Supplementary Material

1. Description

Due to the page limited, we will clarify the quantitative and qualitative analysis on the scale-level data augmentation strategies, and more ablative experiments and architectures on context-enhanced modules of HCAM in the supplementary material.

2. Analysis on Scale-level Data Augmentation

In this section, we firstly implement Random Square Crop (RSC), Data Anchor Sampling (DAS) and Multi Scale Training (MST) on the proposed baseline and find the former two strategies have better performance than the last one. Secondly, based on this experiment result, we further analyze the reason of why does the MST strategy perform poorly on resolving extreme scale variance challenge? Thirdly, in order to utilize the scale information of the training data, we analyze the relationship between the scale information and the detector performance.

Implementation Detail of Scale-level Data Augmentation strategies.

- Multi-scale-training: For each image in the training data, we resize it by reshaping the short side of image into a scale selected from predefined scale range [640, 1280] randomly.
- Data-anchor-sampling: 1) Randomly select a face and compute its scale fs 2) Choose the nearest scale with this face scale from the set $\{16, 32, 64, 128, 256, 512\}$ 3) Uniformly random select the scale sr from the set $\{16, 32, ..., nearest\_scale\}$ and compute the target ratio tr by sr / fs . 4) Resize the image with this target ratio tr. 5) If the resolution of resized image is over 640 $\times$ 640, we crop 640 $\times$ 640 area randomly as the input image and pad zero pixel if it is less than 640 $\times$ 640.
- Random Square Crop: Random crop a square area from the image with the scale that equals to multiplying the short side scale and a factor selected from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ randomly.

2.1. Difference among RSP, DAS and MST.

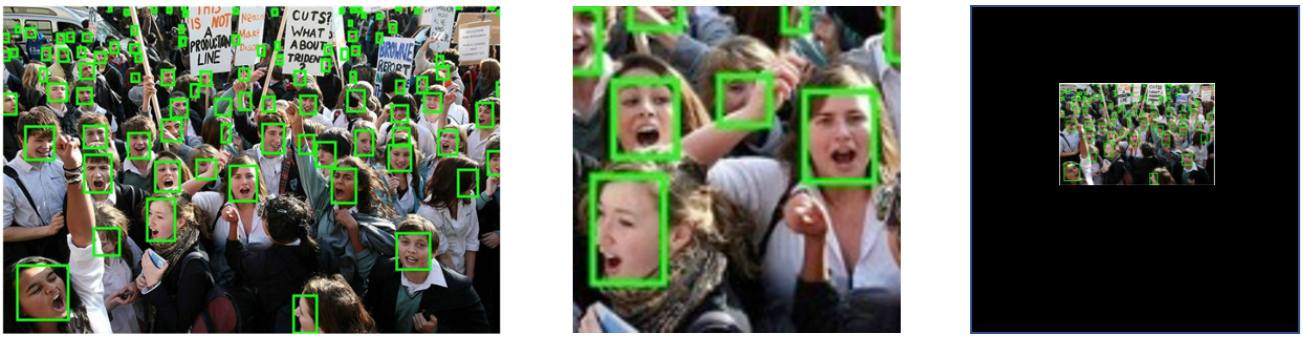
As shown in Table 1, we display the performance of RSP, DAS and MST on our baseline detector, where RSP and DAS achieve almost consistent performance while the MST only achieves 83.60% AP on the Wider Face validation hard subset. Considering the tremendous performance gap , we summary the two differences among them: (1) MST brings more abundant scale information than RSP and DAS. (2) As shown in 2. 1, RSP and DAS both reduce the native image information on the each image of training data by cropping the local square patch from the original image patch and bringing a large amount of padding area separately. These two differences can be further interpreted as that comparing with RSP and DAS, MST introduces more scale information and native image information.

Method	Easy	Medium	Hard
Baseline	92.2	90.5	81.4
Baseline + MST	93.3	91.5	83.6
Baseline + DAS	94.6	93.4	86.5
Baseline + RSC	94.8	94.1	86.4
Baseline + MST + RSC	94.8	94.2	86.5

Table 1: Results of scale-level data augmentation strategies on the Wider Face validation subsets.

Then, we conduct a experiment to explore whether scale information or native image information [1] causes the significant performance gap. As shown in Table 1, we combine MST with RSC to help the detector embrace more scale information and few native image information. Comparing with the detector only with MST, the performance increases 4.3% AP on the Wider Face validation hard subset, that demonstrates the less native image information can provide appropriate knowledge for the detector. Simultaneously, this experiment result is almost consistent with the detector adopting RSC, which explicitly demonstrates the scale information provided by the MST strategy is hard for the detector to absorb. It can be concluded from the experiment results that simplex (less) native image information is conducive to the face detector facing extreme scale variance and the scale information can not be assimilated by the detector effectively.

