RetinaFace: Single-shot Multi-level Face Localisation in the Wild

Jiankang Deng * [1,2,3] Jia Guo * [2] Evangelos Ververas[1,3]
Irene Kotsia[4] Stefanos Zafeiriou[1,3]
[1]Imperial College [2]InsightFace [3]FaceSoft [4]Middlesex University London
{j.deng16, e.ververas16, s.zafeiriou}@imperial.ac.uk
guojia@gmail.com, i.kotsia@mdx.ac.uk

Abstract

Though tremendous strides have been made in uncontrolled face detection, accurate and efficient 2D face alignment and 3D face reconstruction in-the-wild remain an open challenge. In this paper, we present a novel single-shot, multi-level face localisation method, named RetinaFace, which unifies face box prediction, 2D facial landmark localisation and 3D vertices regression under one common target: point regression on the image plane. To fill the data gap, we manually annotated five facial landmarks on the WIDER FACE dataset and employed a semi-automatic annotation pipeline to generate 3D vertices for face images from the WIDER FACE, AFLW and FDDB datasets. Based on extra annotations, we propose a mutually beneficial regression target for 3D face reconstruction, that is predicting 3D vertices projected on the image plane constrained by a common 3D topology. The proposed 3D face reconstruction branch can be easily incorporated, without any optimisation difficulty, in parallel with the existing box and 2D landmark regression branches during joint training. Extensive experimental results show that RetinaFace can simultaneously achieve stable face detection, accurate 2D face alignment and robust 3D face reconstruction while being efficient through single-shot inference.

1. Introduction

Automatic face localisation is a prerequisite for facial image analysis in many applications such as facial attribute analysis(*e.g.* expression [64] and age [41, 39]) and facial identity recognition [18, 12, 56]. A narrow definition of face localisation may refer to traditional face detection [54, 62], which aims at estimating the face bounding boxes without possessing any scale and position prior. Nevertheless, in this paper we refer to a broader definition of face localisation which includes face detection [43], face pose estimation [48, 60, 31, 5], face alignment [14, 57, 17, 16, 15, 58, 23], face segmentation [50, 34] and 3D face reconstruction [72, 1, 19, 70]. In Fig. 1, we show face localisation tasks with multiple levels of detail, from coarse to fine.

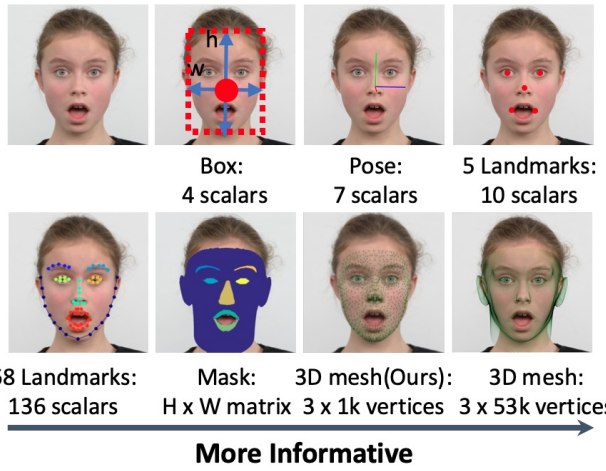


Figure 1. Face localisation tasks from coarse to fine. Face detection only predicts one center point and scales. Face pose estimation calculates the scale, 3D rotation and translation parameters. Sparse face alignment localises more semantic points. Face segmentation computes pixel-wise label maps for different semantic components (*e.g.* mouth, eyes). 3D face reconstruction can establish dense 3D correspondence for every pixel of a face, which is the most informative and demanding face localisation technique.

Typically, face pose estimation, face alignment, face segmentation and 3D face reconstruction are steps subsequent to face detection. These fine-grind face localisation tasks are performed on individual face crops and the computational complexity increases linearly with the number of faces in the input image. Since all face localisation tasks from face detection to 3D face reconstruction aim at establishing the semantic correspondence between different face images, with their main difference being only in the information level, the question that arises is if can we combine them into a united framework by jointly training and make

* Equal contributions.
InsightFace is a nonprofit Github project for 2D and 3D face analysis.

different tasks benefit from each other.

The training process for face detection usually contains classification and box regression losses [21]. Chen *et al*. [8] proposed to combine face detection and alignment in a joint cascade framework based on the observation that aligned face shapes can provide better features for face classification. Inspired by [8], MTCNN [66] and STN [7] simultaneously detected faces and five facial landmarks. Due to training data limitation, JDA [8], MTCNN [66] and STN [7] have not verified whether detection of tiny faces can benefit from the extra supervision of five facial landmarks.

In Mask R-CNN [25], the detection performance is significantly improved by adding a branch for predicting an object mask in parallel with the existing branch for bounding box classification and regression. This confirms that dense pixel-wise annotations are also beneficial for improving detection. In FAN [55], an anchor-level attention map is proposed to improve the occluded face detection. Nevertheless, the proposed attention map is quite coarse and does not contain semantic information. In MFN [6], a single end-to-end network is presented to jointly predict the bounding box locations and 3DMM parameters for multiple faces. This contributes to more precise face detection in-the-wild by leveraging both 2D information from bounding boxes and 3D information from 3DMM parameters. However, 3DMM parameter prediction constitutes an indirect regression target when compared to semantic point prediction (*e.g*. box center) on the image plane. In this paper, we explore joint learning for different face localisation tasks (face detection, 2D face alignment and 3D face reconstruction) based on the single-shot [40, 69, 51] framework. To overcome the limitation of training data [6], we have manually annotated five facial landmarks for $84.6k$ faces from the WIDER FACE training dataset [59]. In addition, we set up a semi-automatic annotation pipeline to generate 1k 3D vertices for $22k$ faces from the WIDER FACE dataset [59], $27.1k$ faces from the AFLW dataset [30], and $39.3k$ faces from the FDDB full image set [28]. Based on these training data, we propose an innovative, straightforward and effective 3D mesh regression method. More specifically, we directly regress x, y and z coordinates in the image space and add a regularization term to control the edge distance of triangles in the mesh for more accurate prediction of z coordinates.

