ACFD: Asymmetric Cartoon Face Detector

Bin Zhang*, Jian Li*, Yabiao Wang, Zhipeng Cui
{* Equal Contribution}

Youtu Lab, Tencent

Southeast University

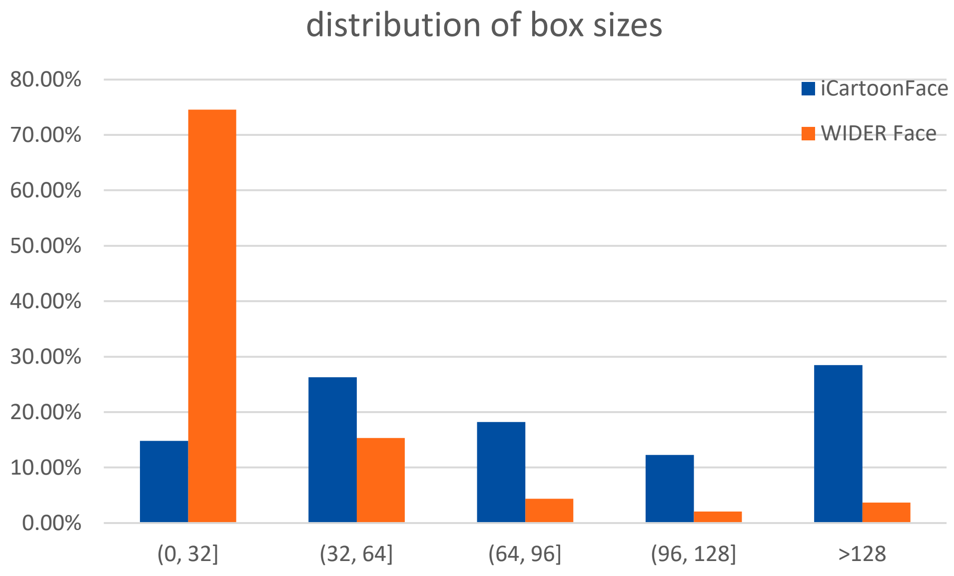
腾讯优图

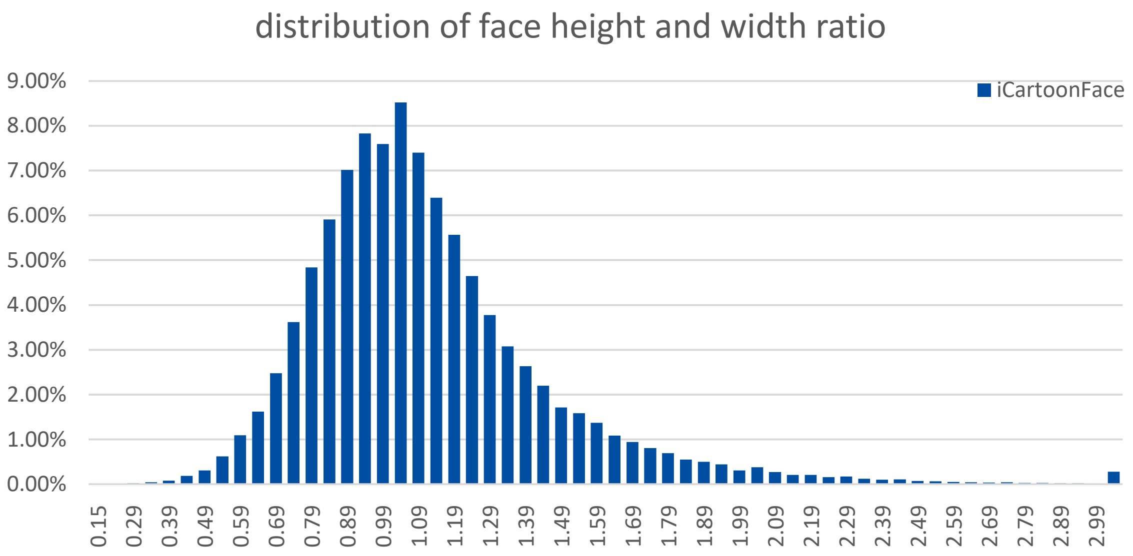
東南大學
SOUTHEAST UNIVERSITY

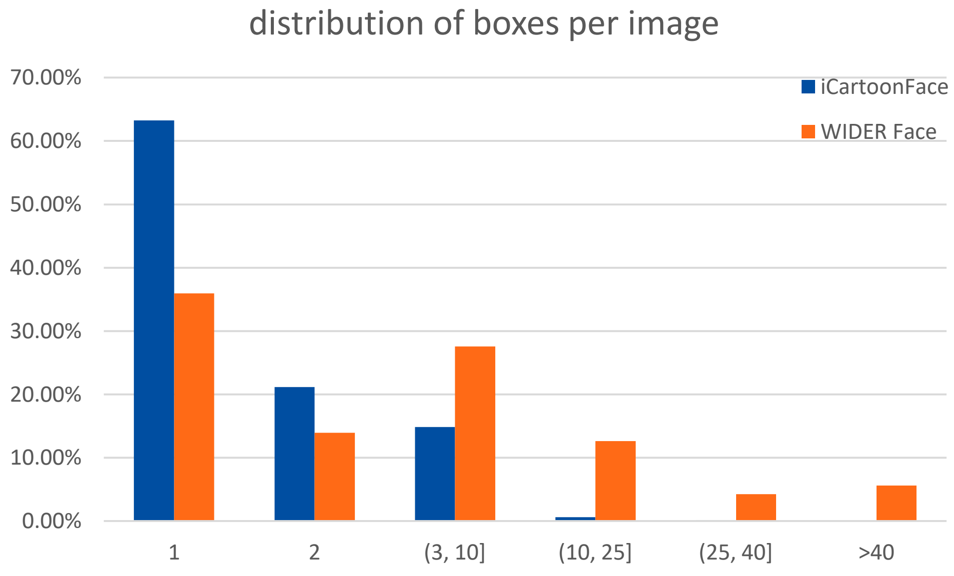
Task Analysis

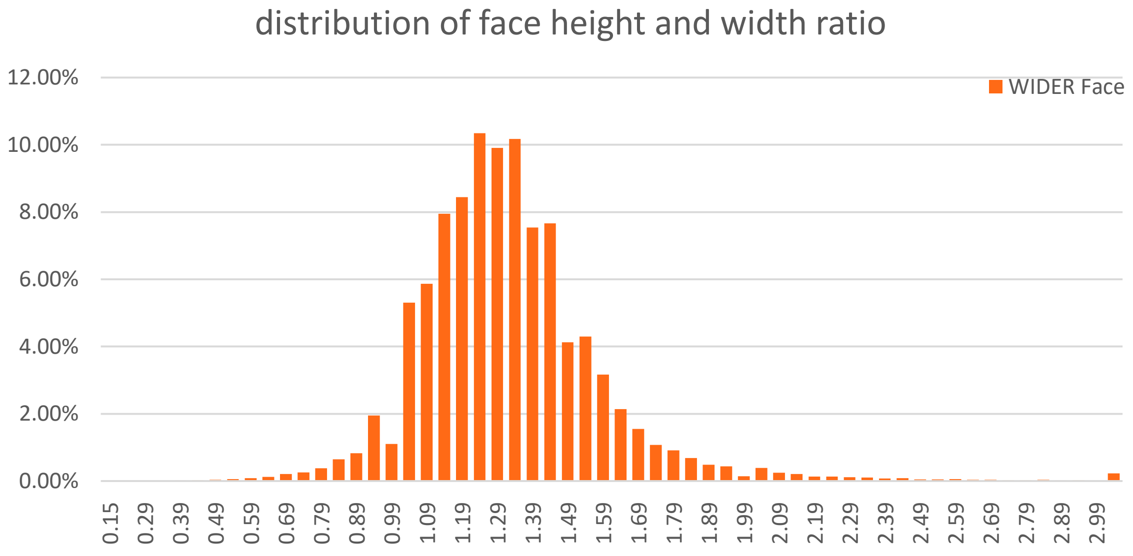
- Model size should not exceed 200M.
- Inference time of a single picture (1920x1080) should not exceed 50ms.
- Multi-scale and multi-model ensembles are allowed, in this way, inference time is the sum of multi-scale multi-models.
- Pretrained model can not be utilized.
- WIDER Face can be used as the training dataset.

Dataset Analysis









