MOS: A Low Latency and Lightweight Framework for Face Detection, Landmark Localization, and Head Pose Estimation

Yepeng Liu[1]
ericyepeng123@gmail.com

Zaiwang Gu[1]
guzaiwang01@gmail.com

Shenghua Gao[2]
gaoshh@shanghaitech.edu.cn

Dong Wang[1]
dong.wang@ubtrobot.com

Yusheng Zeng[1]
bill.zeng@ubtrobot.com

Jun Cheng[†31]
sam.j.cheng@gmail.com

[1] UBTech Research,
UBTech Robotics Corp Ltd, China

[2] Shanghaitech University,
China

[3] Institute for Infocomm Research,
Agency for Science, Technology
and Research, Singapore
[†] Corresponding Author

Abstract

With the emergence of service robots and surveillance cameras, dynamic face recognition (DFR) in wild has received much attention in recent years. Face detection and head pose estimation are two important steps for DFR. Very often, the pose is estimated after the face detection. However, such sequential computations lead to higher latency. In this paper, we propose a low latency and lightweight network for simultaneous face detection, landmark localization and head pose estimation. Inspired by the observation that it is more challenging to locate the facial landmarks for faces with large angles, a pose loss is proposed to constrain the learning. Moreover, we also propose an uncertainty multi-task loss to learn the weights of individual tasks automatically. Another challenge is that robots often use low computational units like ARM based computing core and we often need to use lightweight networks instead of the heavy ones, which lead to performance drop especially for small and hard faces. In this paper, we propose online feedback sampling to augment the training samples across different scales, which increases the diversity of training data automatically. Through validation in commonly used WIDER FACE, AFLW and AFLW2000 datasets, the results show that the proposed method achieves the state-of-the-art performance in low computational resources.

1 Introduction

With the popularization of service robots and surveillance cameras, dynamic face recognition (DFR) in the wild has become widely used. DFR is different from static face recognition

(SFR). In SFR such as Apple face ID, the aim is to recognize faces within a specific area or range with relatively high requirements for the angle, distance, and position, which often requires human cooperation. However, in DFR applications such as surveillance, the aim is to automatically recognize persons walking in natural forms, generally requiring no human cooperation. The traditional pipeline for DFR is to: 1) face detection and tracking; 2) head pose estimation and face image quality evaluation; 3) image selection based on the head pose and quality; 4) face identity comparison. The face recognition can be deployed on device with low computing power or cloud side. However, cloud service is often interrupted by unstable network connection and limited bandwidth. Therefore, real-time face detection and head pose estimation in mobile side become necessary for DFR in the wild.

Face detection is a prerequisite step of facial image analysis such as facial attribute, e.g., expression [43], age [61] and face identity [7, 26, 56]. With the recent development of deep learning, significant improvements have been achieved in face detection by utilizing CNN-based object detectors. Among these methods, single-stage based deep learning methods have shown to be promising [50, 58, 46, 47]. The single-stage methods detect the faces of different scales without significant increment of time-consumption even when there is an increased number of faces in the images. These methods densely sample face locations and scales on feature pyramids [23, 25], demonstrating promising performance and yielding high speed. Very often, the five facial landmarks are localized simultaneously [9]. However, landmark localization is still insufficient for pose estimation in DFR due to poor accuracy for face with large angles. In this work, we propose a novel low latency and lightweight framework for real-time face detection, landmark localization and pose estimation. To avoid a large increment in computation, we propose to detect face, locate the landmarks and estimate the head pose simultaneously.

There are three major challenges. Firstly, it is not an easy task to learn face detection, facial landmark localization and head pose estimation simultaneously and get accurate results for all tasks with low computational resources [9, 52]. Secondly, the existing datasets for face detection, facial landmark localization and head pose estimation are non-unified and inconsistent of the scales for joint training while it is difficult to obtain accurate pose annotations, especially for small faces [53]. Lastly, false positive in face detection affects the user experience and it is very challenging to reduce it with low computational resources.