Joint learning of face bounding box locations, five facial landmarks and 1k 3D vertices forces the network to learn exclusive facial features that characterize face pose, shape, and expression, in addition to differentiating face regions from the background. As five facial landmarks localisation and 3D vertices regression both target on predicting semantic points on the image plane, face box prediction benefits from joint learning and becomes more accurate and stable, producing less false positives. Also, as the annotated but challenging face detection data [59] are employed in the joint training of face detection and the rest fine-grind face localisation tasks (for which usually less challenging datasets are employed, *e.g*. [30]), they directly contribute to robust 3D mesh regression.

To summarise, our key contributions are:

- We integrate face bounding box prediction, 2D facial landmark localisation and 3D vertices regression under a unified multi-level face localisation task with a common goal: point regression on the image plane.
- Based on a single-shot inference, we propose a mutually beneficial learning strategy to train a unified multi-level face localisation method that simultaneously predicts face bounding boxes, five 2D facial landmarks, and 1k 3D vertices.
- Our method achieves state-of-the-art performance in face detection and 2D face alignment, as well as robust 3D face reconstruction with single-shot inference.

2. Related Work

Face Detection. Inspired by generic object detection methods [21, 46, 38, 44, 45, 35, 36], face detection has recently achieved remarkable progress [27, 40, 69, 10, 51]. Different from generic object detection, face detection features smaller ratio variations (from 1:1 to 1:1.5) but much larger scale variations (from several pixels to thousands of pixels). The most recent state-of-the-art methods [40, 69, 51] focus on single-shot design [38, 36] which densely samples face locations and scales on feature pyramids [35], demonstrating promising performance and yielding faster inference compared to two-stage methods [46, 63, 10]. Following this route, we improve the performance of single-shot face detection by exploiting extra-supervisions from multi-level face localisation tasks.

3D Face Reconstruction. Building dense pixel-to-pixel correspondence is one of the most fundamental problems in 3D face reconstruction from 2D images. Recently, a lot of works follow the approach of regressing 3DMM parameters from 2D images using CNNs [29, 72, 53, 47, 24, 52]. Jourabloo *et al*. [29] employs a cascade of CNNs to alternately regress the shape and pose parameters. 3DDFA [72] utilizes cascade iterations on a single CNN to jointly regress the shape and pose parameters. However, as pose and 3DMM parameters are indirect information for a 2D face image, the variations from network prediction can exert huge visual error. Most recently, model parameter regression methods have changed into dense correspondence regression approaches [1, 19]. By using the intermediate UV representation, DenseReg [1] predicts the UV coordinates and PRN [19] forecasts 3D coordinates rearranged in the UV space. However, UV transformation is still an indirect representation for a 2D image. In this paper, we resort the most straightforward 3D representation: 3D vertices pro-

jected on the image plane. This representation is consistent with the regression targets of face detection and 2D facial landmark localisation, and easy to optimise in a single-shot, multi-level face localisation framework. Due to the parallel training with face detection and 2D face alignment, our 3D face reconstruction branch is very robust under in-the-wild scenarios.

3. Proposed Approach

3.1. 3D Face Reconstruction

In Fig. 3, we show a fixed number of N vertices ($\mathbf{V} = [x_1, y_1, z_1; x_2, y_2, z_2; \cdots; x_N, y_N, z_N]$) on a predefined topological triangle context. These corresponding vertices share the same semantic meaning across different faces. With the fixed triangle topology, every pixel on the face can be indexed by the barycentric coordinates and the triangle index, thus there exists a pixel-wise correspondence with the 3D face. Comparing Mesh68 and Mesh1k in Fig. 3, it becomes obvious that more vertices make the mesh more informative and smooth. As the parameters of the last layer increase linearly with the size of the regression output, we choose to regress 1k + 68 vertices, which is a subset of $53,215$ vertices [42], carefully sampled to sufficiently retain the 3D faces structure.

In this paper, we directly regress 3D vertices on the 2D image plane. As each densely aligned 3D face is represented by concatenating its N vertex coordinates, we employ the following vertex loss to constrain the location of vertices:

$$\mathcal{L}_{vert}=\frac{1}{N}\sum_{i=1}^{N}||\mathbf{V}_i(x,y,z)-\mathbf{V}_i^*(x,y,z)||_1, \tag{1}$$

where $N = 1103$ is the number of vertices, $\mathbf{V}$ is the prediction of our model and $\mathbf{V}^*$ is the ground-truth.

The x and y coordinates of visible vertices in the image space can be directly learned from input face images. However, predicting the z coordinates and the x and y coordinates of invisible vertices is challenging due to the information loss occurring when projecting a face from 3D to 2D. By taking advantage of the 3D triangulation topology, we consider the edge length loss [37]:

$$\mathcal{L}_{edge}=\frac{1}{3M}\sum_{i=1}^{M}||\mathbf{E}_i-\mathbf{E}_i^*||_1, \tag{2}$$

where $M = 2110$ is the number of triangles, $\mathbf{E}$ is the edge length calculated from the prediction and $\mathbf{E}^*$ is the edge length calculated from the ground truth. The edge graph is a fixed topology as shown in Fig. 3.

