Robust FEC-CNN: A High Accuracy Facial Landmark Detection System

Zhenliang He[1,2] Jie Zhang[1,2] Meina Kan[1,3] Shiguang Shan[1,3] Xilin Chen[1]
[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
{zhenliang.he,jie.zhang,meina.kan,shiguang.shan,xilin.chen}@vipl.ict.ac.cn

Abstract

Facial landmark detection, as a typical and crucial task in computer vision, is widely used in face recognition, face animation, facial expression analysis, etc. In the past decades, many efforts are devoted to designing robust facial landmark detection algorithms. However, it remains a challenging task due to extreme poses, exaggerated facial expression, unconstrained illumination, etc. In this work, we propose an effective facial landmark detection system, recorded as Robust FEC-CNN (RFC), which achieves impressive results on facial landmark detection in the wild. Considering the favorable ability of deep convolutional neural network, we resort to FEC-CNN as a basic method to characterize the complex nonlinearity from face appearance to shape. Moreover, face bounding box invariant technique is adopted to reduce the landmark localization sensitivity to the face detector while model ensemble strategy is adopted to further enhance the landmark localization performance. We participate the Menpo Facial Landmark Localisation in-the-Wild Challenge and our RFC significantly outperforms the baseline approach APS. Extensive experiments on Menpo Challenge dataset and IBUG dataset demonstrate the superior performance of the proposed RFC.

1. Introduction

Widely used in vision applications like face recognition and facial animation, facial landmark detection plays an essential part in computer vision. Impressive works [5, 4, 6, 3, 23, 14, 20, 27, 21, 22, 9] and benchmarks [11, 29, 2, 16, 17, 10, 26] were proposed to tackle this task in the the past few decades.

Early works are ASMs [5, 7] and AAMs [4, 13], which employ Principal Component Analysis (PCA) to statistically capture the major factors that influence the variation of shape or appearance. Since the linearity of PCA, ASM and AAM like methods are difficult to model complicated variations due to extreme poses, exaggerated facial expression, unconstrained illumination, *etc*.

Figure 1. Exemplar results of RFC prediction on Menpo testing dataset. The first two columns are the results of profile faces and the last two columns are the results of semi-frontal faces.

Cascade method is another traditional and popular one in facial landmark research. Via a greedy scheme, most cascade methods [6, 23, 14] refine the shape stage-by-stage by modeling the shape residual . CPR [6], SDM [23] and LBF [14] employ shape-indexed feature and cascade several forest based regressors or linear regressors to model the shape residual. To reduce the shape prediction sensitivity to the initial shape, CPR [6] predicts the shape several times with different initial shapes and takes the shape in the highest density region of shape-space as the final prediction. However, inaccurate face bounding box makes the shape initialization worse, which may not be tackled by the initialization strategy that CPR used. To reduce the influence of the inaccurate face bounding box, Yan *et al*. [24] propose to

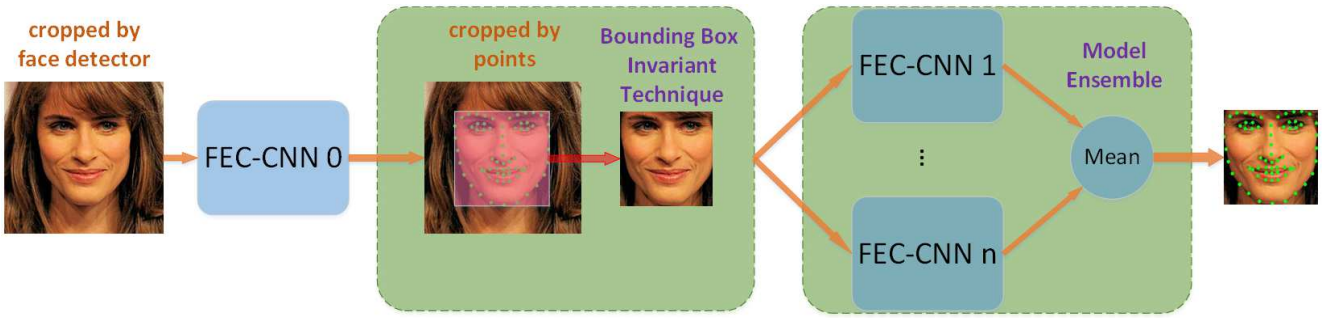


Figure 2. An overview of RFC which consists of three essential parts. FEC-CNN is the basic model for facial landmark detection. Bounding box invariant technique is used to produce a suitable face bounding box, which is less affected by the face detector. Model ensemble technique further improves the landmark detection accuracy.

generate multiple prediction shapes with different bounding boxes and learns to rank or combine these shapes.

