

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

Jiankang Deng, Jia Guo, Evangelos Ververas,
Irene Kotsia, Stefanos Zafeiriou

IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR) 2020

Bernardo Biesseck

Ph.D. student at the PPGInf (UFPR)

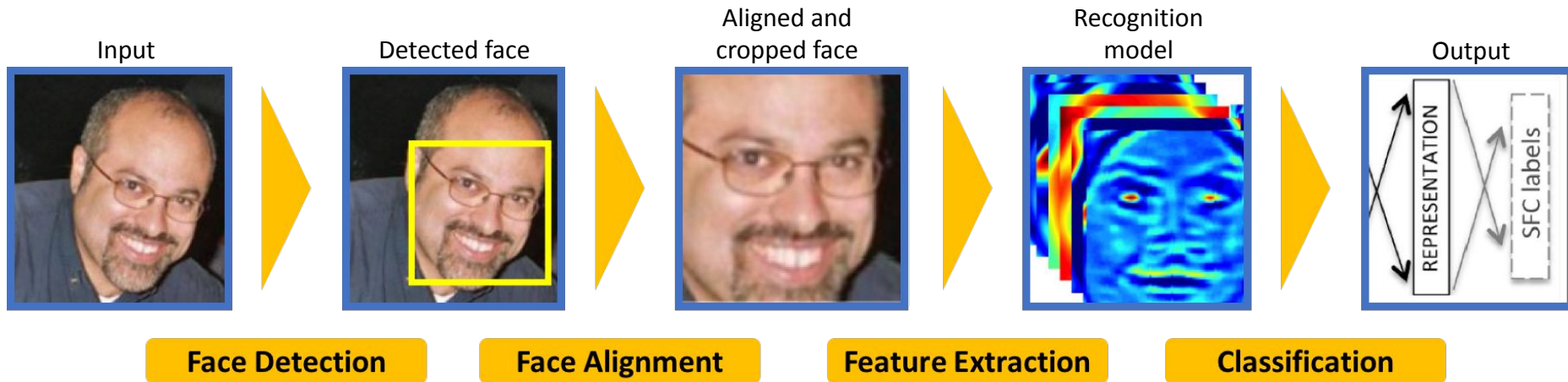
Advisor: David Menotti Gomes

SUMMARY

1. Introduction
2. Background concepts
3. Proposed model
4. Datasets and training
5. Experimental results
6. Conclusion

Introduction

- Recognition systems usually works on cropped images.



Source: <https://wiki.tum.de/display/Ifdv/Face+Recognition>

Introduction

- Recognition systems usually works on cropped images.

This presentation

Input



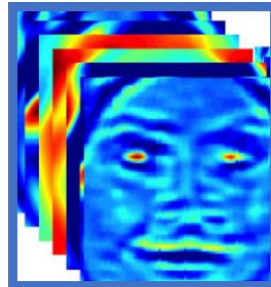
Detected face



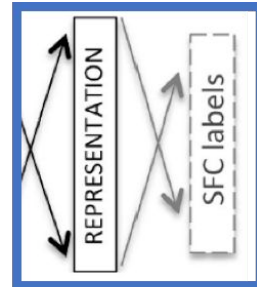
Aligned and
cropped face



Recognition
model



Output



Face Detection

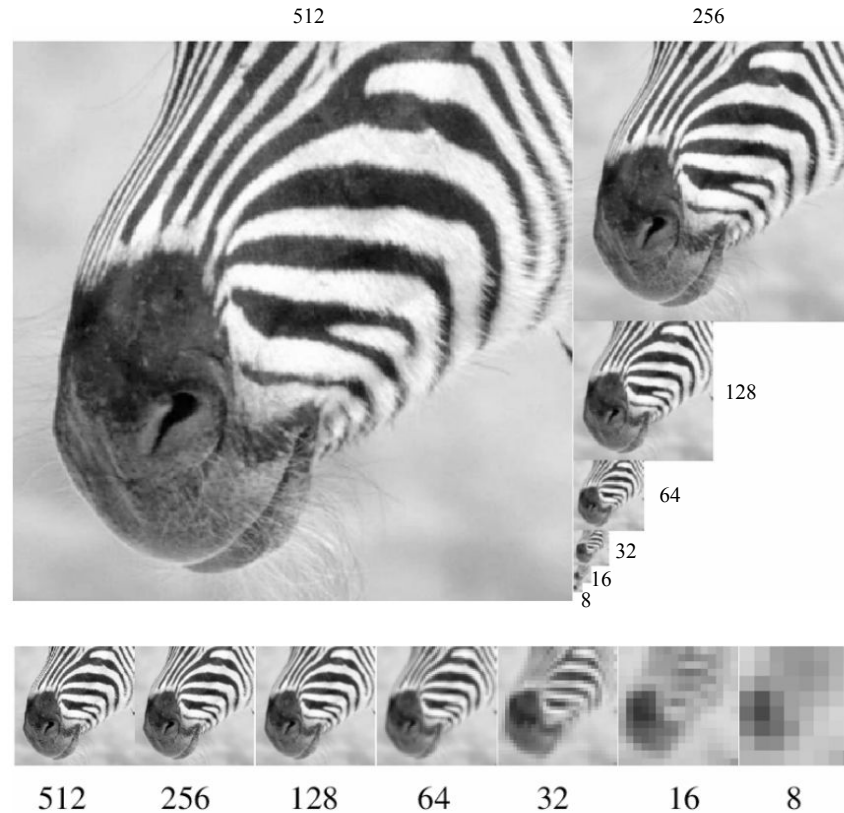
Face Alignment

Feature Extraction

Classification

Source: <https://wiki.tum.de/display/Ifdv/Face+Recognition>

Background concepts - Image Pyramid



Source: Kris Kitani, Carnegie Mellon University, 2020

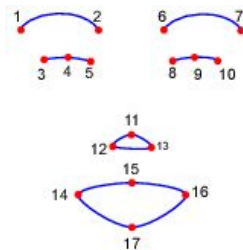
https://www.cs.cmu.edu/~16385/s17/Slides/3.1_Image_Pyramid.pdf

