R-MTCNN: Joint Face detection and Alignment Via Region-based Multi-task Convolution Networks

BMVC 2018 Submission # ??

Abstract

We proposed R-MTCNN, a new method to simultaneous detect face and facial landmarks. Current methods jointly learning these two tasks mainly based on cascade structure which means the time consumed by the algorithm is related to the number of faces in the image. R-MTCNN is region-based method. That is, it can simultaneous detect multiple faces including the facial landmark without consuming more time. R-MTCNN is an excellent multi-task face detection and alignment framework, robust to extreme conditions such as large pose, light and occlusion, and both the performance of face detection and face alignment achieve state-of-the-art performance. Our experiments show that the multi-task learning has a novel improvement on single task. We evaluate face detection's performance on WIDER FACE and FDDB benchmark and alignment's performance on AFLW benchmark.

1 Introduction

Face detection and alignment, as important steps in face recognition have been a very important research topic in the field of computer vision. There have been many literature about these two tasks [20, 36, 50]. The tasks of face detection, landmark localization have generally been solved as separate problems, and is used in many engineering practices. Recently, it has been shown that learning correlated tasks simultaneously can boost the performance of individual tasks [29, 41, 60].

Consider face aliment do well on a precise face location, current methods jointly learning these two tasks mainly based on cascade structure represented by MTCNN [60]. Cascade structure performs landmark localization on well-cropped face images which generated by previous face detector. This pipeline do have a good landmark detection result because a precise face location in the image is provided. However, when the number of faces in the image increased and then the consumed time grows obviously. This is difficult to meet real word's requirements in many surveillance scenes where has lots of faces.

This paper proposed a region-based method to simultaneous detecting face and facial landmarks, called R-MTCNN. We generate proposals that may contain faces in the first stage and then fine regressing the face boxes and do landmark localization. The two stages of R-MTCNN shares the same basic network, the second stage only pass a single fully connected layer with 2048 channels. Hence the second stage have a low time spent. We add three new convolution layers of conv6_1,conv6_2,conv6_3 on VGG-16 [44] with 1024 channels for

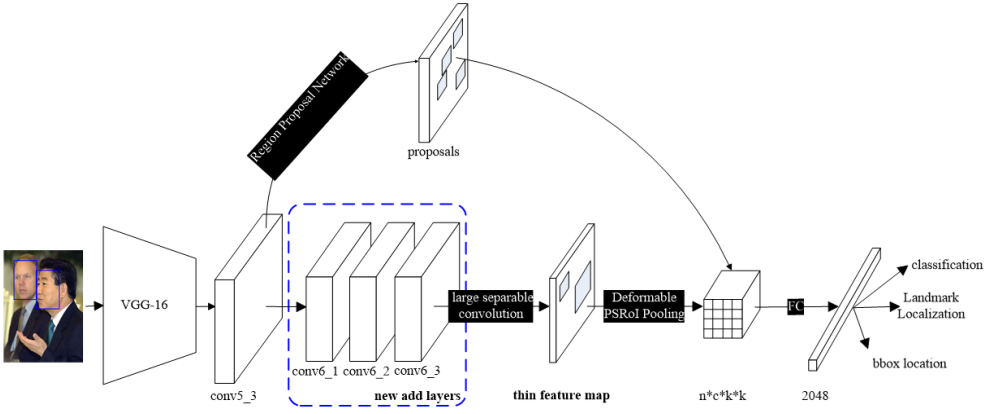


Figure 1: The R-MTCNN framework for joint face detection and alignment. Current methods jointly learning these two tasks mainly based on cascade structure which means the time consumed by the algorithm is related to the number of faces in the image. To solve this problem, we joint training these two task via region-based method and got novel results.

each. Then we apply large separable convolution layers [26, 38, 46] on conv6_3 to reduce channels from 1024 to 490. For an entire image, R-MTCNN firstly pass conv layers from conv1_1 to conv6_3,the conv5_3 layer follow a RPN [42] head to produce face regions. Then we use deformble PSRoI Pooling [10] to pool the face proposals to fixed size on the conv6_3 feature map. At last, we process the fixed size face proposals feature with a fully connected layer with 2048 channels.

