Joint Multi-view Face Alignment in the Wild

Jiankang Deng, *Student Member, IEEE,* George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou, *Member, IEEE*

***Abstract*—The de facto algorithm for facial landmark estimation involves running a face detector with a subsequent deformable model fitting on the bounding box. This encompasses two basic problems: i) the detection and deformable fitting steps are performed independently, while the detector might not provide best-suited initialization for the fitting step, ii) the face appearance varies hugely across different poses, which makes the deformable face fitting very challenging and thus distinct models have to be used (e.g., one for profile and one for frontal faces). In this work, we propose the first, to the best of our knowledge, joint multi-view convolutional network to handle large pose variations across faces in-the-wild, and elegantly bridge face detection and facial landmark localization tasks. Existing joint face detection and landmark localization methods focus only on a very small set of landmarks. By contrast, our method can detect and align a large number of landmarks for semi-frontal (68 landmarks) and profile (39 landmarks) faces. We evaluate our model on a plethora of datasets including standard static image datasets such as IBUG, 300W, COFW, and the latest Menpo Benchmark for both semi-frontal and profile faces. Significant improvement over state-of-the-art methods on deformable face tracking is witnessed on 300VW benchmark. We also demonstrate state-of-the-art results for face detection on FDDB and MALF datasets.**

***Index Terms*—Joint multi-view face alignment, Cascade face detection**

I. Introduction

OBJECT detection in computer vision has seen a huge amount of attention in recent years [1], [2], [3]. The advances in deep learning and the use of more elaborate models, such as Inception [4] and ResNet [5], have allowed for reliable and fine-scale non-rigid object detection even in challenging scenarios. Out of all the objects probably the most studied one is the human face. Face detection, although having embedded in our everyday lives through the use of digital cameras and social media, is still an extremely challenging problem as shown by the recent survey [6].

Human face in images captured in unconstrained conditions (also referred to as "in-the-wild") is a challenging object, since facial appearance can change dramatically due to extreme pose, defocus, low resolution and occlusion. Face detection "in-the-wild" is still regarded as a challenging task. That is, considerable effort was needed in order to appropriately customize a generic object methodology, e.g. Deformable Part-Based Models [7] and Deep Convolutional Neural Networks (DCNNs) [1], in order to devise pipelines that achieve very good performance in face detection [8], [7], [9].

Fig. 1: Facial landmark response maps generated by Multi-view Hourglass Model (MHM). The profile and frontal faces are trained jointly, and the model is robust under large pose variations.

Specifically, when dealing with human face we are also interested in detailed face alignment, that is, localizing a collection of facial landmarks on face images. This step plays an important role in many face analysis task, such as face recognition [10], [11], [12], [13], [14], expression recognition [15], [16], and face animation [17]. Due to the importance of the problem, a large number of facial landmark localization methods have been proposed in the past two decades [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], and the previous works can be categorized as parametric fitting based [18], [19], [20], [29] and non-parametric regression based [21], [22], [23], [24], [25], [26], [28]. The former aims at minimizing the discrepancy between the model appearance and the input image. The latter extracts features from the image and directly regresses to the ground truth landmarks. With the increasing number of training data [30], the performance of regression-based methods is generally better than that of parametric fitting based methods.

Recently, it was shown that it is advantageous to perform jointly face detection and facial landmark localization [31], [9]. Nevertheless, due to the high cost of facial landmark localization step, only few landmarks were detected [9]. Furthermore, in [9] the method made use of extra 400K facial images from the web which are not publicly available. To avoid this, we propose a coarse-to-fine joint multi-view landmark localization architecture. In the coarse step, few landmarks are localized, while in the fine stage, we detect a large number of landmarks (e.g., 68/39). In our methodology, for reproducibility, we made use of publicly available data only.

Face alignment and tracking across medium poses, where all the landmarks are visible, has been well addressed [23],

Manuscript received on August 18, 2017; revised on June 10, 2018; accepted on February 11, 2019. This work was partially funded by EPSRC project EP/N007743/1 (FACER2VM). The associate editors coordinating the review of this manuscript and approving it for publication were Prof. Chunming Li and Prof. Xiaochun Cao.

J. Deng, G. Trigeorgis, Y. Zhou and S. Zafeiriou are with the Intelligent Behaviour Understanding Group (IBUG), the Department of Computing, Imperial College London, London SW7 2AZ, UK. J. Deng and S. Zafeiriou are also with Facesoft.io.

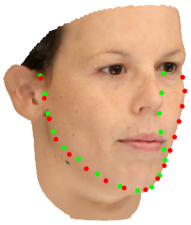


Fig. 2: Inconsistent landmark annotation on face contour between 2D and 3D views. Red annotation is from 2D view, and green annotation is from 3D view.

[24], [25]. However, face alignment across large poses is still a challenging problem with limited attention. There are two main challenges: Firstly, there is a controversy on landmark definition, from 2D view or 3D view? As is shown in Figure 2, facial landmarks are always located at the visible face boundary in the 2D annotation. Faces which exhibit large facial poses are extremely challenging to annotate, because the landmarks on the invisible face side stack together. Since the invisible face contour needs to be always guessed to be consistent with 3D face models, labelling the self-occluded 3D landmarks is also ambiguous for annotators. Secondly, since occlusions can occur on both frontal and profile face images, designing a single shape constraint is hard for large pose face alignment. As view variation is continuous, view-specific modelling [32], [33] inevitably brings the problem of view classification and increases the computation cost.

In this work we present the first, to the best of our knowledge, method for deformable face modelling which jointly detects the face and localizes a large amount of landmarks.

1) We employ a joint face detection and alignment strategy where a face detector is first applied to find a coarse estimate of the facial shape using a small subset of landmarks. After removing the similarity transformation, a refining step is performed to estimate the dense facial shape of each person. Based on accurate face alignment, the face/non-face classification problem is simplified because the aligned shape finds corresponding parts between faces and makes them directly comparable.
2) We formulate a novel Multi-view Hourglass Model (MHM) which tries to jointly estimate both semi-frontal and profile facial landmarks. Different from the other methods which employ distinct models, we try to capitalize on the correspondences between the profile and frontal facial shapes.
3) We demonstrate huge improvement over the state-of-the-art results in the latest benchmarks for deformable face fitting such as IBUG, 300W, COFW and the latest Menpo Benchmark. We demonstrate state-of-the-art results for the deformable face tracking on the 300VW benchmark and face detection on FDDB and MALF.

II. Related work

To better understand the problem of deformable face fitting, we review some related works.

Besides traditional models (such as AAMs [19], CLMs [20] and regression models [23], [34], [35], [36], [37], [38]), recently DCNNs has been employed in face alignment [26], [39], [40]. The resolution loss within the pooling step in DCNN was compensated by the image enlargement in a global to local way. Zhang et al. [41] adopted the similar coarse-to-fine framework with auto-encoder networks. Ranjan et al. [42] combined outputs of multi-resolution convolutional layers to predict the landmark locations. After the presentation of the fully-convolutional network (FCN) [40], which takes input of arbitrary size, produces a correspondingly-sized dense label map and shows convincing results for semantic image segmentation, direct landmark coordinated prediction changed to the landmark response map prediction. Lai et al. [43], Xiao et al. [44] and Bulat et al. [45] employed the convolutional and de-convolutional network to generate the response map for each facial landmark, and added a refinement step by utilizing a network that performs regression. In the area of articulated human pose estimation, Alejandro et al. [46] proposed a novel stacked hourglass model, which repeated bottom-up and top-down processing in conjunction with intermediate supervision and obtained state-of-the-art result. Bulat et al. [47] further explored binarized Hourglass-like convolutional network for face alignment with limited resources.

