# Single-stage Joint Face Detection and Alignment

Jiankang Deng * [1,2,3] Jia Guo * [2] Stefanos Zafeiriou[1,3]
[1]Imperial College London [2]InsightFace [3]FaceSoft
{j.deng16, s.zafeiriou}@imperial.ac.uk , guojia@gmail.com

## Abstract

*In practice, there are huge demands to localize faces in images and videos under unconstrained pose variation, illumination change, severe occlusion and low resolution, which pose a great challenge to existing face detectors. This challenge report presents a single-stage joint face detection and alignment method. In detail, we employ feature pyramid network, single-stage detection, context modelling, multi-task learning and cascade regression to construct a practical face detector. On the Wider Face Hard validation subset, our single model achieves state-of-the-art performance (92.0% AP) compared with both academic and commercial face detectors for detecting unconstrained faces in cluttered scenes. In the Wider Face AND PERSON CHALLENGE 2019, our ensemble model achieves 56.66% average AP (runner-up) in the face detection track. To facilitate further research on the topic, the training code and models have been provided publicly available.*

## 1. Introduction

Automatic face detection underpins facial image analysis and includes, but is not limited to, face alignment [6, 11, 23, 7, 5], face tracking [22], facial attribute recognition, facial expression recognition, and face recognition [3, 8] and verification. There has been considerable interest in face detection research over last the two decades, and a number of accurate and efficient algorithms have been proposed to detect faces in largely uncontrolled settings. However, face detection remains a challenge in uncontrolled situations (*e.g*. the Wider Face dataset [24]) due to complex backgrounds, pose variations, occlusions, extreme illuminations, out-of-focus blurring, and low resolution.

In this technical report, we propose a single-stage joint face detection and alignment method, a conceptually simple and effective approach based on valid existing techniques (*e.g*. feature pyramid network [15], single-stage detector [25], context modelling [18, 2], multi-task learning [12, 4] and cascade regression [14]). Extensive experiments on the Wider Face dataset show that the proposed method achieves state-of-the-art performance compared with both academic and commercial face detectors for detecting unconstrained faces in cluttered scenes. In the WIDER FACE AND PERSON CHALLENGE 2019, our ensemble model achieves 56.66% average AP (runner-up) in the face detection track.

## 2. Our Method

**Overview.** In Fig. 1, we show the framework of the proposed single-stage joint face detection and alignment approach. There are three main components: feature pyramid network, context head module and cascade multi-task loss. First, The feature pyramid network gets the input face images and outputs five scale feature maps. Then, the context head module gets a feature map of a particular scale and calculates the multi-task loss (*e.g*. classification loss, box regression loss and five facial landmark regression loss). The first context head module predicts the bounding box from the regular anchor. Afterwards, the second context head module predicts more accurate bounding box from the regressed anchor, which is generated by the first context head module. The proposed single-stage joint face detection and alignment method employs fully convolutional neural networks, thus it can be easily trained in an end-to-end way.

**Feature Pyramid.** Our method employs feature pyramid levels from $P_2$ to $P_6$, where $P_2$ to $P_5$ are computed from the output of the corresponding ResNet residual stage ($C_2$ through $C_5$) using top-down and lateral connections as in [15, 16]. $P_6$ is calculated through a $3\times3$ convolution with stride=2 on $C_5$. $C_1$ to $C_5$ is a pre-trained ResNet-152 [13] classification network on the ImageNet-11k dataset while $P_6$ are randomly initialised with the “Xavier” method [10].

**Context Module.** Inspired by SSH [18] and PyramidBox [20], we also apply independent context modules on five feature pyramid levels to increase the receptive field and enhance the rigid context modelling power. Drawing lessons from the champion of the Wider Face Challenge 2018 [17], we also replace all $3 \times 3$ convolution layers within the lateral connections and context modules by the deformable convolution network (DCN) [2, 26], which further strengthens the non-rigid context modelling capacity.