Difficulties of iCartoonFace

Related Face Detectors

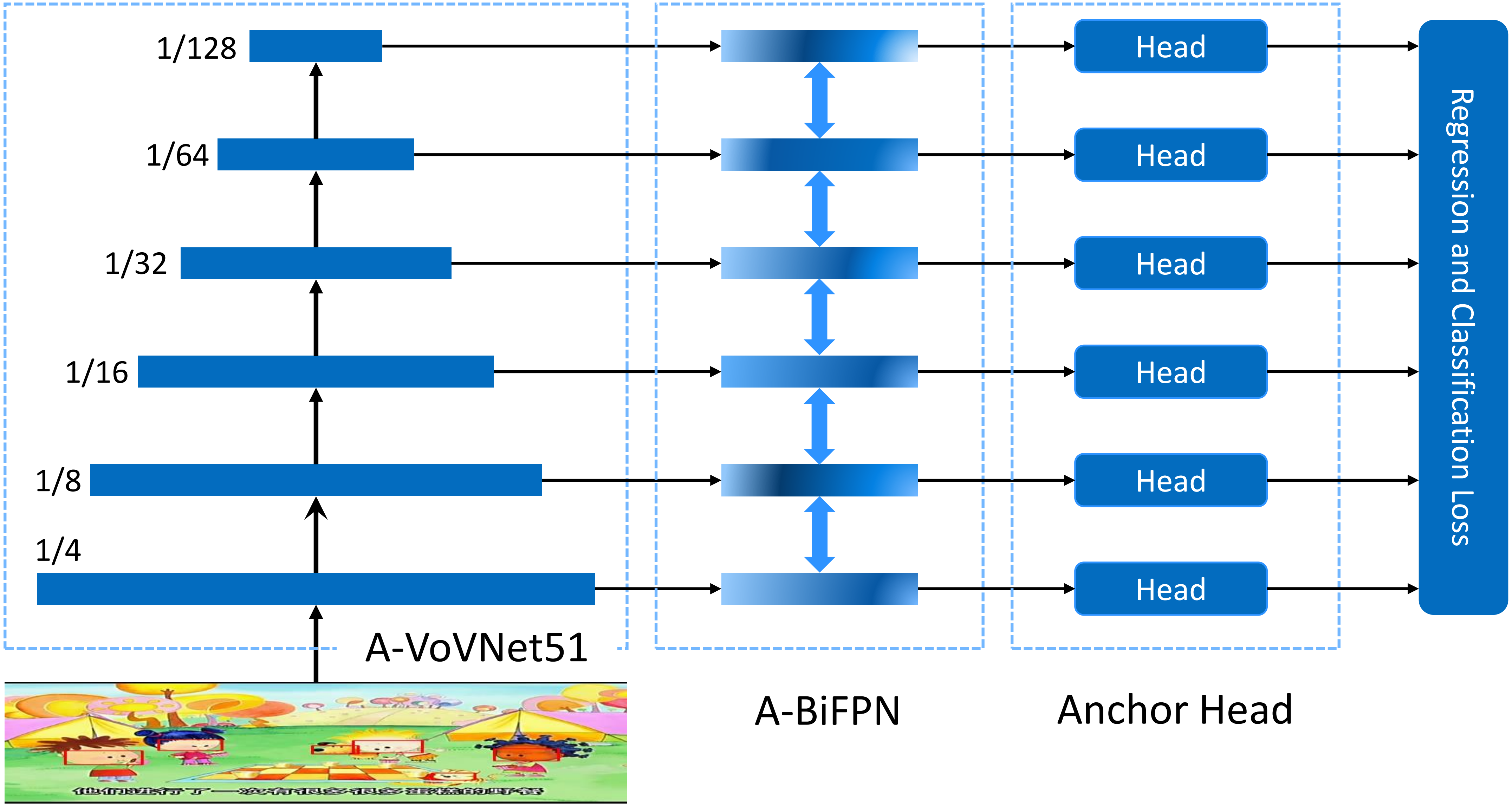
- State-of-the-art methods: SFD[1], PyramidBox[2], SRN[3], DSFD[4], RefineFace[5].
- **Pipeline:**
  - One-stage anchor-based detectors are widely used.
  - Take features of stride from 4 to 128 for predicting. (6 layers of pyramid features, for handling the small faces)
  - Modules follow the backbone to fuse the pyramid features and enhance the semantic information sequentially.
- **Match Strategy:** Match anchors and faces with a smaller IoU threshold (0.35-0.4, it is 0.5 in generic object detection). (to assign enough anchors for each faces)
- **Data Augmentation:** Random crop for multi-scale training.