Despite the large volume of work on semi-frontal face alignment, literature on the large-pose scenario is rather limited. This is attributed to the fact that large-pose face alignment is a very challenging task, until now there are not enough annotated facial images in arbitrary poses (especially with a large number of landmarks). A step towards this direction is the data presented in the new facial landmark competition [48]. The most common method in large-pose image alignment is the multi-view AAMs framework [32], which uses different landmark configurations for different views. However, since each view has to be tested, the computation cost of multi-view method is always high. In [8], [49] the methods utilized the DPM framework to combine face detection and alignment, and the best view fitting was selected by the highest possibility. Since non-frontal faces are one type of occlusions, Wu et al. [50] proposed a unified robust cascade regression framework that can handle both images with severe occlusion and images with large head poses by iteratively predicting the landmark visible status and the landmark locations.

To solve the problem of large pose face alignment, 3D face fitting methodologies have been considered [51], [52], [27], which aims to fit a 3D morphable model (3DMM) [53] to a 2D image. [51] aligned faces of arbitrary poses with the assist of a sparse 3D point distribution model. The model parameter and projection matrix are estimated by the cascaded linear or nonlinear regressors. [52] extended [51] by fitting a dense 3D morphable model, employing the CNN regressor with 3D-enabled features, and estimating contour landmarks. [27] fitted a dense 3D face model to the image via CNN and synthesized large-scale training samples in profile views to solve the problem of data labelling. 3D face alignment methods model the 3D face shape with a linear subspace and achieve fitting by minimizing the difference between image and model appearance. Although 3D alignment methods can cover arbitrary poses, the accuracy of alignment is bounded by the linear parametric 3D model, and the invisible landmarks

are predicted after the visible appearance are fitted. In this paper, we focus on non-parametric visible landmark localization.

Finally, we assess our methodology for facial landmark tracking. The current state-of-the-art around face deformable tracking boils down to a pipeline which combines a generic face detection algorithm with a facial landmark localization method [54]. Variants of this pipeline with different detectors or deformable models appear in the related paper [54]. The pipeline is quite robust since the probability of drifting is reduced due to the application of the face detector at each frame. We demonstrate that by applying the proposed methodology, large improvements over the state-of-the-art can be achieved.

III. Our method

In this paper, multi-view response maps are used to bridge face detection and face alignment. In Sec. III-A, the five facial landmarks from face detector are used to normalize the face region, decrease the variance of regression target for face alignment, thus improve the accuracy of facial landmark localization. In Sec. III-B, we propose the multi-view training for face alignment under large pose variations (Figure 3). Based on the multi-view response maps, we design a light-weighted classifier to remove high score false positives for face detector. In our method, joint face detection and alignment is proved to be better than isolated design. Precise face detector can provide stable and accurate initialization for the following face alignment. Meanwhile, accurate face alignment finds corresponding parts between faces, makes them directly comparable, and simplifies the face/non-face classification problem.

A. Face Region Normalization

The training of our face detection module follows the exact design of three cascade network and sampling strategies in [55]. In that, we minimize an objective function with the multi-task loss. For each face box i, its loss function is defined as:

$$L = L_1(p_i, p_i^*) + \lambda_1 p_i^* L_2(t_i, t_i^*) + \lambda_2 p_i^* L_3(l_i, l_i^*), \quad (1)$$

where p_i is the probability of box i being a face; p_i^* is a binary indicator (1 for positive and 0 for negative examples); the classification loss L_1 is the softmax loss of two classes (face / non-face); $t_i = \{t_x, t_y, t_w, t_h\}_i$ and $t_i^* = \{t_x^*, t_y^*, t_w^*, t_h^*\}_i$ represent the coordinates of the predicted box and ground truth box correspondingly. $l_i = \{l_{x_1}, l_{y_1}, \cdots, l_{x_5}, l_{y_5}\}_i$ and $l_i^* = \{l_{x_1}^*, l_{y_1}^*, \cdots, l_{x_5}^*, l_{y_5}^*\}_i$ represent the predicted and ground truth five facial landmarks. The box and the landmark regression targets are normalized by the face size of the ground truth. We use $L_2(t_i, t_i^*) = R(t_i - t_i^*)$ and $L_3(l_i, l_i^*) = R v_i^*(l_i - l_i^*)$ for the box and landmark regression loss, respectively, where R is the robust loss function (smooth-L_1) defined in [2]. In Figure 4, we give the network structure of the cascade networks with multi-task loss.

One core idea of our method is to incorporate a spatial transformation [56] which is responsible for warping the original image into a canonical representation such that the later alignment task is simplified. Recent work (e.g., [57]) has explored this idea on face recognition and witnessed an improvement on the performance. In Figure 5, the five facial landmark localization network (Figure 4(c)) as the spatial transform layer is trained to map the original image to the parameters of a warping function (e.g., a similarity transform), such that the subsequent alignment network is evaluated on a translation, rotation and scale invariant face image, therefore, potentially reducing the trainable parameters as well as the difficulty in learning large pose variations. Since different training data are used in face region normalization (CelebA [58] and AFLW [59]) and multi-view alignment (300W [30] and Menpo Benchmark [48]), end-to-end training of these two networks with intermediate supervision on the face region normalization step is equal to step-wise training. In this paper, we employ step-wise cascade structure, and the face region normalization step benefits from larger training data as annotation of the five facial landmarks is much easier than dense annotation.

B. Multi-view Hourglass Model

Hourglass [46] is designed based on Residual blocks [5], [60], which can be represented as follows:

$$x_{n+1} = H(x_n) + F(x_n, W_n), \quad (2)$$

where x_n and x_{n+1} are the input and output of the n-th unit, and F is the stacked convolution, batch normalization, and ReLU non-linearity. Hourglass is a symmetric top-down and bottom-up full convolutional network. The original signals are branched out before each down-sampling step and combined together before each up-sampling step to keep the resolution information. n scale Hourglass is able to extract features from the original scale to $1/2^n$ scale and there is no resolution loss in the whole network. The increasing depth of network design helps to increase contextual region, which incorporates global shape inference and increases robustness when local observation is blurred.

Based on the Hourglass model [46], we formulate the Multi-view Hourglass Model (MHM) which tries to jointly estimate both semi-frontal (68 landmarks) and profile (39 landmarks) face shapes. Unlike other methods which employ distinct models, we try to capitalize on the correspondences between the profile and frontal facial shapes. As shown in Figure 6, for each landmark on the profile face, the nearest landmark on the frontal face is regarded as its corresponding landmark in the union set, thus we can form the union landmark set with 68 landmarks (U-68). Considering that the landmark definition varies in frontal and profile data, we also enlarge the union set to 86 landmarks (U-86) by dissimulating two landmarks from eyebrow and seven landmarks from the lower part of face contour for profile annotation. During the training, we use the view status to select the corresponding response maps for the loss computation.

$$L = \frac{1}{N} \sum_{n=1}^{N} (v_n^* \sum_{ij} \|m_n(i,j) - m_n^*(i,j)\|_2^2), \quad (3)$$

where $m_n(i,j)$ and $m_n^*(i,j)$ represent the estimated and the ground truth response maps at pixel location (i,j) for the n-th