By combining the vertex loss and the edge loss, we define the mesh regression loss as: $\mathcal{L}_{mesh}=\mathcal{L}_{vert}+\lambda_0\mathcal{L}_{edge}$, where λ_0 is set to 1 according to our experimental experience.

3.2. Multi-level Face Localisation

For any training anchor i, we minimise the following multi-task loss:

$$\begin{aligned}\mathcal{L} = &\mathcal{L}_{cls}(p_i, p_i^*) + \lambda_1 p_i^* \mathcal{L}_{box}(t_i, t_i^*) \\ &+ \lambda_2 p_i^* \mathcal{L}_{pts}(l_i, l_i^*) + \lambda_3 p_i^* \mathcal{L}_{mesh}(v_i, v_i^*).\end{aligned} \tag{3}$$

where t_i, l_i, v_i are box, five landmarks and 1k vertices predictions, t_i^*, l_i^*, v_i^* is the corresponding ground-truth, p_i is the predicted probability of anchor i being a face, and p_i^* is 1 for the positive anchor and 0 for the negative anchor. The classification loss $\mathcal{L}_{cls}$ is the softmax loss for binary classes (face/not face).

For a positive anchor with center coordinates x^a_{center}, y^a_{center} and scale s^a, we have the box size regression targets: $log(w^*/s^a)$ and $log(h^*/s^a)$ [21], where w^* and h^* are the width and height of the ground-truth face box. In addition, we have the following unified point regression targets for multi-level face localisation tasks:

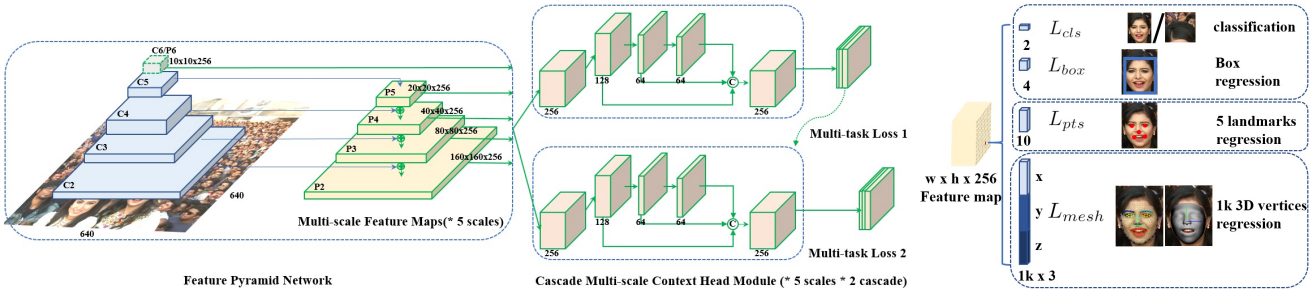
$$\begin{aligned}&({x_j}^* - x^a_{center})/s^a, \\ &({y_j}^* - y^a_{center})/s^a, \\ &({z_j}^* - z^*_{nose-tip}))/s^a,\end{aligned} \tag{4}$$

where ${x_j}^*$ and ${y_j}^*$ are the ground-truth coordinates of the two box corners, five facial landmarks and 1k 3D vertices in the image space, and ${z_j}^*$ is the ground-truth z coordinates of the 1k 3D vertices. As we use orthographic projection to generate the ground-truth 3D meshes, we translate all vertices so that the z coordinate of the nose tip is 0. Afterwards, we normalise the z coordinates by the anchor scale. We follow [21] and use the smooth-L_1 loss for all the above regression targets. As these three localisation tasks are homogeneous, the loss-balancing parameters λ_1-λ_3 are all set to 1.

As illustrated in Fig. 4, face detection, five 2D facial landmark localisation and 3D face reconstruction are three face localisation tasks aiming at different levels of localisation detail which however, share the same target: accurate point regression on the image plane. Integrating direct regression of 3D vertices in the single-shot face detection design, induces no optimisation difficulty as it is compatible with box center regression and five facial landmark regression. Each task can benefit from other tasks such that: (1) localisation of more semantic points contributes to more accurate box prediction, and (2) more challenging training scenarios in the face detection dataset result in more robust point prediction.

3.3. Single-shot Multi-level Face Localisation

In Fig. 2, we present the framework of the proposed single-shot, multi-level face localisation approach. As can be seen, our model consists of three main components: the



(a) Network Structure (b) Multi-task Loss

Figure 2. (a) An overview of the proposed single-shot multi-level face localisation approach. (b) Detailed illustration of our loss design. RetinaFace is designed based on the feature pyramids with five scales. For each scale of the feature maps, there is a deformable context module. Following the context modules, we calculate a joint loss (face classification, face box regression, five facial landmarks regression and 1k 3D vertices regression) for each positive anchor. To minimise the residual of localisation, we employ cascade regression.

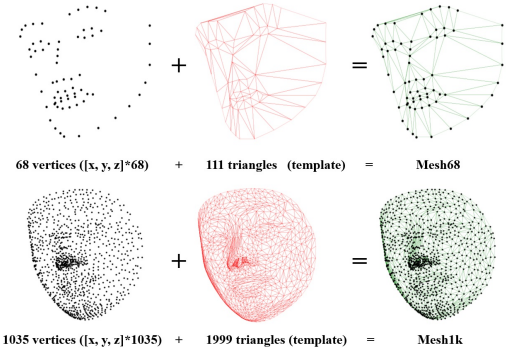


Figure 3. A mesh consists of vertices plus triangles. Mesh68 is a coarse version used for quantitative evaluation and Mesh1k is a more elaborate version which includes facial details. In this paper, we regress Mesh68 and Mesh1k simultaneously.

