequential order would lead to very low efficient performance. To address this issue, we follow [2] to decouple the dependence between successive regressors by directly sample $\mathbf{p}_k$ from a norm distribution:

$$\mathbf{p}^k \sim \mathcal{N}(\mathbf{p}^\star,\ \Lambda^k), \tag{11}$$

where Λ^k is the variation which can be learned offline. Once we flat the cascade of regressors, all mappings can be simultaneously updated in parallel.

We first compute $\tilde{\mathbf{x}}_A$, $\tilde{\mathbf{r}}_A$ and $R_A = \left[(\tilde{\mathbf{x}}_A)^T\tilde{\mathbf{r}}_A + \lambda I\right]^{-1}$ offline by Equation 6. After the representation model is updated, we sample $\Delta\mathbf{p}_B$ based on Equation 11 and recalculate the new representation $\tilde{\mathbf{x}}_B$. Then, we adapt $\tilde{\mathbf{r}}_A$ to $\tilde{\mathbf{r}}_{AB}$ using *incremental Least Square* proposed in [2]

$$\begin{aligned}P_{AB} &= \left[\tilde{\mathbf{x}}_B R_A \tilde{\mathbf{x}}_B^T + I\right]^{-1}, \\ Q_{AB} &= R_A\tilde{\mathbf{x}}_{AB}^T P_{AB}\tilde{\mathbf{x}}_B, \\ R_{AB} &= R_A - Q_{AB}R_A, \\ \tilde{\mathbf{r}}_{AB} &= \tilde{\mathbf{r}}_A - P_{AB}\tilde{\mathbf{r}}_A + R_{AB}(\tilde{\mathbf{x}}_B)^T\Delta\mathbf{p}_B.\end{aligned} \tag{12}$$

By decoupling the dependence of cascade regression, given the fact that $d \gg n$, the cost of matrix inversion in Equation 12 is significantly reduced from $O(d^3)$ to $O(n^3)$ compared with Equation 6. Besides, instead of storing the entire $\tilde{X}_A$, it only needs to maintain a small $\tilde{X}_B$, which is very efficient for online application.

3.5. Hybrid Fitting Evaluation

Blind person-specific adaption using erroneous fittings will inevitably lead to model drifting. To overcome this issue, we leverage both *part* and *holistic* constraints for misalignment detection.

The part fitting evaluation is designed based on the Bayesian inference framework. Since every landmark is supported by pixels in its neighborhood, we can evaluate the fitting quality using a Mixture of Gaussian (MoG) model:

$$t = \sum_{z \in w(\hat{s};\mathbf{I})} p(t_l = 1|z;\mathbf{I}) p(z|\hat{s}), \tag{13}$$

where $\hat{s}$ is the estimation of the landmark location, $w(s;\mathbf{I})$ is the local window centered at the landmark s. The first term is the element-wise value of local response maps defined in Equation 2, where $t_l = 1$ denotes the l-th landmark is correctly aligned and $t_l = 0$ denotes misalignment. The second term depicts the Euclidean distance between $\hat{s}$ and z, which is modeled using a Gaussian distribution:

$$p(z|\hat{s}) \sim \mathcal{N}(0, ||z - \hat{s}||^2), \tag{14}$$

where $||z - \hat{s}||_2$ is the Euclidean distance. Let $\phi_l = p(t_l = 1|z;\mathbf{I})$, we have:

$$y_l = \sum_{z \in w(\hat{s}_l, \mathcal{I})} \phi_l \mathcal{N}(0, ||z - \hat{s}_l||_2). \tag{15}$$

The landmark is correctly aligned when $y_l < \tau_l$, where τ_l is a threshold tuned offline by performing cross-validation.

The holistic evaluation is performed by applying a linear SVM classifier on the holistic facial texture. More specifically, we warp training images to the mean face according to ground-truth annotations to get shape-free textures [9]. These textures are labeled as positive samples while the negative samples are generated in the same way but with shape perturbations introduced. The fitting results need to be qualified by both part and holistic evaluation to become candidates used for incremental model adaptation.

To summarize, we illustrate the testing flow of our approach in Algorithm 1. We use two parallel thread to process face alignment and model adaptation, respectively. We also use parfor-loops to further accelerate the speed. More specifically, the update thread maintains a candidate queue $\mathcal{Q}$. The fitted shape of each frame is pushed into $\mathcal{Q}$ if it passed the hybrid evaluation. Once $\mathcal{Q}$ is full, the representation subspace and cascade mappings are updated simultaneously using all candidates in the queue. We then empty $\mathcal{Q}$ to prepare for next batch update.