[1] S. Zhang, X. Zhu, and et al. S3FD: Single Shot Scale-invariant Face Detector. ICCV, 2017.
[2] X. Tang, D. K. Du, and et al. PyramidBox: A Context-assisted Single Shot Face Detector. ECCV, 2018.
[3] C. Chi, S. Zhang, and et al. Selective Refinement Network for High Performance Face Detection, AAAI, 2019.
[4] J. Li, Y. Wang, and et al. DSFD: Dual-Shot Face Detector. CVPR, 2019.
[5] S. Zhang, C. Chi, and et al. RefineFace: Refinement Neural Network for High Performance Face Detection, TPAMI, 2020.

Our ACFD

- A one-stage anchor-based pipeline with the ability to extract diverse features by employing **asymmetric conv.** layers.
- Data augmentation for better handling the faces hard to detect, e.g., too small and too large faces, blur and occluded faces, etc.
- **Dynamic match strategy** to sample high-quality anchors for training, providing enough anchors for each face.
- **Margin loss** for enhancing the power of discrimination especially for those faces similar to the background, e.g., robot face.

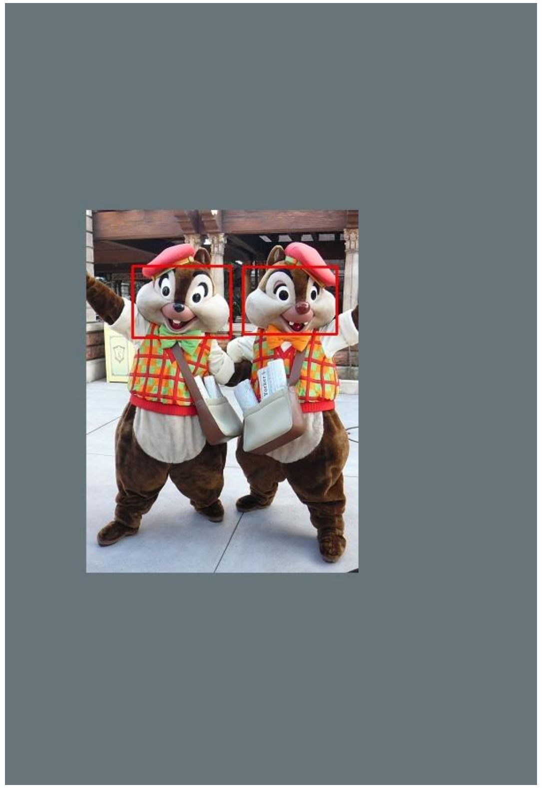
Pipeline


1/128
1/64
1/32
1/16
1/8
1/4
A-VoVNet51
A-BiFPN
Head
Head
Head
Head
Head
Head
Anchor Head
Regression and Classification Loss