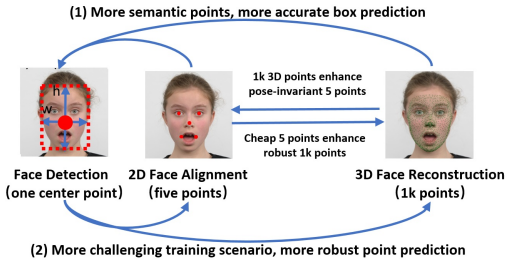


Figure 4. Three face localisation tasks have different levels of detail but share the same target: accurate point prediction on the image plane. Each task can benefit from other tasks.

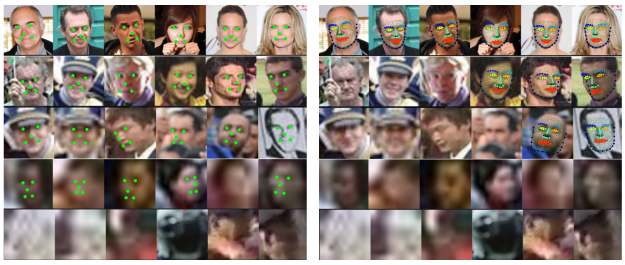
feature pyramid network, the context head module and the cascade multi-task loss. First, the feature pyramid network gets the input face images and outputs five feature maps of different scale. Then, the context head module gets a feature map of a particular scale and calculates the multi-task loss (Eq. 3). In more detail, the first context head module predicts the bounding box from the regular anchor, while subsequently, the second context head module predicts a more accurate bounding box using the regressed anchor, which is generated by the first context head module. The proposed RetinaFace employs fully convolutional neural networks, thus it can be easily trained in an end-to-end way.

Feature Pyramid RetinaFace 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 [35, 36]. P_6 is calculated through a 3×3 convolution with stride=2 on C_5. C_1 to C_5 is a pre-trained classification network on the ImageNet-11k dataset while P_6 are randomly initialised using the "Xavier" method [22].

Context Module Inspired by SSH [40] and PyramidBox [51], we also apply independent context modules on five feature pyramid levels to increase the receptive field and enhance the rigid context modelling power. We replace all 3×3 convolution layers within the lateral connections and context modules with the deformable convolution network (DCN) [11, 71], which further strengthens the non-rigid context modelling capacity.

Cascade Multi-task Loss To further improve the performance of face localisation, we employ cascade regression [4, 65] with multi-task loss (Sec. 3.2). The loss head is a 1×1 convolution across different feature maps of dimension $H_n \times W_n \times 256, n \in \{2, \dots, 6\}$. The first context head module predicts the bounding box from the regular anchor. Subsequently, the second context head module predicts a more accurate bounding box from the regressed anchor.

Anchor Settings and Matching Strategy We employ scale-specific anchors on the feature pyramid levels from P_2 to P_6 similarly to [55]. Here, P_2 is designed to capture tiny faces by tiling small anchors at the cost of more computational time and with 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 640×640, the anchors can cover scales from 16×16 to 406×406 across the feature pyramid levels. In total, there are 102,300 anchors, and 75% of these anchors are from P_2. For the first head module, anchors are matched to a ground-truth box when Intersection over Union (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



(a) Five Landmarks Annotation (b) 1k 3D Vertices Annotation

Figure 5. We annotate (a) five facial landmarks and (b) 1k 3D vertices on faces that can be annotated from the WIDER FACE dataset.

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. We employ OHEM [49, 69] to balance the positive and negative training examples.

4. Experiments

4.1. Dataset

The WIDER FACE dataset [59] consists of $32,203$ images and $393,703$ face bounding boxes with a high degree of variability in scale, pose, expression, occlusion and illumination. As illustrated in Fig. 5, we define five levels of face image quality [13] according to how difficult it is to annotate landmarks on the face. We have manually annotated five facial landmarks (*i.e*. eye centres, nose tip and mouth corners) for $84.6k$ faces on the training set and $18.5k$ faces on the validation set. To obtain accurate ground-truth 3D vertices from 2D faces, we employ a semi-automatic annotation pipeline. That is, for each face we automatically recover 68 3D landmarks [15] and employ them to drive a 3DMM fitting algorithm [3] in order to reconstruct a dense 3D face with 53K vertices, projected on the image plane. To ensure high quality in the 3DMM fitting results, we recover UV texture maps from the fitted faces and ask annotators to inspect them for artifacts. If the fitting is not accurate, the annotator manually updates the 68 landmarks and gets another automatic 3DMM fitting. If the fitting is still not accurate, the face is discarded. Finally, we get $22k$ accurate 3D face annotations from the training set.

Following [6], we also completed $27.1k$ and $39.3k$ 3D face annotations from the AFLW dataset [30] and the full FDDB image set [28] using our semi-automatic annotation method. For the face detection task, only the training data of the WIDER FACE dataset were used. For the rest of the tasks, face annotations from both the AFLW and FDDB datasets were employed, except for the subset of faces that overlaps with faces of the AFLW2000-3D dataset [72].