Background concepts - Facial landmarks

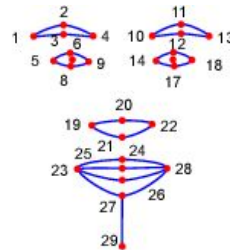
- Landmarks
 - 5, 17, 21, 29, 51, 68 points



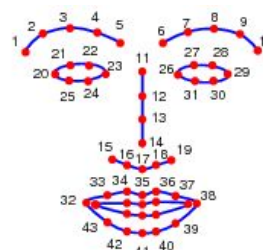
5 points



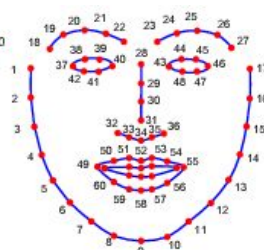
(a) 17 points



(b) 29 points



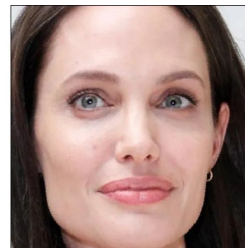
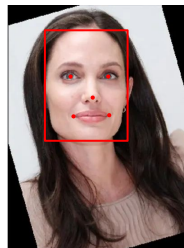
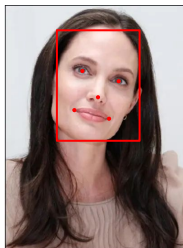
(c) 51 points



(d) 68 points

Source: Benjamin Johnston, Philip de Chazal. A review of image-based automatic facial landmark identification techniques. EURASIP Journal on Image and Video Processing, 2018.

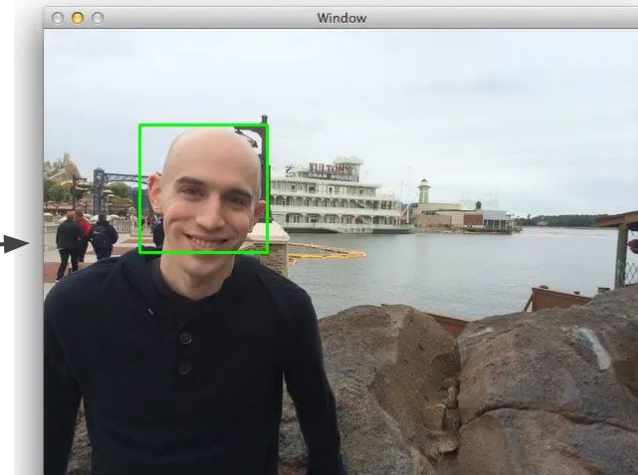
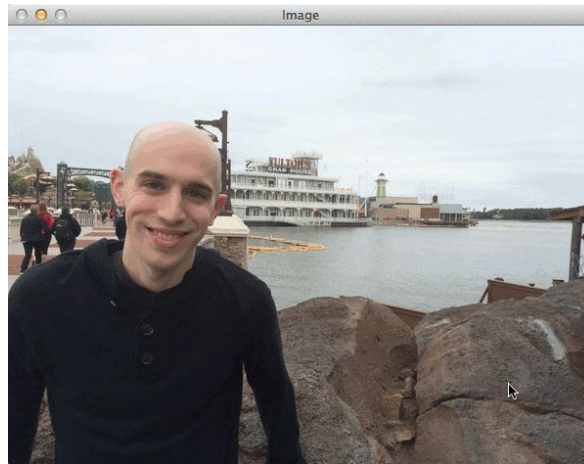
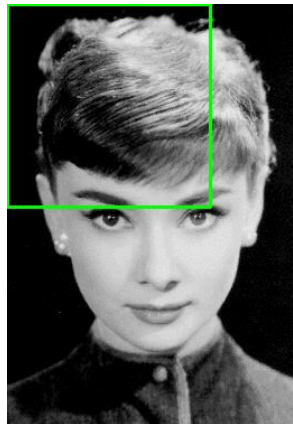
Source: Junliang Xing, Zhiheng Niu, Junshi Huang, Weiming Hu, Xiaoping Zhou, Shuicheng Yan. Towards Robust and Accurate Multi-View and Partially-Occluded Face Alignment. IEEE TPAMI, 2018.