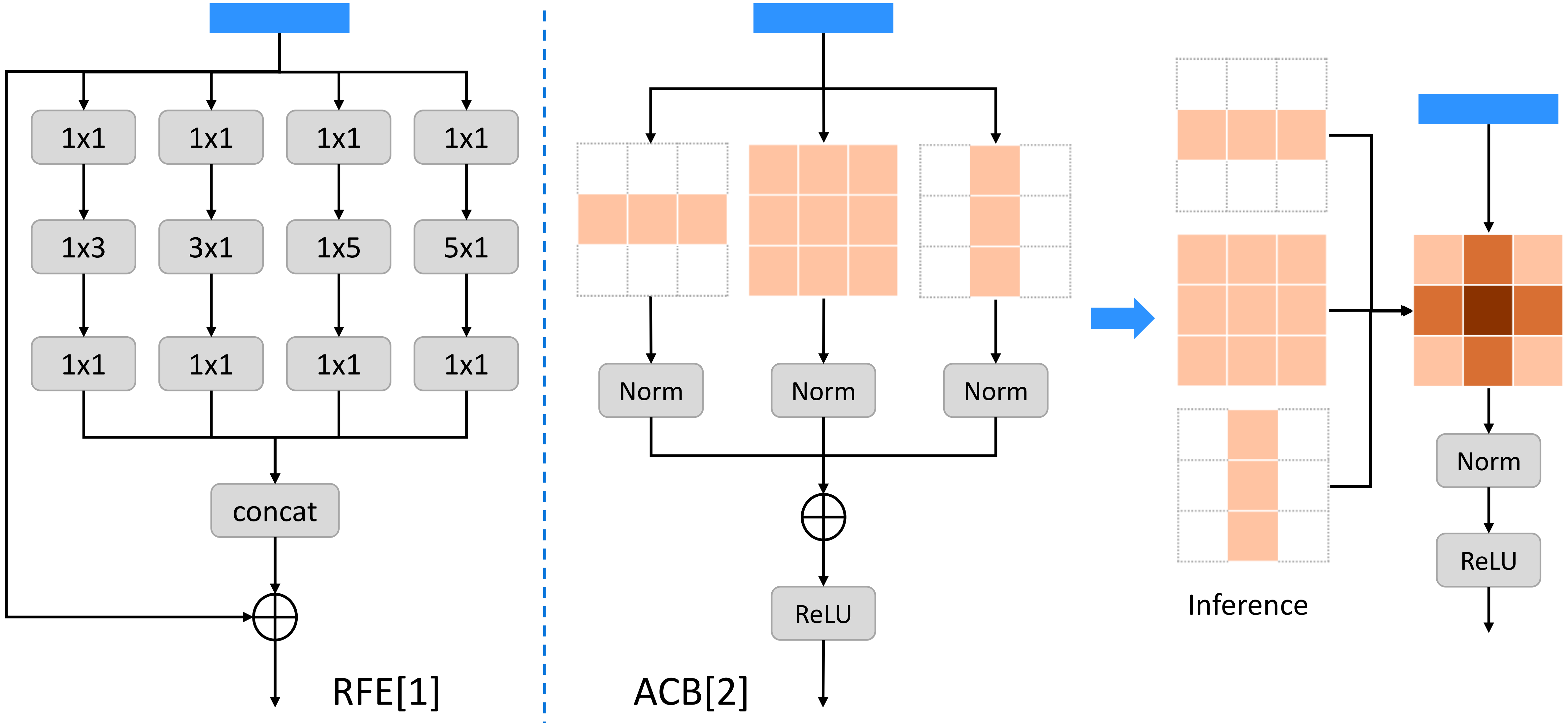
Data Augmentation

- Random crop to generate **large** training samples. (zoom in)
- Random expansion to generate **small** samples. (zoom out)
- Random tile faces to anchor scale for better align the receptive field.






Asymmetric Conv Block (ACB)

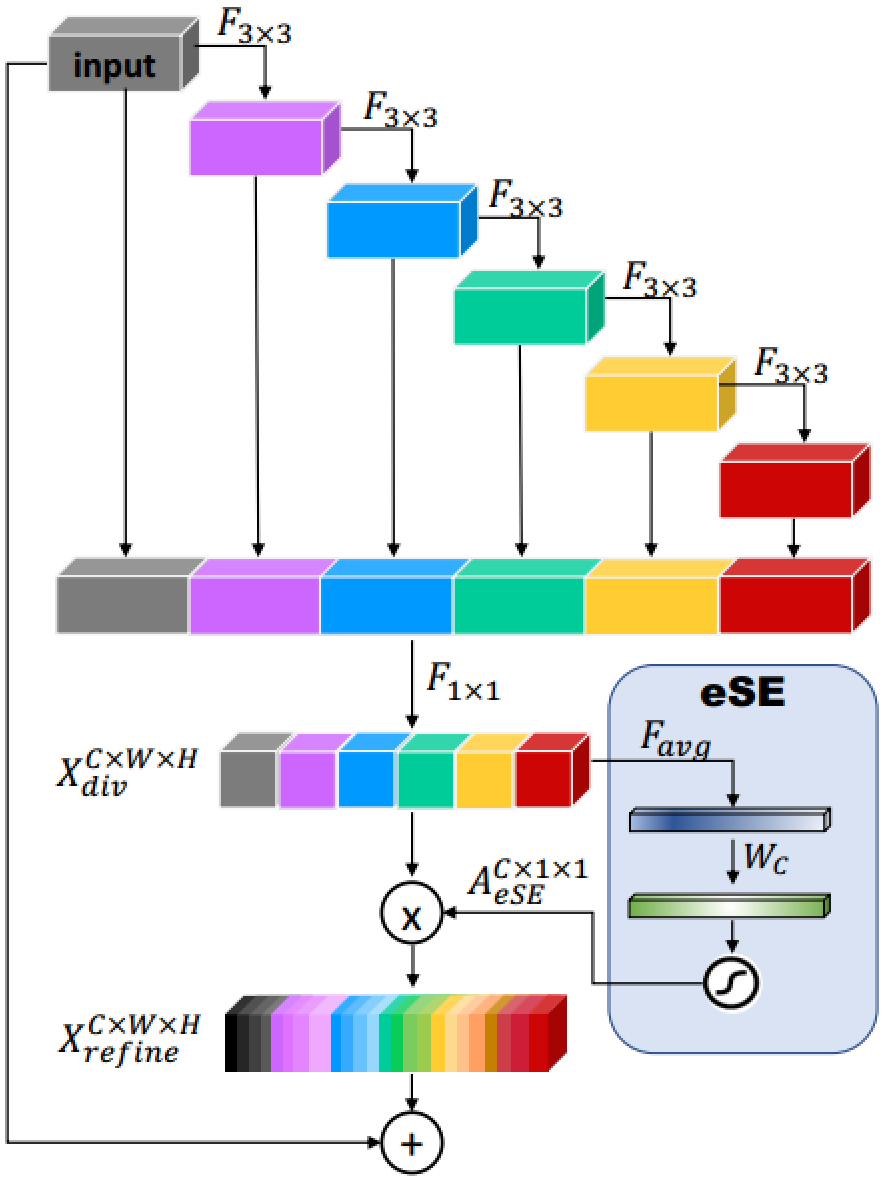


[1] S. Zhang, C. Chi, and et al. RefineFace: Refinement Neural Network for High Performance Face Detection, TPAMI, 2020.
[2] X. Ding, Y. Guo, and et al. Acnet: Strengthening the Kernel Skeletons for Powerful CNN via Asymmetric Convolution Blocks. ICCV, 2019.