Method	Easy	Medium	Hard	average AP
Baseline	95.832	95.243	89.875	52.65
+DCN	96.149	95.568	90.286	53.36
+Cascade	96.233	95.679	90.642	54.20
$\mathcal{L}_{pts}$	96.570	95.913	91.161	54.73
$\mathcal{L}_{vert}$	96.512	95.805	90.983	54.55
$\mathcal{L}_{mesh}$	96.528	95.829	90.991	54.62
$\mathcal{L}_{5pts+mesh}$	**96.713**	**96.082**	**91.447**	**55.02**

Table 1. Ablation experiments of RetinaFace (ResNet-50) on the WIDER FACE validation subset and the Hard test subset.

4.2. Implementation Details

Data Augmentation As there are around 20% tiny faces in the WIDER FACE training set, we follow [69, 51] to randomly crop square patches from the original images and resize them to 640×640 resolution during training. Besides random crops, we also augment the training dataset by applying random horizontal flip and photo-metric colour distortion [69].

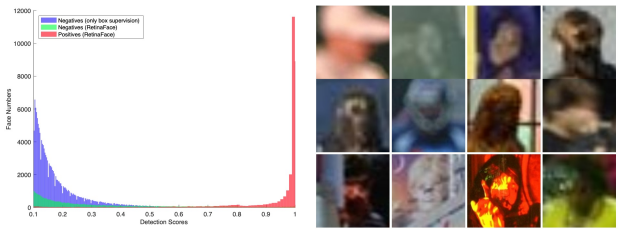
Training and testing Details We train our method using a SGD optimizer (momentum at 0.9, weight decay at 0.0005, and batch size of 8×4) on four NVIDIA Tesla P40 (24GB) GPUs. The learning rate starts from 10^{-3}, rising to 10^{-2} after 5 epochs, then divided by 10 at 55 and 68 epochs. The training process is terminated at 80 epochs. Our implementation is on MXNet [9]. For the evaluation on WIDER FACE, we follow the standard practices of [40, 69] and employ flipping as well as multi-scale (the shorter image size at $[500, 800, 1100, 1400, 1700]$) strategies. Box voting [20] is applied on the union set of predicted facial boxes using an IoU threshold of 0.4. For the evaluation of other tasks, RetinaFace employs single-scale inference using the ResNet-50 backbone (model size: 155MB, speed: 22.3ms on the P40 GPU).

4.3. Face Detection

Besides the standard evaluation metric of Average Precision (AP), we also employ a more strict evaluation metric of average AP for IoU=$0.5 : 0.05 : 0.95$, rewarding more accurate face detectors. 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. Here, we use Resnet-50 [26] as the backbone and focus on the metric of average AP. By applying the Deformable Context Module (DCM) and cascade regression, we improve the average AP of the baseline to 54.20%.

Five facial landmarks regression ($\mathcal{L}_{pts}$) and 1k 3D vertices regression ($\mathcal{L}_{vert}$) improve the average AP by 0.53% and 0.35%, respectively. The difference in improvement percentage is due to the annotation of five facial landmarks being much easier than annotating 3D vertices, and thus more training data with 5 facial points annotations can be

https://competitions.codalab.org/competitions/20146



(a) False Positive Suppression (b) False Positive Examples

Figure 6. Joint five facial landmarks regression and 1k 3D vertices regression can (a) significantly decrease the scores of false positives, and (b) effectively suppress high-score false positives.

included, aiding to achieve a higher performance. Adding topology constraints into 1k 3D vertices regression only slightly improves face detection. However, they are beneficial for predicting z coordinates and pose (Sec. 4.5). We therefore employ the mesh loss ($\mathcal{L}_{mesh}$). Combining five facial landmark regression and mesh regression significantly improves the performance by 0.82%.

Besides improving the accuracy of facial boxes, joint five facial landmark regression and 1k 3D vertices regression can effectively suppress the scores of false positives as illustrated in Fig. 6(a). In Fig. 6(b), we show some high-score false positives (> 0.9) produced from the baseline. RetinaFace assigns much lower scores (< 0.3) to these crops. Moreover, for the baseline the class information for a face is only a binary label and no information exists about the image quality. In contrast, we annotate 1k 3D vertices for easy-to-medium level faces, and five facial landmarks for faces that can be annotated. These annotations implicitly indicate the information level of faces, which can be learned by our model. Therefore, RetinaFace only gives high confidence scores for very informative faces and low scores for less informative faces.

To obtain the evaluation results from the WIDER FACE leader-board, we submitted the detection results of RetinaFace (ResNet-152) to the organisers. As shown in Fig. 7, we compared RetinaFace with other 29 state-of-the-art face detection algorithms (*e.g.* SSH [40], SFD [69], PyramidBox [51], DSFD [33], SFDet [68], RefineFace [67] *etc.*). Our approach sets up a new impressive record in terms of AP (91.7%) and outperforms these state-of-the-art methods on the Hard subset which contains a large number of tiny faces.

4.4. Five Facial Landmark Localisation

Face detection datasets are more challenging [59] than the face alignment datasets, which are usually collected by a pre-trained face detector with a high threshold. In the following experiments, we explore the benefits of training for point regression and face detection simultaneously.