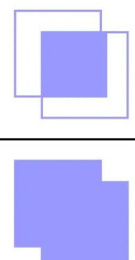
Source: <https://sefiks.com/2020/02/23/face-alignment-for-face-recognition-in-python-within-opencv/>

Background concepts - Region proposal

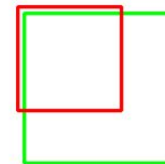
- How to find a face? Searching everywhere



Source: Adrian Rosebrock. <https://pyimagesearch.com/2015/03/23/sliding-windows-for-object-detection-with-python-and-opencv/>

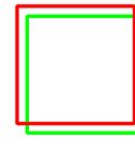
$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$


IoU: 0.4034



Poor

IoU: 0.7330



Good

IoU: 0.9264



Excellent

Background concepts - Region proposal

- How to find a face? Searching everywhere

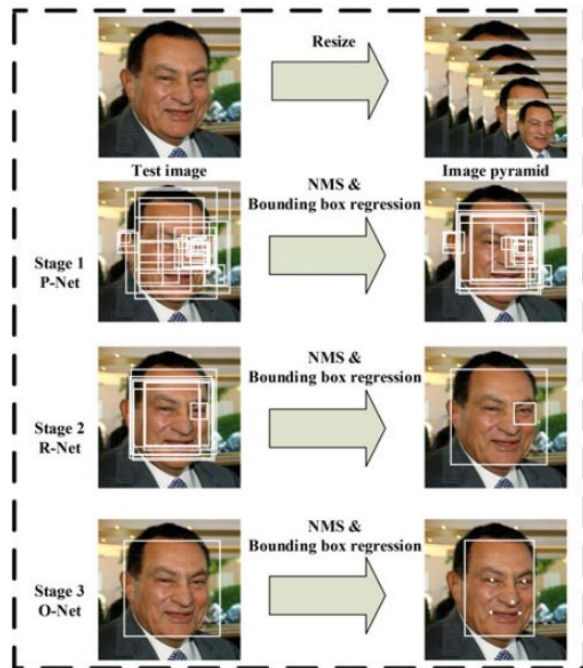


Fig. 1. Pipeline of our cascaded framework that includes three-stage multitask deep convolutional networks. First, candidate windows are produced through a fast P-Net. After that, we refine these candidates in the next stage through a R-Net. In the third stage, the O-Net produces final bounding box and facial landmarks position.

RetinaFace

- A single-shot, multi-level face localisation method that unifies:
 - Face box prediction
 - 2D facial landmark localisation and
 - Head pose estimation
 - Face parts segmentation
 - 3D vertices regression

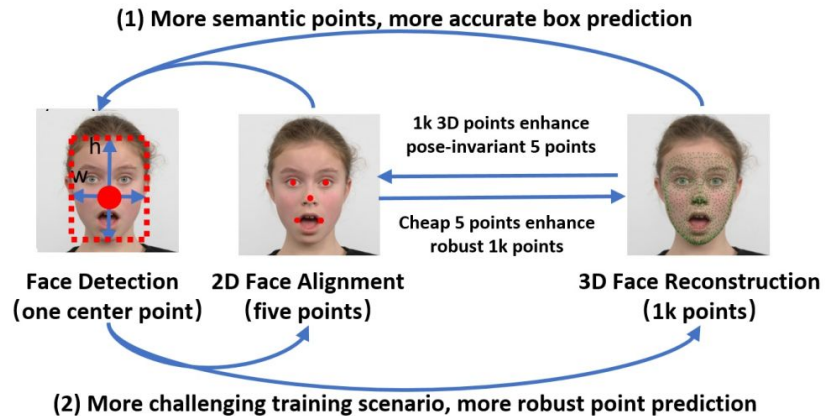


Figure 4. Three face localisation tasks have different levels of detail but share the same target: accurate point prediction on the image plane. Each task can benefit from other tasks.

Definition

- Traditional face detection → bounding box prediction
- Face localisation → face detection, pose estimation, alignment, segmentation and 3D reconstruction

