

HP-FLDet: A High Performance Multi-task Joint Face and Landmark Detector

Xueli Liu

liuxueli@fudan.edu.cn

ENT Institute and Department of Otolaryngology, Eye & ENT Hospital of Fudan University

Feng Wang

Department of Urology, General Hospital of Ningxia Medical University

Yinlong Liu

Faculty of Data Science, City University of Macau

Kexue Fu

Academy for Engineering and Technology, Fudan University

Xiao Chen

ENT Institute and Department of Otolaryngology, Eye & ENT Hospital of Fudan University

Xiaobo Yang

Department of Urology, General Hospital of Ningxia Medical University

Xinrong Chen

Academy for Engineering and Technology, Fudan University

Research Article

Keywords: Face detection, Landmark detection, Scale-aware, Multi-task, Single-stage, mixture dense regression, high resolution heatmap

Posted Date: March 28th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-3926744/v2>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

HP-FLDet: A High Performance Multi-task Joint Face and Landmark Detector

Abstract

Face detection and landmark localization are usually the fundamental and important steps of facial analysis. The accuracy degradation of any task has impacts on the accuracy and robustness of the downstream analysis. In most facial analysis systems, face detection and landmark detection, as two independent tasks, are predicted sequentially with single-task detectors respectively, which suffer from the higher complexity and poorer consistency. In contrast, multi-task joint face and landmark detectors are more efficient and have more shared features. However, when two tasks are predicted in one model, it is difficult to improve the accuracy of all tasks simultaneously, due to the different requirements of the different targets to be detected. In our opinion, the multi-task detection model, integrating the perception ability of both scale variation and detailed texture information for face and landmark detection, is arduous but important in unconstrained settings.

In this paper, a High-Performance joint Face and Landmark detector(HP-FLDet) is proposed, which can predict facial bounding boxes and landmarks in single stage and improve the accuracy of both face detection and landmark localization simultaneously. In this detector, the context inception block and scale-aware multi-task strategies are designed to resolve extreme scale variance and improve the perception ability of detailed texture information. It demonstrates superior performance on both face and landmark detection datasets including AFW, PASCAL Face, FDDB, Wider Face and AFLW. Especially, it achieves SOTA results on the most challenging WIDER FACE dataset with 0.970(Easy), 0.963(Medium), 0.921(Hard) mAP on the test subset, which outperforms all state-of-the-art rivals.

Keywords: Face detection, Landmark detection, Scale-aware, Multi-task, Single-stage, mixture dense regression, high resolution heatmap.

1. Introduction

Face detection and facial landmark detection are two essential and fundamental steps in many facial analysis applications, such as recognition [1, 2], generation [3], manipulation [4], tracking [5], attributes analysis [6], etc. The purpose of face detection is to locate facial bounding boxes on images, while facial landmark detection seeks to predict landmarks (i.e., nose, eye, etc.) in the facial regions. The accuracy of them are vital import for the follow-up tasks.

In typical facial analysis systems, face detection and landmark detection are two separate tasks. Their process typically begins with the face detector [7, 8, 9] followed by the landmark detector [10, 11]. Although these approaches are historically successful, there are still limitations. Firstly, the overall process is complicated as there are two detectors and the landmarks need to be detected on each face separately. Secondly, detectors can't benefit from the shared supervision during the training. Thirdly, it is difficult to jointly optimize the face detectors and the landmark detectors. In most applications, landmark detectors are trained based on the detected faces of specific face detector. When the corresponding face detector is updated, the landmark detector is usually need to be reoptimized [12, 13].

Different from single-task approaches, multi-task approaches can predict facial bounding boxes and landmarks

with single model. Most multi-task detectors adopt the multi-stage architecture [14, 15, 13], in which the face proposals are obtained in the first stage and then the bounding boxes are refined or the landmarks are detected on the facial regions in the succeeding stages. Recently, some multi-task detectors with single-stage architectures have been proposed. Compared to the rivals with multi-stage architectures, they are usually more practical with the competitive face detection accuracy [16, 17, 18, 19]. However, they usually take landmark as a by-production and have the lower landmark detection accuracy.

For all concerns above, a single-stage multi-task face and landmark detector(HP-FLDet) is proposed by addressing two key questions:

- 1.1. 1: How to integrate the global and local information for both face and landmark detection in single stage.

As there are many tiny faces and large scale variations in unconstrained settings, it is hard to achieve ideal accuracy for both face classification and landmark localization simultaneously. In terms of face detection and facial landmark localization tasks respectively, the regression-based detector usually achieves better results on face detection task as it can get more global information for scale variation, while the heatmap-based detector has higher accuracy on facial landmark detection task as it can get more

local detailed texture information for landmark localization [13]. Therefore, a unified framework, integrating both global and local information, is important for single-stage detectors to improve the accuracy of both the face and the landmark detection.

In this paper, a multi-task detection framework is proposed with the heatmap-based and the regression-based detection heads as in Fig 2, inspiring by the bottom-up pose estimation approaches [20, 21, 22] with two kinds of detection heads (the heatmap-based head and the landmark-grouping head). For the face detection task, The heatmap-based head can evaluate the probability of face centers on the grid points, and the regression-based head can predict the faces' offset and size against the grid points. For the landmark detection task, heatmap-based head gets detailed texture information for the landmarks' localization, while the regression-based head provide the supplemented global information.

1.2. 2: How to improve the accuracy of heatmap-based anchor-free joint face detector.

From the previous work, there are some observations: (1) Anchor-based face detectors [23, 7, 17, 24, 25] have demonstrated the most promising accuracy, particularly on the most challenging dataset Wider Face [26]. (2) Almost all the state-of-the-art(SOTA) anchor-based face detectors adopt multi-scale detection heads and deploy many facial specific modules for dense anchor design [9] and anchor mining [8, 23, 24], which may lead to anchor misalignment and poor design generality easily. (3) Although heatmap-based anchor-free detectors can achieve SOTA results on the generic object detection [27, 28, 29], yet most heatmap-based face detectors [30, 15] suffer from low face detection accuracy [13]. By comparing widely used face detection datasets, such as Wider Face, with the generic object detection datasets, like MS-COCO [31], it can be found that the face detection has more rigorous tiny/small scale objects and larger scale variation, as shown in Fig 1. Then, the drawbacks of existing heatmap-based detectors are analysis comparing with anchor-based detectors when addressing tiny faces. Firstly, heatmap-based detectors usually adapt heatmap of 1/4 resolution for object detection, which are sparse for the localization of tiny faces and their landmarks, while anchor-based face detectors can generate more face proposals with multi-scale detection heads and divide-and-conquer solution. Secondly, most heatmap-based detectors pay less attention to scale variation, while anchor-based face detectors proposed many modules to preserve multi-scale context feature like FEM [32] and PA [33], H/LFGP [34], and RFPP [35] etc. In addition, the distance losses of heatmap-based detectors are usually L1/L2 distances, which are not consistent with the evaluation metric IoU and not friendly for tiny faces.

Based on the analysis above, we try to address the rigorous tiny faces and scale variation problems with heatmap-based method adaptively. The UCM module is proposed

to generate the higher-resolution heatmap with denser proposals and more fine-grained localization of tiny faces and their landmarks. Meanwhile, the context inception block and adaptive scale-aware strategies are developed on the detection heads for the scale variation problems. Therefore, the proposed framework unifies two distinct detection tasks, seamlessly integrating global and local information. This integration enhances the models' perceptual capabilities across scale variations and detailed texture details in both face and landmark detection tasks. Experiments are conducted on both the face detection datasets and the landmark detection datasets. In face detection task, PFLDet demonstrates superior performance on all widely-used datasets, including Wider Face, FDDB, AFW and PASCAL Face, and achieve SOTA results on the most challenging Wider Face dataset. In face landmark detection task, HP-FLDet outperform SOTA single-stage joint detectors on sparse landmark detection task, and achieve competitive results with single-task face landmark detectors on dense landmark detection task.