To evaluate the accuracy of five facial landmark localisation, we compare RetinaFace with MTCNN [66] and

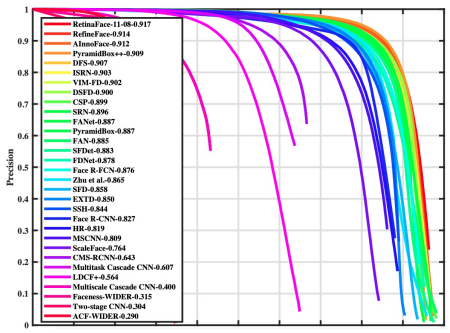


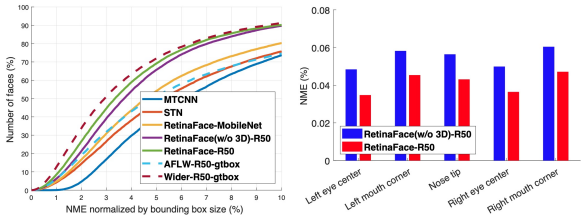
Figure 7. Precision-recall curves of RetinaFace (ResNet-152) on the WIDER FACE test Hard subsets.

Method	AUC (%)	Failure Rate (%)
MTCNN [66]	36.63	26.31
STN [7]	42.63	24.23
RetinaFace-MobileNet0.5	47.12	19.72
RetinaFace-R50	**58.54**	**9.82**
RetinaFace(w/o 3D)-R50	55.66	10.25
AFLW-R50-gtbox	44.91	25.40
Wider-R50-gtbox	**61.55**	**8.78**

Table 2. Summary of five facial landmark localisation results on the WIDER FACE dataset. Accuracy is reported as the Area Under The Curve (AUC) and the Failure Rate (threshold at 10%). "-gtbox" refers to crop-based face alignment based on ground-truth facial boxes.

STN [7] on the WIDER FACE validation set (18.5k faces). Here, we employ the face box size ($\sqrt{W \times H}$) as the normalisation distance. In Fig. 8(a), we show the Cumulative Error Distribution (CED) curves on the WIDER FACE validation set. As shown in Tab. 2, RetinaFace-MobileNet0.5 outperforms the baselines and decrease the failure rate to 19.72%. By employing a deeper backbone (ResNet-50), RetinaFace-R50 further reduces the failure rate to 9.82%. After removing the 3D mesh regression branch from RetinaFace, the AUC obviously decreases from 58.54% to 55.66%. This is because 3D mesh regression is pose-invariant and a joint training framework can improve the accuracy of 2D five facial landmarks. In Fig. 8(b), we test RetinaFace on the AFLW2000-3D profile subset (232 faces) [72], and we confirm that 3D mesh regression can significantly improve five facial landmark localisation under the large-pose scenario.

We further train two crop-based five facial landmark regression networks (ResNet-50) on the AFLW dataset (24,386 faces) [30] and the WIDER FACE dataset, separately. Even with the ground-truth facial boxes, the model trained on AFLW still has a high failure rate (25.40%), which indicates the difference in the difficulty level between the face alignment dataset (AFLW) and the face detection dataset (WIDER FACE). Even though the model trained on WIDER FACE achieves the highest performance, the computational complexity increases linearly with the number of faces in the input image. However, RetinaFace achieves



(a) CED on WIDER FACE (b) NME on AFLW Profile

Figure 8. Qualitative comparison between baselines and RetinaFace on five facial landmark localisation. (a) CED on the WIDER FACE validation set (18k faces) (b) NME on the AFLW2000-3D profile subset [72].

Method	[0°,30°]	[30°,60°]	[60°,90°]	Mean
SDM [72]	3.67	4.94	9.67	6.12
3DDFA [72]	3.43	4.24	7.17	4.94
Yu *et al*. [61]	3.62	6.06	9.56	6.41
3DSTN [2]	3.15	4.33	5.98	4.49
PRN [19]	2.75	3.51	4.61	3.62
FAME [5]	3.11	3.84	6.60	4.52
SS-SFN [6]	3.09	4.27	5.59	4.31
MS-SFN [6]	2.91	3.83	4.94	3.89
$\mathcal{L}_{vert}$	2.77	3.70	4.95	3.81
$\mathcal{L}_{mesh}$	2.72	3.65	4.81	3.72
$\mathcal{L}_{5pts+mesh}$	**2.57**	**3.32**	**4.56**	**3.48**

Table 3. Comparison of NME(%) for 68 points on the AFLW2000-3D dataset.

only slightly lower performance while its computational complexity remains fixed independently of the number of faces in each image.

4.5. 3D Vertex Regression

Following [19], we evaluate the accuracy of 3D vertex regression on the AFLW2000-3D dataset [72] considering 68 points with the 2D projected coordinates. The mean error is normalised by the bounding box size [72]. In addition, our 3D vertex prediction can be used for (1) pose-invariant facial component segmentation [1], and (2) 3D pose estimation by utilizing a least squares solution [19] instead of solving a PnP problem.

In Tab. 3, we compare the proposed RetinaFace (ResNet-50) with other state-of-the-art dense regression algorithms (*e.g*. PRN [19] and MS-SFN [6] *etc*.) for 68 landmark localisation under yaw angle variations. The proposed direct 3D vertex regression is more than able to handle 68 facial landmark localisation under pose-variations, achieving an impressive NME of 3.81%. 3D topology constraints help to slightly decrease the NME to 3.72%. After training the 2D landmark regression and 3D mesh regression jointly with the face detector, we achieve better performance than state-of-the-art methods. PRN [19] employs UV position map regression and MS-SFN [6] uses 3DMM parameter regression, both of which are indirect regression methods. In contrast, our method directly regresses the projected 3D vertices on the image plane, which can benefit from challenging training scenarios of face detection and large-scale five facial landmark annotations.