- 3D face reconstruction
 - Facial template

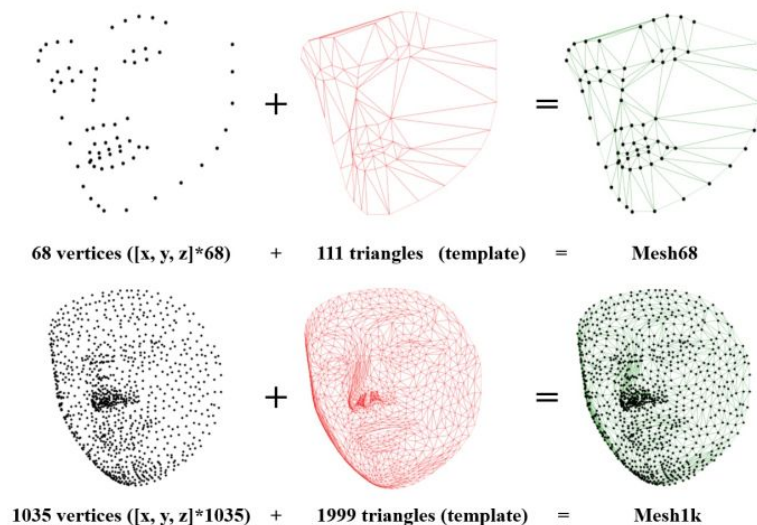
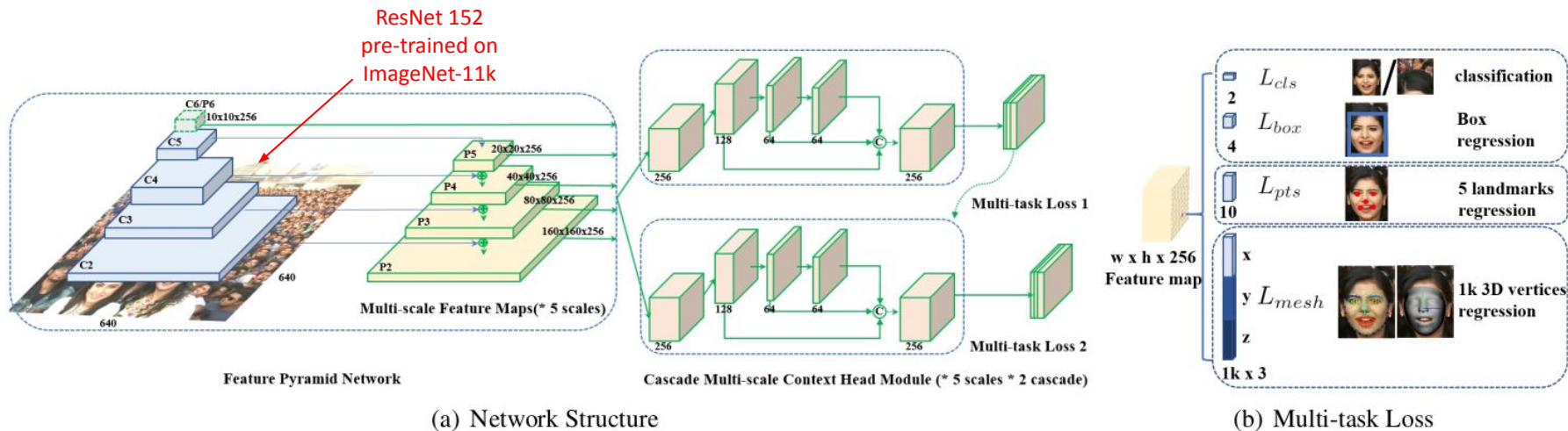


Figure 3. A mesh consists of vertices plus triangles. Mesh68 is a coarse version used for quantitative evaluation and Mesh1k is a more elaborate version which includes facial details. In this paper, we regress Mesh68 and Mesh1k simultaneously.

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss



(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.

RetinaFace

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss

Face detection $\rightarrow \mathcal{L}_{cls}(p_i, p_i^*) = \text{Cross-Entropy Loss}$

predicted \rightarrow gt

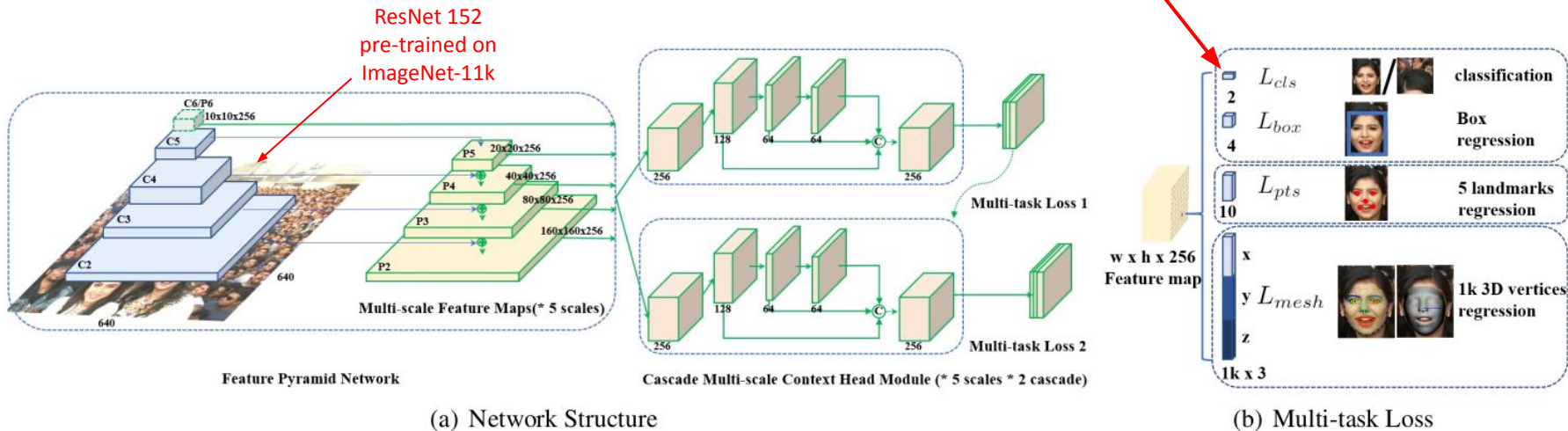
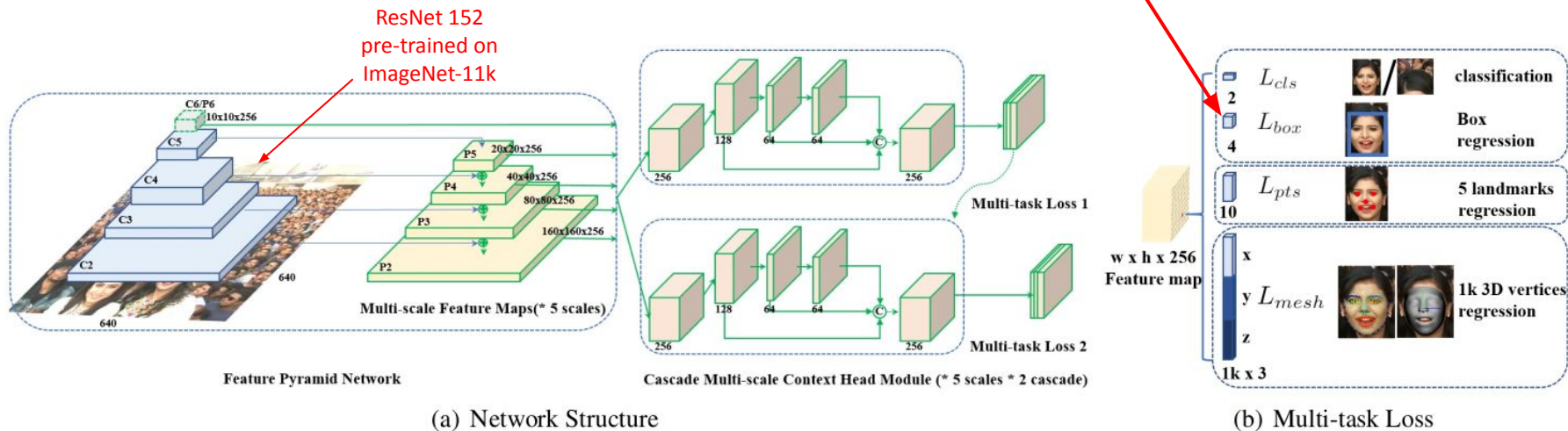