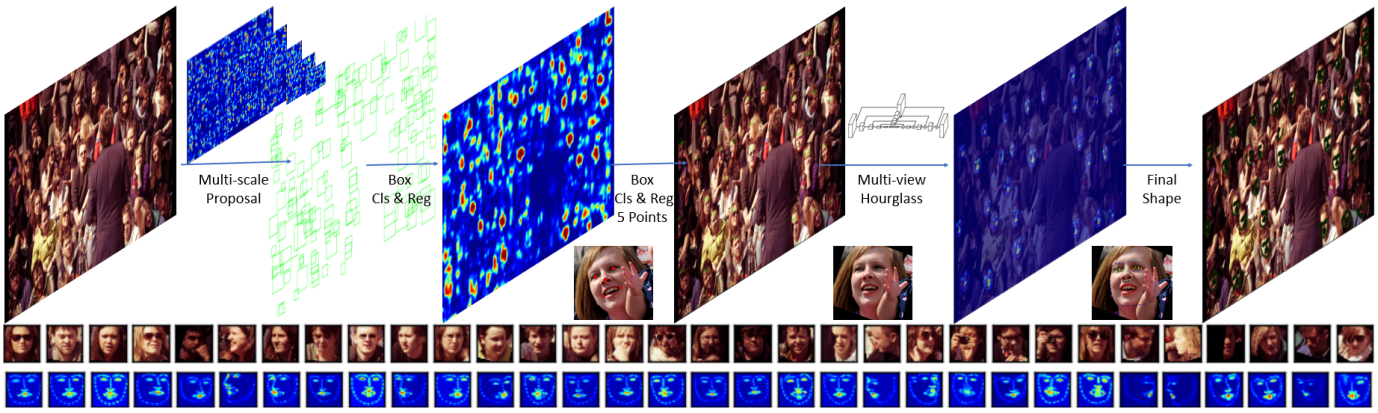
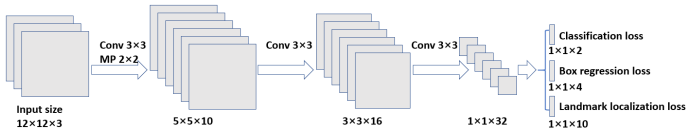
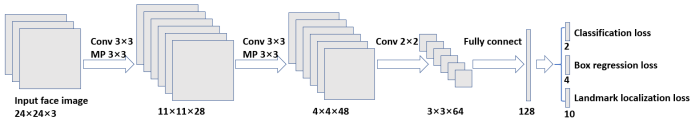


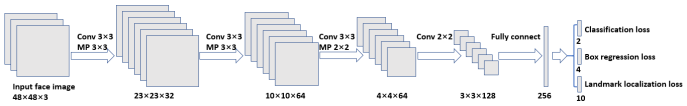
Fig. 3: Proposed coarse-to-fine joint multi-view face alignment. Face regions are generated by the multi-scale proposal, then classified and regressed by the following network. Five facial landmarks are predicted to remove the similarity transformation of each face region. Multi-view Hourglass Model is trained to predict the response map for each landmark. The second and third rows show the normalized face regions and the corresponding response maps, respectively.



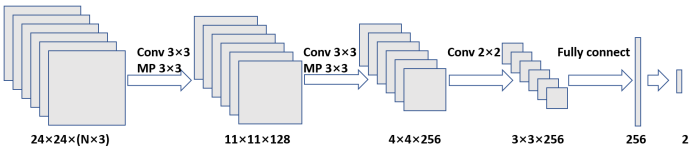
(a) Face Region Proposal



(b) 2nd Face Region Classifier



(c) 3rd Face Region Classifier



(d) 4th Face Region Classifier

Fig. 4: The architectures of face region proposal and classification networks. "Conv" means convolution, "MP" means max pooling, and N is the number of landmarks. The step size in convolution and pooling is 1 and 2 respectively.

Fig. 5: Face Region Normalization. The five facial landmark localization network acts as the spatial transform layer and the subsequent alignment network is evaluated on a translation, rotation and scale invariant face image, therefore, potentially reducing the trainable parameters as well as the difficulty in learning large pose variations.

landmark correspondingly, and $v_n \in \{0, 1\}$ is the indicator to select the corresponding response map to calculate the final loss. We can see from Figure 6 that the semi-frontal response maps (second and forth examples in third row) benefit from the joint multi-view training, and the proposed method is robust and stable in a range of poses.

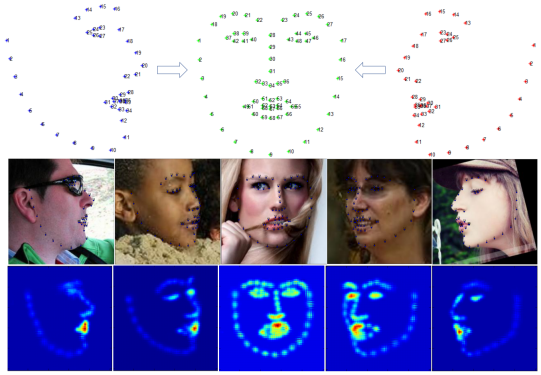


Fig. 6: Multi-view Hourglass Model. First row: facial landmark configuration for frontal (68 landmarks) and profile (39 landmarks) faces [48]. We define a union landmark set with 68 landmarks for frontal and profile shape. For each landmark on the profile face, the nearest landmark on the frontal face is selected as the same definition in the union set. Third row: landmark response maps for all view faces. The response maps for semi-frontal faces (2nd and 4th) benefit from the joint multi-view training.

Based on the multi-view response maps, we extract shape-indexed patch (24×24) around each predicted landmark from the down-sampled face image (128×128). As shown in Figure 4(d), a small classification network is trained to classify face / non-face. This classifier is not only used to remove high score false positives for face detection, but also can be employed as a failure checker for deformable face tracking.

IV. Experiments

A. Experiment Setting

1) Training Data: **Face Detection Model:** The face detection module before the multi-view face alignment step follows the cascaded network design and sampling strategies as in [55]. We crop positive faces (IoU > 0.6), negative faces (IoU < 0.3) and part faces (IoU $\in (0.4, 0.65)$) from Wider Face [61] training set. To guarantee a high accuracy in predicting five facial landmarks, we employ additional labelled faces from the AFLW [59] dataset besides labelled faces from CelebA [58].

For the additional classifier after the multi-view alignment step, the positive (IoU $>$ 0.5) and negative samples (IoU $<$ 0.3) are generated from the previous cascaded face detector. **Multi-view Hourglass Model:** We train the face alignment module MHM on the 300W database [30], and the Menpo Benchmark database [48], [62], where faces are manually annotated with either 68 (semi-frontal face) or 39 (profile face) landmarks. The training set of the 300W database (we denote as *300W-68*) consists of the LFPW trainset [63], the Helen trainset [64] and the AFW dataset [8], hence, a total of 3148 images are available. The Menpo Benchmark database [48] (denoted as *Menpo-39-68*) consists of 5658 semi-frontal face images and 1906 profile face images. In this paper, we defined two training sets (*300W-68-Menpo-39* and *300W-68-Menpo-39-68*) for different evaluation purposes. *300W-68-Menpo-39* includes the *300W-68* data and the profile faces of *Menpo-39*, while *300W-68-Menpo-39-68* groups all the available training images in *300W-68* and *Menpo-39-68*.

2) Testing data: **Face detection:** We evaluate the performance of our face detection module in two challenging datasets, FDDB and MALF. FDDB consists of 5171 faces in 2845 images from the unconstrained environment. MALF is a fine-grained evaluation dataset, in total, there are 5250 images with 11931 annotated faces. The "hard" subset contains faces (larger than 60×60) with huge variations in pose, expression, or occlusion. In particular, we give detailed pose-specific evaluations on MALF. **Face alignment in images & videos:** Evaluations of single face alignment and face tracking are performed in several *in-the-wild* databases. For alignment in static image, we test on *IBUG* dataset, *300W* testset [30], *COFW [65], [66]*, and *Menpo-test [48]*. All these databases are collected under fully unconstrained conditions and exhibit large variations in pose, expression, illumination, etc. In particular, *Menpo-test [48]* collects faces of all different poses, which are categorized into 5535 semi-frontal faces and 1946 profile faces based on [48]. For face tracking experiment, *300VW* is the only publicly available *in-the-wild* benchmark. It consists of 114 videos (about 218k frames in total), captured in the wild with large pose variations, severe occlusions and extreme illuminations.

3) Evaluation Metric: Given the ground truth, the landmark localization performance can be evaluated by Normalized Mean Error (NME), and the normalization is typically carried out with respect to face size.

$$err = \frac{1}{M}\sum_{i=1}^{M}\frac{\frac{1}{N}\sum_{j=1}^{N}|p_{i,j}-g_{i,j}|_2^2}{d_i}, \tag{4}$$

where *M* is the number of images in the test set, *N* is the number of landmarks, *p* is the prediction, *g* is the ground truth, and *d* is the normalize distance. According to the protocol of difference facial alignment benchmarks, various normalize distances are used in this paper, such as eye centre distance [24], outer eye corner distance [30] and diagonal distance of ground truth bounding box [54]. The permissible error (localization threshold) is taken as a percentage of the normalize distance.