[1] Native Image Information refers to the fore-ground and back-ground information of an image



(a) (b) (c)

Figure 1: An image is augmented by multi-scale training and data-anchor-sampling respectively. (a) An original image. (b) The image in (a) is augmented by the random crop strategy. The native image information is less than the original image. Note that this image may blur the image when the expand ratio is large. (c) The image in (a) is augmented by the data-anchor-sampling strategy. This brings a large amount of padding area, that reduces the learning difficulty of the detector on negative anchors. Thus, the fore-ground information in native image information is reduced remarkably.

2.2. Analyze the Relationship between the Scale Information and the Detector Performance.

In our perspectives, to help the detector absorb the scale information effectively, we need answer the following two questions firstly: (1) What is the relationship between the performance of each pyramid layer and the number of ground-truths it matches? (2) Can the larger shrink or expand ratio of the image provide reliable scale information?

In order to assist investigating the first question, we propose a scale control strategy based on the dichotomy that can control the ratio r_i of the ground-truth matched in the target pyramid layer p_i. (1) Select a middle scale s_i from the interval $[start\_s, end\_s]$. (2) For each image in the training data, random sample a ground-truth from it and get a shrink ratio sr by s_i / its scale. (3) Resize the image with shrink ratio sr. (4) Compute the ratio r_c of the ground-truth matched in the pyramid layer p_i under current setting. (5) If $|r_c - r_i| < 0.05$, training the detector with scale-level data augmentation strategy like in the step (2) and (3), break; if $r_c > r_i$, $end\_s = r_c$, restart from step 1; if $r_c < r_i$, $start\_s = r_c$, restart from step 1. Based on this strategy, we further train the detector by controlling the ratio of ground-truths matched on a certain pyramid layer. For instance, when s_i equals to 21, the p_2 can match 80% ground-truth and the detector achieves 81.79 % AP on the Wider Face validation hard subset. Note that the hard subset contains a large amount of small faces, so it is appropriate for the evaluation of the p_2,p_3 learning capacity. Similarly, medium (easy) subset is appropriate for p_4, p_5, p_6 (p_5, p_6). As results reported in the table 2, we get a new appreciation that it is not accurate that the more ground-truths that is matched in a single pyramid layer, the greater performance of this pyramid layer.

	20%	40%	60%	80%
p2 (hard)	81.79	**82.17**	77.85	73.23
p3 (hard)	74.82	76.52	**77.15**	71.94
p4 (easy)	67.01	75.28	81.28	**87.31**
p4 (med)	75.28	82.60	**84.48**	83.62
p5 (easy)	81.79	**82.17**	77.85	73.23
p5 (med)	**85.44**	85.16	84.97	83.68
p6 (easy)	**86.17**	83.12	84.78	83.34
p6 (med)	**85.24**	80.47	82.17	81.22

Table 2: The results of scale control strategy on the Wider Face validation subsets.

To investigate the second question, we revise the step 4 in data anchor sampling as follows: if $tr > r\_th$, $tr = r\_th$. if $tr < 1/r\_th$, $tr = 1/r\_th$. Thus, r_th controls the maximum shrink ratio. We show the results on the detector with different r_th in table 2. The performance is almost consistent among different r_th. Thus, the large expand/shrink ratio of the image can also provide reliable scale information absolutely. Thus, in our SSE, we neglect to add any constraints on the maximum expand/shrink ratio of the image.

2.3. Ablative Experiments and Architectures on Context-Enhanced Modules

As described in the step 4 of Hierachical Context-Aware Module, we introduce the Context-Enhanced Module to explicit encode context information on the backbone feature map with 3 types, atrous spatial pyramid pooling (aspp), detection head module (SSH-DH)in SSH and single 3x3 convolution layer (SCL). As shown in the table 4, single 3x3 convolution layer can bring 1.1 % enhancement on

r_th	Easy	Medium	Hard
2	94.8	93.7	86.5
4	94.7	93.6	86.4
8	94.7	93.4	86.2
16	94.9	93.5	86.4
32	94.6	93.8	86.5
64	94.7	93.6	86.4

Table 3: The results of the detector with different r_th on the Wider Face validation subsets.

the Wider Face hard subset. Comparing with another two Context-Enhanced modules, 3x3 convolution layer achieves the best trade-off between accuracy and computation cost. Thus, our HCAM adopt single 3x3 convolution layer as Context-Enhanced Module.

Method	Easy	Medium	Hard
Baseline	94.6	93.4	86.5
Baseline + ASPP	**95.5**	**94.7**	**87.9**
Baseline + SSH-DH	95.2	94.4	87.8
Baseline + SCL	95.3	94.4	87.6

Table 4: Results of different Context-Awared Module on the Wider Face validation subsets.