Figure 2. (a) An overview of the proposed single-shot multi-level face localisation approach. (b) Detailed illustration of our loss design. RetinaFace is designed based on the feature pyramids with five scales. For each scale of the feature maps, there is a deformable context module. Following the context 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.

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss

Bbox regression $\rightarrow \mathcal{L}_{box}(t_i, t_i^*) = \text{Mean Absolute Error}$



(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.

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss

5 landmarks regression $\rightarrow \mathcal{L}_{pts}(l_i, l_i^*) = \text{Mean Absolute Error}$

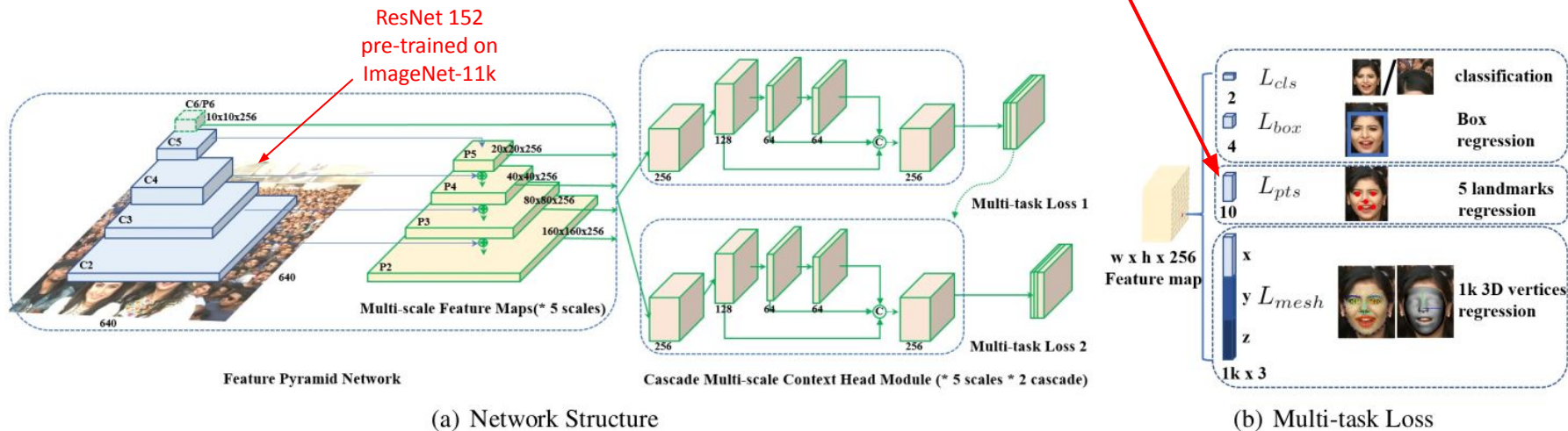


Figure 2. (a) An overview of the proposed single-shot multi-level face localisation approach. (b) Detailed illustration of our loss design. RetinaFace is designed based on the feature pyramids with five scales. For each scale of the feature maps, there is a deformable context module. Following the context 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.