Intuitively, attention area by pose estimation shall be consistent with that by face classification. Therefore, we train a multi-task network and utilize the feature maps to validate our assumption visually. The result shows that the attention area for pose estimation and classification are highly overlapped. This motivates us to integrate the pose estimation into face detection such that it may maintain or even improve the accuracy while reducing extra step for pose estimation. In addition, we observe that the landmark localization is more difficult in faces with large angles and propose an additional constraint called pose loss to regularize the training. In order to train a model for simultaneous face detection and head pose estimation, we further label the pose of each face for the WIDER FACE dataset, which will be released.

The main contributions are summarized as follows:

1. We propose a low latency and lightweight architecture for face detection, facial landmark localization and head pose estimation simultaneously.

2. We propose uncertainty multi-task loss to make the face detection, landmark localization and head pose estimation more accurate.
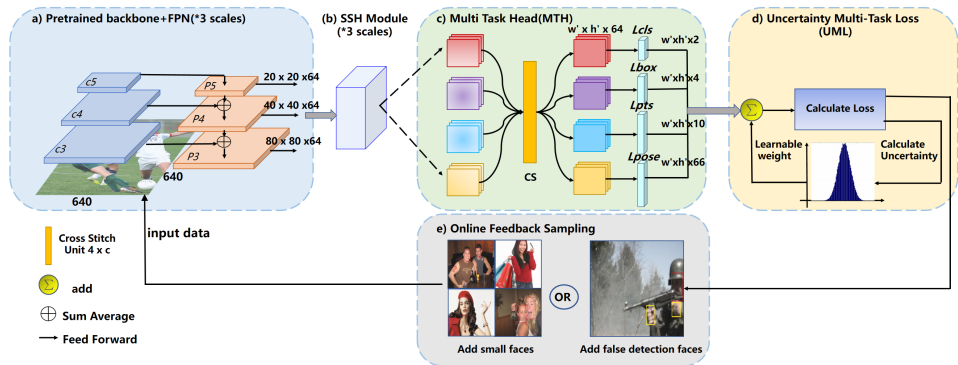
Figure 1: Overview of the proposed method. MOS adopts the feature pyramid from *P3* to *P5* followed by SSH context head module and a multi-task head module. A cross stitch unit ($n \times c$, n denotes task number and c denotes channel number) is used in the multi-task head to combine the feature maps linearly for face detection, landmark localization and head pose estimation.

3. We propose an online data-feedback augmentation strategy, which improves the data balance in face detection.

4. The results show that our method outperforms other lightweight methods, especially for the *hard* subset.

2 Related Work

Face detection: Face detection [17, 40] has been a hot topic in the past decades. Earlier methods [39] for face detection are mostly based on hand-crafted features. Existing methods are divided into two categories: two-stage methods (e.g. Faster R-CNN [33]) and single-stage methods (e.g. SSD [25]). Single-stage methods have great advantages in inference speed, though these detectors often lead to higher false positive. There are some methods [4, 27] that reduce the false positive detection by improving the quality of the matching box. Another problem is how to detect small faces. S3FD [46] used multiple strategies to improve the performance of small faces. ProgressFace [50] proposed a novel scale-aware progressive training mechanism to address large scale variations(across faces.

Head pose estimation: Head pose estimation has been widely studied. Early method uses the facial landmarks [3, 18, 21, 52] to estimate the head pose. The methods based on facial landmarks and Perspective-n-Point (PnP) [10] are very popular because there is no need to include a pose estimation model. However, the error in landmark localization will be propagated. Recently, CNN based methods for direct pose estimation made some progress. Hopenet [34] combined ResNet50 with a multi-loss function, which is composed of a classification loss and regression loss. FSAnet [41] proposed to learn a fine grained structure mapping for spatially grouping features before aggregation. Although some of these methods [6, 16, 49] achieve excellent results on pose estimation, none of them is combined with face detection.