4) Training of Multi-view Hourglass Model: The training of the proposed method follows a similar design as in the Hourglass Model [46]. Before the training, several pre-processing steps are undertaken. We firstly remove scale, rotation and translation differences by five facial landmarks among the training face images (referred as the spatial transformer step), then crop and resize the face regions to 256×256. We augment the data with rotation (+/- 30 degrees), scaling (0.75-1.25), and translation (+/- 20 pixels) that would help simulate the variations from face detector and five landmark localization. The full network starts with a 7×7 convolutional layer with stride 2, followed by a residual module and a round of max pooling to bring the resolution down from 256 to 64, as it could save GPU memory while preserving alignment accuracy. The network is trained using Tensorflow [67] with an initial learning rate of 1e-4, batch size of 12, and learning steps of 100k. The Mean Squared Error (MSE) loss is applied to compare the predicted heatmaps to the ground-truth heatmaps. Each training step takes 1.2s on one NVIDIA GTX Titan X (Pascal) GPU card. During testing, face regions are cropped and resized to 256×256, and it takes 12.21ms to generate the response maps.

B. Ablation Study

We consider different training strategies and validate these setting on the challenging IBUG dataset in Table I. (1) Hourglass Model (HM) trained on *300W-68*. (2) HM trained on *300W-68*, with spatial transformer step based on five facial landmarks. (3) HM trained on *300W-68* with simulated response maps from the output five landmarks. The input channel increases from 3 to 8, and this Hourglass model is trained with the spatial facial clue from face detector. The result of Method 3 is worse than that of Method 2, which indicates that the spatial transformer step for each face region is better than the spatial indication. (4) Multi-view Hourglass Model (MHM) trained on *300W-68-Menpo-39* with 68 union landmarks. (5) MHM trained on *300W-68-Menpo-39* with 86 union landmarks. (6) MHM trained on *Menpo-39-68* with 68 union landmarks. (7) MHM trained on *300W-68-Menpo-39-68* with 68 union landmarks. (8) Two-stage Multi-view Hourglass with intermediate supervision. This model barely improves the performance but doubling the computation cost.

Method	AUC	FR (%)	NME (%)
1	0.4470	8.14	6.09
2	0.4737	1.48	5.30
3	0.4629	2.96	5.49
4	0.5076	0.74	4.92
5	0.5141	0.74	4.86
6	0.5226	0.74	4.78
7	0.5324	0.74	4.68
8	0.5409	0.74	4.59

TABLE I: Landmark localization results on the IBUG dataset using 68 landmarks. Accuracy is reported as the Area Under the Curve (AUC) of the Cumulative Error Distribution curve, the Failure Rate (FR) at threshold 0.1, and out eye corner distance Normalized Mean Error (NME).

From the ablation experiments, we could conclude that by integrating the spatial transformer step, joint multi-view training and feeding more quality training data, the robustness and accuracy of proposed method improve hugely. As shown in Figure 7, although responses are more evident on facial organs than those on face contour, owing to more available profile training data, the proposed joint Multi-view Hourglass Model is able to deal with large pose variation.

Fig. 7: Demo results with large pose variation on IBUG predicted by Method (7). The score is higher on the inner facial organs than on the face contour.

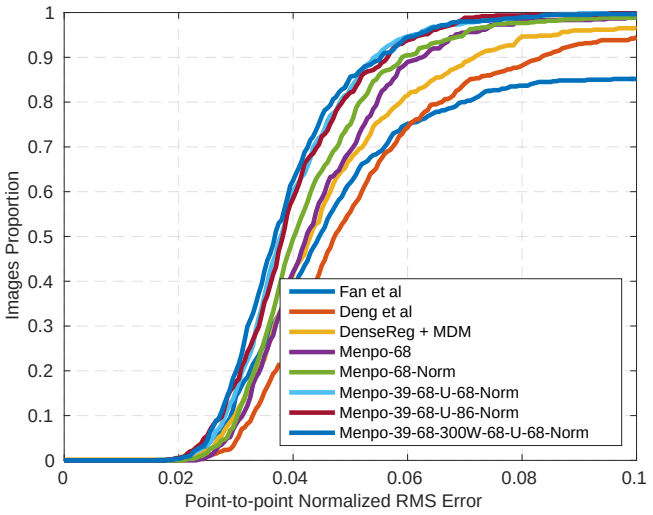


Fig. 8: Landmark localization results on the 300W dataset. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalized with the out eye corner distance.

C. Face Alignment on Images

We present experimental results on three face image databases, 300W database [30], COFW [65], [66] dataset and Menpo Benchmark [48]. The alignment method we evaluate here is the proposed Multi-view Hourglass Model (MHM), where the **-Norm** means the spatial transformer, and the **-U-86** means the union 86 landmarks. Experiment results on 300W database are shown in Figure 8, where we compared the proposed methods with the best results in the 300W competition [30], such as Deng et al. [33] and Fan et al. [68]. Besides, we also compare with the state-of-the-art face alignment method "DenseReg + MDM" [69]. It is obvious that our model (Menpo-39-68-300W-68-U-68-Norm) outperforms those methods by a large margin. Table II reports the area under the curve (AUC) of the CED curves, as well as the failure rate for a maximum error of 0.1. Apart from the accuracy improvement shown by the AUC, we believe that the reported failure rate of 0.33% is remarkable and highlights the robustness of our MHM. Additionally, we found that the union landmark definition only has little influence on semi-frontal face alignment accuracy. Thus we stick to the union 68 landmarks definition to avoid any confusion.

Method	AUC	FR (%)
Fan et al.	0.4802	14.83
Deng et al.	0.4752	5.5
DenseReg + MDM	0.5219	3.67
Menpo-68	0.5485	1.00
Menpo-68-Norm	0.5656	1.17
Menpo-39-68-U-68-Norm	0.5973	0.17
Menpo-39-68-U-86-Norm	0.5987	0.17
Menpo-39-68-300W-U-68-Norm	0.6071	0.33

TABLE II: Landmark localization results on the 300W (indoor and outdoor) testing dataset using 68 landmarks. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the RMS point-to-point error normalized with out eye corner distance. "Norm" stands for the spatial transformer step from five facial landmarks. "U" stands for the union set number of profile and frontal data.

We also present the performance of the MHM on the COFW [65], [66] dataset. Robust face alignment under occlusion and occluded landmark prediction are coupled problem that could be resolved simultaneously. Given the landmark occlusion status, local observation noise can be removed and the occluded landmark location can be predicted by shape context or constraint. Given a good fitting result, exploiting the fact that appearance of occluded region is quite different from the normal face appearance, even the simplest binary classifier could achieve excellent performance on occlusion classification. In Figure 9, we show the result of the proposed method comparing with state-of-the-art methods on COFW [65], such as HPM [70], SAPM [71], CFSS [25], TCDCN [72], and RCPR [65]. It can be clearly seen that even the baseline Hourglass model obtains a much better result because the bottom-up and top-down processing steps model the scale variations that would benefit the context inference. Moreover, by adding the spatial transformer, joint multi-view training and combined training data step-by-step, we gradually improved the alignment result, with the final success rate approaching 97.44%. Based on our best result, we employ the adaptive exemplar dictionary method [38] to predict occlusion status and refine the occluded landmarks. The normalized mean error decreases from 5.69% to 5.58%, and the occlusion prediction obtains a recall rate of 70.36% at the precision rate of 85.97%. In Figure 10, we give some fitting examples on COFW under heavy occlusions. To our surprise, responses of the occluded parts are still very clear and evident, which would prevent weird fitting results. This suggests that the proposed method captures and consolidates information across whole face images under different conditions, and incorporates local observation and global shape context in an implicit data-driven way, and thus improves the model's robustness under occlusions.