RetinaFace

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss

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

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

$$\mathcal{L}_{mesh} = \mathcal{L}_{vert} + \lambda_0 \mathcal{L}_{edge}$$

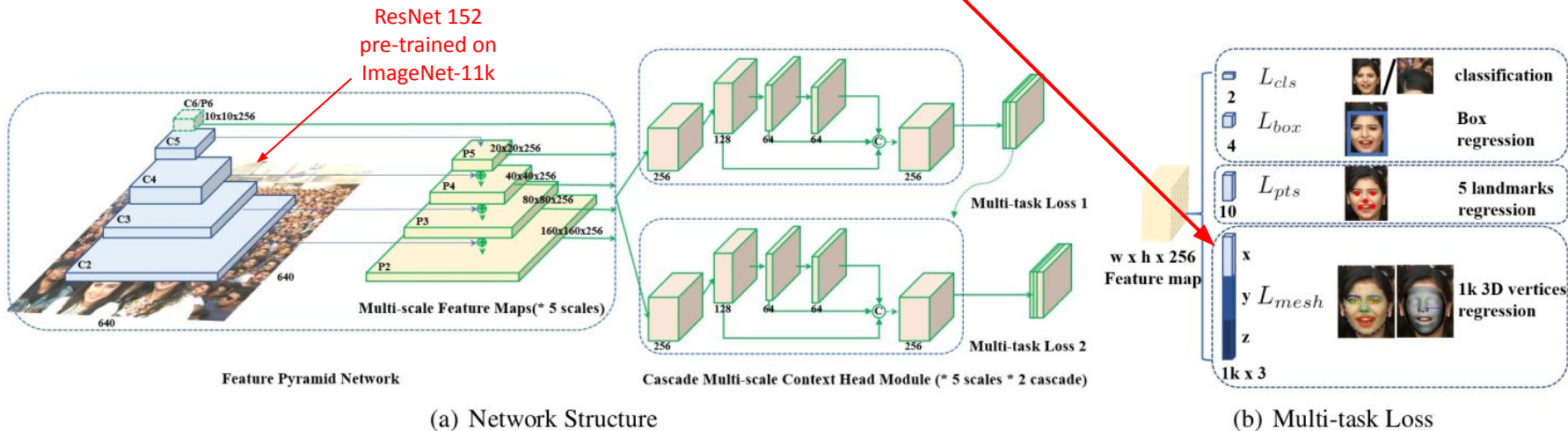


Figure 2. (a) An overview of the proposed single-shot multi-level face localisation approach. (b) Detailed illustration of our loss design. RetinaFace is designed based on the feature pyramids with five scales. For each scale of the feature maps, there is a deformable context module. Following the context 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.

RetinaFace

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss

$$\text{Multi-task Loss} \rightarrow \mathcal{L} = \mathcal{L}_{cls}(p_i, p_i^*) + \lambda_1 p_i^* \mathcal{L}_{box}(t_i, t_i^*) + \lambda_2 p_i^* \mathcal{L}_{pts}(l_i, l_i^*) + \lambda_3 p_i^* \mathcal{L}_{mesh}(v_i, v_i^*)$$