4. Experiments

In this section, we first introduce datasets and settings, and then conduct experiments in two aspects: **(1)** Comparison with previous work. **(2)** Algorithm validation and discussion.

Algorithm 1 Sequential Face Alignment

Alignment Thread:

Given $\mathbf{I}$, $\{\mathcal{M}^s, \mathcal{V}^s\}$, $\{\mathcal{M}^a_l, \mathcal{V}^a_l\}_{l=1}^L$, $\{\mathbf{r}^k, \mathbf{b}^k\}_{k=1}^K$
for $k \leftarrow 1 \rightarrow K$ **do**
  Compute $\mathbf{s}$ using $\{\mathcal{M}^s, \mathcal{V}^s\}$
  parfor $l \leftarrow 1 \rightarrow L$ **do**
    Compute $\mathbf{c}_l$ using $\{\mathcal{M}^a_l, \mathcal{V}^a_l\}$ by Eqn 3
  end parfor
  Compute $\mathbf{p}^{k+1}$ using $\{\mathbf{r}^k, \mathbf{b}^k\}$ by Eqn 4
end for
Evaluate $\{\mathbf{I}, \mathbf{p}^K\}$ using hybrid fitting evaluation
if Success **then**
  Push $\{\mathbf{I}, \mathbf{p}^K\}$ into $\mathcal{Q}$
end if

Update Thread:

Initialize $\mathcal{Q}$ to empty
while 1 **do**
  if $\mathcal{Q}$ is full **then**
    Update $\{\mathcal{M}^s, \mathcal{V}^s\}$ using $\mathcal{Q}$ by Eqn 10
    parfor $l \leftarrow 1 \rightarrow L$ **do**
      Update $\{\mathcal{M}^a_l, \mathcal{V}^a_l\}_{l=1}^L$ using $\mathcal{Q}$ by Eqn 10
    end parfor
    parfor $k \leftarrow 1 \rightarrow K$ **do**
      Update $\{\mathbf{r}^k, \mathbf{b}^k\}$ using $\mathcal{Q}$ by Eqn 12
    end parfor
    Empty $\mathcal{Q}$
  end if
end while

4.1. Datasets and Settings

The public benchmark image datasets are utilized to train the offline model, and the tracking performance is evaluated in both experimental sequences and in-the-wild videos from YouTube.

MultiPIE [11] contains images of 337 subjects with different poses, expressions and illumination. We collected 1300 images with landmark annotations, which include 13 different poses from different subjects.

LFPW [3], ***Helen*** [16] and ***AFLW*** [14] are image datasets collected in wild conditions, which present challenges in different aspects. We downloaded 1035 images from LFPW, 2330 images from Helen and 4050 images from AFLW. These images together with images from MultiPIE are used to compose an ***all-in-one*** static image datasets which contains 8715 images in total.

FGNET [10] and ***ASLV*** [18] are experimentally recorded videos with head movements and expression variations. FGNET contains 5 sequences of a male subject with totally 5000 frames. while we select 10 sequences of a female subject with totally 2178 frames from ASLV.

YtbVW contains 6 videos downloaded from YouTube.

Table 1: Frome left to right: average Norm RMSE on (a) FGNET and ASLV, (b) YtbVW. The proposed method has the best performance except on Ytb06, where the serious image blurring impeded the online adaptation of our approach.

(10^{-2})	FGNET	ASLV
RLMS [30]	5.27	7.53
SDM [43]	4.41	5.77
ESR [6]	4.11	5.36
IFA [2]	4.77	6.13
RLB [26]	3.94	5.25
OURS	**3.81**	**4.21**

(10^{-2})	Ytb01	Ytb02	Ytb03	Ytb04	Ytb05	Ytb06
RLMS [30]	9.22	10.9	7.85	13.7	11.9	10.2
SDM [43]	8.54	7.62	6.22	8.60	8.11	7.69
ESR [6]	7.17	6.03	5.56	9.21	6.85	6.93
IFA [2]	7.70	7.91	6.05	9.72	8.46	8.15
RLB [26]	6.79	6.22	5.41	8.27	6.36	**6.74**
OURS	**5.03**	**5.15**	**3.68**	**6.58**	**6.11**	7.53