In Figure 11, we also report the test results of our model on the Menpo Benchmark by comparing with the best three entries (Jing Yang [73], Zhenliang He [74], Wenyan Wu [75])

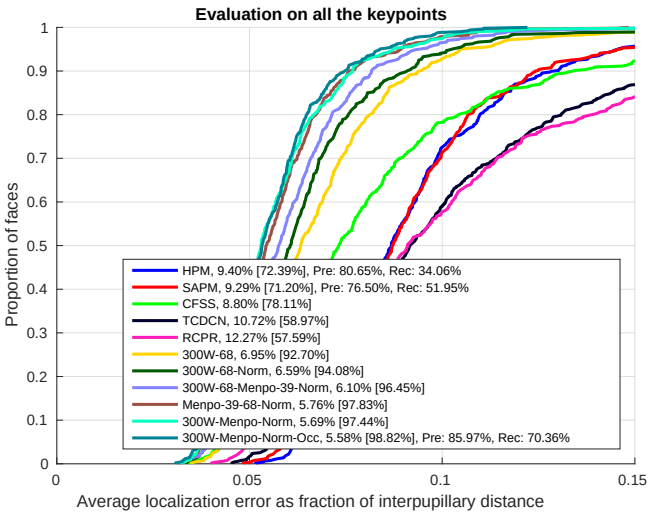


Fig. 9: Landmark localization results on the COFW dataset. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalized with the eye centre distance.

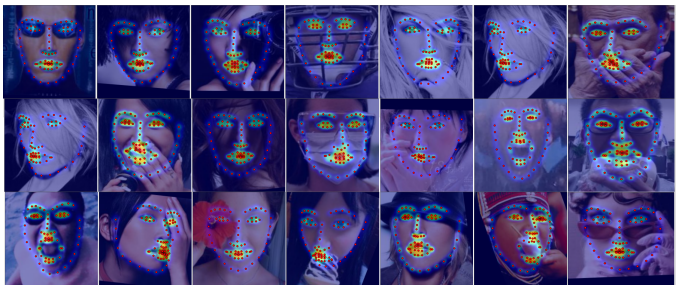
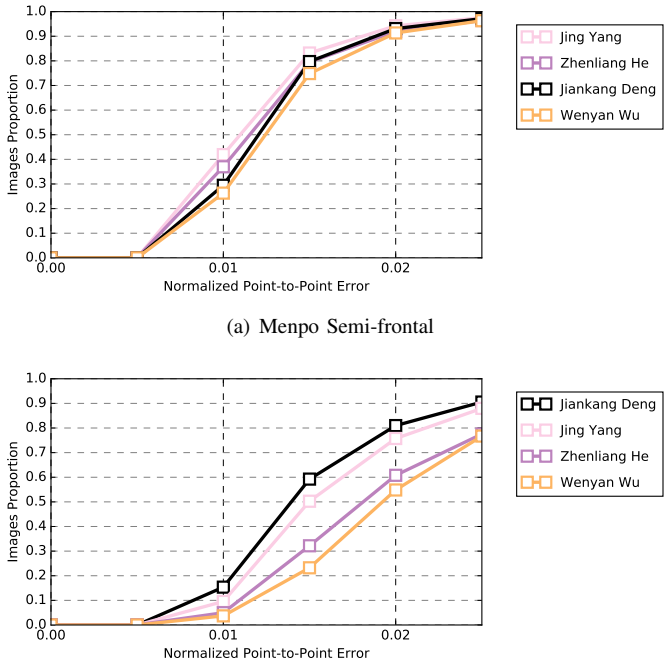


Fig. 10: Example results by MHM on COFW. Response maps on the occluded parts are still very clear and evident.



(a) Menpo Semi-frontal

(b) Menpo profile

Fig. 11: Landmark localization results on the Menpo Benchmark. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalized with the diagonal of the ground truth bounding box.

of the competition [48]. We draw the curve of cumulative error distribution on semi-frontal and profile test data separately. The proposed method has similar performance to the best performing methods in semi-frontal faces. Nevertheless, it outperforms the best performing method in profile faces. Despite that result on profile data is worse than that on semi-frontal data, both of their normalized (by diagonal length of bounding box) fitting errors of our method are remarkably small, approaching 1.48% and 1.27% for profile and semi-frontal faces respectively. In Figure 12, we give some fitting examples on the Menpo test set. As we can see from the alignment results, the proposed multi-view hourglass model is robust under pose variations, exaggerate expressions and occlusions on both semi-frontal and profile subset.

D. Face Alignment on Videos

We employ the 300VW challenge [76] testset for the challenging task of deformable face tracking on videos. Using our joint MHM method, we perform a frame-by-frame tracking on the video, and we initialize the next frame by the previous facial bounding box. The classifier based on the multi-view response maps is used as the failure checker during tracking. The face detector will be called if the fitting fails. The MHM takes 12.21 ms per face, and the classifier takes 2.32ms per face. The proposed multi-view face alignment and tracking method can run at about 50 FPS on the 300VW testset. We compare our method against the winners of the 300VW challenge: Yang et al. [77] and Xiao et al. [78]. Figure 14 reports the CED curves for all three video scenarios, and Table III reports the AUC and Failure Rate measures. The proposed MHM achieves the best performance, by a large margin compared to the winner of the 300VW competition ($\geq$15% at RMSE = 0.02 in Scenario1&2, $\approx$10% at RMSE = 0.02 in Scenario3) as well as the best setting for CFSS method [25], [54] ($\approx$15% at RMSE = 0.02 in Scenario1&2, $\approx$10% at RMSE = 0.02 in Scenario3), despite the fact that our approach is not fine-tuned on the training set of 300VW, while the rest of the methods were trained on video sequences and sometimes even with temporal modelling. Besides, our frame-by-frame tracking result is good enough that additional smoothing step (Kalman Filter) might be unnecessary.

In Figure 13, we select some frames from most challenging videos in Scenario3 and show their corresponding response maps for visualization purpose. The response maps of proposed method is very robust under large pose variation (yaw + pitch angles) and occlusion. In addition, response maps of invisible face parts are also reasonable, which indicates an implicit facial shape constraint within our method.

E. Face Detection

We evaluate the effectiveness of the multi-view response maps to remove high score false positives and obtain a state-of-the-art result on the FDDB dataset. As in [9], we review the annotation of FDDB [79], and add 67 unlabelled faces in FDDB dataset to make sure all the false alarms are correct.

(a) Menpo Semi-frontal set

(b) Menpo profile set

Fig. 12: Example landmark localization results on the test set of the Menpo Benchmark.

Method	Scenario1		Scenario2		Scenario3	
	AUC	Failure Rate (%)	AUC	Failure Rate (%)	AUC	Failure Rate (%)
Yang et al.	0.791	2.400	0.788	0.322	0.710	4.461
Xiao et al.	0.760	5.899	0.782	3.845	0.695	7.379
MDNET + CFSS + Kalman	0.784	1.754	0.783	0.341	0.713	7.466
MTCNN + CFSS + Kalman	0.734	8.507	0.725	8.518	0.726	5.685
MTCNN + CFSS + previous	0.748	6.055	0.760	2.717	0.726	4.388
Our method	0.847	0.290	0.838	0.033	0.769	0.972
Kalman smooth	0.849	0.285	0.842	0.030	0.7734	0.889

TABLE III: Landmark localization results on three categories of the 300VW test sets using 68 landmarks. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the RMS point-to-point error normalized with the diagonal of the ground truth bounding box [54].