Multi-task learning: The recent progress in multi-task learning mainly focuses on the

shared network architecture [29] and the loss function weights of the tasks [19]. In MTAN [24], Liu *et al.* proposed a new network based on SegNet [2] and obtained the state-of-the-art performance on the tasks of semantic segmentation and depth estimation on the outdoor CityScapes dataset [5]. There are some methods attempting to solve the problem of face detection and alignment in one model. In MTCNN [44], Zhang *et al.* used a cascaded architecture with three stages of shallow networks to predict face and landmark locations in a coarse-to-fine manner. In RetinaFace [8], Deng *et al.* manually annotated five facial landmarks on the WIDER FACE dataset and observed significant improvement in *hard* subset with the assistance of this extra supervision. There are also some efforts to include head pose estimation and other face attributes [32, 48]. Ranjan *et al.* [32] presented an algorithm for simultaneous face detection, landmark localization, head pose estimation and gender recognition, which shows that the multi-task learning can get better results. However, these methods use hard parameter connection, which limits the feature sharing among different tasks.

3 Method

In this paper, we propose a low latency and lightweight framework including the following sub-tasks: face classification, bounding box regression, landmark regression and head pose estimation. Fig. 1 shows the overview of the proposed method.

3.1 Multi-Task Head

Most of the multi-task branches of face detection [9, 32, 44] are forked from the last layer directly, and therefore these sub-tasks actually share all the previous features. Fig. 2(a) shows the baseline head module [9, 32] with hard parameter connection. Although some correlation exists between them, there are still many differences among the sub-tasks. For example, landmark regression pays more attention to the location of each landmark while bounding box regression pays more attention to the edge of face area [53].

Inspired by the above, we propose a novel head sharing unit called multi-task head (MTH). The MTH units try to find the best shared representations for multi-task learning. Different heads share representations through linear combinations and learn the optimal combinations for these tasks. Fig. 2(a) shows the structure of MTH Different from the baseline head module [9, 32] with hard parameter connection, we further include a cross stitch unit size of $n \times c$ in Fig. 2(a), where n denotes the number of tasks and c denotes the number of channels. It computes a linear combination of feature maps followed by additional 1×1 convolutions. It shall be noted that the cross stitch unit here is different from that in [29]. In our unit, we use channel-wise weights in each channel while shared values were used in [29].

3.2 Multi-task Loss

3.2.1 Loss functions used for the tasks

The classification loss is a softmax loss for binary classification denoted as $L_{cls}(p_i, p_i^*)$, where 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.

(a) The different connection method of head modules
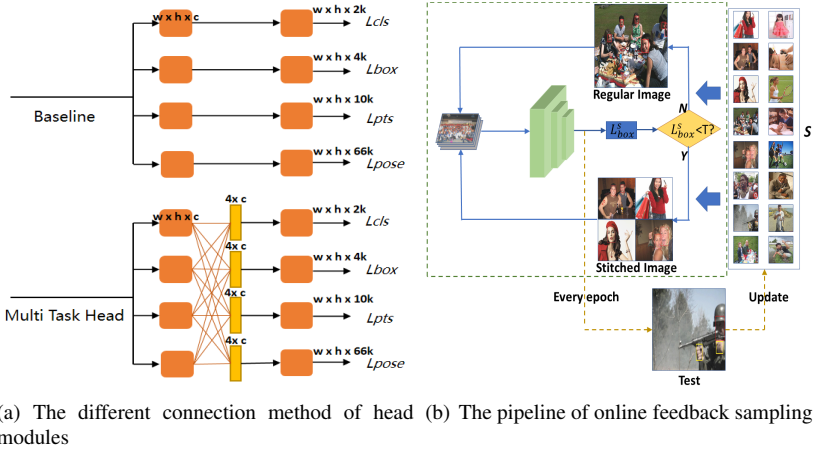
(b) The pipeline of online feedback sampling

Figure 2: (a) The top figure is the baseline head module with hard parameters connection; The bottom is the proposed MTH module with cross-connection. (b) Online Feedback Sampling: Compute L_{box}^s for faces no larger than 25×25 and apply stitching accordingly after every iteration, and value T is set to 0.35; Update the set S after every epoch. The yellow and red boxes in the test image refer to false detection and correct detection respectively. The images will be added into the set S according to the number of false detection boxes.