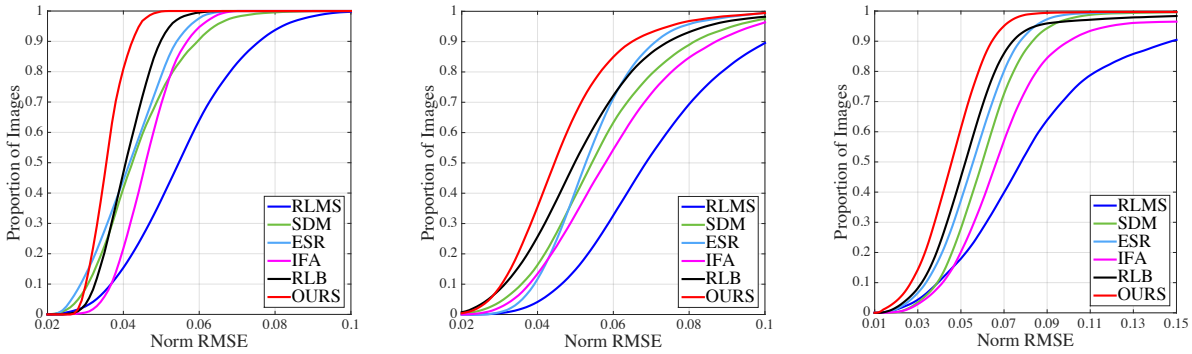

Figure 3: Frome left to right: cumulative norm RMSE distribution curves on (a) FGNET, (b) ASLV and (c) YtbVW in batch.

These videos are extremely challenging due to the unpredictable variations in pose, expression, illumination, and occlusion. We manually labeled totally 2150 frames with 49-landmarks for the purpose of qualitative analysis.

We train multi-view models for our approach based on yaw intervals: left $[-90°, -30°)$, frontal $[-30°, 30°]$ and right $(30°, 90°]$. The image registration [30] is performed by warping all images to a reference 2D shape with an interocular distance of 50 pixels to remove any 2D rigid movement [23]. HoG feature is used to best balance the performance and efficiency. We empirically set the size of the patch expert and the local support window as 11×11 and 21×21 respectively. We sample 10 perturbations for each training image with the standard deviations of ± 0.1 for scaling, $\pm 10°$ for rotation, ± 10 pixels for translation and 1.5 for non-rigid deformations. The length of $\mathcal{Q}$ is set to 5 for the batch update. Normalized Root Mean Square Error (Norm RMSE) is used to measure the tracking accuracy.

4.2. Comparison with Previous Work

Five approaches that report achieving state-of-arts performance are employed for quantitative comparisons:

- Regularized Landmark Mean-Shift (RLMS) [30].
- Supervised Descent face alignment (SDM) [43].
- Explicit Shape Regression face alignment (ESR) [6].
- Incremental Face Alignment (IFA) [2].
- Regressing Local Binary Features (RLB) [26].

For our method, we use the all-in-one dataset to train the representation and fitting models. For RLMS, ESR and RLB, we implement them in a multi-view tracking scenario and perform the training on the same dataset. For IFA and SDM, we use the pre-trained models provided by the authors to achieve the best performance.

Comparison on FGNET and ASLV we first compare different methods on constrained videos. The comparisons in Table 1a show that our approach has the average fitting errors of 0.035 and 0.042 on the two datasets, which are lowest among all the methods. The cumulative error distributions in Figure 3a and 3b prove again that the proposed method outperforms others by a substantial margin. We also notice that RLB and SDM have better performance than RLMS and IFA. A possible reason is the explicit 2D shape used in RLB and SDM is more flexible than the constrained 3D shape used in RLMS and IFA, which enables more accurate fittings when large pose and violent expression exist. However, they still cannot provide results as robust and accurate as ours, for they totally rely on offline models and lack the capability to follow the intensive online changes.

Comparison on YtbVW we then test all the methods on the extremely challenging unconstrained videos. From

Figure 4: Examples of the person-specific tracking results of the proposed method on Ytb01,02,03,04,05,06 from left to right top to down. The results are very robust under extreme violent variations in pose, expression and illumination condition.