Theoretically, we can directly using RoI Pooling [13] or RoI Align [16] without using deformable PSRoI Pooling. However, the experiments' results in this paper show that the performance of landmark localization by directly using RoI Pooling or RoI Align is not ideal, this may due to the feature got by RoI Pooling or RoI Align that can not represent the information required by landmark localization. The defromable PSRoIPooling will learn an offset for every bin in the pooling process. This offset will be more inclined to learn more sensitive feature information for the alignment task under the supervision of landmark localization. We have ablation study about different RoI cropping methods' influence on landmark localization in section 4.5.

Contributions. We propose a region-based method called R-MTCNN to simultaneous detect face and facial landmarks. Unlike cascade-based methods, R-MTCNN's detection time has nothing to do with the number of faces. In addition, we find the multi-task learning effectively improve the single task's performance. Deformable PSRoI Pooling is designed for enhancing CNNs' capability of modeling geometric transformations in object detection. In this paper we find it also gets significant improvements on landmark localization. Finally, our joint deep architecture achieves state-of-the-art result on the challenging WIDER FACE dataset [58] as well as FDDB dataset [17] and AFLW dataset [21].

2 Related Work

Multi-task learning has a wide range of concerns recently. A joint cascade-based method was recently proposed in [**?**] for simultaneously detecting faces and facial landmarks on a given image. This method yields improved detection performance by incorporating a face alignment step in the cascade structure. Technical report [41] also combine face detection with the tasks of locating facial landmarks and recognizing gender. MTCNN [60] proposes a multi-task cascaded CNNs based framework for joint face detection and alignment. These work parts have achieved good results in practical applications. However, they still have to firstly detect face regions which get a time-consumed problem. Mask R-CNN [16] uses a region-based method to efficiently detect objects in an image while simultaneously generating a high-quality segmentation mask for each instance. Driven by the novel results of Mask R-CNN, we try joint face detection and alignment in a region-based method.

Face detection. Viola-Jones detector [49] combines adaboost, Haar feature and cascade structure which gets huge success in face detection field. This method has a profound effect on many follow-up work such as new boosting algorithms [2, 39], new hand-craft features [28, 56] and new cascade structures [1, 25]. Deformable part model [12] for face detection trait face as a collection of parts. These methods based on structure models achieve better performance. However, it has been shown that hand-craft features can not represents the discriminative information of face when come across to large pose, illumination and other extreme conditions. CNN-based method comes up to overcome these problems. The earliest CNN-based face detector can be traced back to 1994. Vaillant et al. [48] trained CNN for detecting face in a sliding windows manner. Until the end of 2013, R-CNN [14] achieves landmark progress in the field of object detection, and then CNN-based face detection technology has been developing by leaps and bounds. CascadeCNN [24] gets powerful discriminative capability and high performance built on CNNs. UnitBox [18] introduces a new intersection-over-union loss function. CMS-RCNN [64] uses Faster R-CNN [42] in face detection with body contextual information. S^3FD [61] proposing a scale-equitable face detection framework to handle different scales of faces well. SSH [36] simultaneously detect faces with different scales in a single forward pass of the network.

landmark localization. Facial landmark detection has been proved an important step in face recognition system, it has been shown that a aligned face can effective boost recognition accuracy. Current landmark detection method mainly divided into two categories. One is regression-based methods [5, 19, 47, 51, 52, 54] and another is model-based [8, 27, 35] methods.While the former learns the shape increment given a mean initial shape, the latter trains an appearance model to predict the keypoint locations. CNN-based landmark localization methods have also been proposed in recent years [22, 45, 53] and have achieved remarkable performance.