The bounding box regression loss is a Smooth ℓ_1 loss [[11]] denoted as $L_{box}(t_i, t_i^*)$, where $t_i = (t_x, t_y, t_w, t_h)$ and $t_i^* = (t_x^*, t_y^*, t_w^*, t_h^*)$ represent the coordinates of the predicted box and ground-truth box associated with the positive anchors respectively.

The landmark regression loss $L_{pts}(l_i, l_i^*)$ is also based on Smooth ℓ_1, where $l_i = \{l_{x_1}, l_{y_1}, \cdots, l_{x_5}, l_{y_5}\}$ and $l_i^* = \{l_{x_1}^*, l_{y_1}^*, \cdots, l_{x_5}^*, l_{y_5}^*\}$ represent the predicted landmarks and ground truth with the positive anchors respectively.

Head Pose estimate loss: Previously, a regression loss L_{pose} based on cross entropy and mean square error is used for pose estimation [[34]].

$$L_{pose} = H(a, \hat{a}) + \beta \cdot MSE(a, \hat{a}), \tag{1}$$

where $H(a, \hat{a})$ denotes the cross entropy, MSE denotes the mean square error, a and $\hat{a}$ denote the annotated bins and the predict bins. Similar to that in [[34]], we use 66 bins, β is set to 0.001.

3.2.2 Uncertainty Multi-task Loss

The weights in multi-task learning are set empirically in RetinaFace [[9]] and MTCNN [[44]]. Recently, some work utilizes the uncertainty to estimate the output in the generic object detection [[13, 19]]. In this work, we derive a uncertainty multi-task loss (UML) function by maximizing the overall likelihood with the network.

Let $f^W(x)$ be the output of a neural network with input x and weights W. The multi-task likelihood is defined as the multiplication of single-task probabilities. In this work, we have one classification task and three regression tasks.

$$p(y|f^W(x)) = p(y_1|f^W(x))...p(y_4|f^W(x)), \tag{2}$$

where $y = (y_1, ..., y_4)$ denotes the distribution of the four tasks.

In maximum likelihood estimation, we compute the log likelihood of the model:

$$\log(p(y|f^W(x))) = \sum_{k=1}^{4} \log(p(y_k|f^W(x))) \tag{3}$$

Following [19], the classification likelihood is defined as Boltzmann distribution function:

$$p(y_{cla}|f^W(x)) = \texttt{Softmax}(\frac{1}{T^2}f^W(x)), \tag{4}$$

where the subscript cla denotes the classification task. T is temperature of the system.

The uncertainty loss of classification task $L_{cla}^*(W, T)$ is computed as:

$$L_{cla}^*(W, T) = \frac{1}{T^2}L_{cla}(W) + \log(T), \tag{5}$$

The cross entropy loss without coefficient T can be defined as $L_{cla}(W)$, and it represents the loss function of the classification task without the temperature factor T.

The likelihood of a regression task can be defined as a Gaussian with the mean given by the model output $N(f^W(x), \sigma^2)$, where σ is the model's observation noise parameter. The uncertainty loss of the regression task $L_{reg}^*(W, \sigma)$ is computed as :

$$L_{reg}^*(W, \sigma) \propto \frac{1}{2\sigma^2}L_{reg}(W) + \log(\sigma), \tag{6}$$

Where $L_{reg}(W)$ denotes the L_2 norm of the regression task, including two bounding box regression tasks, a landmark regression task and pose estimation task in this work.

Applying (2)-(6) on the losses of individual tasks, we compute the overall loss function L^* as:

$$L^* = \frac{1}{T_1^2}p_i^*L_{cls} + \frac{1}{2\sigma_1^2}p_i^*L_{box} + \frac{1}{2\sigma_2^2}p_i^*L_{pts} + \frac{1}{2\sigma_3^2}p_i^*L_{pose} + \log(T_1\sigma_1\sigma_2\sigma_3), \tag{7}$$

where T_1 is the temperature of the classification task, $\sigma_1, \sigma_2, \sigma_3$ are the observation noise of the bounding box detection, landmark regression and pose estimation tasks. These parameters are obtained through learning.