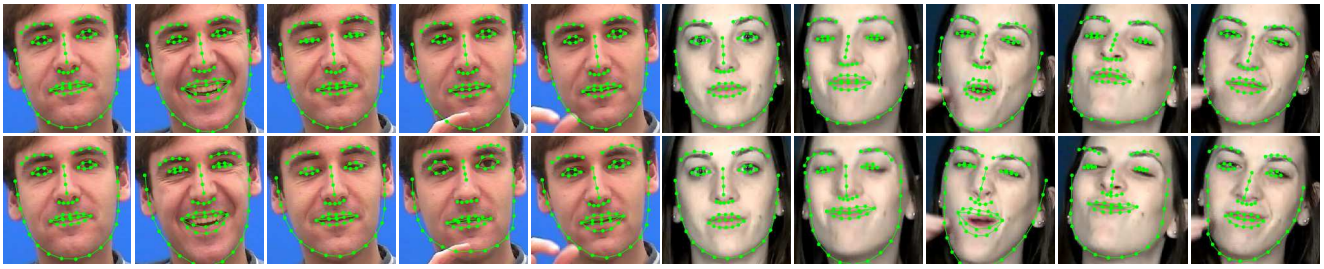

Figure 5: Examples of the tracking results on FGNET (left) and ASLV (right). The first row shows person-specific tracking with the online adaptation of both the representation and fitting models while the second row shows tracking without any model update. Notice the substantial improvement of the fitting accuracy especially around eyebrows, mouth and face contour.

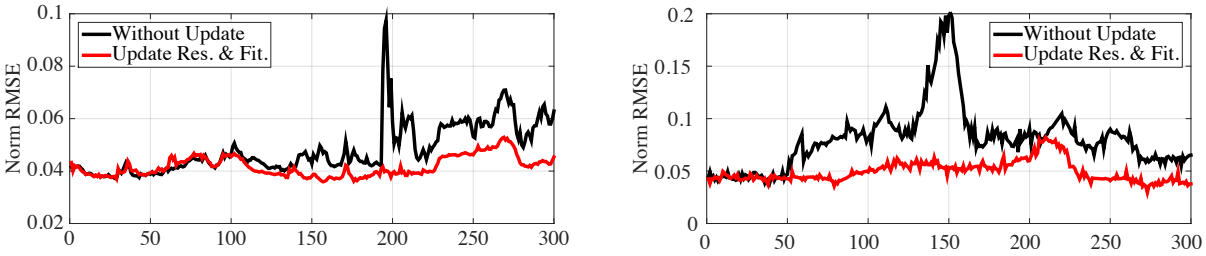

Figure 6: Frame-wise Norm RMSE with and without the model update on FGNET (left) and ASLV (right).

the experiments, we can observe: First, Table 1b shows that our approach achieves the best performance in all sequences except Ytb06 where serious image blurring occurs. The reason is the proposed hybrid evaluation could not get enough credit from the blurring frames for selective model update, which impedes the process to construct the person-specific model and deteriorates the accuracy. Second, compared with the performance on FGNET and ASLV (Figure 3a and 3b), the advantage of our approach is more significant on YtbVW shown by Figure 3c, where exists highly dynamic head movements, expression variations, illumination changes and partial occlusions. This result proves that our approach can better handle wild data than others due to person-specific modeling. Figure 4 shows some examples.

4.3. Algorithm Validation and Discussion

We verify the effectiveness of different components of the proposed approach by the following set of experiments.

Table 2: Percentages of frames with Norm RMSE less than given levels on FGNET and ASLV under four settings: (1) turn off model update; (2) update only the rep. model; (3) update only the fit. model; and (4) update both models.

Dataset	Update	< 0.04	< 0.06	< 0.08
FGNET	Off	69.3%	88.7%	95.2%
	Rep.	78.2%	93.1%	95.6%
	Fit.	84.0%	96.4%	98.5%
	Rep. & Fit.	**91.8%**	**97.0%**	**99.4%**
ASLV	Off	36.7%	57.9%	78.1%
	Rep.	62.9%	78.0%	89.4%
	Fit.	58.2%	73.3%	91.2%
	Rep. & Fit.	**68.7%**	**84.6%**	**93.5%**