To summarize our contributions:

- A multi-task single-stage face and landmark detector is proposed, which can detect face and their landmarks with single model in single stage. Meanwhile, it achieves high performance on both face detection and landmark localization tasks, enhancing the detector's versatility and applicability across various scenarios.
- By systematically studying how to integrate the face and landmark detection in single-stage with higher accuracy, an adaptive multi-task detection framework are proposed. The framework can integrate the global and local information for both face detection and facial landmark detection tasks with targeted solutions for tiny/small faces and large scale variation.
- For face detection, HP-FLDet achieve SOTA results on common face detection datasets, especially on the challenging Wider Face Dataset. For landmark detection, HP-FLDet can achieve competitive results on both sparse and dense facial landmark detection datasets.

2. Related Work

2.1. Multi-stage and single-stage face detectors

Based on the detection pipeline, current detectors can be divided into two categories: multi-stage and single-stage. In this subsection, we will discuss the differences of multi-stage and single-stage methods on both single-task face detection methods and multi-task face and landmark detection ones.

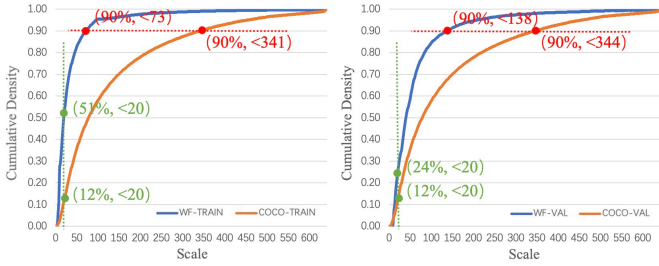


Figure 1: Cumulative density of object size(the minimum length/width) on face detection dataset Wider Face and generic object detection dataset MS-COCO. Face detection has severer tiny-scale problem. There are around 51%/24% faces smaller than 20 pixels on train/val set of Wider Face, while only 12% objects smaller than 20 pixels on MS-COCO.

Single-task Face detectors. Inspired by multi-stage generic object detectors, many multi-stage face detectors are proposed, such as MTCNet [14] and MaskFace [13]. Although with high face detection accuracy, multi-stage detectors suffer from high complexity and latency. Meanwhile, single-stage face detectors [8, 24, 7] have sprung up and been dominated recently. With Feature Pyramid Network(FPN), context enhancement and anchor related strategies, they can achieve high face detection accuracy at higher inference speed compared with multi-stage detectors.

Multi-task face and landmark detectors. For joint face and landmark detection, multi-stage joint detectors offer a general and flexible architecture with the ‘proposal and refinement’ mechanism in which the faces are detected in the first stage and then the landmarks are detected or refined in the following stages. For example, BB-FCN [15] follows a coarse-to-fine cascaded pipeline. MOS [16] fuses separate CNN with a landmark feature learning branch. MaskFace [13] adopt a RoIAlign layer for landmark localization. By contrast, There are much fewer single-stage joint face and landmark detectors as landmarks are hard to be localized with facial context information in the single-stage setting, especially for tiny faces. Almost all of them [17, 18, 19] adapt the architecture of single-task face detector and add a simple landmark regression branch. These methods usually have higher face detection accuracy, but lower landmark localization accuracy compared with multi-stage rivals.

2.2. Anchor-based and anchor-free detectors

Based on the design of detection heads, current detectors can be categorized into anchor-based and anchor-free detectors. Most SOTA face detectors belongs to anchor-based detectors. Practically, the performance of anchor-based detectors drop dramatically when objects becoming smaller as the discrete anchor scales and anchor misalignment. Therefore, most SOTA anchor-based face detectors focus on effective anchor design and sample strategies, such as scale compensation anchor match-

ing strategy [23], PyramidAnchors(PA) [33], Data-Anchor-Sampling [33], Improved Anchor Matching(IAM) and Progressive Anchor Loss(PAL) [32] and IoU-aware Branch [7], to address these problems. Nonetheless, even with empirical settings, proper anchor selection is still challenging with the following misalignment phenomenon [8].

Compared with anchor-based methods, anchor-free methods [27, 36, 37] can reduce anchor-related hyper-parameters with more potentials of generalization ability, and achieve high performance in generic object detection as their superior definition of positive and negative samples. However, there are much fewer anchor-free detectors in the domain of face detection. LFFD [38] follows center-based settings and regards the RFs as natural anchors, which can cover continuous face scales. Actually, LFFD is still an anchor-based method with another way to define anchors. CenterFace [18] transform face detection as a center estimation problem following [36]. Both LFFD and CenterFace focus on light-weighted face detection whose accuracy are lower compared with SOTA anchor-based face detectors.

In this paper, a single-stage anchor-free joint face and landmark detector is proposed, which can improve the landmark detection accuracy in single stage by combining structure and contextual information. We then analyze and mitigate the problems of heatmap-based face detectors when addressing tiny faces, and achieve competitive face detection accuracy with anchor-based face detectors.

3. Methods

The framework of HP-FLDet consists of three parts including backbone, context inception block, and multi-task heads as in Fig 2. In this section, the overall structure will be described with two parts: 1) context inception block, 2) multi-task face and landmark detection, and the scale-aware strategies of both face and landmark detection tasks accordingly.

3.1. Context Inception Block

Different from generic object detectors, HP-FLDet pays more attention to tiny faces and scale variation. As indicated in the introduction section, high resolution heatmap and context inception are import for high performance face and landmark detection. Therefore, Context Inception Block (CIB) is proposed as in the left bottom side of Fig 2. Given an input image I of size $H \times W \times C$, the basic featuremap $F \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C'}$ is generated from the backbone of the feature extractor, then the Context Inception Block extracts task-specific featuremaps $F_{det} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C''}$ and $F_{kps} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C'}$. The Context Inception Block mainly consists of two modules including the Context-aware Module (UCM) and the task-specific SPP modules. Firstly, a Context-aware Module (UCM) is designed for the generation of higher resolution heatmap. The UCM module up-samples the basic featuremap F from $1/4$ to $1/2$ resolution

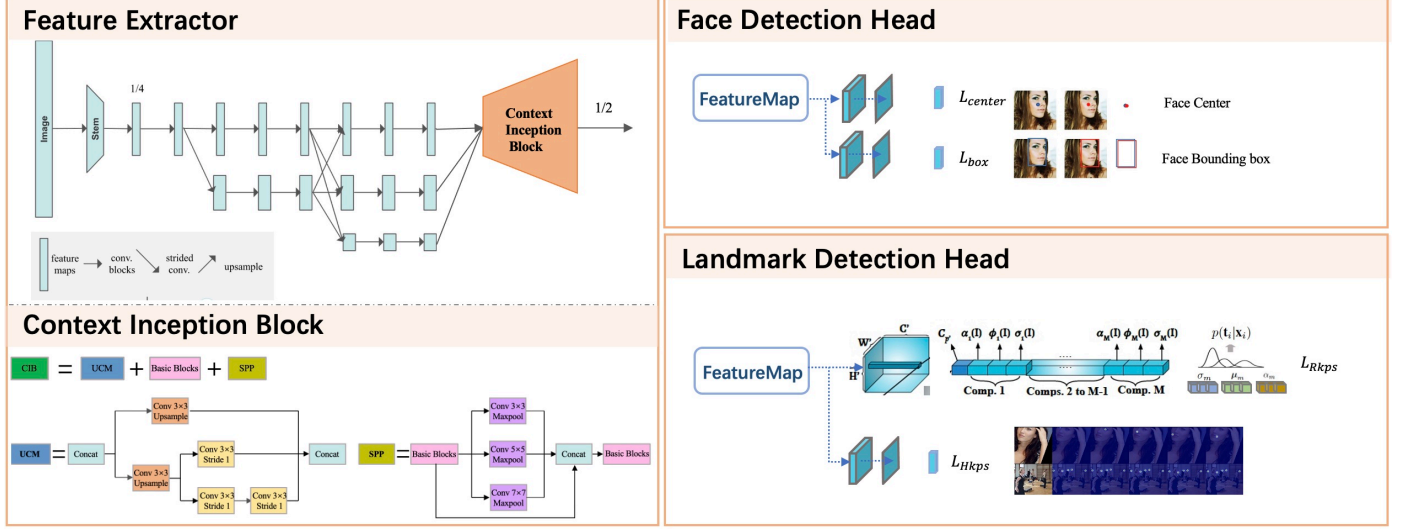


Figure 2: An overview of the proposed single-stage joint face detector HP-FLDet. The framework consists of one feature extractor on the left side and two detection heads on the right side. The feature extractor generate feature maps, while the detection heads detect faces with bounding boxes and landmarks. The context inception block(CIB) of the feature extractor is detailed on the left bottom side.