3 Our Approach

We propose a new method to simultaneous detection face and facial landmarks called R-MTCNN as show in Figure 1. R-MTCNN adopts a region-based method, it generates face regions in the fist stage and in the second stage it does face alignment task. The two stage shares the same basic network where the computation in the second stage is relatively small. R-MTCNN is easy to implement by extent to current region-based detection methods [26, 42]. Next we will go into details of our approach in this section.

Method	VGG-16 backbone	VGG-19 backbone
Recall with 100 false positive on FDDB	90.85%	92.25%
Normalized Mean Error on AFLW	19.72%	9.72%

Table 1: Compare different backbone architecture's effect on face detection and alignment.

3.1 Overview

R-MTCNN adopts current popular region-based object detection strategy [42] that consists of two stages where shares a same network. The fist stage is Region Proposal Network(RPN). Given an entire image of arbitrary size. RPN generate face regions takes a sliding-window way that use features from conv5_3. After the face regions given in RPN, the R-MTCNN architecture is designed to classify wether it is a face, fine regressing the face bounding box and regressing the landmark location. In the last convolution layer conv6_3, if we directly perform RoI Pooling as mentioned in [26] on this layer it will get a heavy head and lead a low inference speed. We apply large separable convolution layers [26] on conv6_3 where reduce the channels of conv6_3 by 1024 to 490. While reducing the dimension, we apply a deformable PSRoI Pooling [10] to pool the face regions to a fixed-length feature vector. Each feature vector is further processed by a single fully-connected layer with 2048 channels(no drop out), followed by three sibling fully connected layer to predict RoI classification, regression and landmark location.

3.2 Backbone architecture

We do some modifiers on VGG-16 [44]. Firstly, in the conv5_3(stride=16) we apply a max pooling layer with stride=2. Following the max pooling layer, we connect three convolution layers of conv6_1,conv6_2,conv6_3. These new add layers all set with kernel_size=3, pad=1 and channels=1024. Follow by conv6_3 we connect two fully connected layers with 4096 channels. After every convolution or a fully connected layer, we deploy the Rectified Layer Unit (ReLU) non-linearity. Then we pre-trained this new network(VGG-19) on ImageNet [43] and get Top-1 accuracy with 73.5%.

We do not use VGG-19 directly as our basic network. We reduce VGG-19 effective stride from 32 pixels to 16 pixels, increasing the score map resolution. All layers before and on the conv5_3 (stride=16) are unchanged; We remove the max pooling layer followed by conv5_3. All convolutional filters on the new added convolution layers are modified by the "hole algorithm" [7, 30]("*Algorithme à trous*" [32]) to compensate for the reduced stride. The RPN is computed on top of the conv5_3 (that are shared with R-MTCNN). In R-MTCNN, we replace the two 4096 fully connected layer with a single 2048 fully connected layer. Because there are no parameters pre-trained in this layer, we use the "xavier" method [15] to randomly initialize the parameters with standard deviation 0.01. Table 1 shows the ablation results of R-MTCNN using different backbone. R-MTCNN adopts VGG-19 as backbone architecture got 1.4 points' improvement on face detection while amazing 10 points' improvement on AFLW.

3.3 RoI cropping

RoI cropping is a critical step in region-based object detection. It converts an input rectangular region of arbitrary size into fixed size features. Follow we will introduce currently popular RoI cropping methods. In this paper, we detailed explore the different RoI cropping method's influence on our multi-task learning.

RoI Pooling [13] RoI Pooling is used in all region proposal based object detection methods for cropping RoI. Given the input feature map and a RoI of size $w \times h$, RoI Pooling divides the RoI into $k \times k$(here we adopt 7×7) bins of approximate size $\frac{w}{k} \times \frac{h}{k}$ and then max-pooling or average-pooling the values in each bin into corresponding output grid cell. RoI Pooling gets a intensive computation especially when the number of object proposals is large because pooling is applied independently to each feature map channel.