A-VoVNet51[1]

- 6 stages with strides from 4 to 128

Layer	Output	Stride	Repeat	Channel
Image	640×640	-	-	-
Conv1	320×320	2	1	64
Conv2	320×320	1	1	64
Conv3	160×160	2	1	128
Stage1	160×160	1	1	256
Down-sampling	80×80	2	1	256
Stage2	80×80	1	1	512
Down-sampling	40×40	2	1	512
Stage3	40×40	1	2	768
Down-sampling	20×20	2	1	768
Stage4	20×20	1	2	1024
Down-sampling	10×10	2	1	1024
Stage5	10×10	1	1	128
Down-sampling	5×5	2	1	128
Stage6	5×5	1	1	128



[1] Y. Lee, J. Park, and et al. CenterMask: Real-Time Anchor-Free Instance Segmentation. CVPR, 2020.

A-VoVNet51[1]

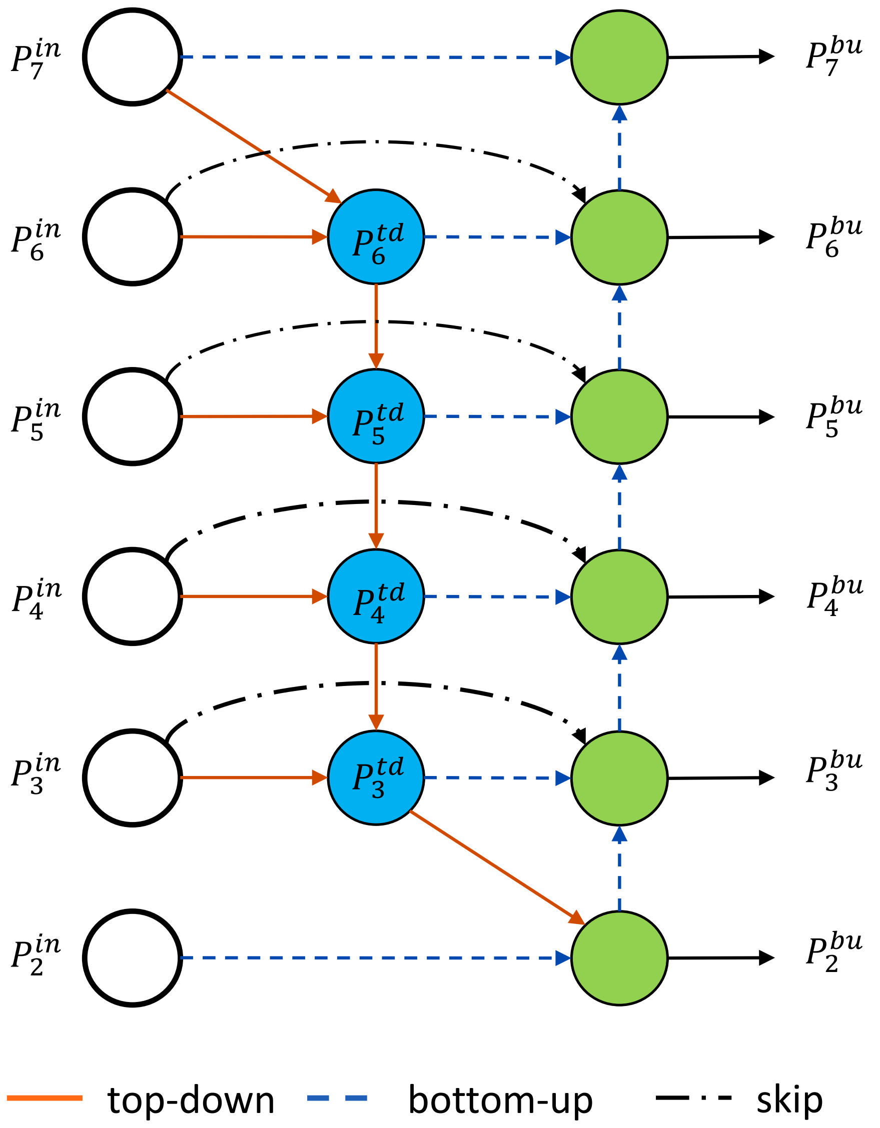
- 6 stages with strides from 4 to 128

Layer	Output	Stride	Repeat	Channel
Image	640×640	-	-	-
Conv1	320×320	2	1	64
Conv2	320×320	1	1	64
Conv3	160×160	2	1	128
Stage1	160×160	1	1	256
Down-sampling	80×80	2	1	256
Stage2	80×80	1	1	512
Down-sampling	40×40	2	1	512
Stage3	40×40	1	2	768
Down-sampling	20×20	2	1	768
Stage4	20×20	1	2	1024
Down-sampling	10×10	2	1	1024
Stage5	10×10	1	1	128
Down-sampling	5×5	2	1	128
Stage6	5×5	1	1	128

backbone	AP
ResNet50	0.9018
SE-ResNet50	0.9023
Res2Net50	0.8959
ResNeSt50	0.8997
EfficientNet-B3	0.8863
VoVNet51	0.9037
A-VoVNet51	**0.9074**

[1] Y. Lee, J. Park, and et al. CenterMask: Real-Time Anchor-Free Instance Segmentation. CVPR, 2020.

A-BiFPN[1]



- Top-down path:

$$P_i^{td} = Conv\left(\frac{\omega_1 \cdot P_i^{in} + \omega_2 \cdot Resize(P_{i+1}^{td})}{\omega_1 + \omega_2 + \epsilon}\right)$$

- Bottom-up path:

$$P_i^{bu} = Conv((\omega_1 \cdot P_i^{in} + \omega_2 \cdot P_i^{td} + \omega_3 \cdot Resize(P_{i-1}^{td}))/(\omega_1 + \omega_2 + \omega_3 + \epsilon))$$