Recently, more and more works tend to use deep method due to its powerful ability for modeling nonlinearity. DCNN [20], CFAN [27] use deep neural networks as cascade stages in the cascaded regression framework. Furthermore, MDM [21] and RAR [22] propose to model the cascade process by the recurrent scheme, which enables the end-to-end training of the cascaded model. On the other hand, FEC-CNN [9] enables the end-to-end training of the cascaded model by employing a differentiable shape-indexed feature extracting function to connect the cascade stages.

In this paper, we propose a facial landmark detection system named Robust FEC-CNN (RFC), which employs FEC-CNN [9] as basic method. To alleviate the prediction sensitivity of the basic FEC-CNN to the bounding box produced by the face detector, we employ a bounding box invariant technique. We also ensemble several complementary FEC-CNN models trained under different conditions to further improve the performance of RFC.

The rest of this paper is organized as follows: Section 2 reviews the related works. Section 3 illustrates the details about RFC. Section 4 gives the experimental results and analysis on Menpo dataset and IBUG dataset. Section 5 makes conclusions of this work.

2. Related Works

Cascaded regression methods achieve significant success in facial landmark detection. As a novel work that first tackles the pose estimation problem in a cascaded regression framework, CPR [6] uses shape-indexed control point feature and cascades several random fern regressors to predict the shape residual. Instead of the forest based regressor, SDM [23] employs simpler linear regressor in each cascade stage. Furthermore, shape-indexed SIFT [12] feature is used in SDM for each regressor. SDM achieves huge success because of its simple framework and good performance.

Afterwards, deep cascaded regression framework [27, 20] raises more attention. In a coarse-to-fine way, CFAN [27] uses stacked auto-encoder and shape-indexed SIFT feature in each stage to refine the shape. DCNN [20] is another deep cascaded regression method which cascades several deep convolution neural network to predict the shape stage-by-stage. Different from the previous methods which use handcraft feature, DCNN uses more powerful deep convolutional feature which can be learnt in a data driven manner.

Recurrent framework [21, 22] is proposed to tackle the facial landmark detection problem more recently. MDM [21] employs a recurrent neural network which takes shape-indexed convolutional features and the previous hidden states as input to recurrently refine the shape. RAR [22] employs an attentive-refinement mechanism, in which an attention LSTM sequentially locates a reliable landmark while a refinement LSTM sequentially refines the landmarks near that reliable one. Similar to the traditional cascaded framework, the recurrent framework refines the shape stage-by-stage. Furthermore, the recurrent framework also enables the whole network to be trained end-to-end while enhancing the information flow among the recurrent stages.

FEC-CNN [9] is also proposed to enable the end-to-end training of the cascaded framework. Different from MDM [21] and RAR [22], without any recurrent process, FEC-CNN employs a differentiable shape-indexed feature extracting function among the cascaded CNN networks, which makes the whole cascaded CNN framework differentiable. Therefore, the cascaded CNN networks in FEC-CNN can be trained end-to-end by gradient descent under a single objective. RFC adopts FEC-CNN as the basic method for facial landmark detection. More details about FEC-CNN is demonstrated in Section 3.

