EResFD: Rediscovery of the Effectiveness of Standard Convolution for Lightweight Face Detection
(Supplementary Material)

1. Implementaion details

1.1. Anchor Box Settings

As shown in Table 1, anchor boxes whose sizes are ranged from 16 to 512 are assigned from detection layers from D1 to D6, respectively, following [2]. The ratio of width to height for each anchor box is set to 1:1.25.

Table 1. Anchor box settings: the stride size within the network, anchor box size assigned to 6 detection layers in Figure 2 of the manuscript.

Detection Layer	Stride	Anchor Size
D1	4	16
D2	8	32
D3	16	64
D4	32	128
D5	64	256
D6	128	512

1.2. Training

During training[1], we adopt data pre-processing steps, augmentation strategies, and loss functions used in [2]. Specifically, we used color distortion by changing the hue, saturation, and value (brightness) of an image. Also, for generating various scales of faces, zoom-in and out operations are applied to the image along with its face bounding boxes. Consequently, the resultant image is resized to 640 × 640 after horizontal flipping. Also, max-out background label [2] is applied on D1 detection head for reducing false positives with regard to small faces. We also employed multi-task loss, where both classification and regression loss are normalized by the number of positive anchors. We set the balancing parameter between classification and regression loss as 1:1.

Table 2. Latency and detection accuracy according to the stem design. We only changed the stem layer of EResFD with that of ResNet and our proposed stem design of EResNet.

Stem	ResNet	EResNet
Stem FLOPs	180.6 M	11.5 M
Stem Latency (Ratio)	24.1ms (42%)	4.5ms (12%)
Overall Latency	56.8ms	37.7ms
Easy mAP (%)	87.5	89.0
Medium mAP (%)	86.3	87.9
Hard mAP (%)	77.6	80.4

For optimization hyper-parameters, we used ADAM [1] optimizer with initial learning rate 0.001, weight decay 5e-4, and batch size 32. The maximum number of iterations is 330k and the learning rate is decayed at [250k, 300k, 320k] with the decaying factor of 0.1.

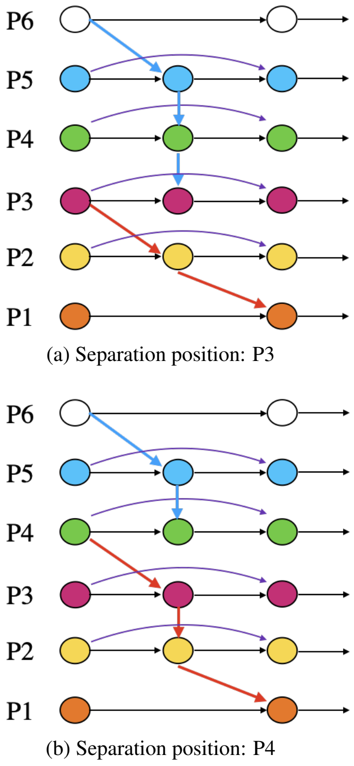
2. Stem Modification

In Table 2, we compare detection performance depending on the design of the stem layer. It is worth notable that our proposed stem layer requires much more reduced computational costs with significantly improved detection accuracy scores compared to that of ResNet.

3. SepFPN

In Figure 1, we visualize the architecture of variations for our proposed SepFPN, where its separation position is varied from P3 to P4.

[1] We used source code from https://github.com/yxlijun/S3FD.pytorch



(a) Separation position: P3

(b) Separation position: P4

Figure 1. Architectures of SepFPN with various separation positions (P3, P4).

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[2] 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.