Position-Sensitive (PS) RoI Pooling [9] In R-FCN [9], The last convolutional layer produces a bank of k^2 position-sensitive score maps for each category, and thus has a $k^2(C+1)$-channel output layer with C object categories (+1 for background). The bank of k^2 score maps correspond to a $k \times k$ spatial grid describing relative positions. PSRoI Pooling operates on this conv layer, it aggregates the outputs of the last convolutional layer and generates scores for each RoI. Unlike RoI Pooling, position-sensitive RoI layer conducts selective pooling, and each of the $k \times k$ bin aggregates responses from only one score map out of the bank of $k \times k$ score maps. With end-to-end training, this RoI layer shepherds the last convolutional layer to learn specialized position-sensitive score maps. In our implementation, we refer to [26] and set the (C +1) as 10 without limited to the number of categories.

Deformable PSRoI Pooling [10] It adds an offset to each bin position in the regular bin partition of the previous PSRoI pooling, the offsets are learned from the preceding feature maps and the RoIs, enabling adaptive part localization for objects with different shapes. While the offset is typically fractional, we get the pixels via bilinear interpolation.

RoI Align [16] RoI Pooling has quantitative problems. It first quantizes a floating-number RoI to the discrete granularity of the feature map, this quantized RoI is then subdivided into spatial bins which are themselves quantized, and finally feature values covered by each bin are aggregated (usually by max pooling). To address this, RoIAlign layer removes the harsh quantization of RoI Pooling, properly aligning the extracted features with the input. RoI Align avoid any quantization of the RoI boundaries or bins by using bilinear interpolation to compute the exact values of the input features at four regularly sampled locations in each RoI bin, and aggregate the result.

In our experiments, we find deformable PSRoI Pooling gets novel results on landmark localization. The defromable PSRoIPooling learn an offset for every bin in the pooling process. We think the offset will be more inclined to learn more sensitive feature information for the alignment task under the supervision of landmark localization. Figure 2 shows the visual results of different RoI cropping methods. Deformable PSRoI Pooling is more robust to extreme environment such as large pose, low resolution etc.

3.4 Training

The RPN is trained end-to-end by back-propagation and stochastic gradient descent (SGD). We follow the "image-centric" sampling strategy from [13] to train this network. Each mini-batch arises from a single image that contains many positive and negative example anchors. We set seven aspect ratios $\{1:3, 1:2, 1:1.5, 1:1, 1.5:1, 2:1, 3:1\}$ and six scales $\{2^2, 3^2, 5^2, 9^2, 16^2, 32^2\}$ to control the anchor for covering faces of different shapes.

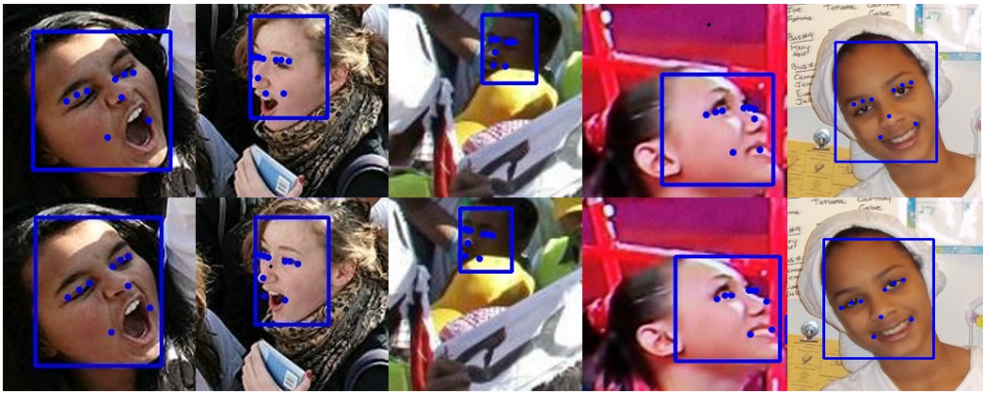


Figure 2: Visual results of deferent RoI cropping methods. The top row is the effect of RoI Align. The bottom row is the effect of deformable PSRoI Pooling. Deformable PSRoI Pooling is more robust to extreme environment such as large pose, low resolution etc.