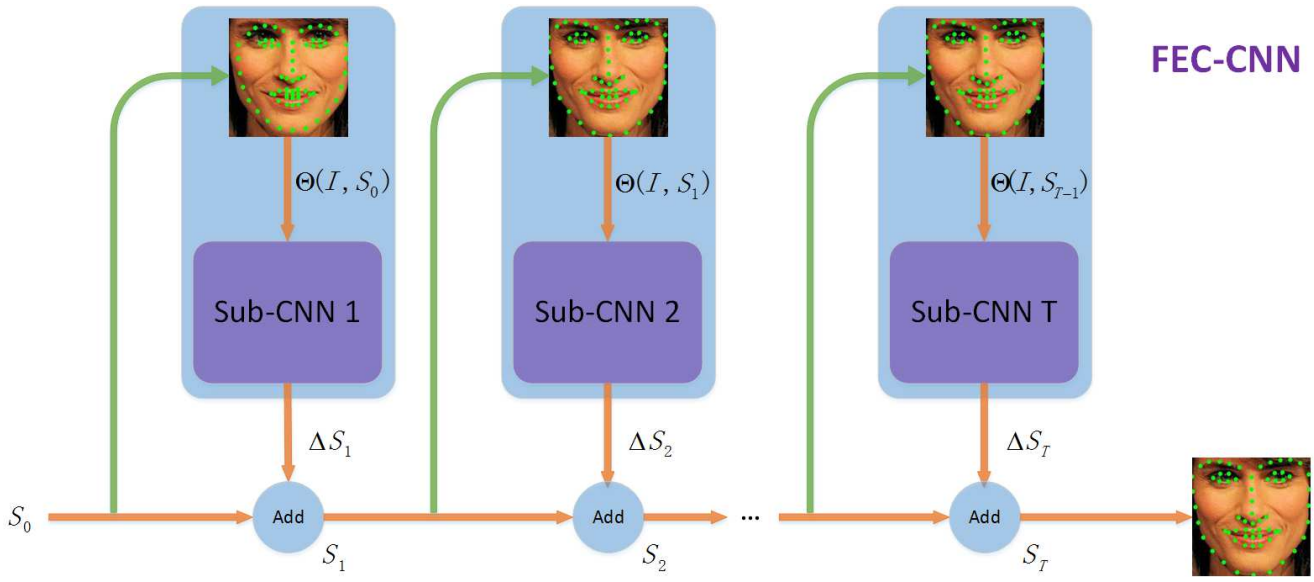


Figure 3. An overview of FEC-CNN framework. FEC-CNN cascades several sub-CNNs each of which takes the shape-indexed patches $\Theta(I, S_t)$ as input and predicts the shape residual ΔS_t. The shape-indexed patch extracting function Θ is designed by a differentiable bilinear interpolation process, which enables the end-to-end training of the whole FEC-CNN framework.

3. Robust FEC-CNN

As shown in Figure 2, RFC consists of three essential parts, including basic FEC-CNN for facial landmark detection, bounding box invariant technique for reducing the landmark localization sensitivity to the face detection bounding box, and model ensemble technique for further performance improvement. We first review the FEC-CNN [9] in brief.

3.1. FEC-CNN

3.1.1 Formulation

FEC-CNN formulates the facial landmark detection problem as a nonlinear mapping H from face image I to shape S. As shown in Figure 3, FEC-CNN models H via cascading several sub-CNNs. The landmark refinement process is formulated as follow:

$$S = H(I) = \sum_{t=1}^{T} F_t(\Theta(I, S_{t-1})) + S_0 \tag{1}$$

with

$$S_t = S_{t-1} + F_t(\Theta(I, S_{t-1})), t = 1, \cdots, T \tag{2}$$

where F_t denotes the t^{th} sub-CNN which outputs the shape residual, S_t denotes the shape predicted at the t^{th} stage which is the sum of last predicted shape and the current predicted shape residual, S_0 is the initial shape given by a mean shape or an initial network, and Θ is a differentiable function that maps the input image I and shape S_t to the shape-indexed patches.

The overall objective of FEC-CNN is formulated as follow:

$$\{F_t^*\}_{t=1}^T = \underset{\{F_t\}_{t=1}^T}{argmin} \sum_{i=1}^{N} ||\hat{S}^i - \sum_{t=1}^{T} F_t(\Theta(I^i, S_{t-1}^i)) - S_0^i||_2^2 \tag{3}$$

where $\hat{S}^i$ is the ground truth shape of the i^{th} training sample. Since each part of this objective function is differentiable, it can be optimized by gradient descent in an end-to-end scheme.

Since the most common metric for evaluating the landmark prediction error is the normalized root mean squared error (NRMSE) [3, 2, 27, 22], in RFC, we modify the objective in Equation 3 to the mean of NRMSE over all samples:

$$\{F_t^*\}_{t=1}^T = \underset{\{F_t\}_{t=1}^T}{argmin} \frac{1}{N} \sum_{i=1}^{N} NRMSE_i \tag{4}$$

with

$$NRMSE_i = \sum_{j=1}^{n} \frac{\sqrt{(\hat{y_{ij}} - y_{ij})^2 + (\hat{x_{ij}} - x_{ij})^2}}{nd_i} \tag{5}$$

for the i^{th} sample, $(\hat{y_{ij}}, \hat{x_{ij}})$ is the j^{th} ground truth point, (y_{ij}, x_{ij}) is the j^{th} prediction point and d_i is the normalization factor, such as the inter-ocular distance.