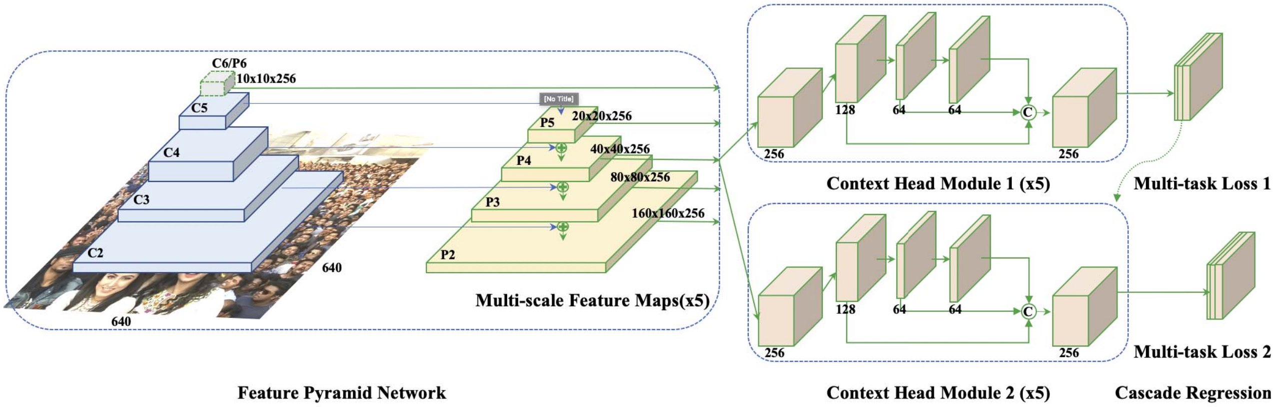


Figure 1. An overview of the proposed single-stage joint face detection and alignment approach. Our method is designed based on the feature pyramid network with independent context head module for each scale. "C" in the context head module denotes concatenation operation and the number around the blob indicates the channel number. Following the context module, we calculate a multi-task loss for each anchor. Cascade regression is used to further improve the detection results. The first context head module predicts the bounding box from the regular anchor, while the second context head module predicts more accurate bounding box from the regressed anchor, which is generated by the first context head module.

In our cascade regression framework, the proposed context head module with DCN can solve the mis-alignment problem in the single-stage face detector to some extend.

**Anchor Settings.** We employ scale-specific anchors on the feature pyramid levels from $P_2$ to $P_6$ like [21]. Here, $P_2$ is designed to capture tiny faces by tiling small anchors at the cost of more computational time and at the risk of more false positives. We set the scale step at $2^{1/3}$ and the aspect ratio at 1:1. With the input image size at $640 \times 640$, the anchors can cover scales from $16 \times 16$ to $406 \times 406$ on the feature pyramid levels. In total, there are 102,300 anchors, and $75\%$ of these anchors are from $P_2$.

**Multi-task Loss Head.** For any training anchor $i$, we minimize the following multi-task loss:

$$\begin{aligned} L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) \\ + \lambda_2 p_i^* L_{pts}(l_i, l_i^*). \end{aligned} \tag{1}$$

(1) Face classification loss $L_{cls}(p_i, p_i^*)$, (2) Face box regression loss $L_{box}(t_i, t_i^*)$, and (3) Facial landmark regression loss $L_{pts}(l_i, l_i^*)$. The loss-balancing parameters $\lambda_1$ and $\lambda_2$ are set to 0.25 and 0.1, respectively.

**Anchor Matching.** For the first head module, anchors are matched to a ground-truth box when IoU is larger than $0.7$, and to the background when IoU is less than $0.3$. For the second head module, anchors are matched to a ground-truth box when IoU is larger than $0.5$, and to the background when IoU is less than $0.4$. Unmatched anchors are ignored during training. Since most of the anchors ($> 99\%$) are negative after the matching step, we employ standard OHEM [19, 25] to alleviate significant imbalance between the positive and negative training examples. More specifically, we sort negative anchors by the loss values and select the top ones so that the ratio between the negative and positive samples is at least 3:1.

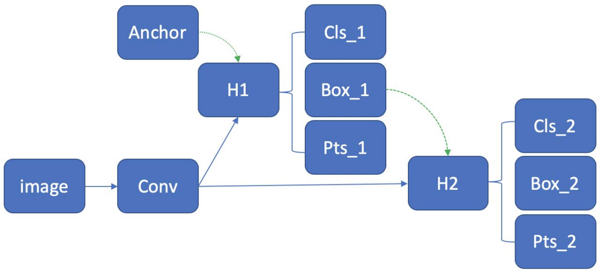


Figure 2. Cascade regression for single-stage face detector. "H1" and "H2" denotes the context head modules in Fig. 1, respectively. The first context head module predicts the bounding box from the regular anchor, while the second context head module predicts more accurate bounding box from the regressed box predicted by the first context head module.