by paralleling with different depth of 3×3 convolution layers. With higher resolution featuremap, the UCM module helps the detection heads locate tiny faces and their landmarks more precisely. Meanwhile, the paralleled multi-context branches help the UCM module separate out the more important contextual information from higher resolution featuremap and get features from different receptive fields with multi-scale context. Then, we employ two SPP [39, 19] modules to enhance the scale-aware ability and extract task-specific features by pooling multi-scale versions of itself. Featuremaps F_{det} and F_{kps} are generated from separate SPP module for face and landmark detection tasks respectively. In addition, considering the large amount of tiny faces in face and landmark detection task, the convolution kernels in SPP modules are revised to smaller size adaptively to increase the overall face detection performance.

3.2. Multi-task face and landmark detection

In this subsection, we will describe the overall design of the multi-task joint detector, and the scale-aware strategies of both face and landmark detection tasks accordingly.

Multi-task detection settings. For multi-task face and landmark detection, each face is detected with a bounding box and K landmarks. After feature extractor, featuremaps F_{det} and F_{kps} are generated for face and landmark detection, respectively. With the face detection heads on the top right side in Fig 2, each cell f_{det} in featuremap F_{det} generates two kinds of embeddings with channel size $C_i^{det} (i \in [0, 1])$: $C_0^{det} = 1$ for center localization, $C_1^{det} = 4$ for bounding box localization. With the face detection heads on the bottom right side in Fig 2, each cell f_{kps} in featuremap F_{kps} generates two kinds of embeddings

with channel size $C_i^{kps} (i \in [0, 1])$: $C_0^{kps} = K + 2$ for heatmap-based landmark localization, and $C_1^{kps} = M \times (2K + 3)$ channels for regression-based landmark localization. K indicates the number of predicted landmarks of each face. M indicates the number of Gaussian distribution of mixture dense network in equation (6). Therefore, the overall multi-task loss is composed of 4 components including the face classification, the facial bounding box localization, the heatmap-based landmark localization and the regression-based landmark localization as in equation (1).

$$L = L_{center} + \lambda_{box} L_{box} + \lambda_{Hkps} L_{Hkps} + \lambda_{Rkps} L_{Rkps} \quad (1)$$

where L_{center} is the binary loss of face classification, L_{box} is the regression loss of facial bounding boxes, L_{Hkps} is the binary loss of landmarks localization, L_{Rkps} is the regression loss of landmarks with Mixture Dense Regression, λ_{box} , λ_{Hkps} and λ_{Rkps} are the corresponding coefficients. The detailed explanations of both face detection and landmark detection tasks in equation (1) are as follows.

Face detection. As in Figure 2, face detection includes two branches: face classification branch and bounding box regression branch. Face classification is performed on the high-resolution face detection heatmap instead of multi-scale detection heads based on the feature pyramid. In face detection head, cells on the heatmap with peak values are selected as faces. As tiny faces have smaller perceived area, cells with tiny faces are smoothed with adaptive Gaussian radius on the heatmap. We set Gaussian radius with the overlaps of predicted bounding box and the ground-truth adaptively instead of same standard deviation. Face classification loss L_{center} is calculated by the focal loss [29] as

in equation (2).

$$L_{cls} = -\frac{1}{N_{pos}} * \left[\alpha \sum_{i \in Pos} (1 - p_i)^\gamma \log p_i + (1 - \alpha) \sum_{i \in Neg} p_i^\gamma \log (1 - p_i) \right] \quad (2)$$

where p_i is the cells' value on the heatmap, which indicates their probability of face centers, α and γ are the hyperparameters of the focal loss, which are set to 2 and 4 in our experiments. N_{pos} indicates the number of faces in the image.

facial bounding box is generated by a face center with corresponding height, weight and offset on the bounding box regression branch. In most anchor-free object detectors, the box regression loss are smooth L1 or L2 Loss, which are not consistent with the regression evaluation metric (IoU) and not friendly for tiny faces. Therefore, we turn to GIoU Loss for the bounding box regression as in equation (3), where A_c and U are the intersections of the detected bounding boxes and the ground-truth. GIoU Loss is selected as the loss function of face detection instead of others, such as DIoU and CIoU, as it is invariant for face scales and there are no initial anchor-matching problems, which is friendly for heatmap based anchor-free face detection.

$$GIoU = IoU - \frac{|A_c - U|}{|A_c|} \quad (3)$$

In addition, the offset loss is added to adjust the center's localization slightly when remapping its position from the original resolution to the lower resolution as in equation (4), where x_f and y_f are the x and y coordinate of cell f , n is the down sampling ratio of the face detection heatmap compared with the original image.

$$o_f = \left(\frac{x_f}{n} - \left\lfloor \frac{x_f}{n} \right\rfloor, \frac{y_f}{n} - \left\lfloor \frac{y_f}{n} \right\rfloor \right) \quad (4)$$

The overall loss of facial boxes L_{box} is in equation (5).

$$L_{box} = 1 - GIoU + \frac{1}{N} \sum_{n=1}^N \text{Smooth L1 Loss}(o_f, \hat{o}_f) \quad (5)$$

Landmark detection. Face landmarks are detected with two branches: the heatmap-based landmark localization branch and the regressed-based landmark localization branch. The regression-based branch pays more attention to the structural information, while the heatmap-based one can get more local contextual information. In heatmap-based head, points are selected as landmarks when they fall in the facial bounding boxes with peak values. In most heatmap-based landmark detectors, the ground truth heatmaps are smoothed by fixed Gaussian kernels. It is not friendly for large scale variation, and may

introduce confusion when there are tiny/small faces, especially in the unconstrained settings. Therefore, since we can get facial bounding boxes simultaneously, the Gaussian kernel radius is set adaptively by fixed IOU ratio (eg. 0.7) of the facial bounding box in our method. The heatmap-based landmark detection loss L_{Hkps} is calculated similarly with the face classification loss L_{center} in equation (2).

In the cases when heatmap-based head cannot form valid facial landmarks (all cells in the bounding box has lower probability value), especially with large pose, expression, makeup, occlusion and low resolution, landmarks will be generated with the regression-based branch alone. The regression-based loss L_{Rkps} is calculated according to the Mixture Dense Regression [41], which is the first time that MDN is adapted to the facial landmark detection task. For the detected landmarks $T \in \mathbb{R}^{2K}$ and its ground truth $\hat{T}$, The target of MDN is to formulate the probability density of landmarks conditioned on the input. For a M component mixture model (MDN_m), the probability density of the ground truth landmarks $\hat{T}$ conditioned on input image I is generated in equation (6) with component item $m \in M$.

$$p(\hat{T} | I) = \sum_{m=1}^M \alpha_m(I) \phi_m(\hat{T} | I) \quad (6)$$

where α_m is the mixing coefficient and indicates the probability of component m being responsible for generation of landmarks. ϕ_m denotes the density function of m -th mixture component, which can be calculated in equation(7).

$$\phi_m(\hat{T} | I) = \frac{1}{(2\pi)^{c/2} \sigma_m(I)^c} \exp \left\{ -\frac{\|\hat{T} - T_m(I)\|^2}{2\sigma_m(I)^2} \right\} \quad (7)$$

where $T_m(I)$ is the input dependent pose parameters of the component m . $\delta_m(I)$ is the standard deviation of the component m in the horizontal and vertical axes. In our implementation, each point get the same $\delta_m(I)$ and summed up to 1. The regression-based loss L_{Rkps} is generated in a dense fashion where density estimation is done independently at each output cell $f \in F$ as in equation (8).

$$L_{Rkps} = \sum_{i=1}^N -\ln \sum_{m=1}^M \alpha_m(I_i) \phi_m(\hat{T}_i | I_i) \quad (8)$$

where N is the number of images, while I_i is the i -th input image.