With and without model update The goal of this experiment is to investigate the relation between the model adaptation and the tracking accuracy. We use 100 images from MultiPIE to train the offline model. The reason is that the less well-trained model with limited representation and fitting power, can reveal the performance variations in a better way than a well-trained one. Two clips of 300 frames with most intense changes from both FGNET and ASLV are used for the experiments. We test the trained models with two settings: (1) with both the representation and fitting models adapted, and (2) without any model adaptation. The frame-wise Norm RMSE in Figure 6 shows that, both methods have comparable accuracy at the beginning. The online method begins to outperform the offline method as model adaptation is effective. The superiority becomes more significant when intensive variations and partial occlusions exist (around frame 200 of FGNET and frame 150 of ASLV). Examples of the comparison are shown in Figure 5. The result demonstrates that our approach can achieve robust and accurate tracking even with less well-trained offline models.

Update either representation or fitting model Then we carried out experiments on the full datasets of FGNET and ASLV with four online settings: (1) turn off model update; (2) update only the representation model; (3) update only the fitting model; and (4) update both models. The average tracking errors are recorded in Table 2. It confirms the validity of the proposed method from two aspects. First, updating both models has the best accuracy compared with the rest settings, which reclaims the effectiveness of the proposed model adaptation in person-specific tracking. Second, there is a substantial performance gain in accuracy between updating both models and updating either one while keep the other one fixed, which proves the necessity to simultaneously update both models in a uniform framework.

Local and global fitting evaluation We validate the proposed hybrid fitting evaluation on MultiPIE, LFPW, Helen and the all-in-one dataset. As illustrated in Figure 2, we label images with gound-truth shapes as positive and images with perturbed shaped as negative. 10-fold cross-validations are performed on each dataset for quantitative analysis. The percentage of images that are correctly classified as correct (ground-truths) or erroneous (perturbations) by local and hybrid evaluators are reported in Table 3. It shows that the local fitting evaluator performed well on the experimental dataset (MultiPIE), but deteriorated drastically ($> 10\%$) on wild datasets (LFPW and Helen). However, the hybrid evaluation can significantly boost the accuracy especially in unconstrained conditions, which highlights its capability to distinguish well fittings from outliers to alleviate drifting.

Table 3: Average fitting evaluation accuracy.

Evaluator	MultiPIE	LFPW	Helen	All-in-One
Local	87.9%	76.3%	71.6%	75.3%
Hybrid	**93.7%**	**85.5%**	**79.0%**	**84.7%**

Comparison of running time We compare the speed of different methods on YtbVW in Table 4. RLMS [30] and IFA [2] are used for the comparison since they are implemented in Matlab, the same as ours. It shows that when turning the model adaptation off, the proposed method and IFA, both of which use regression-based fitting models, are 3 to 4 times faster than RLMS which uses optimization based fitting model. Besides, our method is faster than IFA because the cascade of linear regression framework is more efficient than the boosted regressors. Moreover, even equipped with the online model adaptation and the hybrid fitting evaluation, our approach still has a comparable speed with the very efficient RLMS (142 ms to 116 ms). It is reasonable to expect our method to achieve real-time performance with better implementation other than Matlab.

Table 4: Average running time per frame in ms on YtbVW.

RLMS [30]	IFA [2]	OURS (off)	OURS (on)
116 ± 22	43.9 ± 14	32.4 ± 11	142 ± 27

5. Conclusion

In this paper, we exploit incremental learning for sequential face alignment in wild conditions. Our approach achieves person-specific modeling by incrementally updating the representation subspace and simultaneously adapting the cascade mappings in parallel. Both part and holistic descriptors are used for erroneous detection, which significantly alleviate the drifting issue. Experimental results on multiple datasets have validated our approach in different aspects and demonstrated its superior performance compared with the state of the arts.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451, 2013. 2, 4

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 1, 2, 3, 4, 6, 8

[3] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 35, pages 2930–2940, December 2013. 2, 5

[4] M. Black and Y. Yacoob. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 374–381, 1995. 1

[5] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(9):1063–1074, Sep 2003. 2

[6] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 1, 2, 4, 6

[7] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, pages 954–962, 2015. 1

[8] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38 – 59, 1995. 2, 3

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, June 2001. 2, 5

[10] FGNet. Talking face video, 2004. 5

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Computing (IVC)*, 28(5):807–813, May 2010. 5

[12] Y. Guo, G. Zhao, and M. Pietikinen. Dynamic facial expression recognition with atlas construction and sparse representation. *IEEE Transactions on Image Processing (TIP)*, 25(5):1977–1992, 2016. 1