(a) Scenario3-411

(b) Scenario3-557

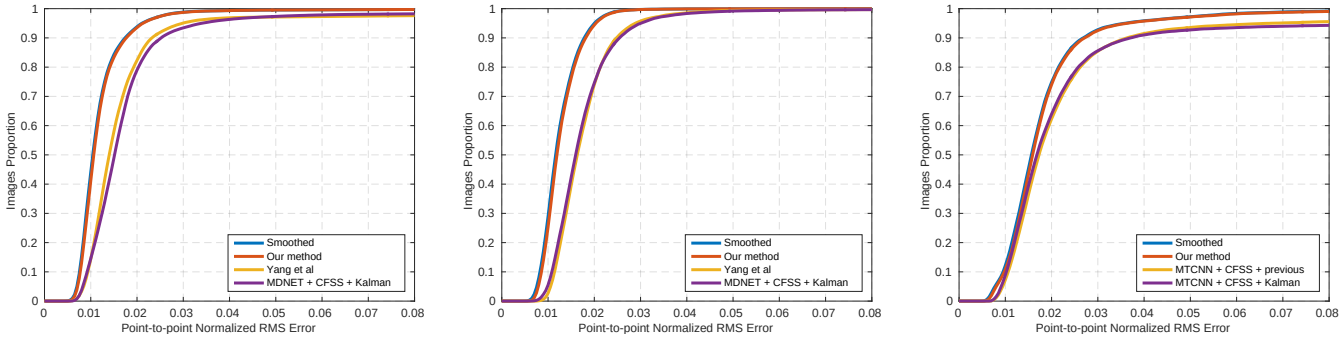
Fig. 13: Response maps generated by MHM on two challenging videos from 300VW Scenario3 (Video ID: 411 and 557). The response maps are invariant to large pose variation and robust under occlusion and fast motion.

False Positives	5	50	150
Precision Rate	99.9%	99%	97.1%
Our method *th1*	84.3	90.4	90.5
Our method *th2*	84.5	90.5	94.8
Baseline	65.1	89.9	92.4
MTCNN [55]	64.2	88.8	91.8
HR-ER,CVPR17 [80]	73.1	87.9	93.1
Conv3D,ECCV16 [81]	66.1	81.6	86.2
STN,ECCV16 [9]	88.3	90.3	91.5
Xiaomi [82]	78.6	90.8	94.6
DeepIR [83]	82.7	91.2	94.7

TABLE IV: Recall rate comparison with the state-of-the-art face detectors on FDDB within the high precision rate section (150 false positives, 97.1%).

We enlarge FDDB images by 1.6, and the average resolution is about 639×604. We test the model on a single NVIDIA GTX Titan X (Pascal) GPU setting minimum face as 20. As shown in Table IV and Figure 15(a), we observe the improvement of recall within the high precision section (150 false positives, precision rate 97.1%). The baseline method refers to our re-implementation of MTCNN [55], due to adopting additional labelled faces from AFLW, our implementation is slightly better than the original MTCNN. Our method *th1* sets a higher thresholds (0.6, 0.7, 0.7, 0.7) for cascaded classifiers, while our method *th2* employs a lower thresholds (0.5, 0.5, 0.3, 0.7). As can be seen from Table IV and Table V, the setting of *th2* is slightly better than *th1*, but increases the running time from 49.8 ms to 62.9ms per image. The proposed joint multi-view response maps contribute to removing high score false positives from previous cascade classifiers. At the precision rate of 99.9%, the proposed method improves the recall from 65.1% to 84.5%. At the precision rate of 99%, the proposed method improves the recall from 89.9% to 90.5%. The result is obviously higher than HR-ER [80] and Conv3D [81], and comparable with the best academic face detectors, e.g. STN [9], Xiaomi [82], and DeepIR [83]. After investigating our false positives, we surprisingly find some tiny regions (shown in Figure 15(b)) that can hardly be removed by our method, since they have very similar appearance and structure of the face, and may only be resolved by context-based model.

We also submitted our face detection results to http://www.cbsr.ia.ac.cn/faceevaluation/ and obtained the true positive vs. false positive curve on MALF. In Figure 16, our submission is named "sub_v1" and the threshold setting is (0.5,0.5,0.3,0.7). We compared with the off-the-shelf face detectors including HeadHunter [7], ACF [84], DPM [7], JDA [31], and DenseBox [85]. The proposed method obtains the best performance on MALF compared to the best academic algorithms including cascade models (HeadHunter [7], ACF [84], JDA [31]), structure models (DPM, JDA) and the structure-constrained deep model (Densebox). We also outperform the big data driven commercial models such as the FacePP-v2 and Picasa algorithms. Compared to the state-of-the-art method DenseBox, our joint multi-view response



(a) Scenario1 (b) Scenario2 (c) Scenario3

Fig. 14: Deformable face tracking results on 300VW. We only compare with the best two results evaluated by [54] on each scenario.

	proposal	CLS2	CLS3	**CLS4**
threshold	0.6	0.7	0.7	0.7
output boxNum	194.29	15.72	1.84	1.776
time(ms)	11.45	2.41	1.74	4.27
recall	97.76	95.17	90.97	90.60
precision	0.91	11.03	89.82	92.72
threshold	0.5	0.5	0.3	0.7
output boxNum	265.27	25.16	2.65	1.784
time(ms)	11.89	2.88	2.11	6.15
recall	98.30	97.87	95.44	95.10
precision	0.67	7.07	65.51	96.89

TABLE V: Output face box number and computation time of each step of our detector under two different threshold setting. Time consumption on image resize and Non-Maximum Suppression (NMS) take about 7.5ms. **CLS4** is after multi-view face alignment step (12.21ms per face). For *th1*, the mean running time is about 49.8ms per image. For *th2*, the mean running time is about 62.9ms per image.

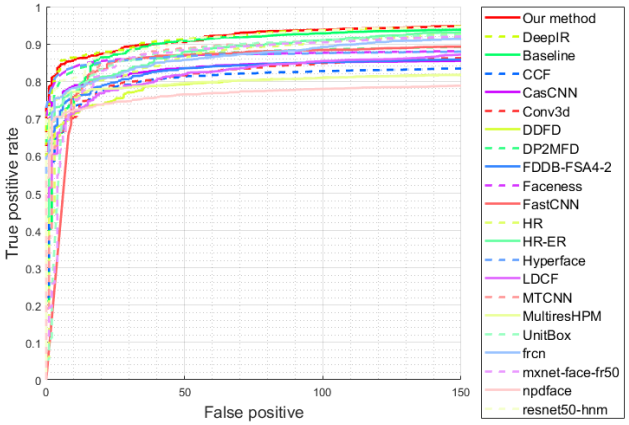
maps achieve a significantly better detection result in large pose data (yaw angle $>$ 40 degrees). A similar improvement could also be observed on the "hard" subsets.

V. Conclusion

In this paper, we proposed a joint multi-view face detection and alignment method where a face detector is used to estimate a coarse estimate of the facial shape using a small subset of landmarks and then after removing similarity transformations a refining subsequent step is performed that estimates the high-resolution facial shape of each person. We formulate a novel multi-view hourglass model which tries to jointly estimate both semi-frontal and profile facial landmarks, and the joint training model is stable and robust under continuous view variations. We demonstrate huge improvement over the state-of-the-art results in the latest benchmarks for face alignment such as 300W, COFW and the latest Menpo Benchmark. We also demonstrate state-of-the-art results for the deformable face tracking on the 300VW benchmark and face detection on FDDB and MALF datasets.

VI. Acknowledgements

Jiankang Deng was supported by the President's Scholarship of Imperial College London. This work was partially funded by the EPSRC project EP/N007743/1 (FACER2VM), the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 688520 (TeSLA), and a Google Faculty Fellowship to Dr. Zafeiriou. We thank the NVIDIA Corporation for donating several GPUs used in this work.