4. Experiments

In this section, experiments are conducted on both face detection datasets (Wider Face [26], FDDB [42], AFW [43], PASCAL Face [44]) and landmark detection datasets (Wider Face and AFLW [45]) to evaluate the performance of our method on face and landmark detection tasks respectively.

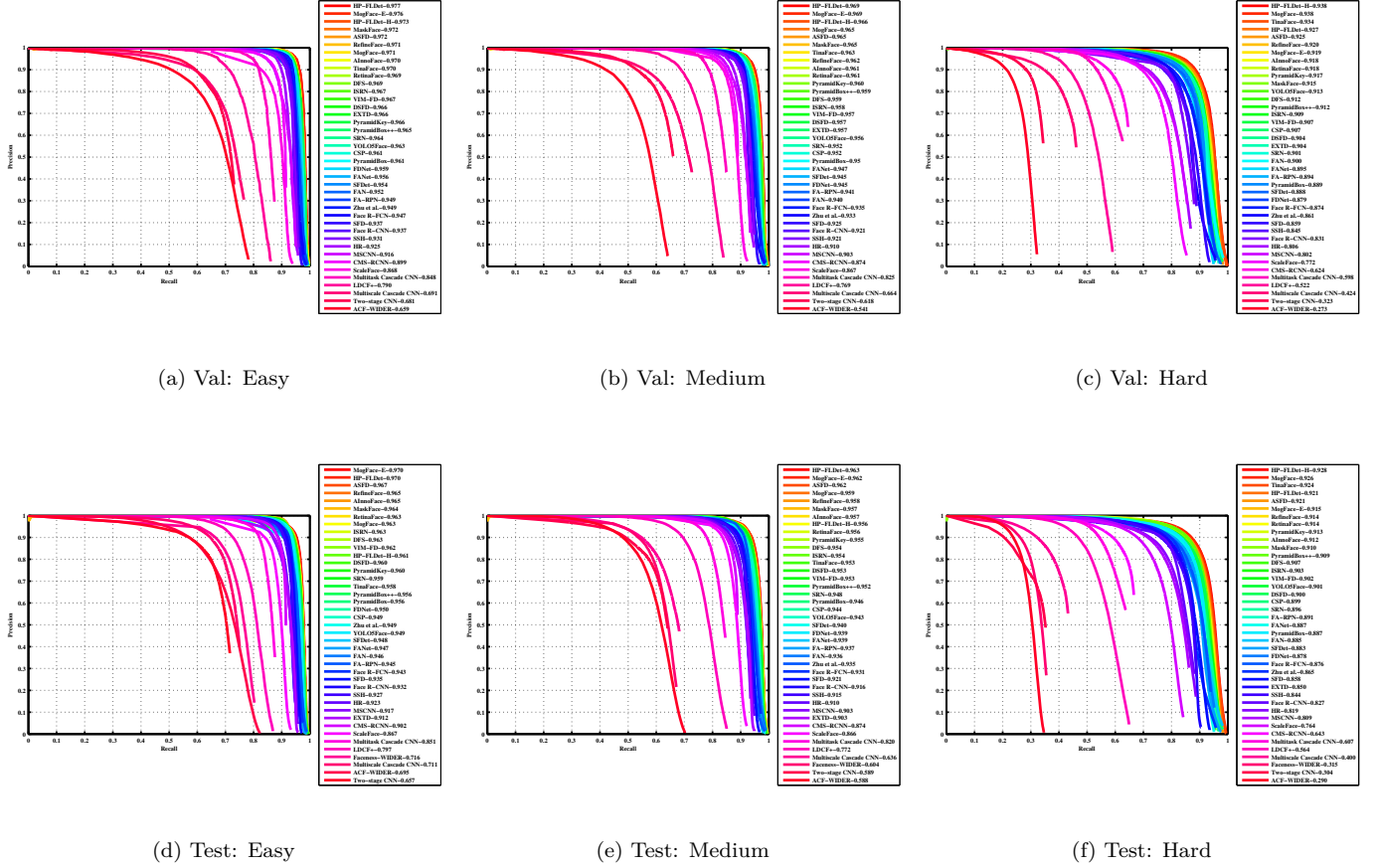


Figure 3: Precision-recall curves on the Wider Face validation and test subsets. The higher the better.

Wider Face contains 32203 images and 393703 labeled face boxes with large variability of scale, pose and occlusion. It is generated by three subset (train, validation and test) with three levels of difficulty (easy, medium and hard). [17] added 5 facial landmarks on the train subset for joint face and landmark detection.

Fddb has 5171 faces in 2845 images. Original Fddb is partly marked by ellipses. Therefore, the annotation with additional labeled faces are adopted in [46] in this section. AFW has 205 images with 473 faces with large variations in both face viewpoint and appearance.

PASCAL Face consists of 851 images and 1335 faces from PASCAL person layout dataset.

AFLW contains 21997 images with 25993 annotated faces. Each face annotation includes a bounding box and 21 visible landmarks. the annotation with 5 landmarks are selected for sparse landmark detection and 19 landmarks for dense landmark detection.

4.1. Implementation Details

HRNet [47, 48] are selected as backbone, which is initialized by the pre-trained models of MS-COCO dataset. The newly additional layers are initialized with 'xavier'. For data augmentation, training patches are generated

by bounding boxes (generated by SSD-sampling [49] and data-anchor-sampling [33]) resized by different scales referring to the SSE strategies [23]. During training, four different schedules are set on the 120, 240, 360 and 400 epochs with SGD optimizer. Learning rate is initialized with $2.5e-4$ and drops by factor 10. For scale-aware supervision, we get the ground truth of face center and landmarks on heatmap with the revised Gaussian kernel setting, and take $M = 3$ as the number of Gaussian distribution in mixture dense network in equation (6,7,8) referring to [41].

4.2. Evaluation on face detection.

In this subsection, HP-FLDet is compared with SOTA face detectors on common face detection benchmarks including AFW, PASCAL Face, Fddb and Wider Face. All of our models are trained only with Wider Face and evaluated with single-scale and multi-scale fusion strategies [8, 33]. For the comparison on the Wider Face and Fddb datasets, the official evaluation toolboxes are adopted. For the comparison on the AFW and PASCAL Face datasets, the evaluation toolbox from [50] are adopted.

Table 1: Performance on the validation subset and test subset of Wider Face.