3.1.2 Differentiable Shape-Indexed Patch

Different from most traditional cascaded regression methods which choose non-differentiable feature like SIFT, FEC-CNN extracts differentiable shape-indexed patches and feeds them into the corresponding sub-CNNs outputting differentiable shape-indexed features. The differentiable function Θ in Equation 1 maps I and S_t to several independent shape-indexed patches corresponding to different landmarks, therefore for simplicity, Θ can be reformulated as for one landmark (y, x):

$$\Theta : I, y, x \rightarrow V \tag{6}$$

where I is the input image with height H and width $W$and V is (y, x)-indexed patch with height h and width w.

To let Θ differentiable, FEC-CNN adopts bilinear interpolation to generate patch V, which is formulated as follow:

$$V_{qp} = \sum_{n=0}^{H-1} \sum_{m=0}^{W-1} \lambda_{nm} I_{nm} \tag{7}$$

with

$$\lambda_{nm} = max(0, 1 - |y_q - n|)max(0, 1 - |x_p - m|) \tag{8}$$

$$y_q = y + q - (h - 1)/2 \tag{9}$$

$$x_p = x + p - (w - 1)/2 \tag{10}$$

3.2. Bounding Box Invariant Technique

The performance of facial landmark detection algorithm can be highly affected by the face detector. Suitable face bounding boxes guarantee the landmark prediction performance as well as the model training quality.

One straightforward way to reduce the prediction sensitivity to the face bounding boxes is to provide more stable bounding boxes instead of the unstable face detection bounding boxes either in training or testing phase. To achieve this goal, we first obtain the face detection results with high variance among different faces, and a FEC-CNN model A is trained with these cropped faces. Although the face detection bounding boxes has high variance, the variance of the landmarks predicted by FEC-CNN model A is lower due to its robustness. Therefore, the minimum enclosing rectangle of the predicted landmarks provides a more stable face bounding box as shown in the central part of Figure 2. Using these minimum enclosing rectangles as new face bounding boxes to crop faces, we train another FEC-CNN model B, which is much less affected by the face detector.

3.3. Model Ensemble

Model ensemble is proven to be effective to improve the performance in practice [19, 8]. RFC also employs model ensemble technique in a simple average scheme. Specifically, we train several FEC-CNN models under different conditions including different data augmentations, network structures and face cropping manners, which might be complementary. Then we simply predict the facial landmarks by each model and take the average result. Moreover, since not all FEC-CNN models help improve the performance, we adopt a greedy scheme to select models for ensemble as shown in Algorithm 1.

Algorithm 1 Greedy Model Ensemble.

Input:
$models$: the set of all well trained FEC-CNN models;
$error(set)$: function that returns the ensemble error of a set of models on the validation set;
Output:
$selected\_models$: the set of models for ensemble;

$selected\_models = models$
$e = error(selected\_models)$
$found = true$
while $found$ **do**
  $found = false$
  for m in $selected\_models$ **do**
    if $error(selected\_models - m) \leq e$ **then**
      $selected\_models = selected\_models - m$
      $e = error(selected\_models)$
      $found = true$
      $break$
    end if
  end for
end while
return $selected\_models$

4. Experiments

4.1. Dataset

We employ 300W [16, 18], 300W Competition [16, 17] and Menpo dataset [26] for evaluating RFC. The 300W dataset consists of LFPW [2], HELEN [11], AFW [29] and IBUG [17]. The 300W Competition dataset consists of indoor and outdoor subset. Proposed recently for the Menpo Challenge, the Menpo dataset consists of a 68 point semi-frontal subset and a 39 point profile subset.

For semi-frontal facial landmark detection, the above datasets are divided into two parts. LFPW, HELEN, AFW, 300W Competition and Menpo 68 point subset are used as training set while IBUG is used as testing set. For profile facial landmark detection, the Menpo 39 point subset is randomly divided into a training set and a testing set. To distinguish our division of the 39 point testing set from the testing set released by Menpo official, we denote our 39 point testing set as 39TestA.

4.2. Implementation Details

Face Detector We adopt a Faster R-CNN [15] which is trained on WIDER FACE [25] as face detector.