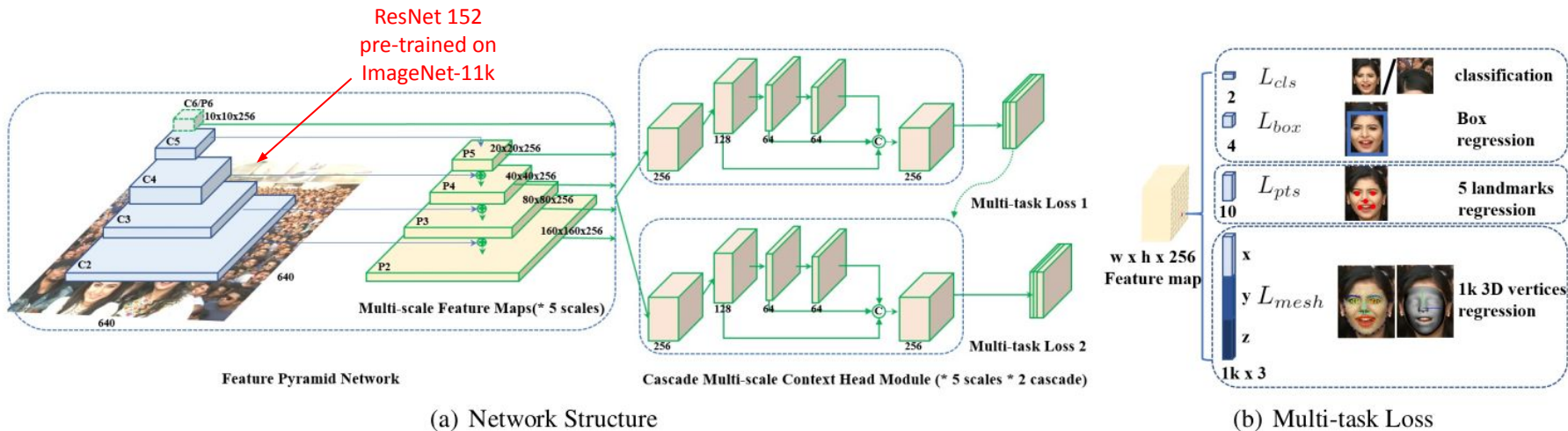


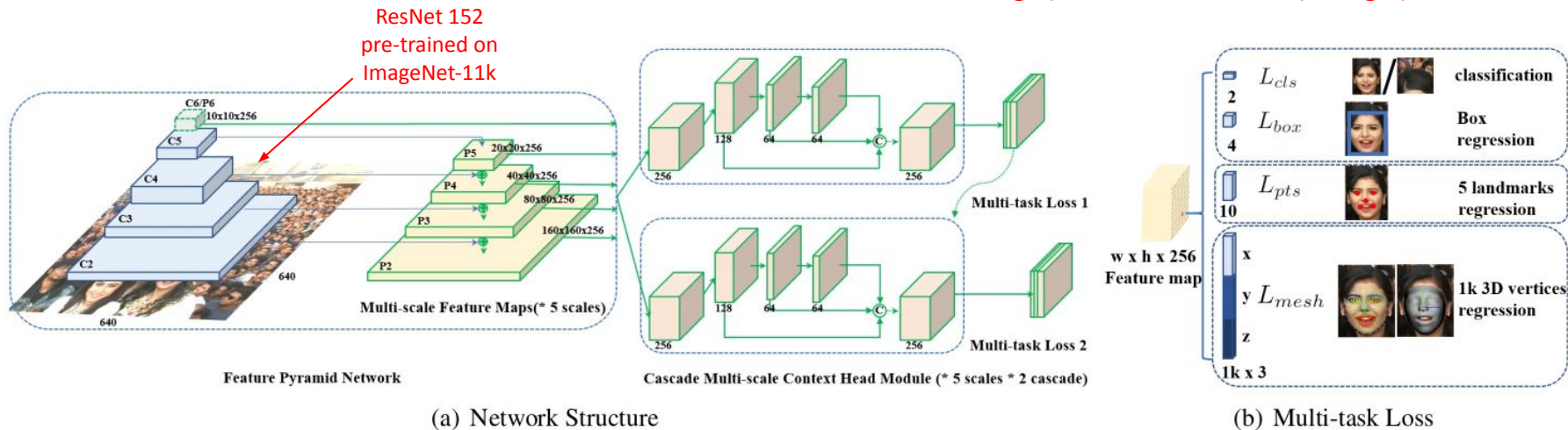
Figure 2. (a) An overview of the proposed single-shot multi-level face localisation approach. (b) Detailed illustration of our loss design. RetinaFace is designed based on the feature pyramids with five scales. For each scale of the feature maps, there is a deformable context module. Following the context 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.

RetinaFace

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss

$$\text{Multi-task Loss} \rightarrow \mathcal{L} = \mathcal{L}_{cls}(p_i, p_i^*) + \lambda_1 p_i^* \mathcal{L}_{box}(t_i, t_i^*) + \lambda_2 p_i^* \mathcal{L}_{pts}(l_i, l_i^*) + \lambda_3 p_i^* \mathcal{L}_{mesh}(v_i, v_i^*)$$

$\text{gt (1=face; 0=nonface)}$
 $\text{gt (1=face; 0=nonface)}$
 $\text{gt (1=face; 0=nonface)}$



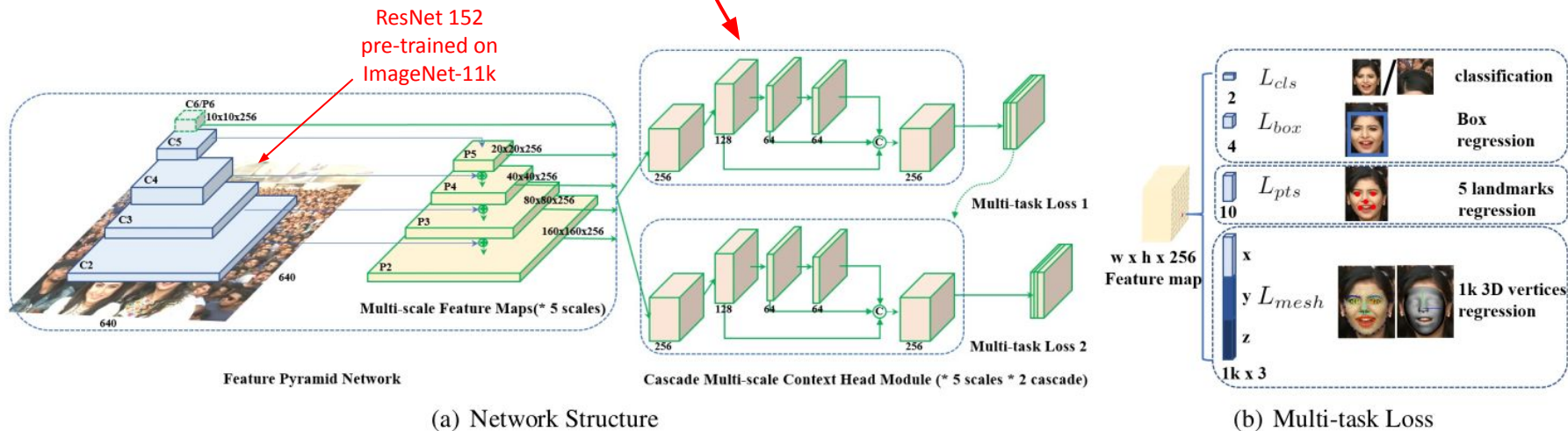
(a) Network Structure

(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.

RetinaFace

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss
- Input image size: 640×640
 - Anchors (face candidates): 16×16 to 406×406
 - 102,300 anchors
 - First head module:
 - Positive anchors = $\text{IoU} \geq 0.7$ to ground-truth
 - Negative anchors = $\text{IoU} \leq 0.3$ to ground-truth
 - Others are ignored



(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.