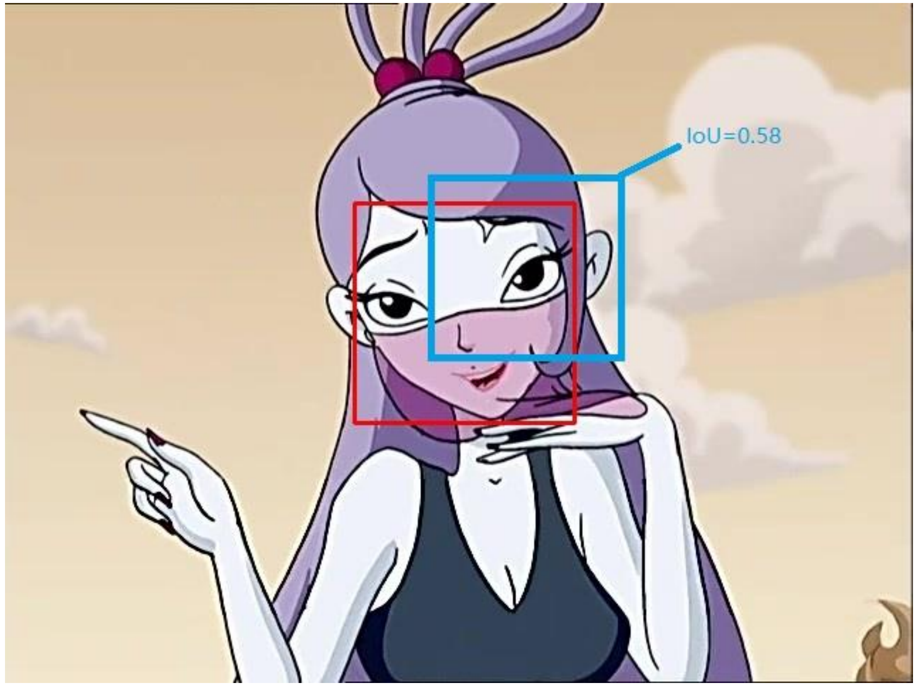
Exp:

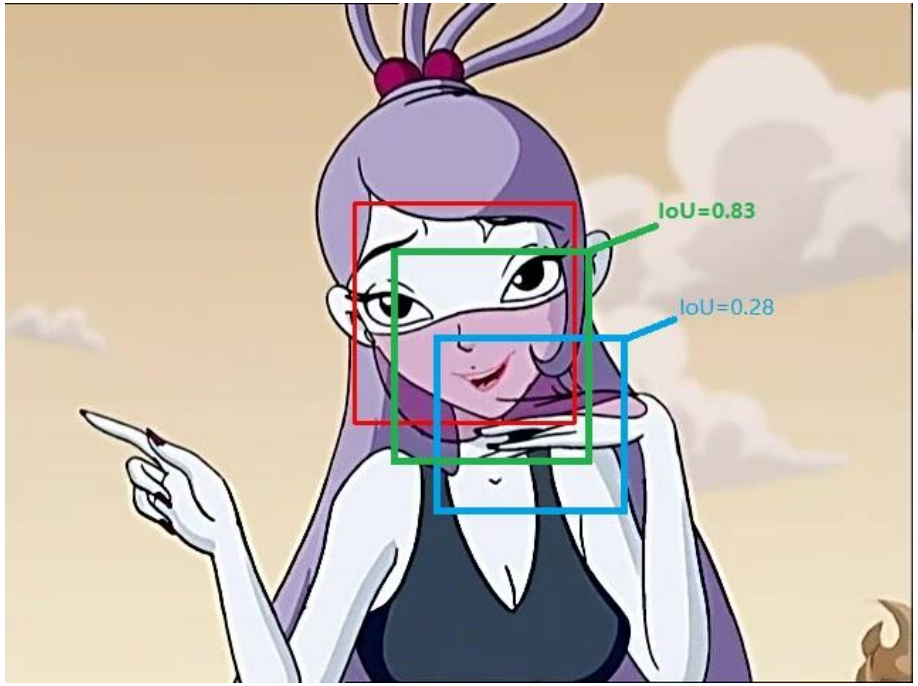
Feature Module	AP
A-BiFPN	**0.9036**
BiFPN	0.9018
SEPC (CVPR'20)	0.8880
FPN	0.8830

[1] M. Tan, R. Pang, and et al. EfficientDet: Scalable and Efficient Object Detection. CVPR, 2020.

Dynamic Match Strategy

- First step: the faces match anchors with IoU higher than a small threshold, usually 0.35~0.45.
- Second step: a anchor would be matched when IoU of its regressed box with any box greater than a large threshold, usually 0.7~0.8.





[1] T. Kong, F. Sun, and et al. Consistent optimization for single-shot object detection. arXiv, 2019.
[2] Y. Liu, X. Tang, and et al. HAMBox: Delving Into Mining High-Quality Anchors on Face Detection. CVPR, 2020.

Loss Design

- Regression loss

$$\ell_{reg} = \frac{1}{N_1} \sum_{i \in \psi_1} \mathcal{L}_{smoothL1}(x_i, x_i^*) + \frac{\lambda_{reg}}{N_2} \sum_{i \in \psi_2} \mathcal{L}_{smoothL1}(x_i, x_i^*)$$

- Classification loss

$$\ell_{cls} = \frac{1}{N_1} \sum_{i \notin \psi_2} \mathcal{L}_{focal}(f_{margin}(p_i), p_i^*) + \frac{\lambda_{cls}}{N_2} \sum_{i \in \psi_2} \mathcal{L}_{focal}(f_{margin}(p), p_i^*)$$