3.3 Online Feedback Sampling

Some methods improve the accuracy of small face by changing the distribution of face samples. In PyramidBox++ [22], Li *et al.* propose to resize the images into different scales evenly. These strategies are beneficial for multi-scale sample training, but it needs empirical parameters to prevent the network from over-fitting.

In this work, we propose an online feedback method for data augmentation. An illustration of the algorithm is shown in Fig. 2(b). In the method, the training data is adjusted dynamically according to the actual effect in each iteration, which is beneficial to keep a balance on face scales. In our method, we construct a set S which starts from all training samples for face detection. In each iteration, we calculate the loss L_{box}^s of face size less than 25×25. The stitching strategy is used if the ratio of L_{box}^s/L_{box} is less than value T. After each epoch, we test the current model on the training data and obtain all images with false positive and false negative boxes. Then we update S with these difficult images. In this way, the proposed method improves the balance of the data automatically.

4 Experiments

4.1 Implementation details

Training Dataset: The first dataset we use is the WIDER FACE dataset [40]. It consists of 32,203 images and 393,703 face bounding boxes. As the original dataset does not contain pose labels, we first annotate the pose for the dataset. Head pose annotation from 2D images is an extremely challenging and time-consuming task, especially for very small faces. In our annotation, we annotate in a semi-supervised way and skip faces smaller than 35×35 pixels. We first pre-label the data through FSAnet [41] and Hopenet [34], and then two trained experts adjust the pre-labelled annotations manually after careful examination and mutual agreement. In our annotation, we mainly adjust the Yaw, which is most important for DFR. To facilitate the research in this area, we are preparing the data for release and the data will be available at https://github.com/lyp-deeplearning/MOS-Multi-Task-Face-Detect.

Evaluation Metrics: The AFLW [20] dataset is used to test the landmark localization accuracy. It contains 25993 faces with up to 21 landmarks per image. The evaluation is based on the mean errors measured by the distances between the estimated landmarks and the ground-truth, normalized with the face box size $\sqrt{w \times h}$.

We conduct testing of Head Pose estimate on AFLW2000 [52] dataset. It provides ground-truth 3D faces with pose angles for the first 2,000 images of the AFLW dataset. We follow the protocol of Hopenet [34].

Model architecture and training protocols: We implement MOS-S based on lightweight backbone ShuffleNet V2 [28]. MOS-S employs three feature pyramid levels from $P3$ to $P5$. The backbones are pretrained on the Imagenet. We use online hard example mining [37] and constrain the ratio of positive and negative anchors to 1:3. The SSH [30] module is added to increase the receptive field. The anchor aspect ratio is set to 1:1. In MOS-S, we train the model with 640×640 input images, the feature pyramid is set to $P3(80 \times 80)$, $P4(40 \times 40)$, $P5(20 \times 20)$, which leads to 16800 anchors. Although we mainly focus on the lightweight backbones, we also implement MOS-L with heavy backbone ResNet152 [12] to explore the performance of the proposed modules in this situation.

We train the face detection networks with a batch size of 64 using 1 NVIDIA Tesla V100 GPUs. We use SGD optimizer with momentum at 0.9, weight decay at 0.0005. The initial learning rate is set to 0.001 and then divided by 10 at 150 and 180 epochs. The training terminates at 200 epochs. For inference, we use multi-scale testing strategy as in [46].

4.2 Ablation Study

To justify the effectiveness of each component in MOS-S, we conduct the following ablation studies. The baseline approach is based on ShuffleNet V2 and SSH without MTH, as shown in Fig. 1 and Fig. 2(a). Then we add the four modules including MTH, Pose loss, UML, Online Feedback and the commonly used deformable convolution network (DCN) [51] one by one. The results are shown in Table 1.

Multi-task head. To justify the effectiveness of MTH, we first add MTH into the baseline, denoted as +MTH. As shown in Table 1, it obtains 0.79%, 0.86%, and 1.44% improvement on the *easy*, *medium* and *hard* subsets respectively for lightweight backbones.