Method	Validation Set			Test Set		
	Easy	Medium	Hard	Easy	Medium	Hard
RefineFace [40]	0.972	0.962	0.920	0.966	0.958	0.914
TinaFace [7]	0.970	0.963	0.934	0.958	0.953	0.924
ASFD [24]	0.972	0.965	0.925	0.967	0.962	0.921
MaskFace [13]	0.972	0.965	0.915	0.964	0.957	0.910
MogFace-E [23]	0.976	0.969	0.919	0.970	0.962	0.915
MogFace [23]	0.972	0.965	0.938	0.963	0.959	0.926
HP-FLDet	0.977	0.969	0.927	0.970	0.963	0.921
HP-FLDet-H	0.973	0.967	0.938	0.961	0.956	0.928

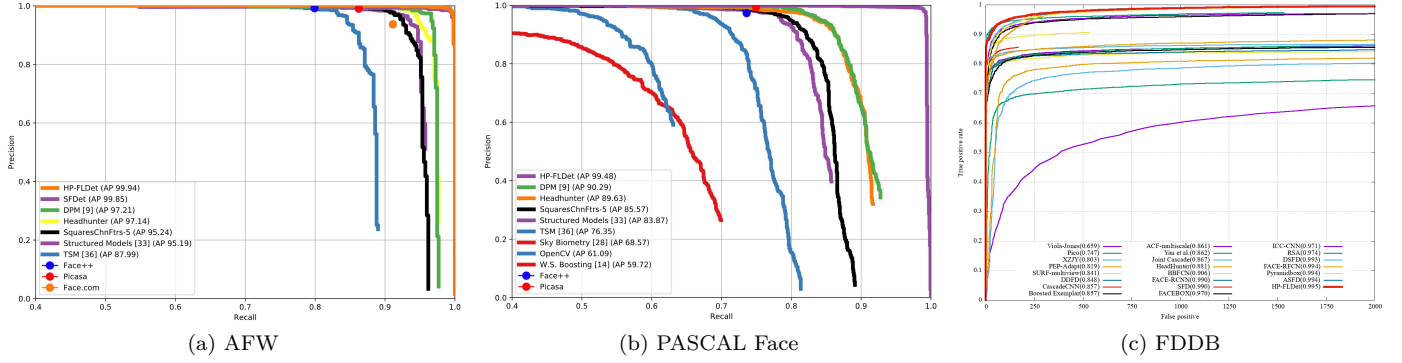


Figure 4: Precision-recall curves for the AFW, PASCAL and FDDB datasets. The higher the better

The precision-recall curves of HP-FLDet are shown in Fig 3, comparing with SOTA face detectors, on the most challenging face detection dataset Wider Face. For a better comparison, detectors with average accuracy of top-5 are selected and listed in Table 1. As in Fig 3 and Table 1, HP-FLDet achieves SOTA results with 0.977(Easy), 0.969 (Medium), 0.927(Hard) mAP on the validation subset and 0.970(Easy), 0.963(Medium), 0.921(Hard) mAP on the test subset. Furthermore, as the scale-level data augmentation influence the models' performance on Easy/Med and Hard subset [23], we trained the hard-preferred model referring to SSE [23] and show the hard-preferred results as HP-FLDet-H in Fig 3 and Table 1. The results of HP-FLDet-H demonstrate the ability of our method to deal with tiny faces and scale variation.

The precision-recall curves of HP-FLDet and other SOTA face detectors on the AFW, PASCAL and FDDB datasets are shown in Fig 4. Methods are evaluated with AP performance on the AFW and PASCAL datasets and TPR@FP2000 (True Positive rate when the number of False Positives Equals to 2000) on FDDB dataset. As we could not generate the precision-recall curves of the recent SOTA detectors on AFW and PASCAL Face datasets, we additionally show the AP metrics collected from re-

lated papers in Table 2 and Table 3. As demonstrated in Fig 4(a) and Table 2, our method achieves 99.94% AP, which outperforming other rivals 0.1% AP at least on the AFW dataset. As in Fig 4(b) and Table 3, our method achieves SOTA results of 99.48% AP on the PASCAL Face dataset. As shown in Fig 4(c), our method achieve 99.5% TPR@FP2000 and outperform other rivals on the FDDB dataset.

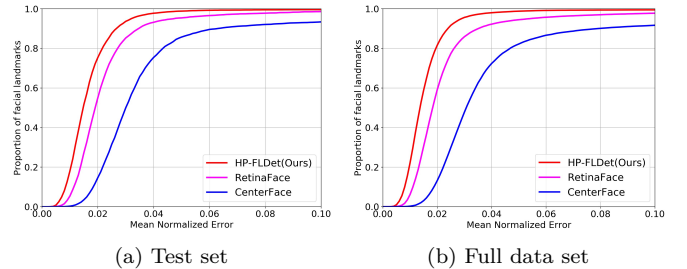


Figure 5: Cumulative distribution for NME of 5 facial landmarks on the AFLW. The higher the better.

Table 2: Performance on AFW (ranked by performance on Wider Face)

Method	Performance (AP)
S3FD [46]	99.85
SRN [51]	99.87
HAMBox [8]	99.90
RefineFace [40]	99.90
MaskFace [13]	99.90
MogFace [23]	99.85
HP-FLDet	99.94

4.3. Evaluation on landmark detection.

In this subsection, we first compare our method with SOTA single-stage joint face and landmark detectors on sparse landmark dataset. Then, our method are compared with SOTA single-task facial landmark detectors on dense landmark dataset. We claim that we do not apply any test time image augmentations as well as any kind of predictions averaging for landmark evaluation.

For the evaluation on the sparse landmark detection, HP-FLDet is compared with the single-stage joint face and landmark detectors RetinaFace [17] and CenterFace [18] on the AFLW dataset [45] with 5 facial landmarks. For fair comparison, all of our models are trained only with the Wider Face training set. All the compared detectors are conducted with the officially pretrained models or models with the best results. We plot the cumulative error distribution (CED) curves of NME on the AFLW test set (4385 images) and AFLW full data set (21997 images) in Figure 5. If any of the methods cannot detect some faces, it will affect only the upper part of the corresponding CED curve to relieve the domain discrepancy of face detection and landmark detection datasets. For qualitative comparison, we demonstrate the detection results of Centerface, Retinaface and HP-FLDet in Fig 6. Both Fig 5 and Fig 6 show that our method outperforms these single-stage joint face detectors by large margins. Additionally, we illustrate the detection result on a selfie of dense faces with score threshold 0.4 in Fig 7. As shown in Fig 7, the facial bounding boxes and the facial landmarks can be stably predicted under the variations of pose, occlusion and resolution.

For the evaluation on the dense landmark detection, we compare HP-FLDet with SOTA facial landmark detectors on both full and frontal sets of the AFLW dataset with 19 facial landmarks in Table 4. During training and validation, we use the same data settings with HRNet. For validation, the obtained values of NME are given in the case when faces are detected. The visualized results of our method are shown in Fig 8.

As shown in Table 4 and Fig 8, our method can achieve competitive results with single-task face landmark detec-

Table 3: Performance on PASCAL Face (ranked by performance on Wider Face).

Method	Performance (AP)
S3FD [46]	98.49
SRN [51]	99.09
HAMBox [8]	99.50
RefineFace [40]	99.45
MaskFace [13]	99.48
MogFace [23]	99.32
HP-FLDet	99.48

tors on dense landmark dataset. Note that even though PDB(w/DA) and LAB(w/B) has higher accuracy, LAB uses extra boundary information(w/B), PDB uses stronger data augmentation strategies(w/DA).

Table 4: NME of 19 facial landmarks on AFLW test subset.

Method	Full	Front
TSR [52]	2.17	—
CPM+SBR [53]	2.14	—
SAN [54]	1.91	1.85
LAB(w/oB) [55]	1.85	1.62
HRNet [56]	1.57	1.46
LUVLiFace [57]	1.39	1.19
PIPNet [58]	1.42	-
PDB(w/DA) [59]	1.47	—
LAB(w/B) [55]	1.25	1.14
HP-FLDet	1.33	1.25

Table 5: Ablation experiments of face detection on the Wider Face validation subset.

No.	Modules	Easy	Medium	Hard
1	Baseline	0.905	0.898	0.841
2	Baseline+CIB	0.952	0.944	0.895
3	Baseline+CIB+MRS	0.953	0.944	0.902

4.4. Ablation Studies

To achieve a better demonstration of the proposed method, extensive ablation experiments are conducted to examine how the proposed modules affect the performance of face and landmark detection.