Method	Eyebrow	Eye	Nose	Lip
DenseReg [1]	47.62	74.29	87.71	72.35
$\mathcal{L}_{5pts+vert}$	71.3	76.85	90.90	75.43
$\mathcal{L}_{5pts+mesh}$	**72.23**	**78.51**	**92.21**	**77.55**

Table 4. Semantic segmentation accuracy on the Helen test set [32] measured using IoU ratio.

Method	[0°,30°]	[30°,60°]	[60°,90°]
DenseReg [1]	4.14 ± 3.93	5.96 ± 4.74	6.38 ± 4.90
PRN [19]	3.96 ± 3.43	5.75 ± 4.42	6.08 ± 4.41
$\mathcal{L}_{5pts+vert}$	3.79 ± 3.08	5.28 ± 3.83	5.60 ± 3.81
$\mathcal{L}_{5pts+mesh}$	3.69 ± 2.99	5.11 ± 3.73	5.41 ± 3.57

Table 5. Yaw angle estimation on the AFLW2000-3D dataset.

Besides the evaluation on facial landmark localisation, we can also transfer our 3D vertex predictions into pixel-wised segmentation maps for different semantic components [1]. After an additional linear regression from 1k vertices to 53k vertices, we directly employ the segmentation mask (right/left eyebrow, right/left eye, nose, and upper/lower lip) defined in [1]. Tab. 4 reports evaluation results on the Helen test set [32] using the IoU ratio. Note that the ground-truth here is generated by deformation-free coordinates [1]. The results indicate that the proposed RetinaFace (ResNet-50) outperforms DenseReg (ResNet-101), which is based on indirect UV coordinate regression. In contrast, RetinaFace employs direct vertex regression on the image plane, which is beneficial for more accurate localisation on the image, *e.g*. the substantial improvement for eyebrows. Given that RetinaFace is not optimized for the segmentation task, we believe that the pose-invariant facial component segmentation results in Fig. 9 are impressive.

As we can directly predict 3D vertices, the pose estimation is only the estimation of a rotation matrix obtained as the least squares solution between the regressed landmarks and the landmarks of a template face in frontal pose. In Tab. 5, we compare RetinaFace (ResNet-50) with DenseReg [1] and PRN [19] on the yaw angle estimation. Both DenseReg [1] and PRN [19] use an intermediate UV representation, while RetinaFace employs direct vertex regression on the image plane. As we can see from Tab. 5, RetinaFace can predict more accurate yaw angles with low variance, while topology constraints can further improve pose estimation. Both z coordinates regression and pose estimation try to predict indirect 3D information from 2D images, thus the inclusion of topology constraints can boost both tasks by building the connections between direct image clues and indirect information estimation. In Fig. 9, we present the pose estimation results at the nose tip. RetinaFace is robust on pose estimation under expression variations, illumination changes and occlusions.

Figure 9. Example results of RetinaFace on the AFLW2000-3D dataset. First row: 1k 3D vertices predicted by RetinaFace (ResNet-50, $\mathcal{L}_{5pts+mesh}$). Second row: 3D pose estimation and mesh render by the Vulkan toolkit. Third row: pose-invariant facial component segmentation.

Figure 10. Testing results of RetinaFace (ResNet-50, $\mathcal{L}_{5pts+mesh}$) compared to MFN [6] (first row). We show both the predicted 1k 3D vertices (Second row) and the 3D meshes rendered by the Vulkan toolkit (Third row). Please zoom in to check the missing faces (Column **2-4**) and obvious mis-alignment (Column **5-6**) by MFN.

4.6. Multi-Face Reconstruction

In Fig. 10, we compare RetinaFace with MFN [6] on multi-face images. MFN employs a single end-to-end network to jointly predict the bounding box locations and 3DMM parameters for multiple faces. However, 3DMM parameter regression is not as straightforward as our vertex regression on the image plane. In Columns 5 and 6, a misalignment problem in MFN can be observed. Even tiny variations in the predicted 3DMM parameters can significantly affect the reconstruction results. Nevertheless, RetinaFace can exactly fit the face boundary. In Columns 2, 3 and 4, it can be seen that several faces are missed by MFN. In contrast, our RetinaFace achieves state-of-the-art performance on WIDER FACE and can easily detect the tiny faces even under low illumination conditions. In the last row of Fig. 10, we render the 3D 1k vertices predicted by RetinaFace.

5. Conclusion

In this paper, we innovatively unify multi-level face localisation tasks under one common target: point regression on the image plane. We directly regress 3D vertices in the image space while constrained by the 3D topology of the employed 3D face template. Also, the proposed 3D mesh regression branch can be easily incorporated in parallel with the existing box and 2D landmark regression branches without any optimisation difficulty during joint training. Finally, extensive experimental results showed that the proposed mutually beneficial design can simultaneously achieve accurate face detection, 2D face alignment and 3D face reconstruction with efficient single-shot inference.

Acknowledgements. We are thankful to Nvidia for the GPU donations and Amazon for the cloud credits. Jiankang Deng acknowledges financial support from the Imperial President's PhD Scholarship. Stefanos Zafeiriou acknowledges support from the EPSRC Fellowship DEFORM (EP/S010203/1), FACER2VM (EP/N007743/1) and a Google Faculty Fellowship. An early version of RetinaFace can be found in [13], where Yuxiang Zhou has contributed to the coloured mesh decoder by graph convolution and Jinke Yu has contributed to the face detection experiments on mobile devices.