Uncertainty Multi-task Loss. The baseline uses heuristic weights ($L_{cls} : L_{box} : L_{pts} : L_{pose} = 2 : 1 : 1 : 0.25$) for multi-tasking. Table. 1 compares the results of face detection and pose estimation tasks using UML. The UML loss further improves the performance by

Table 1: Ablation study of the proposed modules on the WIDER FACE validation set and AFLW2000 [52]. The baseline adopts the the hard connect structure in Fig. 2(a) with MOS-S.

Method	Easy	Medium	Hard	Avg. MAE
Baseline	89.32	88.45	80.20	6.73
+ MTH	90.11	89.31	81.64	6.49
+ UML	91.18	90.26	83.01	6.14
+ Online Feedback	92.41	91.37	85.24	6.09
+ DCN	92.91	91.61	85.72	5.97

Table 2: Comparison of AP with other methods for light networks in the WIDER FACE validation and test set. * indicates the work without peer review. Blue indicates the best result with lightweight backbone.

Method	backbone	easy(val‖test)	medium(val‖test)	hard(val‖test)	Multi-task
MTCNN[42]	Customized	0.851 ‖ −	0.820 ‖ −	0.607 ‖ −	✓
FACEBOXES[45]	Customized	0.879 ‖ 0.881	0.857 ‖ 0.853	0.771 ‖ 0.774	✗
lffd v1*[14]	Customized	0.910 ‖ 0.896	0.881 ‖ 0.865	0.780 ‖ 0.770	✗
ASFD-D0*[13]	Customized	0.901 ‖ −	0.875 ‖ −	0.744 ‖ −	✗
RetinaFace[9]	MobileNet-0.25	0.914 ‖ −	0.901 ‖ −	0.782 ‖ −	✓
img2pose*[1]	ResNet-18	0.908 ‖ 0.900	0.899 ‖ 0.891	0.847 ‖ 0.839	✓
MOS-S	ShuffleNet V2	0.929 ‖ 0.922	0.916 ‖ 0.911	0.859 ‖ 0.857	✓
MOS-L	ResNet-152	0.969 ‖ 0.955	0.961 ‖ 0.952	0.921 ‖ 0.913	✓

1.07%, 0.95% and 1.37%. Simultaneously, this method also improves the result of head pose estimation from 6.49 to 6.14.

Online Feedback data-augmentation. In the online feedback, we add the samples with small losses and false positive. The results are denoted as +Online Feedback. Table 1 show that it obtains an improvement of 1.23%, 1.11% and 2.23% for *easy*, *medium* and *hard* subsets respectively. This shows that the proposed feed-back strategy improves the balance of the data and therefore the accuracy of face detection. In our experiments, we have observed that arbitrary increase of small faces would bring the training into a bottleneck as many false detection come from medium or large faces.

4.3 Benchmark Results

Face Detection Accuracy. We train the model on the training set and test on the WIDER FACE validation and test sets. We follow the standard practices of [30, 46] and employ flip as well as multi-scale strategies. The standard Average Precision (AP) is computed. Table 2 shows the comparison in the validation set. Taking the lightweight ShuffleNet V2 as backbone, MOS-S achieves AP of 92.9%, 91.6% and 85.9% on the three subsets respectively, surpassing other methods. Taking the lightweight Mobilenet V2 [35] as backbone, MOS-Mobilenetv2 achieves AP of 94.19%, 93.25% and 88.34% on the three subsets respectively which confirms that our method is also effective on other lightweight backbones. MOS-S has achieved trade-off between speed and accuracy.

Using the ResNet152 as backbone, MOS-L achieves AP of 96.9%, 96.1%, 92.1% on the three subsets respectively.