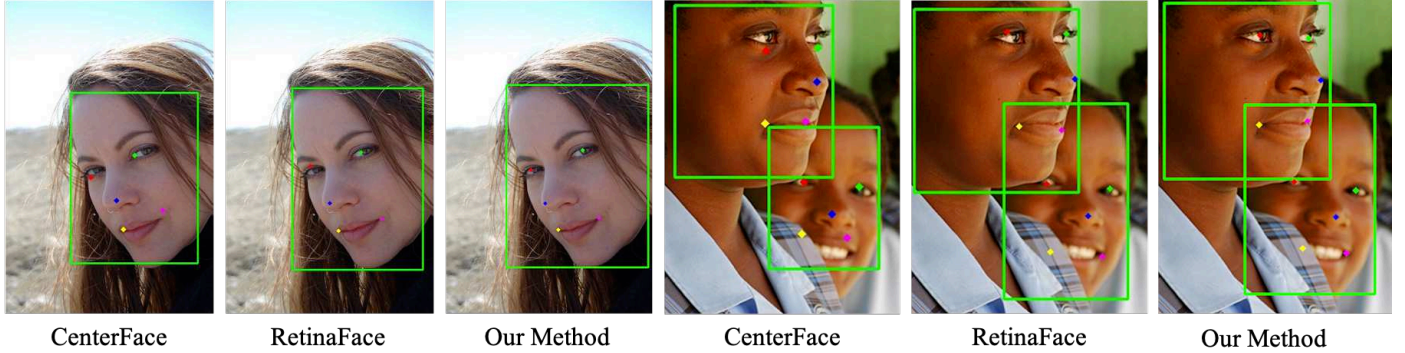


Figure 6: Illustration of differences in localization of 5 facial landmarks for CenterFace, RetinaFace and our method on AFLW test set.

Table 6: Ablation experiments for dense landmark detection with NME for 19 facial landmarks on AFLW test set.

No.	CIB	Regression	MDN	Full	Frontal
1	–	✓	–	1.59	1.52
2	✓	✓	–	1.34	1.26
3	✓	–	✓	1.33	1.25

Firstly, the AP performance of face detection on the Wider Face validation subset are compared with the proposed modules in section 3, including the Context Inception block(CIB) and Multi-resolution Supervision strategies(MRS). For more intuitive comparison of different modules, the ablation studies of face detection are conducted with the Single Model Single Inference on the Original(SIO) schema [38] instead of multi-scale strategies. As shown in Table 5, The context inception block can get gains of nearly 0.05 mAP on Wider Face val subset comparing experiments 1 and 2. Meanwhile, the detection accuracy can be improved by 0.01 mAP on the hard set with multi-resolution strategies comparing experiments 2 and 3.

Then, we demonstrate how the heatmap-based head and MDN-based head influence the performance of facial landmark detection. We adapt the same setting with subsection 4.4 and conduct four experiments with the combination of heatmap, MDN and ordinary regression based heads as in Table 6. By comparing the results of experiments 1, 2 and 3, we can see that both high resolution heatmap and multi-task detection heads play roles in the accuracy of landmark localization. What’s more, MDN outperform ordinary regression head for landmark localization as shown in experiment 3 and 4.

5. Discussion

To enhance the accuracy of both face and landmark detection in a single-stage model, a high performance face and landmark detector HP-FLDet is proposed. This de-

tector, within a unified framework, ensures the consistency of both face and facial landmark detection tasks in a single stage. For the perception capability of the rigorous tiny/small faces, we introduce the UCM module with a higher-resolution heatmap, which can not only generate more anchor proposals for face detection task, but also obtain more local detail information for landmark localization task. For the scale variation issue, the context inception block and scale-aware strategies are proposed, which can mitigate the limitations of scale perception imposed by the single-resolution heatmap. Experimental results demonstrate that our proposed framework achieves high performance on both face detection and landmark localization tasks.

One significant challenge of joint face detectors is to detect the rigorous tiny/small faces. Although existing detectors deploy specific modules for dense anchor design and mining, their recall rates still need to be improved under the challenging conditions. Our method, as illustrated in Fig 3 and Fig 4, exhibits the higher recall rate under the same precision compared with the competitors in face detection tasks, especially on the extreme conditions like the WIDER FACE hard sets (Fig 3 (c) and (f)). This is due to its heightened ability to perceive the tiny targets with more detailed texture information and more anchor proposals. At the same time, the accuracy of facial landmark localization is improved with more detailed facial texture information, especially on the corners of eyes and mouth in occluded, lateral situations (Fig 6) and tiny faces (Fig 7).

Another challenge of joint face detectors is the scale variation problem. Even with FPN and context enhancement, the existing detectors struggle with deep-level feature fusion and misalignment problems. To improve scale awareness on a single feature heatmap, the context inception block and the scale-aware strategies are proposed in HP-FLDet. Results in Table 5 and Table 6 demonstrate their significant contributions to the accuracy improvement in both face and landmark detection tasks, which achieve nearly 5% mAP gains in face detection on the Wider Face dataset (Table 5) and reduce the full landmark NME by about 0.3 in facial landmark detection on the AFLW test set (Table 6). Notably, HP-FLDet maintains balanced ac-

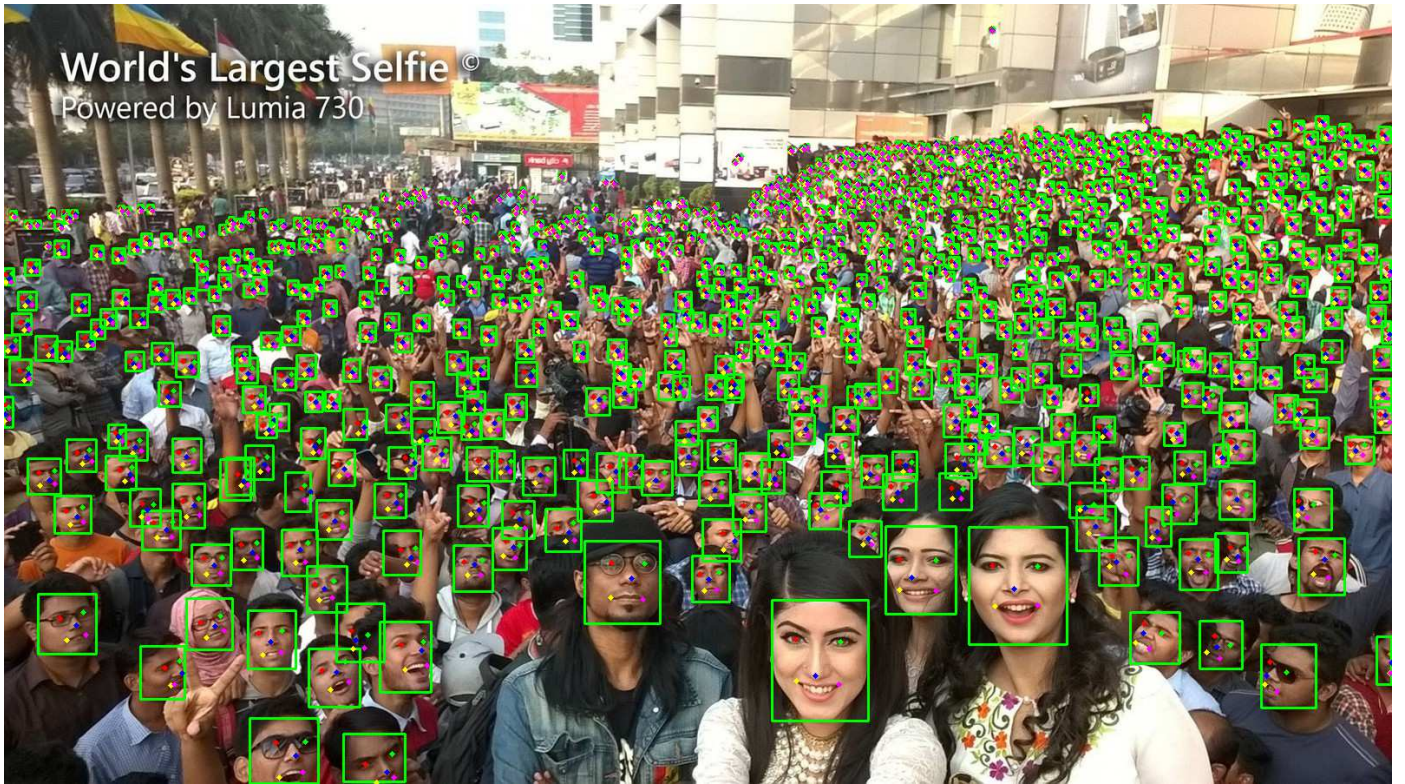


Figure 7: Visualization of face and landmark detection result of our method with single stage.