$$f_{margin}(x, x^o) = [x^o = 1] \cdot (x - m) + [x^o = 0] \cdot x$$

model	AP
Baseline	0.8765
+ dynamic match	0.8890

model	AP
Baseline	0.9048
+ margin loss	0.9073

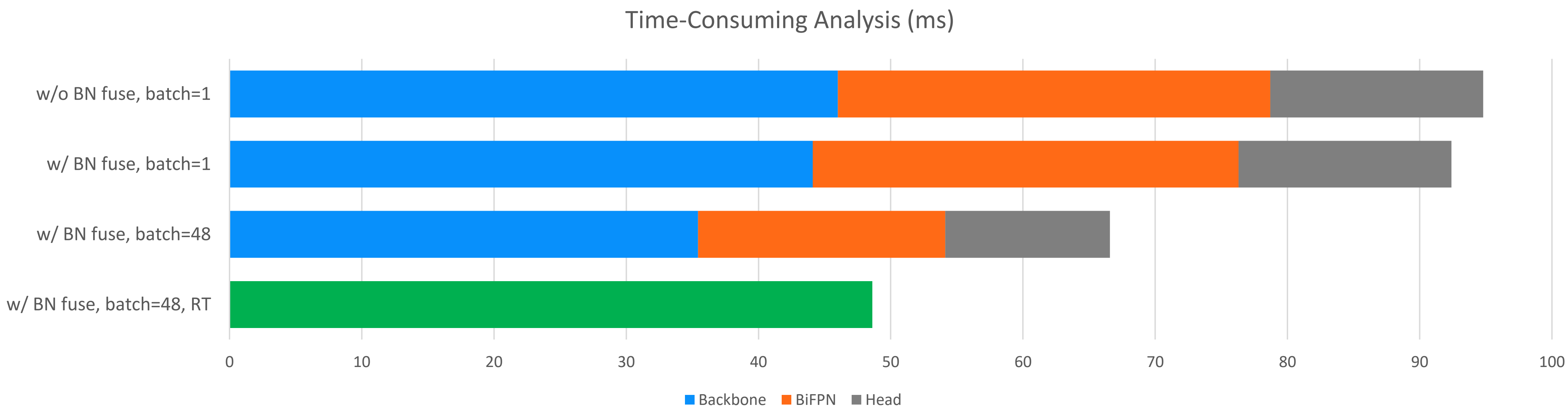
[1] H. Wang, Y. Wang. Cosface: Large Margin Cosine Loss for Deep Face Recognition. CVPR, 2018.

Experimental Details

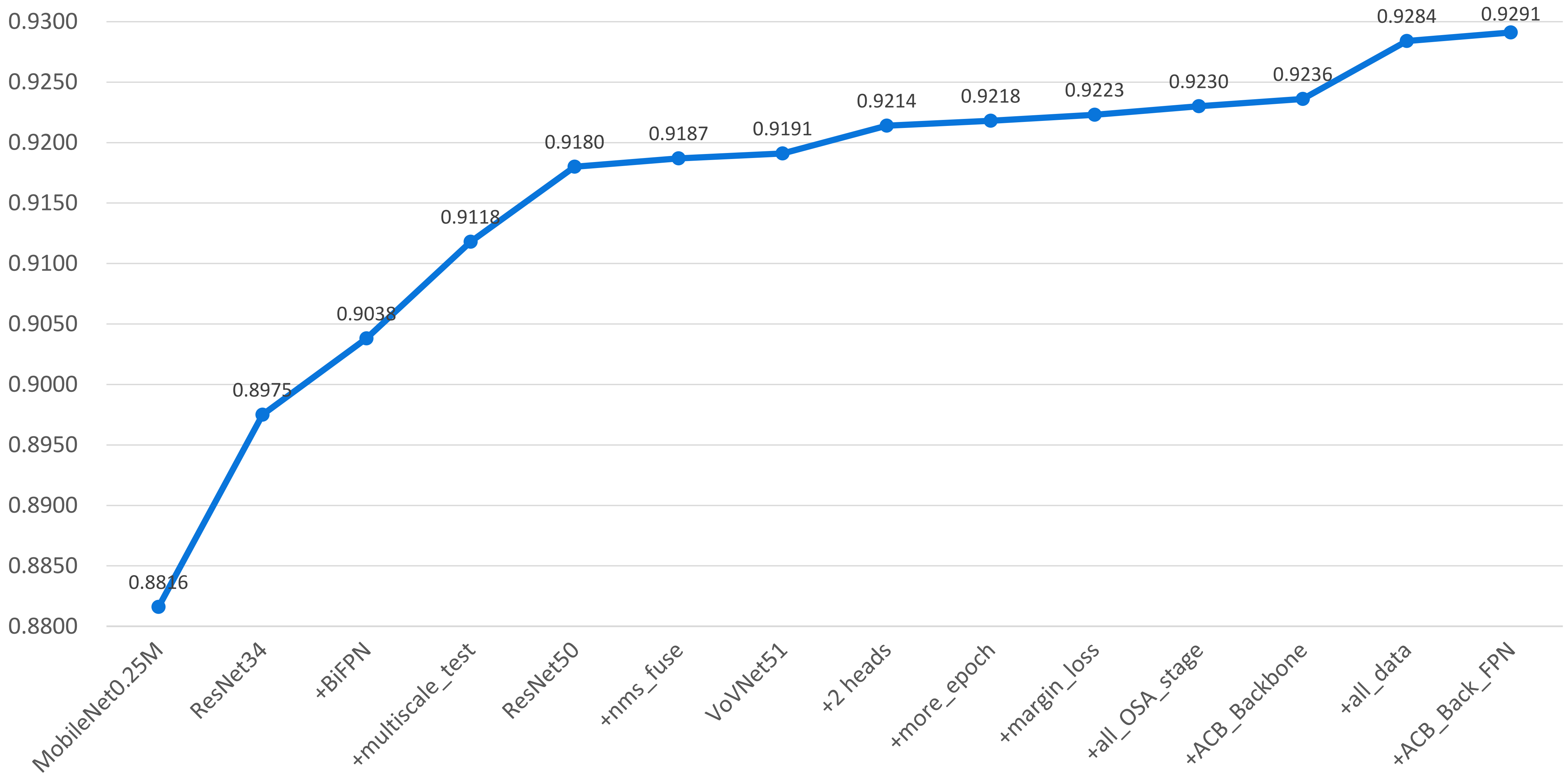
- Training
  - Split 50000 pictures into 45000 for training and 5000 for validating.
  - Sample size: 640×640, batch size: 16×4 Tesla V100 GPUs.
  - SGD with learning rate 0.04, multiplied by 0.1 at 200, 250 and 280 epoch, stop at 300 epoch.
  - IoU thresholds for first and second match: 0.35 and 0.7, $\lambda_{reg} = \lambda_{cl.s} = 0.8$.
  - The margin of classification loss is 0.2.
- Testing
  - Test scale: (480, 645), (640, 860), (800, 1075)
  - Top-1000 predictions with confidence scores higher than 0.08 are processed by NMS with IoU threshold 0.55, top-100 boxes would be preserved as the final results.