With region proposals generated by RPN, it is easy to end-to-end train the R-MTCNN architecture. We define a multi-task loss on each sampled RoI in Eq. 1.

$$L(p,u,t^u,t,l^u,l^*) = L_{cls}(p,u) + \lambda[u \geq 1]L_{box}(t^u,t) + \varphi[u \geq 1]L_{landmark}(l^u,l,t). \quad (1)$$

Here u is the RoI's ground-truth label ($u = 0$ means background). As usual, p is computed by a softmax over the two outputs(face or not face) of a fully connected layer. $L_{cls}(p,u) = -logp_u$ is log loss for true class u. $L_{box}(t^u,t)$ is the bounding box regression loss as defined in [13]. t^u represents the ground truth bounding-box regression target and t is a box predicted tuple. $[u \geq 1]$ is an indicator which equals to 1 if the argument is true and 0 otherwise.

$L_{landmark}(l^u,l,t)$ is defined for landmark location task. l^u is ground-truth landmark coordinates which can be represent by $(x_1^u,y_1^u,...,x_N^u,y_N^u)$ and l is the landmark location predicted tuple $(\hat{x_1},\hat{y_1},...,\hat{x_N},\hat{y_N})$, N represents the number of landmarks, in this paper, we perform nine landmarks detection. The predicted box of t characterize the region by $\{t_x,t_y,t_w,t_h\}$, where (t_x,t_y) are the coordinates of the center of the region and t_w,t_h are the width and height of the region respectively. Each landmark point is shifted with respect to the region center (t_x,t_y), and normalized by (t_w,t_h) as given by Eq. 2.

$$(c_i^x,c_i^y) = \left(\frac{x_i^u - t_x}{t_w}, \frac{y_i^u - t_y}{t_h}\right), i \in [1,N], \quad (2)$$

The (c_i^x,c_i^y)'s are treated as labels for training the landmark localization task using the $smooth_L1$ loss defined in [13]. The landmark localization loss is computed from Eq. 3.

$$L_{landmark} = \sum_{i=1}^{N}(smooth_L1(\hat{x_i} - c_i^x) + smooth_L1(\hat{y_i} - c_i^y)). \quad (3)$$

The hyper-parameter λ,φ in Eq. 1 controls the balance between the three task losses. We set the balance weight $\lambda = \varphi = 1$.We define positive examples as the RoIs that have

intersection-over-union (IoU) overlap with a ground-truth box of at least 0.5, and negative otherwise.

We use a weight decay of 0.0005 and a momentum of 0.9. We use multi-scale training: the short side of image is randomly resized to one of [416, 448, 480, 512, 544, 576, 608, 640, 672, 704, 736, 768, 800, 832, 864] for every iteration. We train the model with 1 GPU which hold 1 image each iteration and selects 128 RoIs for backprop during training. We fine-tune R-MTCNN using a learning rate of 0.001 for 100k mini-batches and 0.0001 for next 20k mini-batches on WIDER FACE. To have R-MTCNN share features with RPN (Figure 1), we adopt the 4-step alternating training in [42], alternating between training RPN and training R-MTCNN.

4 Experiments

In this section we perform a detailed experimental analysis of our R-MTCNN network. Firstly, we evaluate the influence of multitasking on single tasks. Then we compare our algorithm with current state-of-the-art face detection and landmark localization on tree popular benchmarks. Finally we conduct the ablation study to explore different RoI cropping method's influence to the multi-task learning.