[13] Q. Hu, X. Peng, P. Yang, F. Yang, and D. N. Metaxas. Robust multi-pose facial expression recognition. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 1782–1787. IEEE, 2014. 1

[14] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 2, 5

[15] F. D. la Torre and M. H. Nguyen. Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 2

[16] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision (ECCV)*, pages 679–692, 2012. 2, 5

[17] A. Levey and M. Lindenbaum. Sequential karhunen-loeve basis extraction and its application to images. *IEEE Transactions on Image Processing (TIP)*, 9(8):1371–1374, 2000. 4

[18] C. Neidle, J. Liu, B. Liu, X. Peng, C. Vogler, and D. Metaxas. Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in american sign language (asl). *LREC Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, 2014. 5

[19] M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative {RVM} regression for dimensional and continuous emotion prediction. *Image and Vision Computing (IVC)*, 30(3):186 – 196, 2012. Best of Automatic Face and Gesture Recognition 2011. 1

[20] A. Papaioannou and S. Zafeiriou. Principal component analysis with complex kernel: The widely linear model. *IEEE Transactions on Neural Networks and Learning Systems*, 2014. 3

[21] I. Patras and M. Pantic. Particle filtering with factorized likelihoodsfor tracking facial features. In *The IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 97–102, 2004. 1

[22] X. Peng, J. Huang, Q. Hu, S. Zhang, A. Elgammal, and D. Metaxas. From circle to 3-sphere: Head pose estimation by instance parameterization. *Computer Vision and Image Understanding (CVIU)*, 136:92–102, 2015. 1

[23] X. Peng, J. Huang, Q. Hu, S. Zhang, and D. N. Metaxas. Three-dimensional head pose estimation in-the-wild. In *The IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6. IEEE, 2015. 6

[24] X. Peng, S. Zhang, Y. Yang, and D. N. Metaxas. Piefa: Personalized incremental and ensemble face alignment. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[25] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma. RASL: Robust Alignment by Sparse and Low-rank Decomposition for Linearly Correlated Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 2010. 1

[26] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 6

[27] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision (IJCV)*, 77(1-3):125–141, May 2008. 4

[28] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Raps: Robust and efficient automatic construction of person-specific deformable models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1789–1796, June 2014. 1, 2

[29] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2013. 1, 2

[30] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision (IJCV)*, 91(2):200–215, Jan. 2011. 1, 2, 3, 6, 8

[31] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015. 1

[32] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3476–3483, 2013. 1, 2

[33] J. Sung and D. Kim. Adaptive active appearance model with incremental learning. *Pattern Recognition Letters (PRL)*, 30(4):359 – 367, 2009. 1, 2

[34] M. Tang and X. Peng. Robust tracking with discriminative ranking lists. *IEEE Transactions on Image Processing (TIP)*, 21(7):3273–3281, 2012. 1

[35] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. W. Schuller. A Deep Semi-NMF Model for Learning Hidden Representations. In *International Conference on Machine Learning (ICML)*, 2014. 2

[36] G. Trigeorgis, M. Nicolaou, S. Zafeiriou, and B. W. Schuller. Towards Deep Multimodal Alignment. In *NIPS Multimodal Machine Learning Workshop*, 2015. 2

[37] G. Trigeorgis, P. Snape, M. A. Nicolaou, E. Antonakos, and S. Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *IEEE International Conference on Computer Vision Pattern Recognition (CVPR)*, June 2016. 1

[38] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3659–3667, 2015. 1

[39] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Principal component analysis of image gradient orientations for face recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 553–558, 2011. 1

[40] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 149–, 2006. 1

[41] X. Wang, G. Guo, M. Rohith, and C. Kambhamettu. Leveraging geometry and appearance cues for recognizing family photos. In *The IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2015. 1

[42] Y. Wu and Q. Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[43] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 1, 2, 3, 6

[44] S. Zafeiriou, I. Kotsia, and M. Panti. Unconstrained face recognition. *Face Recognition in Adverse Conditions*, 2, 2014. 1

[45] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *European Conference on Computer Vision (ECCV)*, pages 1–16, 2014. 1, 2

[46] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*, pages 94–108, 2014. 1, 2

[47] S. Zhu, C. Li, C. C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4998–5006, 2015. 1, 3

[48] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2