(a) Accuracy of landmark detection

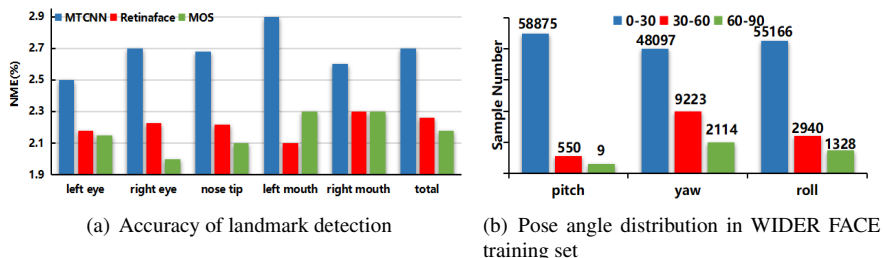(b) Pose angle distribution in WIDER FACE training set

Figure 3: (a) The comparison of MOS-S with other multi-task face detector on AFLW dataset. (b) The pose angle distribution in WIDER FACE training set. We can observe that majority of face angles are concentrated below 30 degrees.

Landmark Localization Accuracy. To evaluate the accuracy of the landmark localization, we compare MOS-S with the commonly used MTCNN [44] and RetinaFace [9] with landmark output. We train MOS-S on the WIDER FACE and test on the AFLW [21] dataset. Fig. 3(a) compares the proposed MOS-S with RetinaFace and MTCNN. As we can see, MOS-S performs the best.

Table 3: Evaluation on AFLW2000.

Method	Yaw	Pitch	Roll	MAE
Dlib(68 points) [13]	23.1	13.6	10.5	15.8
Fan(12 points) [3]	6.36	12.3	8.71	9.12
3DDFA [52]	5.40	8.53	8.25	7.39
Hopenet [34] ($\alpha = 2$)	6.47	6.56	5.44	6.16
FsaNet [41]	4.50	6.08	4.64	5.07
QuatNet [14]	3.97	5.62	3.92	4.50
MOS-S-widerface	4.52	6.91	6.48	5.97
MOS-L-widerface	4.05	6.29	5.87	5.40
MOS-S-300wlp	3.91	5.42	3.98	4.43

Table 4: Avearge inference time in WIDER FACE validation set.

Method	Time (ms)
RetinaFace-M[9]+ PnP [10](GPU)	7.2
RetinaFace-M[9] + Hopenet [34](GPU)	47.9
RetinaFace-M[9] + FSAnet [41](GPU)	41.1
MOS-S(GPU)	10.8
MOS-S(ARM)	68
MOS-L(GPU)	62.2

Head Pose Estimation Accuracy. We evaluate the pose estimation using the AFLW2000 [52]. Of the six methods [3, 16, 18, 34, 41, 52] compared, MOS-S is the only method to predict face location and head pose simultaneously. From Table 3, we observe that MOS-S-widerface performs good in Yaw, but worse in Pitch and Roll. The lower performance of in Pitch and Roll estimation is likely because of few samples with large pitch and roll angles in the training data. Fig. 3(b) plots the distribution of the angles. As we can see, the WIDER FACE dataset has few faces with large pitch and roll angle.

Note that the comparison here is not a completely fair one as MOS is trained using WIDER FACE while others are trained using 300W-LP [52] which include various distribution angle data. For fair comparison, we add another experiment by training the model using only the 300W-LP dataset. MOS-S-300wlp in Table 3 achieves state-of-the-art accuracy and minimal inference time.

Some examples have been given to compare MOS-S with RetinaFace with different pose estimation methods in Fig. 4. As we can see, PnP [10] performs poorly while Hopenet [34] requires more computational time. Our method is able to provide accurate results efficiently.

Comparison with Hyperface. For a fair comparison, we use AFLW as the training set, same as that in Hyperface [32]. We compared MOS-S and Hyperface-alexnet in face detection and pose estimation. For face detection, MOS-S achieves mAP of 93.2% on the FDDB dataset[17], which is better than 90.1% by Hyperface-alexnet. For pose estimation task, MOS-S achieves 4.89 MAE better than 5.88 by Hyperface-alexnet in the AFLW dataset.

(a) RetinaFace+PnP [10]   (b) RetinaFace+Hopenet [54]   (c) MOS-S

Figure 4: Visual comparison of results between MOS-S and RetinaFace (MobileNet-0.25 backbone) with different pose methods. We only show the yaw angle which is most important in DFR, and the yellow circle indicates cases with large errors. It takes 7.6ms, 49.0ms, and 12.1ms to process the images in the first row by the three different methods respectively.