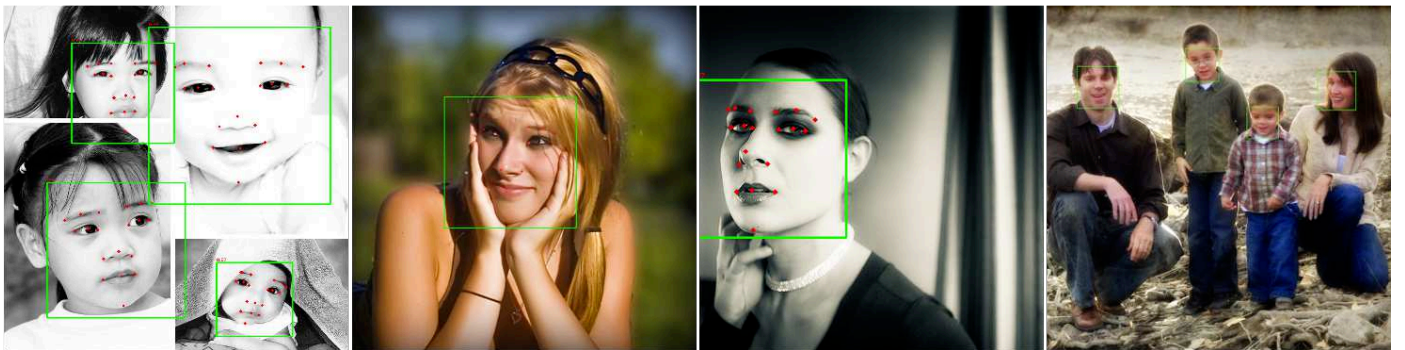


Figure 8: Visualization of 19 facial landmarks predicted by our method on AFLW test set.

curacy across different face sizes, outperforming competitive methods like MogFace-E, particularly on the hard sets (with improvements of nearly 0.7% in Table 1). The higher and balanced accuracy indicates that the model can supervise detection on faces of different sizes, which can achieve higher accuracy not only on the hard sets with more tiny faces, but also on the easy and medium sets with more large and normal sized faces simultaneously.

In the face detection task, HP-FLDet is compared with the SOTA detectors on almost all of the common used face detection datasets including Wider Face, FDDB, AFW and ASCAL Face. HP-FLDet achieve superior performance on all of these datasets compared with the rivals, especially in challenging and difficult situations. The proposed Context Inception Block and all of the Multi-resolution Supervision strategies contribute to these improvements when dealing with tiny faces and scale variation. In the landmark detection task, the performance of HP-FLDet and the state-of-the-art methods are compared on both sparse landmark dataset and dense landmark on ALFW datasets. HP-FLDet shows superior performance on both tasks in the unconstrained challenging settings including large pose, facial expressions, low image quality and occlusions. As in Table 6, heatmap-based multi-task detection framework combined with Context Inception Block and Mixture Dense Regression can contribute to all these challenging situations.

As a single-stage detector, HP-FLDet is more efficient and practical than single-task or multi-stage detectors, offering higher accuracy without the need for a second stage for refinement or landmark detection. However, it still has room for improvement as the up-sample operation in UCM module increases the computational complexity of the detector. In the future, the efficient of HP-FLDet can be further improved by optimizing the UCM module or deploying lighter backbones.

6. Conclusion

In this paper, a joint detection framework HP-FLDet is proposed to perform face and landmark detection simultaneously within single stage. For face detection, the heatmap-based detector can achieve competitive accuracy with anchor-based rivals when dealing with scale variation and tiny faces. For landmark detection, multi-task detection framework combined with high resolution heatmap and mixture dense regression show superior performance on both sparse and dense facial landmark detection tasks in the unconstrained settings. With a single model in single stage, our method can integrate structural and contextual information and achieve SOTA results on both face detection and landmark localization tasks, which makes the face detector more useful, since landmarks are often needed in many facial related applications.

Acknowledgments

This research was funded by the 2022 Special Research Grant for Medical-Engineering Integration from the General Hospital of Ningxia Medical University (Grant No. NYZYYG-007) and the Key Research and Development Program of Ningxia Hui Autonomous Region (Grant No. 2023BEG02035).

Author Contributions

Xueli Liu and Yinlong Liu contributed to the design and implementation of the method. Feng Wang and Kexue Fu prepared the datasets and analyzed the results. Xueli Liu, Xiao Chen and Xinrong Chen wrote the manuscript. Xiaobo Yang revised the manuscript. Xinrong Chen and Xiaobo Yang in charge of overall direction and planning. All authors discussed the results and commented on the manuscript.

Data Availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study. Datasets utilized for the evaluation of this work have been clarified in Section 4. This study does not contain any studies with human or animal subjects performed by any of the authors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Z. Huang, J. Zhang, H. Shan, When age-invariant face recognition meets face age synthesis: A multi-task learning framework and a new benchmark, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (6) (2023) 7917–7932. doi:10.1109/TPAMI.2022.3217882.
- [2] D. Cheng, X. Zhang, X. Xu, Reweighted robust and discriminative latent subspace projection for face recognition, *Information Sciences* (2023) 119947.
- [3] Z. Huang, K. C. Chan, Y. Jiang, Z. Liu, Collaborative diffusion for multi-modal face generation and editing, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6080–6090.
- [4] C. Peng, C. Zhang, D. Liu, N. Wang, X. Gao, Hifisketch: High fidelity face photo-sketch synthesis and manipulation, *IEEE Transactions on Image Processing* 32 (2023) 5865–5876. doi:10.1109/TIP.2023.3326680.
- [5] S. Hariharan, J. D. Pushparaj, M. Malarvel, Computer vision based student behavioral tracking and analysis using deep learning, in: *International Conference on Electronics and Sustainable Communication Systems*, 2022, pp. 772–777.
- [6] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, F. Wen, General facial representation learning in a visual-linguistic manner, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18697–18709.