**Cascade Regression.** As shown in Fig. 2, the first context head module predicts the bounding box from the regular anchor. Then, the second context head module predicts more accurate bounding box from the regressed box, which is generated by the first context head module.

## 3. Experiments

### 3.1. Dataset

The Wider Face dataset [24] consists of $32,203$ images and $393,703$ face bounding boxes with a high degree of variability in scale, pose, expression, occlusion and illumination. The Wider Face dataset is split into training ($40\%$), validation ($10\%$) and testing ($50\%$) subsets by randomly sampling from 61 scene categories. Based on the detection rate of EdgeBox [27], three levels of difficulty (*i.e.* Easy, Medium and Hard) are defined by incrementally incorporating hard samples. In RetinaFace [4], five facial landmarks (*i.e.* eye centres, nose tip and mouth corners) are annotated on the Wider Face training and validation subsets. In total, $84.6k$ faces on the training set and $18.5k$ faces on the validation set are annotated with five facial landmarks.

### 3.2. Implementation details

**Data Augmentation.** Since there are around 20% tiny faces in the Wider Face training set, we follow [25, 20] and randomly crop square patches from the original images and resize these patches into $640 \times 640$ to generate larger training faces. More specifically, square patches are cropped from the original image with a random size between [0.3, 1] of the smaller dimension of the original image. For the faces on the crop boundary, we keep the overlapped part of the face box if its centre is within the crop patch. Besides random crop, we also augment training data by random horizontal flip with the probability of 0.5 and photometric colour distortion [25].

**Training Details.** We train our method using SGD optimizer (momentum at 0.9, weight decay at 0.0005, and batch size of $8 \times 4$) on four NVIDIA Tesla P40 (24GB) GPUs. The learning rate starts from $10^{-3}$, rising to $10^{-2}$ after 5 epochs, then divided by 10 at 55 and 68 epochs. The training process terminates at 80 epochs. Our implementation is on MXNet [1].

**Testing Details.** For testing on Wider Face, we follow the standard practices of [18, 25] and employ flip as well as multi-scale (the minimum image size at $[500, 800, 1100, 1400, 1700]$) strategies. Box voting [9] is applied on the union set of predicted face boxes using an IoU threshold at 0.4. For the challenge submission, we save box locations by the format of float instead of int.

**Model Ensemble.** We train four face detection models and get the multi-scale test results from each model. Then, we get the final ensemble results by box voting.

### 3.3. Ablation Study

To achieve a better understanding of the proposed method, we conduct extensive ablation experiments to examine how the five facial landmarks, the cascade regression, and the ensemble strategy quantitatively affect the performance of face detection. Besides the standard evaluation metric of average precision (AP) when IoU=0.5 on the Easy, Medium and Hard subsets, we also make use of the test server (Hard test subset) of the Wider Face Challenge 2019, which employs a more strict evaluation metric of average AP for IoU=0.5:0.05:0.95, rewarding more accurate face detectors. Please note that the participant can only have five submissions to the test server. Thus, we only report the most important improvements in Tab. 1.

As illustrated in Tab. 1, we evaluate the performance of several different settings on the Wider Face validation set and report the average AP on the Hard test subset. By applying the valid techniques (*i.e.* FPN, context module, and deformable convolution), we set up a strong baseline (54.76%). Adding the branch of five facial landmark regression significantly improves the average AP (0.48%) on the Hard subset, suggesting that landmark localization is crucial for improving the accuracy of face detection. By using cascade regression, average AP on the Hard test subset further improves to 56.05%. After model ensemble, our final average AP approaches 56.66%, ranking 2nd on the challenge leader-board.

| Method | Easy | Medium | Hard | average AP |
|---|---|---|---|---|
| Baseline | 96.349 | 95.833 | 91.286 | 54.76 |
| +Landmark | 96.467 | 96.075 | 91.694 | 55.24 |
| +Cascade | 96.670 | 95.841 | 92.064 | 56.05 |
| +Ensemble | 96.895 | 96.315 | 92.101 | 56.66 |

Table 1. Ablation experiments of the proposed methods on the Wider Face validation subset and the Wider Face Hard test subset.