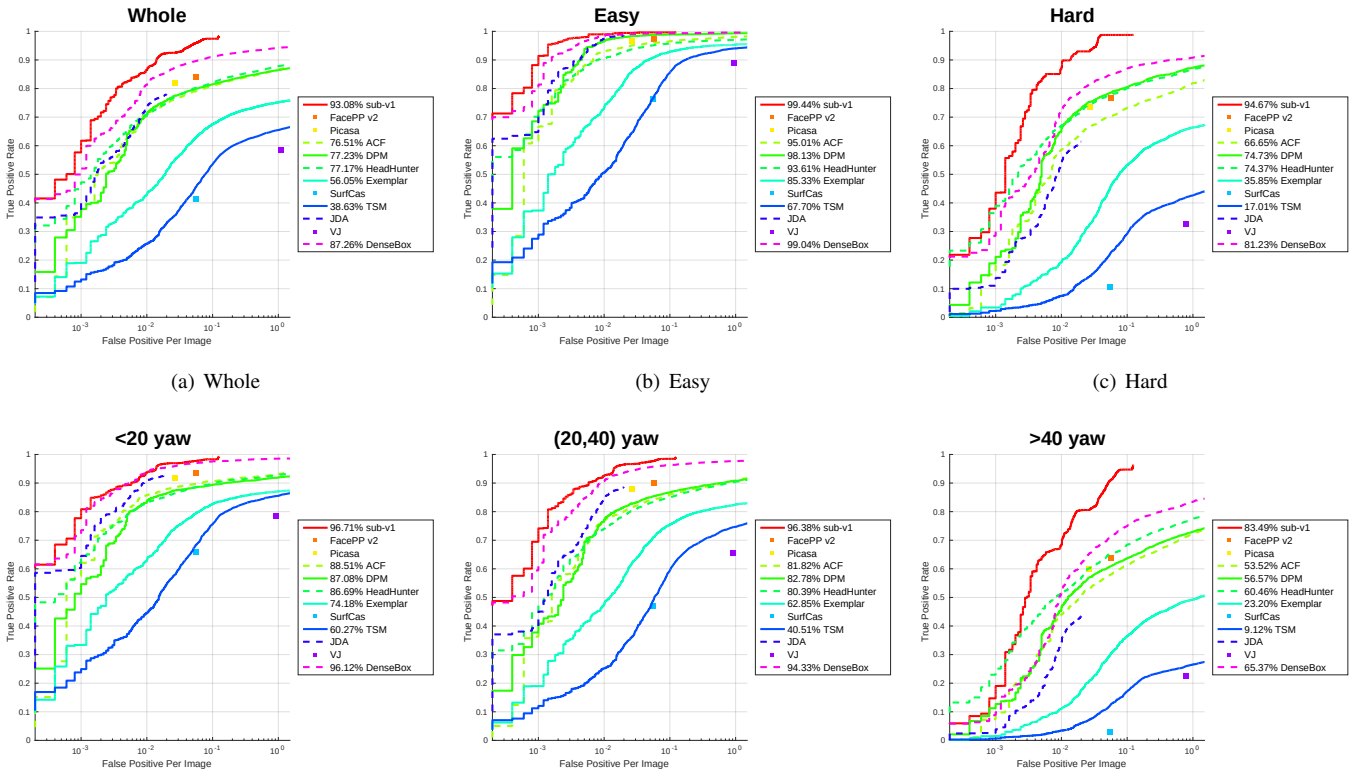
(a) Evaluation on FDDB

(b) Hard False Positives

Fig. 15: (a) Face detection results on FDDB. Our method utilizes the joint multi-view response maps to remove high score false positives. (b) Some interesting hard false positives from FDDB that even can not remove by our classifier. The resolution of these regions are about 20×20, and the structure and contour are very similar to a human face. The first row is covered with the predicted response maps, and the second row is the enlarged image crops.

References

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*, 2014, pp. 580–587.



(a) Whole (b) Easy (c) Hard

(d) Yaw angle <20 (e) Yaw angle $\in$ (20,40) (f) Yaw angle >40

Fig. 16: Fine-grained evaluation on MALF dataset.

[2] Ross Girshick, "Fast r-cnn," in *International Conference on Computer Vision*, 2015, pp. 1440–1448.

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

[4] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[6] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang, "A survey on face detection in the wild: past, present and future," *Computer Vision and Image Understanding*, vol. 138, pp. 1–24, 2015.

[7] M. Mathias, R. Benenson, M. Pedersoli, and G. L. Van, "Face detection without bells and whistles," in *European Conference on Computer Vision*, pp. 720–735. Springer, 2014.

[8] Xiangxin Zhu and Deva Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition*, 2012, pp. 2879–2886.

[9] Dong Chen, Gang Hua, Fang Wen, and Jian Sun, "Supervised transformer network for efficient face detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 122–138.

[10] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[11] Dong Li, Huiling Zhou, and Kin-Man Lam, "High-resolution face verification using pore-scale facial features," *IEEE transactions on image processing*, vol. 24, no. 8, pp. 2317–2327, 2015.

[12] Ying Tai, Jian Yang, Yigong Zhang, Lei Luo, Jianjun Qian, and Yu Chen, "Face recognition with pose variations and misalignment via orthogonal procrustes regression," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2673–2683, 2016.

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *arXiv preprint arXiv:1801.07698*, 2018.

[14] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou, "Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7093–7102.

[15] Vinay Bettadapura, "Face expression recognition and analysis: the state of the art," *arXiv preprint arXiv:1203.6722*, 2012.

[16] Yimo Guo, Guoying Zhao, and Matti Pietikäinen, "Dynamic facial expression recognition with atlas construction and sparse representation," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 1977–1992, 2016.

[17] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[18] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[19] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[20] David Cristinacce and Timothy F Cootes, "Feature detection and tracking with constrained local models.," in *British Machine Vision Conference*, 2006, vol. 1, p. 3.

[21] Piotr Dollár, Peter Welinder, and Pietro Perona, "Cascaded pose regression," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1078–1085.

[22] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2887–2894.

[23] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[24] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun, "Face alignment at 3000 fps via regressing local binary features," in *Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[25] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang, "Face