- [7] Y. Zhu, H. Cai, S. Zhang, C. Wang, Y. Xiong, Tinaface: Strong but simple baseline for face detection (2021). arXiv:2011.13183.
- [8] Y. Liu, X. Tang, J. Han, J. Liu, D. Rui, X. Wu, Hambox: Delving into mining high-quality anchors on face detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13043–13051.
- [9] Z. Chen, C. Yang, Q. Li, F. Zhao, Z.-J. Zha, F. Wu, Disentangle your dense object detector, in: 29th ACM international conference on multimedia, 2021, pp. 4939–4948.
- [10] U. Watchareeruetai, B. Sommana, S. Jain, P. Noinongyao, A. Ganguly, A. Samacoits, S. W. Earp, N. Sritrakool, Lotr: face landmark localization using localization transformer, IEEE Access 10 (2022) 16530–16543.
- [11] H. Li, Z. Guo, S.-M. Rhee, S. Han, J.-J. Han, Towards accurate facial landmark detection via cascaded transformers, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4176–4185.
- [12] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, X.-J. Wu, Face detection, bounding box aggregation and pose estimation for robust facial landmark localisation in the wild, in: IEEE/CVF conference on computer vision and pattern recognition workshops, 2017, pp. 160–169.
- [13] D. Yashunin, T. Baydasov, R. Vlasov, Maskface: multi-task face and landmark detector (2020). arXiv:2005.09412.
- [14] L. Zhou, H. Zhao, J. Leng, Mtcnet: Multi-task collaboration network for rotation-invariance face detection, Pattern Recognition 124 (2022) 108425.
- [15] L. Liu, G. Li, Y. Xie, Y. Yu, Q. Wang, L. Lin, Facial landmark machines: A backbone-branches architecture with progressive representation learning, IEEE Transactions on Multimedia 21 (9) (2019) 2248–2262.
- [16] Y. Liu, Z. Gu, S. Gao, D. Wang, Y. Zeng, J. Cheng, Mos: A low latency and lightweight framework for face detection, landmark localization, and head pose estimation (2021). arXiv:2110.10953.
- [17] J. Deng, J. Guo, E. Ververas, I. Kotsia, S. Zafeiriou, Retinaface: Single-shot multi-level face localisation in the wild, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5203–5212.
- [18] Y. Xu, W. Yan, G. Yang, J. Luo, T. Li, J. He, Centerface: joint face detection and alignment using face as point, Scientific Programming (2020) 1–8.
- [19] D. Qi, W. Tan, Q. Yao, J. Liu, Yolo5face: why reinventing a face detector, in: European Conference on Computer Vision, Springer, 2022, pp. 228–244.
- [20] W. McNally, K. Vats, A. Wong, J. McPhee, Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation, in: European Conference on Computer Vision, Springer, 2022, pp. 37–54.
- [21] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, C. Lu, Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (6) (2022).
- [22] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, Foveabox: Beyond anchor-based object detection, IEEE Transactions on Image Processing 29 (2020) 7389–7398.
- [23] Y. Liu, F. Wang, B. Sun, H. Li, Mogface: Towards a deeper appreciation on face detection, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021) 4083–4092.
- [24] B. Zhang, J. Li, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Xia, W. Pei, R. Ji, Asfd: Automatic and scalable face detector, 29th ACM International Conference on Multimedia (2021) 2139–2147.
- [25] J. Jeong, B. Kim, J. Yu, Y. Yoo, Eresfd: Rediscovery of the effectiveness of standard convolution for lightweight face detection (2023). arXiv:2204.01209.
- [26] S. Yang, P. Luo, C.-C. Loy, X. Tang, Wider face: A face detection benchmark, in: IEEE/CVF conference on computer vision and pattern recognition, 2016, pp. 5525–5533.
- [27] J. Wang, L. Song, Z. Li, H. Sun, J. Sun, N. Zheng, End-to-end object detection with fully convolutional network, in: IEEE/CVF conference on Computer Vision and Pattern Recognition, 2021, pp. 15849–15858.
- [28] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, Q. Ye, Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8792–8801.
- [29] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, J. Yang, Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection, Advances in Neural Information Processing Systems 33 (2020) 21002–21012.
- [30] L. Wang, X. Yu, T. Bourlai, D. N. Metaxas, A coupled encoder-decoder network for joint face detection and landmark localization, Image and Vision Computing 87 (2019) 37–46.
- [31] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Processing of the European conference on computer vision, Springer, 2014, pp. 740–755.
- [32] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, F. Huang, Dsfd: dual shot face detector, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5060–5069.
- [33] X. Tang, D. K. Du, Z. He, J. Liu, Pyramidbox: A context-assisted single shot face detector, in: European conference on computer vision, 2018, pp. 797–813.
- [34] D. Mamieva, A. B. Abdusalomov, M. Mukhiddinov, T. K. Whangbo, Improved face detection method via learning small faces on hard images based on a deep learning approach, Sensors 23 (1) (2023) 502.
- [35] J. He, X. Song, Z. Feng, T. Xu, X. Wu, J. Kittler, Etm-face: effective training sample selection and multi-scale feature learning for face detection, Multimedia tools and applications (2023) 1–17.
- [36] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: European conference on computer vision, 2018, pp. 734–750.
- [37] Zhu, Chencheng and He, Yihui and Savvides, Marios, Feature selective anchor-free module for single-shot object detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 840–849.
- [38] D. Xu, L. Wu, Y. He, Q. Zhao, M. Jian, J. Yan, L. Zhao, Os-lfd: a light and fast face detector with ommateum structure, Multimedia Tools and Applications 80 (2021) 34153–34172.
- [39] S. Elfving, E. Uchibe, K. Doya, Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, Neural Networks 107 (2018) 3–11.
- [40] S. Zhang, C. Chi, Z. Lei, S. Z. Li, Refineface: Refinement neural network for high performance face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (11) (2020) 4008–4020.
- [41] A. Varamesh, T. Tuytelaars, Mixture dense regression for object detection and human pose estimation, in: IEEE/CVF conference on Computer Vision and Pattern Recognition, 2020, pp. 13086–13095.
- [42] V. Jain, E. Learned-Miller, Fddb: A benchmark for face detection in unconstrained settings, Tech. rep., UMass Amherst technical report (2010).
- [43] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: IEEE/CVF conference on Computer Vision and Pattern Recognition, 2012, pp. 2879–2886.
- [44] J. Yan, X. Zhang, Z. Lei, S. Z. Li, Face detection by structural models, Image and Vision Computing 32 (10) (2014) 790–799.
- [45] M. Kostinger, P. Wohlhart, P. M. Roth, H. Bischof, Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization, in: 2011 IEEE International Conference on Computer Vision Workshops, 2011, pp. 2144–2151.
- [46] Zhang, Shifeng and Zhu, Xiangyu and Lei, Zhen and Shi, Hailin and Wang, Xiaobo and Li. Stan Z, S3fd: Single shot scale-invariant face detector, in: IEEE international conference on

- computer vision, 2017, pp. 192–201.
- [47] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, L. Zhang, Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, in: IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5386–5395.
 - [48] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, *IEEE transactions on pattern analysis and machine intelligence* 43 (10) (2020) 3349–3364.
 - [49] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, *Lecture Notes in Computer Science* (2016) 21–37.
 - [50] M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, in: *European conference on computer vision*, Springer, 2014, pp. 720–735.
 - [51] Chi, Cheng and Zhang, Shifeng and Xing, Junliang and Lei, Zhen and Li, Stan Z and Zou, Xudong, Selective refinement network for high performance face detection, in: *AAAI conference on artificial intelligence*, Vol. 33, 2019, pp. 8231–8238.
 - [52] J. Lv, X. Shao, J. Xing, C. Cheng, X. Zhou, A deep regression architecture with two-stage re-initialization for high performance facial landmark detection, in: *IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 3317–3326.
 - [53] X. Dong, S.-I. Yu, X. Weng, Y. Yang, Y. Sheikh, Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors, in: *IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2018, pp. 360–368.
 - [54] X. Dong, Y. Yan, W. Ouyang, Y. Yang, Style aggregated network for facial landmark detection, in: *IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2018, pp. 379–388.
 - [55] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, Q. Zhou, Look at boundary: A boundary-aware face alignment algorithm, in: *IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 2129–2138.
 - [56] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.
 - [57] A. Kumar, T. K. Marks, W. Mou, Y. Wang, M. Jones, A. Cherian, T. Koike-Akino, X. Liu, C. Feng, Luvli face alignment: Estimating landmarks’ location, uncertainty, and visibility likelihood (2020).
 - [58] H. Jin, S. Liao, L. Shao, Pixel-in-pixel net: Towards efficient facial landmark detection in the wild, *International Journal of Computer Vision* 129 (12) (2021) 3174–3194.
 - [59] Feng, Zhen-Hua and Kittler, Josef and Awais, Muhammad and Huber, Patrik and Wu, Xiao-Jun, Wing loss for robust facial landmark localisation with convolutional neural networks, in: *IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 2235–2245.