As shown in Fig. 3, we compare our best single model (FPN+context module+DCN+landmark+cascade) with other 25 state-of-the-art face detection algorithms. Our method produces the impressive APs in all subsets of the validation set, *i.e.*, 96.7% (Easy), 95.8% (Medium) and 92.0% (Hard). On the Easy and Medium subsets, our method achieves comparable performance with RetinaFace [4], which has another dense regression branch. On the Hard subset, our approach outperforms all these state-of-the-art methods and set up a new record.

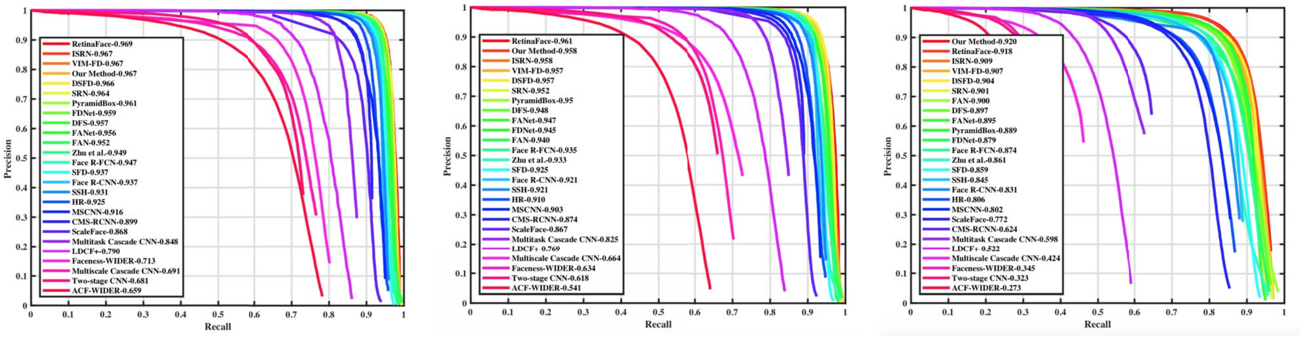
## 4. Conclusions

In this challenge report, we propose a single-stage joint face detection and alignment method. By employing feature pyramid network, single-stage detector, context modelling, multi-task learning and cascade regression, our method achieves state-of-the-art performance for detecting unconstrained faces in cluttered scenes. In the WIDER FACE AND PERSON CHALLENGE 2019, our ensemble model achieves 56.66% average AP (runner-up) in the face detection track.

## References

[1] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv:1512.01274*, 2015. 3

[2] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, 2017. 1



(a) Val: Easy

(b) Val: Medium

(c) Val: Hard

Figure 3. Precision-recall curves on the Wider Face validation subsets.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[4] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv:1905.00641*, 2019. 1, 3

[5] J. Deng, Q. Liu, J. Yang, and D. Tao. M3 csr: Multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47:19–26, 2016. 1

[6] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *IJCV*, 2018. 1

[7] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648, 2019. 1

[8] J. Deng, Y. Zhou, and S. Zafeiriou. Marginal loss for deep face recognition. In *CVPR Workshops*, pages 60–68, 2017. 1

[9] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, 2015. 3

[10] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 1

[11] J. Guo, J. Deng, N. Xue, and S. Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. *BMVC*, 2018. 1

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[14] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi. Consistent optimization for single-shot object detection. *arXiv:1901.06563*, 2019. 1

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1

[17] C. C. Loy, D. Lin, W. Ouyang, Y. Xiong, S. Yang, Q. Huang, D. Zhou, W. Xia, Q. Li, P. Luo, et al. Wider face and pedestrian challenge 2018: Methods and results. *arXiv:1902.06854*, 2019. 1

[18] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis. Ssh: Single stage headless face detector. In *ICCV*, 2017. 1, 3

[19] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 2

[20] X. Tang, D. K. Du, Z. He, and J. Liu. Pyramidbox: A context-assisted single shot face detector. In *ECCV*, 2018. 1, 3

[21] J. Wang, Y. Yuan, and G. Yu. Face attention network: an effective face detector for the occluded faces. *arXiv:1711.07246*, 2017. 2

[22] J. Yang, J. Deng, K. Zhang, and Q. Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 41–49, 2015. 1

[23] J. Yang, Q. Liu, and K. Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshops*, 2017. 1

[24] S. Yang, P. Luo, C.-C. Loy, and X. Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. 1, 2

[25] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, 2017. 1, 2, 3

[26] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. *arXiv:1811.11168*, 2018. 1

[27] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 3