4.1 Training and Testing Setup

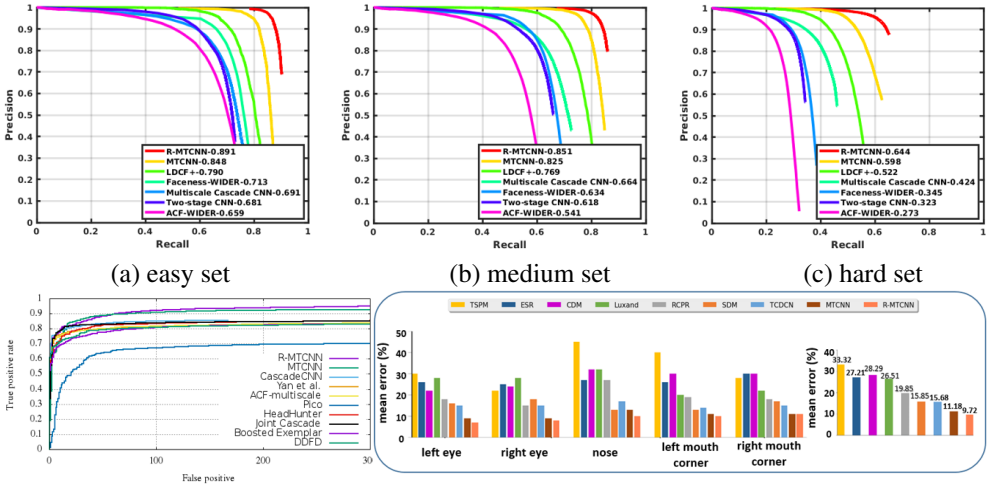
WIDER FACE [53] is by far the largest face detection database. This dataset contains 32, 203 images with 393, 703 annotated faces, 158, 989 of which are in the train set, 39, 496 in the validation set and the rest are in the test set. The validation and test set are divided into "easy", "medium", and "hard" subsets cumulatively (i.e. the "hard" set contains all images). This is one of the most challenging public face datasets mainly due to the wide variety of face scales and occlusion.

In this paper, we label all faces whose short side's pixels is larger than 20 with nine landmarks in WIDER FACE's training set. The faces without landmark labeled will not involved in training. However, this could trigger the issue of false positives. So we mask the faces whose short side's pixels is less than 20 by gaussian initialization. We obey two basic principle to label the faces. (1) For the occlusion situation, such as wearing sunglasses, crowded crowds whose faces are blocked by others etc, we need to estimate the feature points of the occluded locations and mark them. (2) For a large pose face, if the rotation angle is 90 degrees, the position of the landmark cannot be marked by the estimation, we simply mark the landmark's location on the edge of the face contour. The relabeled WIDER FACE's training set as shown in . We train all models on this data set and evaluate the face detection task on the validation sets.

Face Detection Dataset and Benchmark (FDDB) [17] contains 2845 images and 5171 annotated faces. We use this dataset only for testing. Annotated Facial Landmarks in the Wild (AFLW) [21] for evaluating landmark localization.

4.2 Joint face detection and alignment

In order to verify the effect of landmark localization on face detection. We only removed R-MTCNN's landmark localization task and the the network backbone remains the same. We compare this face detection network with R-MTCNN on WIDER FACE validation data sets



(a) easy set (b) medium set (c) hard set

(d) FDDB discrete score (e) Normalized Mean Error

Figure 3: Comparison to state-of-the-arts. (a-c) Comparisons on WIDER FACE. (d) Comparisons on FDDB. (e) Comparisons on AFLW for landmark localization.

and FDDB. Figure[...] shows that we got a beneficial for the multi task learning especially we have a 6 point promotion on WIDER hard datasets.

4.3 Face detection

We evaluate our methods by comparing with the state-of-the-art methods MTCNN [60], LDCF+ [37], Faceness [57], Multitask Cascade CNN, Two-stage CNN [58] and ACF-WIDER [56] on the WIDER FACE validation set. We inference SSH without an image pyramid, we rescale the shortest side of the image up to 1200 pixels while keeping the largest side below 1600 pixels without changing the aspect ratio. Figure 3 (a-c) shows that R-MTCNN outperforms other methods on all the subsets of WIDER FACE by a margin.