Data Augmentation For better generalization, we augment the data by the way that [9] uses. Concretely, the training set is 40 times augmented by random rotation, translation, horizontal flipping and resizing. To train complementary FEC-CNN models for ensemble, different sets of augmentation parameters are used.

Model Transfer Since there are few profile images can be used to train FEC-CNN model, we adopt AFLW [10] dataset which consists of large amount of profile faces with 19 point labels provided by Zhu *et al*. [28]. We pretrain the FEC-CNN with AFLW dataset and finetune the model with Menpo 39 point training set. Instead of pretrain-finetune strategy, the second choice we can use the AFLW data for model transfer is to train the the FEC-CNN model simultaneously using the AFLW dataset and Menpo profile dataset, which avoids the useful information of AFLW data being flushed by Menpo dataset.

Model Ensemble We train several FEC-CNN models under different conditions including different data augmentations, network structures and face cropping manners, which might be complementary.

4.3. Experimental Results

Under the greedy model selection scheme mentioned in Section 3, we finally employ 11 models for semi-frontal facial landmark detection while 8 for profile facial landmark detection. We compare the best single model performance and the ensemble performance in Figure 4, where cumulated error curves are drawn. As seen, model ensemble considerably improves the landmark detection performance.

Furthermore, in Table 1, we present the mean error on IBUG dataset comparing to CFAN [27], RAR [22] and the original FEC-CNN [9]. The comparison is unfair because we use more training data and model ensemble, however, it still shows that RFC makes huge progress in facial landmark detection.

Finally, testing results on Menpo official testing set are shown in Figure 5. As seen, RFC significantly outperforms APS [1] on the landmark prediction performance.

Method	Mean Error
CFAN	15.04
RAR	8.35
FEC-CNN	7.89
RFC (ours)	**6.56**

Table 1. Mean error on IBUG.

5. Conclusion

In this paper, we present the details of RFC, which is used to participate the Menpo Challenge. RFC uses FEC-CNN for basic method for facial landmark detection. Furthermore, bounding box invariant technique is adopted to reduce the prediction sensitivity to face detector while model ensemble is adopted for further performance improvement. RFC is evaluated on IBUG and Menpo dataset and shows significant performance.

Acknowledgments

This work was partially supported by Natural Science Foundation of China under contracts Nos. 61650202, 61402443, 61672496, and the Strategic Priority Research Program of the CAS (Grant XDB02070004).

References

[1] E. Antonakos, J. Alabort-i Medina, and S. Zafeiriou. Active pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5435–5444, 2015.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2930–2940, 2013.

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

[4] T. F. Cootes, G. J. Edwards, C. J. Taylor, et al. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 23(6):681–685, 2001.

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

[6] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1078–1085, 2010.

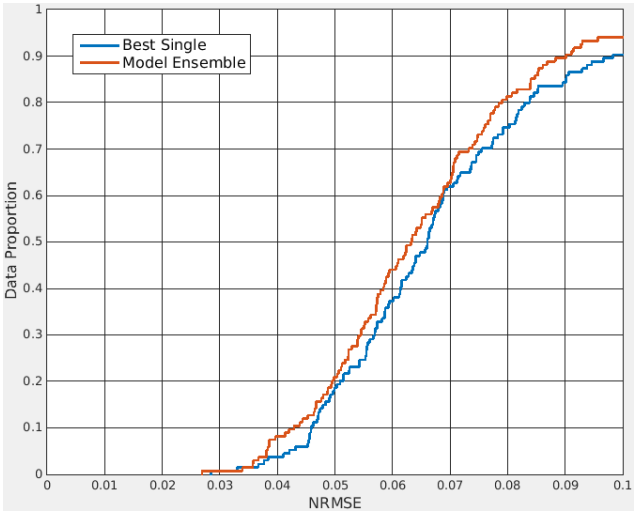
[7] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *European Conference on Computer Vision (ECCV)*, pages 413–426, 2008.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