References

[1] Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*, 2017. 1, 2, 7

[2] Chandrasekhar Bhagavatula, Chenchen Zhu, Khoa Luu, and Marios Savvides. Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In *ICCV*, 2017. 7

[3] James Booth, Anastasios Roussos, Evangelos Ververas, Epameinondas Antonakos, Stylianos Ploumpis, Yannis Panagakis, and Stefanos Zafeiriou. 3d reconstruction of "in-the-wild" faces in images and videos. *TPAMI*, 2018. 5

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 4

[5] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gérard Medioni. Deep, landmark-free fame: Face alignment, modeling, and expression estimation. *IJCV*, 2019. 1, 7

[6] Bindita Chaudhuri, Noranart Vesdapunt, and Baoyuan Wang. Joint face detection and facial motion retargeting for multiple faces. In *CVPR*, 2019. 2, 5, 7, 8

[7] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *ECCV*, 2016. 2, 6

[8] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *ECCV*, 2014. 2

[9] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv:1512.01274*, 2015. 5

[10] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Selective refinement network for high performance face detection. *AAAI*, 2019. 2

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 4

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1

[13] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *arXiv:1905.00641*, 2019. 5, 8

[14] Jiankang Deng, Qingshan Liu, Jing Yang, and Dacheng Tao. M3 csr: Multi-view, multi-scale and multi-component cascade shape regression. *IVC*, 2016. 1

[15] 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. *IJCV*, 2019. 1, 5

[16] Jiankang Deng, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou. Joint multi-view face alignment in the wild. *TIP*, 2019. 1

[17] Jiankang Deng, Yuxiang Zhou, Shiyang Cheng, and Stefanos Zaferiou. Cascade multi-view hourglass model for robust 3d face alignment. In *FG*, 2018. 1

[18] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *CVPR Workshops*, 2017. 1

[19] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 1, 2, 7

[20] Spyros Gidaris and Nikos Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, 2015. 5

[21] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 2, 3

[22] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010. 4

[23] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018. 1

[24] Yudong Guo, Jianfei Cai, Boyi Jiang, Jianmin Zheng, et al. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *TPAMI*, 2018. 2

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[27] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *CVPR*, 2017. 2

[28] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst Technical Report, 2010. 2, 5

[29] Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *CVPR*, 2016. 2

[30] Martin Koestinger, 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 *ICCV workshops*, 2011. 2, 5, 6

[31] Felix Kuhnke and Jorn Ostermann. Deep head pose estimation using synthetic images and partial adversarial domain adaption for continuous label spaces. In *ICCV*, 2019. 1

[32] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *ECCV*, 2012. 7

[33] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. *arXiv:1810.10220*, 2018. 6

[34] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *CVPR*, 2019. 1

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 4

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 4

[37] Feng Liu, Luan Tran, and Xiaoming Liu. 3d face modeling from diverse raw scan data. In *ICCV*, 2019. 3

[38] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

[39] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected in-the-wild age database. In *CVPR Workshops*, 2017. 1

[40] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *ICCV*, 2017. 2, 4, 5, 6

[41] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, 2018. 1

[42] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Advanced Video and Signal Based Surveillance*, 2009. 3

[43] Deva Ramanan and Xiangxin Zhu. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 1

[44] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[45] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 2

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2

[47] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *3DV*, 2016. 2

[48] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPR Workshops*, 2018. 1

[49] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 5

[50] Brandon M Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *CVPR*, 2013. 1

[51] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *ECCV*, 2018. 2, 4, 5, 6

[52] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017. 2

[53] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *CVPR*, 2017. 2

[54] Paul Viola and Michael J Jones. Robust real-time face detection. *IJCV*, 2004. 1

[55] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: an effective face detector for the occluded faces. *arXiv:1711.07246*, 2017. 2, 4

[56] Jing Yang, Adrian Bulat, and Georgios Tzimiropoulos. Fan-face: a simple orthogonal improvement to deep face recognition. In *AAAI*, 2020. 1

[57] Jing Yang, Jiankang Deng, Kaihua Zhang, and Qingshan Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *ICCV Workshops*, 2015. 1

[58] Jing Yang, Qingshan Liu, and Kaihua Zhang. Stacked hourglass network for robust facial landmark localisation. In *CVPR Workshops*, 2017. 1

[59] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. 2, 5, 6

[60] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *CVPR*, 2019. 1

[61] Ronald Yu, Shunsuke Saito, Haoxiang Li, Duygu Ceylan, and Hao Li. Learning dense facial correspondences in unconstrained images. In *ICCV*, 2017. 7

[62] Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: past, present and future. *CVIU*, 2015. 1

[63] Changzheng Zhang, Xiang Xu, and Dandan Tu. Face detection using improved faster rcnn. *arXiv:1802.02142*, 2018. 2

[64] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *CVPR*, 2018. 1

[65] Hongkai Zhang, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cascade retinanet: Maintaining consistency for single-stage object detection. In *BMVC*, 2019. 4

[66] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 2016. 2, 6

[67] Shifeng Zhang, Cheng Chi, Zhen Lei, and Stan Z Li. Refineface: Refinement neural network for high performance face detection. In *arXiv:1909.04376*, 2019. 6

[68] Shifeng Zhang, Longyin Wen, Hailin Shi, Zhen Lei, Siwei Lyu, and Stan Z Li. Single-shot scale-aware network for real-time face detection. *IJCV*, 2019. 6

[69] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *ICCV*, 2017. 2, 5, 6

[70] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *CVPR*, 2019. 1

[71] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *arXiv:1811.11168*, 2018. 4

[72] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 1, 2, 5, 6, 7