On the FDDB benchmark, we resize the shortest side of the input to 600 pixels while keeping the larger side less than 1024 pixels. We compare R-MTCNN with MTCNN [60], CascadeCNN [24], Yan et al. [55], ACF-multiscale [56], Pico [33], HeadHunter [34], Joint Cascade [6], Boosted Exemplar [23], DDFD [11] in FDDB . Figure 3 (d) shows that our methods achieved excellent results. R-MTCNN recalls 92.25% faces with 100 false positive while outperforming the second face detector by 1.45 points.

4.4 Landmark localization

We randomly sample 1000 images in AFLW for testing the ablility for landmark localization. We resize the shortest side of the input to 600 pixels while keeping the larger side less than 1024 pixels. and compare our landmark localization performance with the follow methods: MTCNN [60], PCPR [3], TSPM [40], luxand face SDK [51], ESR [4], CDM [59], SDM [53], and TCDCN [62]. In the testing phase of MTCNN, we directly crop the faces in the image and treat them as the input for O-Net. The mean error is measured by the distances

between the estimated landmarks and the ground truths, and normalized with bounding box. Figure 3 (e) show that our method outperforms all the state-of-the-art methods with a margin.

4.5 Ablation study

...

5 Conclusion

In this paper, we have proposed a region-based multi-task convolution networks for joint face detection and alignment. Unlike cascade-based approaches, R-MTCNN's time consumed is independent with the number of faces. And we find that multi-task learning effectively help single task learning. Deformable PSRoI Pooling is designed for enhancing CNNs' capability of modeling geometric transformations in object detection. We find it also gets significant improvements on landmark localization task. R-MTCNN achieves state-of-the-art performance on the challenging WIDER dataset as well as FDDB and AFLW. In this paper, we only perform nine landmarks detection. In the future, we will further explore more landmarks detection task such as 68 landmarks' localization.

References

[1] L. Bourdev and J. Brandt. Robust object detection via soft cascade. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, pages 236–243, 2005.

[2] S. Charles Brubaker, Jianxin Wu, Jie Sun, Matthew D. Mullin, and James M. Rehg. On the design of cascades of boosted ensembles for face detection. *International Journal of Computer Vision*, 77(1-3):65–86, 2008.

[3] Xavier P. Burgosartizzu, Pietro Perona, and Piotr Dollar. Robust face landmark estimation under occlusion. In *IEEE International Conference on Computer Vision*, pages 1513–1520, 2014.

[4] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression, 2014.

[5] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[6] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122, 2014.

[7] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *Computer Science*, (4):357–361, 2014.

[8] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

Figure 4: Qualitative results of R-MTCNNp on the validation set of the WIDER dataset. R-MTCNN is robust to extreme conditions such as large pose, light and occlusion.

[9] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. 2016.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. pages 764–773, 2017.

[11] Sachin Sudhakar Farfade, Mohammad J. Saberian, and Li Jia Li. Multi-view face detection using deep convolutional neural networks. pages 643–650, 2015.

[12] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[13] Ross Girshick. Fast r-cnn. *Computer Science*, 2015.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[15] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9:249–256, 2010.

[16] Kaiming He, Georgia Gkioxari, Piotr DollÍćr, and Ross Girshick. Mask r-cnn. 2017.

[17] Vidit Jain and Erik Learned-Miller. *FDDB: A Benchmark for Face Detection in Unconstrained Settings*. 2010.

[18] Yuning Jiang, Yuning Jiang, Zhimin Cao, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM on Multimedia Conference*, pages 516–520, 2016.

[19] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, United States, 23 June 2014 through 28 June 2014*, pages 1867–1874. IEEE Computer Society, 2014.

[20] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. pages 2034–2043, 2017.

[21] Martin K?stinger, 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 *IEEE International Conference on Computer Vision Workshops*, pages 2144–2151, 2012.

[22] Amit Kumar, Rajeev Ranjan, Vishal Patel, and Rama Chellappa. Face alignment by local deep descriptor regression. *arXiv preprint arXiv:1601.07950*, 2016.

[23] Haoxiang Li, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Gang Hua. Efficient boosted exemplar-based face detection. In *Computer Vision and Pattern Recognition*, pages 1843–1850, 2014.