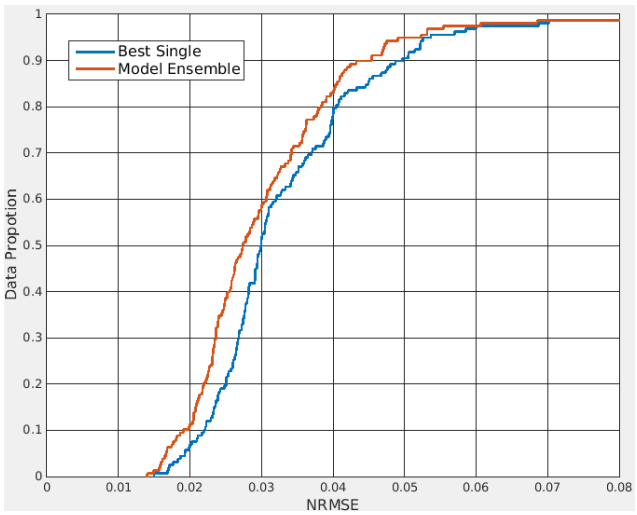
[9] Z. He, M. Kan, J. Zhang, X. Chen, and S. Shan. A fully end-to-end cascaded cnn for facial landmark detection. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2017.

[10] M. Köstinger, 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 *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2144–2151, 2011.

[11] 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.

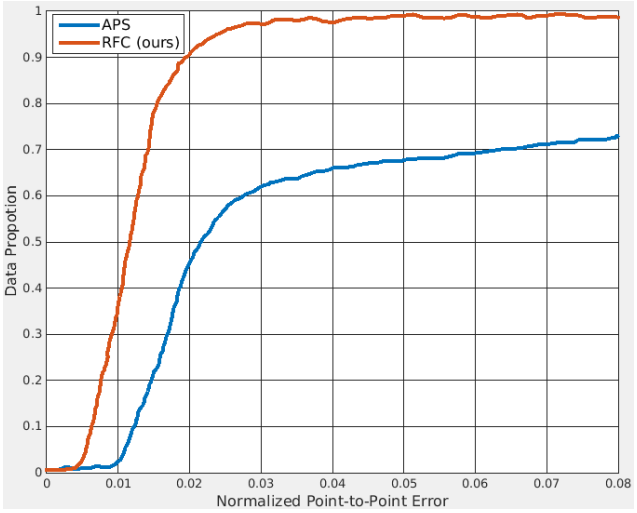


(a) Semi-frontal facial landmark detection on IBUG.

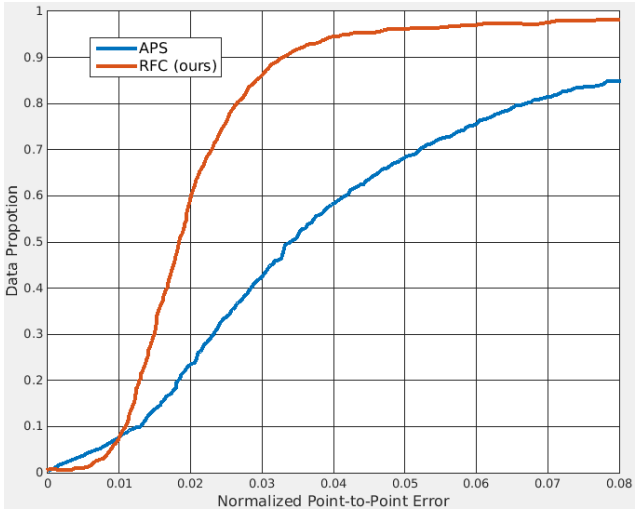


(b) Profile facial landmark detection on 39TestA.

Figure 4. The Improvement of Model Ensemble.



(a) On Menpo frontal testing set.



(b) On Menpo profile testing set.

Figure 5. Prediction Results of Menpo Official Testing Set.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[13] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision (IJCV)*, 60(2):135–164, 2004.

[14] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1692, 2014.

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015.

[16] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing (IVC)*, 47:3–18, 2016.

[17] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 397–403, 2013.

[18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annota-

tion. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 896–903, 2013.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] 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.

[21] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *European Conference on Computer Vision (ECCV)*, pages 57–72. Springer, 2016.

[23] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[24] J. Yan, Z. Lei, D. Yi, and S. Li. Learn to combine multiple hypotheses for accurate face alignment. In *International Conference on Computer Vision Workshops (ICCVW)*, pages 392–396, 2013.

[25] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[26] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen. The menpo facial landmark localisation challenge: A step closer to the solution. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[27] 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.

[28] S. Zhu, C. Li, C.-C. Loy, and X. Tang. Unconstrained face alignment via cascaded compositional learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3409–3417, 2016.

[29] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012.