Time-Consuming Analysis

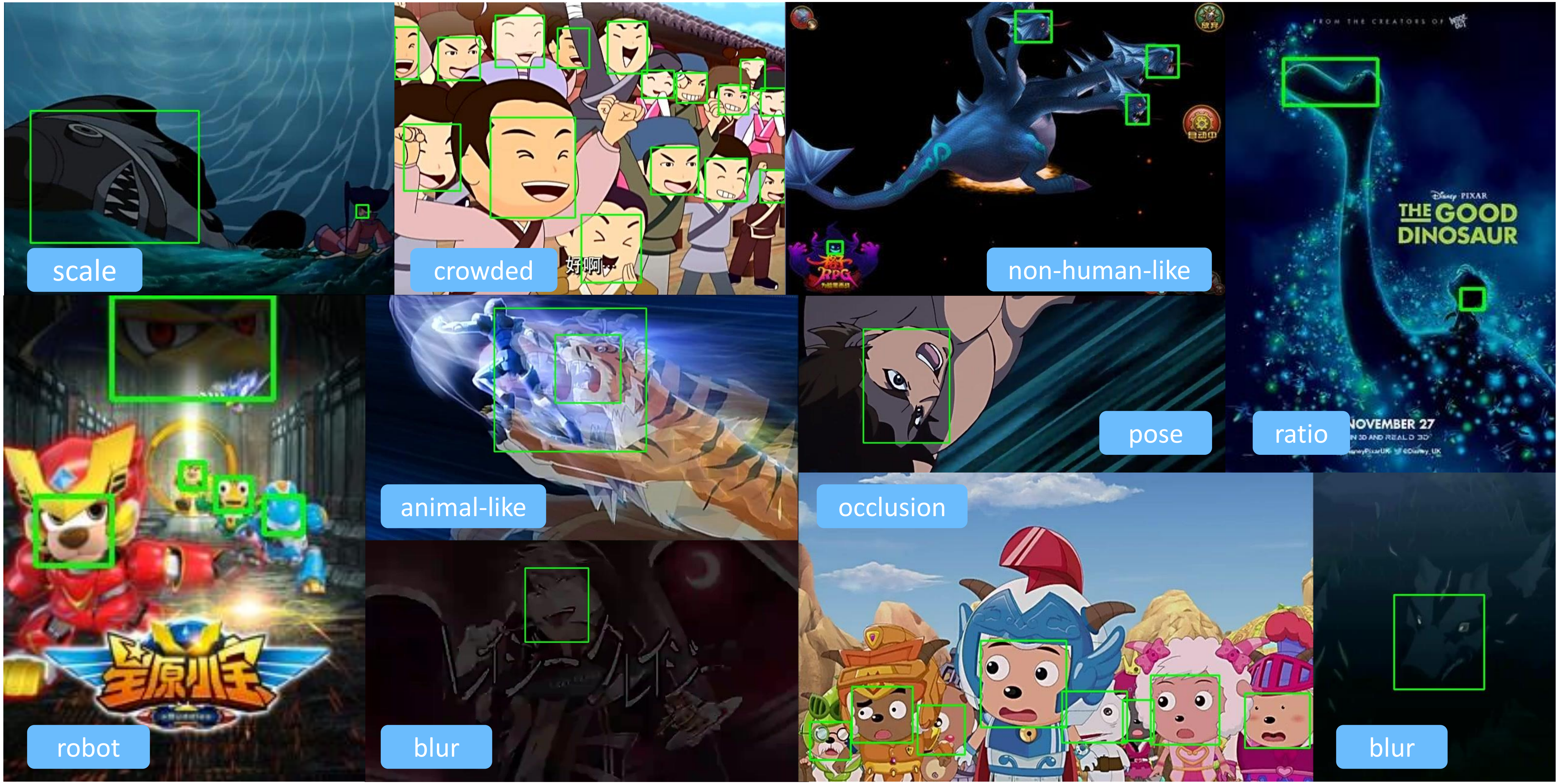
- Time-consuming optimization
  - Fuse conv. and BN layer. (accelerate about 3 ms when batch=1)
  - Batch processing. (speed up to 60+ ms without post-processing)
  - Convert PyTorch model to TensorRT by a simple torch2trt toolbox available at https://github.com/z-bingo/torch2trt. (average runtime: 48~49 ms)



Results on Leaderboard



Visualization



Conclusion

- Asymmetric Conv Block (ACB)
  - Provide the backbone network with the ability to generate features with more diverse receptive fields.
  - BiFPN with ACB could aggregate and enhance multi-scale features at the same time, previous methods achieved these by two separate modules.
- Dynamic Match Strategy
  - Instead of matching anchors and faces by a simple IoU threshold, it is employed for mining enough high-quality anchors for each face.
- Margin Loss
  - Enhance the power of discrimination, especially for those faces who are similar to the background, e.g., blur face, robot face.
- Data Augmentation
  - Augment the faces that are too small and too large to be accurately located.

Bin Zhang*, Jian Li*, Yabiao Wang, Zhipeng Cui
{* Equal Contribution}
Contact e-mail: swordli@tencent.com

Youtu Lab, Tencent

Southeast University

腾讯优图检测技术交流群

该二维码7天内(6月23日前)有效，重新进入将更新

腾讯优图

東南大學 SOUTHEAST UNIVERSITY