[24] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Computer Vision and Pattern Recognition*, pages 5325–5334, 2015.

[25] Stan Z. Li, Long Zhu, Zhen Qiu Zhang, Andrew Blake, Hong Jiang Zhang, and Harry Shum. Statistical learning of multi-view face detection. In *European Conference on Computer Vision*, pages 67–81, 2002.

[26] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-head r-cnn: In defense of two-stage object detector. 2017.

[27] Lin Liang, Rong Xiao, Fang Wen, and Jian Sun. Face alignment via component-based discriminative search. In *European conference on computer vision*, pages 72–85. Springer, 2008.

[28] Shengcai Liao, Anil K. Jain, and Stan Z. Li. *A Fast and Accurate Unconstrained Face Detector*. IEEE Computer Society, 2016.

[29] Yu Liu, Hongyang Li, Junjie Yan, Fangyin Wei, Xiaogang Wang, and Xiaoou Tang. Recurrent scale approximation for object detection in cnn. pages 571–579, 2017.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[31] Inc Luxand. Luxand facesdk. *http://www.luxand.com/facesdk/*.

[32] Stĺęphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press,, 1999.

[33] Nenad Marku?, Miroslav Frljak, Igor S Pand?i?, J?rgen Ahlberg, and Robert Forchheimer. Object detection with pixel intensity comparisons organized in decision trees. *Computer Science*, 14(4):2657–62, 2013.

[34] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. 8692(19):720–735, 2014.

[35] Iain Matthews and Simon Baker. Active appearance models revisited. *International journal of computer vision*, 60(2):135–164, 2004.

[36] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S. Davis. Ssh: Single stage headless face detector. pages 4885–4894, 2017.

[37] Eshed Ohn-Bar and Mohan M Trivedi. To boost or not to boost? on the limits of boosted trees for object detection. pages 3350–3355, 2017.

[38] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters ął improve semantic segmentation by global convolutional network. 2017.

[39] Minh Tri Pham and Tat Jen Cham. Fast training and selection of haar features using statistics in boosting-based face detection. In *IEEE International Conference on Computer Vision*, pages 1–7, 2007.

[40] Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.

[41] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, PP (99):1–1, 2016.

[42] S. Ren, K. He, R Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*, 39(6):1137–1149, 2017.

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *Computer Science*, 2014.

[45] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3476–3483. IEEE, 2013.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *Computer Science*, pages 2818–2826, 2015.

[47] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[48] R Vaillant, C Monrocq, and Y Le Cun. Original approach for the localisation of objects in images. *Vision, Image and Signal Processing, IEE Proceedings* -, 141(4):245 – 250, 1994.

[49] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[50] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces. 2017.

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

[52] Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015.

[53] Xuehan Xiong and Fernando De La Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition*, pages 532–539, 2013.

[54] Junjie Yan, Zhen Lei, Dong Yi, and Stan Z Li. Learn to combine multiple hypotheses for accurate face alignment. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on*, pages 392–396. IEEE, 2013.

[55] Junjie Yan, Zhen Lei, Longyin Wen, and Stan Z Li. The fastest deformable part model for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2497–2504, 2014.

[56] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Aggregate channel features for multi-view face detection. pages 1–8, 2014.

[57] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *IEEE International Conference on Computer Vision*, pages 3676–3684, 2015.

[58] Shuo Yang, Ping Luo, Change Loy Chen, and Xiaoou Tang. Wider face: A face detection benchmark. In *Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

[59] 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 *IEEE International Conference on Computer Vision*, pages 1944–1951, 2014.

[60] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[61] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S^3fd: Single shot scale-invariant face detector. pages 192–201, 2017.

[62] Zhanpeng Zhang, Ping Luo, Change Loy Chen, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108, 2014.

[63] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[64] Chenchen Zhu, Yutong Zheng, Khoa Luu, and Marios Savvides. Cms-rcnn: Contextual multi-scale region-based cnn for unconstrained face detection. 2017.