4.4 Inference Efficiency

In order to better compare the results in dynamic face detection, MOS-S is compared with RetinaFace-M(MobilenetV1[15]) combined with Hopenet or FSAnet. The average time for images (resized to 640×480) from the WIDER FACE validation set are summarized in Table 4. We measure inference time of MOS using RTX 2060 GPU and ARM platform(RK3399). With lightweight backbones, MOS-S outperforms state of the art in face detection and pose estimation with less inference time. With heavy backbones, MOS-L takes about 62.2 ms. Besides the above, we have also deployed MOS-S using ncnn with multi-threads on ARM platform and it achieves 15 FPS for a mobile device.

5 Conclusion

Real-time face detection, landmark localization and head pose estimation with low computational resource are challenging tasks. In this work, we propose a novel low latency and lightweight backbone to learn the three tasks simultaneously, which facilitates the computation for DFR in mobile devices such as robots. Uncertain multi-task loss has been proposed to regularize the learning. Moreover, we propose an online feedback sampling to augment the data according to the performance of the trained models in the training iterations. The experimental results show that the proposed method achieves the state-of-the-art results compared with other methods with similar computation resources. Our codes and annotation will be made publicly available to facilitate further research in the area.

References

[1] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. *arXiv preprint arXiv:2012.07791*, 2020.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[4] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Selective refinement network for high performance face detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8231–8238, 2019.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[6] Donggen Dai, Wangkit Wong, and Zhuojun Chen. Rankpose: Learning generalised feature with rank supervision for head pose estimation. *arXiv preprint arXiv:2005.10984*, 2020.

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[8] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[9] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020.

[10] Martin A. Fischler and Robert Coy Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM.*, 24(6):381–395, 1981.

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2888–2897, 2019.

[14] Yonghao He, Dezhong Xu, Lifang Wu, Meng Jian, Shiming Xiang, and Chunhong Pan. Lffd: A light and fast face detector for edge devices. *arXiv preprint arXiv:1904.10633*, 2019.

[15] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[16] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2018.

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

[18] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[19] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[20] 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 *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.

[21] Amit Kumar, Azadeh Alavi, and Rama Chellappa. Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 258–265. IEEE, 2017.

[22] Zhihang Li, Xu Tang, Junyu Han, Jingtuo Liu, and Ran He. Pyramidbox++: High performance detector for finding tiny face. *arXiv preprint arXiv:1904.00386*, 2019.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] S. Liu, E. Johns, and A. J. Davison. End-to-end multi-task learning with attention. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880, 2019. doi: 10.1109/CVPR.2019.00197.

[25] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[26] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[27] Yang Liu, Xu Tang, Junyu Han, Jingtuo Liu, Dinger Rui, and Xiang Wu. Hambox: Delving into mining high-quality anchors on face detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13043–13051. IEEE, 2020.

[28] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[29] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

[30] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. Ssh: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4875–4884, 2017.

[31] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018.

[32] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (1):121–135, 2017.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[34] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018.

[35] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[37] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[38] Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 797–813, 2018.

[39] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[40] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. pages 5525–5533, 2016.

[41] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[42] Bin Zhang, Jian Li, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Yili Xia, Wenjiang Pei, and Rongrong Ji. Asfd: Automatic and scalable face detector. *arXiv preprint arXiv:2003.11228*, 2020.

[43] Feifei Zhang, Tianzhu Zhang, Qirong Mao, and Changsheng Xu. Joint pose and expression modeling for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3359–3368, 2018.

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

[45] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017.

[46] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017.

[47] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4203–4212, 2018.

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

[49] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020.

[50] Jiashu Zhu, Dong Li, Tiantian Han, Lu Tian, and Yi Shan. Progressface: Scale-aware progressive learning for face detection.

[51] X. Zhu, H. Hu, S. Lin, and J. Dai. Deformable convnets v2: More deformable, better results. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9300–9308, 2019.

[52] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.

[53] Chubin Zhuang, Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Jinqiao Wang, and Stan Z Li. Fldet: A cpu real-time joint face and landmark detector. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.