RetinaFace

- 3 main components:
 - Feature pyramid network
 - Context head module
 - Cascade multi-task loss
- Input image size: 640×640
 - Anchors (face candidates): 16×16 to 406×406
 - 102,300 anchors
 - Second head module:
 - Positive anchors = $\text{IoU} \geq 0.5$ to ground-truth
 - Negative anchors = $\text{IoU} \leq 0.4$ to ground-truth
 - Others are ignored

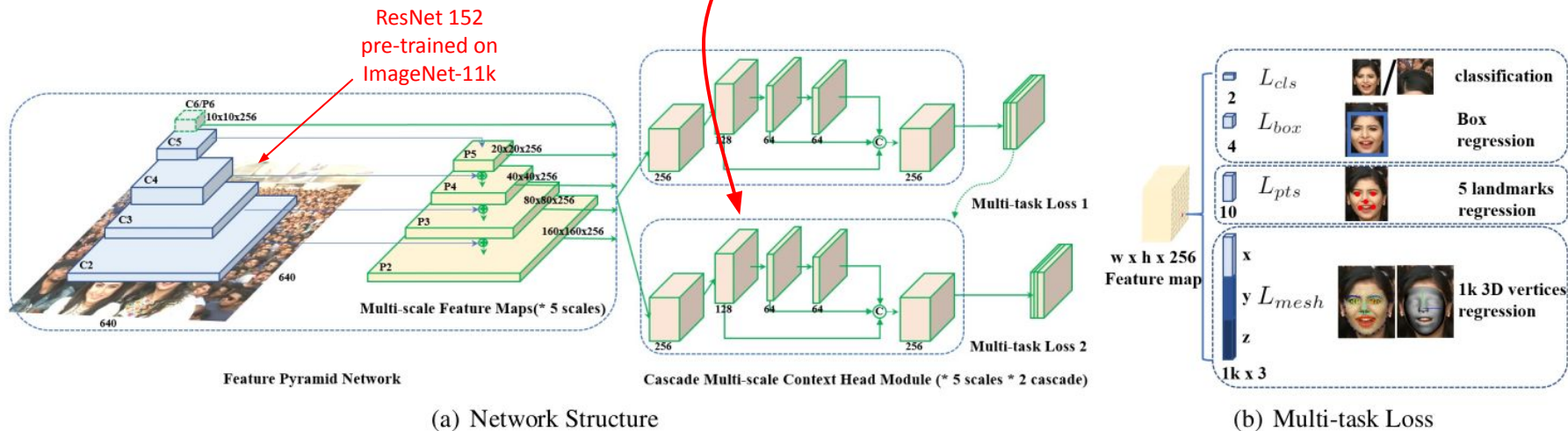
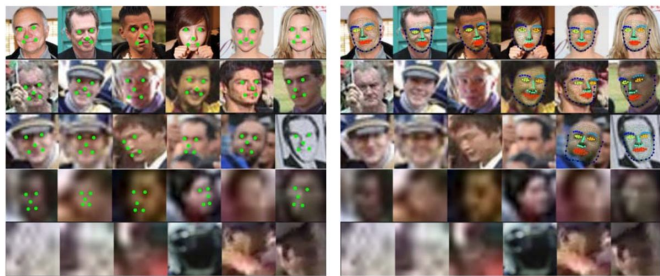


Figure 2. (a) An overview of the proposed single-shot multi-level face localisation approach. (b) Detailed illustration of our loss design. RetinaFace is designed based on the feature pyramids with five scales. For each scale of the feature maps, there is a deformable context module. Following the context 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.

Datasets and Training

- WIDER FACE - <http://shuoyang1213.me/WIDERFACE/>
 - Images are selected from the publicly available WIDER dataset
 - 393,703 labeled faces from 32,203 images
 - They manually annotated 5 facial landmarks for 84.6k faces on the training set and 18.5k faces on the validation set (Figure 5-a).
 - To generate 3D ground-truth, they automatically recover 68 3D landmarks [15] and apply a 3DMM fitting algorithm [3] to reconstruct a dense 3D face with 53K vertices.



(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.

[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.

[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.

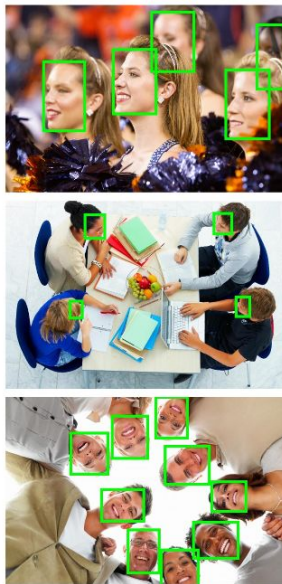
Datasets and Training

- WIDER FACE - <http://shuoyang1213.me/WIDERFACE/>
 - WIDER_train.zip
 - wider_face_split.zip

Scale



Pose



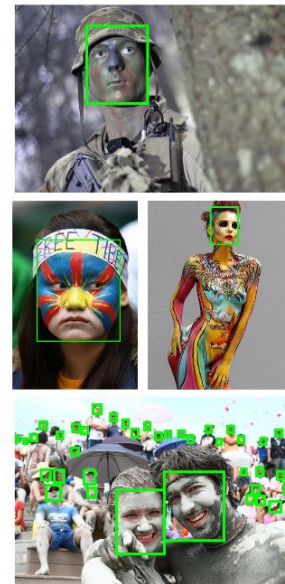
Occlusion



Expression



Makeup



Illumination



Conclusion

- Experimental results showed that the multi-task approach can simultaneously achieve accurate face detection, 2D face alignment and 3D face reconstruction with efficient single-shot inference.
- Official code:
<https://github.com/deepinsight/insightface/tree/master/detection/retinaface>
- Google Colab:
https://colab.research.google.com/drive/1M_K43OyJlbMRCayvkDVydZRS4Cwm_ZfD