alignment by coarse-to-fine shape searching," in *Computer Vision and Pattern Recognition*, 2015, pp. 4998–5006.
[26] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep convolutional network cascade for facial point detection," in *Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.
[27] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li, "Face alignment across large poses: A 3d solution," in *Computer Vision and Pattern Recognition*, 2016, pp. 146–155.
[28] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou, "Mnemonic descent method: A recurrent process applied for end-to-end face alignment," in *Computer Vision and Pattern Recognition*, 2016, pp. 4177–4187.
[29] Georgios Tzimiropoulos and Maja Pantic, "Optimization problems for fast aam fitting in-the-wild," in *International Conference on Computer Vision*, 2013, pp. 593–600.
[30] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic, "300 faces in-the-wild challenge: Database and results," *Image and Vision Computing*, vol. 47, pp. 3–18, 2016.
[31] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *European Conference on Computer Vision*, pp. 109–122. Springer, 2014.
[32] Timothy F Cootes, Gavin V Wheeler, Kevin N Walker, and Christopher J Taylor, "View-based active appearance models," *Image and vision computing*, vol. 20, no. 9, pp. 657–664, 2002.
[33] Jiankang Deng, Qingshan Liu, Jing Yang, and Dacheng Tao, "M 3 csr: multi-view, multi-scale and multi-component cascade shape regression," *Image and Vision Computing*, vol. 47, pp. 19–26, 2016.
[34] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun, "Face alignment via regressing local binary features," *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1233–1245, 2016.
[35] Zhen-Hua Feng, Guosheng Hu, Josef Kittler, William Christmas, and Xiao-Jun Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3425–3440, 2015.
[36] Qingshan Liu, Jiankang Deng, and Dacheng Tao, "Dual sparse constrained cascade regression for robust face alignment," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 700–712, 2016.
[37] Heng Yang, Xuming He, Xuhui Jia, and Ioannis Patras, "Robust face alignment under occlusion via regional predictive power estimation," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2393–2403, 2015.
[38] Qingshan Liu, Jiankang Deng, Jing Yang, Guangcan Liu, and Dacheng Tao, "Adaptive cascade regression model for robust face alignment," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 797–807, 2017.
[39] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *International Conference on Computer Vision Workshops*, 2013, pp. 386–391.
[40] Zhujin Liang, Shengyong Ding, and Liang Lin, "Unconstrained facial landmark localization with backbone-branches fully-convolutional networks," *arXiv preprint arXiv:1507.03409*, 2015.
[41] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *European Conference on Computer Vision*. Springer, 2014, pp. 1–16.
[42] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa, "An all-in-one convolutional neural network for face analysis," *arXiv preprint arXiv:1611.00851*, 2016.
[43] Hanjiang Lai, Shengtao Xiao, Yan Pan, Zhen Cui, Jiashi Feng, Chunyan Xu, Jian Yin, and Shuicheng Yan, "Deep recurrent regression for facial landmark detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.
[44] Shengtao Xiao, Jiashi Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Kassim, "Robust facial landmark detection via recurrent attentive-refinement networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 57–72.
[45] Adrian Bulat and Georgios Tzimiropoulos, "Convolutional aggregation of local evidence for large pose face alignment," in *British Machine Vision Conference*, 2016.
[46] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.
[47] Adrian Bulat and Georgios Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *International Conference on Computer Vision*, 2017.
[48] Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *Computer Vision and Pattern Recognition Workshop*, 2017.
[49] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *International Conference on Computer Vision*, 2013, pp. 1944–1951.
[50] Yue Wu and Qiang Ji, "Robust facial landmark detection under significant head poses and occlusion," in *International Conference on Computer Vision*, 2015, pp. 3658–3666.
[51] Amin Jourabloo and Xiaoming Liu, "Pose-invariant 3d face alignment," in *International Conference on Computer Vision*, 2015, pp. 3694–3702.
[52] Amin Jourabloo and Xiaoming Liu, "Large-pose face alignment via cnn-based dense 3d model fitting," in *Computer Vision and Pattern Recognition*, 2016, pp. 4188–4196.
[53] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3d morphable model," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.
[54] Grigorios G Chrysos, Epameinondas Antonakos, Patrick Snape, Akshay Asthana, and Stefanos Zafeiriou, "A comprehensive performance evaluation of deformable face tracking" in-the-wild"," *International Journal on Computer Vision*, 2016.
[55] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.
[56] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.
[57] Oren Tadmor, Tal Rosenwein, Shai Shalev-Shwartz, Yonatan Wexler, and Amnon Shashua, "Learning a metric embedding for face recognition using the multibatch method," in *Advances In Neural Information Processing Systems*, 2016, pp. 1388–1389.
[58] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *International Conference on Computer Vision*, December 2015.
[59] Martin Köstinger, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *International Conference on Computer Vision Workshops*. IEEE, 2011, pp. 2144–2151.
[60] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.
[61] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "Wider face: A face detection benchmark," in *Computer Vision and Pattern Recognition*, 2016, pp. 5525–5533.
[62] Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou, "The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking," *International Journal of Computer Vision*, pp. 1–26, 2018.
[63] Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar, "Localizing parts of faces using a consensus of exemplars," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2930–2940, 2013.
[64] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*. Springer, 2012, pp. 679–692.
[65] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár, "Robust face landmark estimation under occlusion," in *International Conference on Computer Vision*, 2013, pp. 1513–1520.
[66] Golnaz Ghiasi and Charless C Fowlkes, "Occlusion coherence: Detecting and localizing occluded faces," *arXiv preprint arXiv:1506.08347*, 2015.
[67] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.
[68] Haoqiang Fan and Erjin Zhou, "Approaching human level facial landmark localization by deep learning," *Image and Vision Computing*, vol. 47, pp. 27–35, 2016.
[69] Rıza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos, "Densereg: Fully convolutional dense shape regression in-the-wild," *Computer Vision and Pattern Recognition*, 2017.

[70] Golnaz Ghiasi and Charless C Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Computer Vision and Pattern Recognition*, 2014, pp. 2385–2392.
[71] Golnaz Ghiasi, Charless C Fowlkes, and CA Irvine, "Using segmentation to predict the absence of occluded parts.," in *BMVC*. Citeseer, 2015, pp. 22–1.
[72] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 918–930, 2016.
[73] Jing Yang, Qingshan Liu, and Kaihua Zhang, "Stacked hourglass network for robust facial landmark localisation," in *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017, vol. 3, p. 6.
[74] Zhenliang He, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen, "Robust fec-cnn: A high accuracy facial landmark detection system," in *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017, vol. 3, p. 6.
[75] Wenyan Wu and Shuo Yang, "Leveraging intra and inter-dataset variations for robust face alignment," in *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, 2017, vol. 3, p. 6.
[76] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1003–1011.
[77] Jing Yang, Jiankang Deng, Kaihua Zhang, and Qingshan Liu, "Facial shape tracking via spatio-temporal cascade shape regression," in *International Conference on Computer Vision Workshops*, 2015, pp. 41–49.
[78] Shengtao Xiao, Shuicheng Yan, and Ashraf A Kassim, "Facial landmark detection via progressive initialization," in *International Conference on Computer Vision Workshops*, 2015, pp. 33–40.
[79] Vidit Jain and Erik G Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," *UMass Amherst Technical Report*, 2010.
[80] Peiyun Hu and Deva Ramanan, "Finding tiny faces," 2017.
[81] Yunzhu Li, Benyuan Sun, Tianfu Wu, and Yizhou Wang, "Face detection with end-to-end integration of a convnet and a 3d model," in *European Conference on Computer Vision*. Springer, 2016, pp. 420–436.
[82] Shaohua Wan, Zhijun Chen, Tao Zhang, Bo Zhang, and Kong-kat Wong, "Bootstrapping face detection with hard negative examples," *arXiv preprint arXiv:1608.02236*, 2016.
[83] Xudong Sun, Pengcheng Wu, and Steven CH Hoi, "Face detection using deep learning: An improved faster rcnn approach," *arXiv preprint arXiv:1701.08289*, 2017.
[84] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Aggregate channel features for multi-view face detection," in *International Joint Conference on Biometrics*. IEEE, 2014, pp. 1–8.
[85] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "Densebox: Unifying landmark localization with end to end object detection," *arXiv:1509.04874*, 2015.

Jiankang Deng is a Ph.D. candidate in the Intelligent Behaviour Understanding Group (IBUG), Department of Computing, Imperial College London. He is funded by the Imperial President's PhD Scholarships and his research interest is face analysis.

George Trigeorgis has received his Ph.D degree from the Department of Computing, Imperial College London. He was a recipient of the prestigious Google Ph.D. Fellowship in Machine Perception for 2017. He has regularly published in several prestigious conferences in his field including ICML, NIPS, and CVPR, while he is also a reviewer in IEEE T-PAMI, IEEE CVPR/ICCV/FG.

Yuxiang Zhou is a Ph.D. candidate in the Intelligent Behaviour Understanding Group (IBUG), Department of Computing, Imperial College London. His research interest is statistical deformable model with dense shape.

Stefanos Zafeiriou is currently a Senior Lecturer in Pattern Recognition/Statistical Machine Learning for Computer Vision with the Department of Computing, Imperial College London, London, U.K, and a Distinguishing Research Fellow with University of Oulu under Finish Distinguishing Professor Programme. He was a recipient of the Prestigious Junior Research Fellowships from Imperial College London in 2011 to start his own independent research group. He was the recipient of the Presidents Medal for Excellence in Research Supervision for 2016. He has received various awards during his doctoral and post-doctoral studies. He has been a Guest Editor of over six journal special issues and co-organized over nine workshops/special sessions on face analysis topics in top venues, such as CVPR/ICCV/ECCV/FG (including two very successfully challenges run in ICCV13 and ICCV15 on facial landmark localization/tracking). He has more than 7K citations to his work, h-index 44. He was the General Chair of BMVC 2017.