Multi-Scale Aligned Distillation for Low-Resolution Detection
Supplementary Material

Lu Qi[1†], Jason Kuen[2†], Jiuxiang Gu[2], Zhe Lin[2], Yi Wang[1], Yukang Chen[1], Yanwei Li[1], Jiaya Jia[1]
[1]The Chinese University of Hong Kong [2]Adobe Research

In this supplementary material document, we provide some details on extending our multi-scale feature fusion method to students. Furthermore, we provide additional experimental results on using

- Slimmed backbones and high-resolution students
- Different input resolutions during inference
- Multi-scale fusion students with slimmed backbones,

as well as

- Training details for $3\times$ training schedule
- Visualization results.

A. Multi-Scale Feature Fusion for Students

As shown in Table 4 of the main paper, it is still relatively challenging for the improved low-resolution student models to detect small objects. This is caused by the loss of fine visual details in the low-resolution input images. To counteract this problem, we extend cross feature-level fusion module introduced in Sec. 3.2 to combine two or more students with varying input resolutions and model complexities. The loss of visual details within a low-resolution student can be compensated by dynamically fusing its pyramidal features with the features from another high-resolution student. Due to the high computational requirements of processing high-resolution input images, we let the high-resolution student use a network architecture more compact and lightweight than the low-resolution student's. This strategy allows us to achieve a good balance between model efficiency and detection performance. The results of various multi-scale fusion students are provided in Table 7 of the main paper.

B. Slimmed Backbones and High-resolution Students

Table 1 shows the comprehensive results on using slimmed ResNet-50 backbones[1] [6]. Even when using a quarter of the original network width, the performance of the low-resolution student still reaches 30.4 AP which is reasonably good. This model's backbone runs at merely $\frac{1}{64}$ of the original/vanilla model's backbone FLOPS. Additionally, we demonstrate in this table that the strong teachers trained with our approach can also be used to improve the performance of high-resolution students remarkably well, enabling them to match the performance of the multi-scale fusion teachers (H&L) while merely requiring the computation costs of single-resolution backbones.

C. Inference at Different Input Resolutions

Table 2 shows the performance of student when performing inference across a wide range of different input resolutions. It can be seen that the student models trained with our multi-scale aligned distillation framework perform reasonably well even in the extremely low-resolution regime.

D. Multi-scale Fusion Students with Slimmed Backbones

The cross feature-level fusion module introduced in Sec. 3.2 (main paper) can also be applied to the student models. The loss of fine visual details in a low-resolution student can be compensated by dynamically fusing its pyramidal features with the features from another small-width high-resolution backbone. In Table 3, we show the performance of several such backbone combinations for the multi-scale fusion students that requires less computation costs (FLOPS) than the full-width high-resolution model. It can be seen that the performance on small-sized objects, $\text{AP}^S_{\mathbb{S}}$ has been improved materially compared to the models' single-low-resolution counterparts in Table 1.

With $0.50\times$ width for high-resolution and $1.00\times$ width for low-resolution backbones, the student model (first row) achieves 41.4 AP which is comparable with 41.3 AP of the $0.75\times$-width high-resolution student ($0.75\times$; S; H) in Table 1. These two models have comparable back-

†Equal contribution.

[1]As mentioned in the main paper, we do not adopt Slimmable training [6] for the detection models. We merely initialize the detection models with different-width backbones pretrained using Slimmable training on ImageNet dataset.

Width	Role	Input	AP	AP_{50}	AP_{75}	$AP_{\mathbb{S}}$	$AP_{\mathbb{M}}$	$AP_{\mathbb{L}}$
1.0×	T	H	39.7	57.9	43.2	27.2	44.0	49.3
		L	37.8	55.7	40.6	18.9	40.5	54.4
		H&L	42.5	61.1	46.3	28.0	45.7	55.7
	S	H	42.3	61.3	45.9	28.2	45.9	53.5
		L	39.7	58.0	42.5	21.7	42.9	55.0
0.75×	T	H	38.3	56.1	41.5	25.7	42.5	47.3
		L	36.8	54.4	39.4	18.4	39.1	53.2
		H&L	41.1	59.2	44.5	26.7	44.2	54.3
	S	H	41.3	60.0	44.6	27.4	44.7	51.9
		L	38.6	56.7	41.3	20.8	41.2	53.8
0.5×	T	H	36.0	53.5	39.0	23.2	39.8	44.7
		L	33.8	50.8	36.4	16.3	35.4	49.1
		H&L	38.5	56.3	41.7	23.5	40.9	50.6
	S	H	38.8	57.1	42.0	25.3	41.9	48.9
		L	36.1	53.7	38.8	18.4	38.2	50.4
0.25×	T	H	30.5	46.5	32.5	18.1	33.2	39.1
		L	28.5	43.9	30.2	12.4	29.2	42.8
		H&L	33.2	49.9	35.4	19.0	34.7	44.4
	S	H	33.4	50.5	35.6	20.0	35.5	43.0
		L	30.4	46.5	32.3	15.1	31.5	44.1

Table 1: Performance evaluation on using slimmed ResNet-50 [6] with different widths as the detector backbones. H&L is the multi-scale fusion teacher that distills knowledge to student S with either H (800px) or L (400px) input resolution. At a particular network width, there is only a single teacher that is evaluated on several resolution types, and there are two distinct students trained respectively for H and L input resolutions. $1\times$ training schedule is used here.

bone FLOPS which are about half of the full-width high-resolution model's. However, the dual-resolution model comes with separate dual-resolution backbones that can fully run in parallel before the feature fusion happens. The inference runtime efficiency can be significantly boosted by running the dual-resolution backbones separately on different hardware accelerators, in a similar spirit to parallelizing large-scale neural network training [2]. This is not applicable to the single-resolution $0.75\times$-width model that has only sequentially-dependent layers. Our multi-scale feature fusion approach can be seen as a way to perform effective *model separation* that can potentially benefit from advances in model parallelism.

E. Training Details for $3\times$ Training Schedule

For FCOS [5], RetinaNet [4] and MEInst [7], λ, γ are set to 0.4 and 0.8. Whereas, for Mask R-CNN [1], λ and γ are set to 0.2 and 0.6. Compared to the experiments with $1\times$ training schedule, we adopt smaller γ values here for $3\times$ training schedule. λ balances our proposed aligned knowledge distillation (KD) loss and original detection loss. In a prolonged training process, it is beneficial to provide a stronger emphasis to KD loss due to the fact that the original detection loss converges sooner than KD loss does. A greater KD loss weight allows the training to better focus on minimizing the KD loss after the original detection loss converges.

Resolution Type	Inference Resolution	AP^S	AP^S_{50}	AP^S_{75}	$AP^S_{\mathbb{S}}$	$AP^S_{\mathbb{M}}$	$AP^S_{\mathbb{L}}$
H	800	42.3	61.3	45.9	28.2	45.9	53.5
	768	42.4	61.3	45.8	27.4	45.9	53.6
	736	42.2	61.1	45.7	27.2	45.9	54.3
	704	42.1	61.0	45.6	26.9	45.8	54.5
	672	41.9	60.7	45.1	25.2	45.7	54.3
	640	41.7	60.3	44.7	24.7	45.6	55.3
L	400	39.7	58.0	42.5	21.7	42.9	55.0
	384	39.3	57.8	42.0	20.9	42.3	55.3
	368	38.9	57.2	41.7	19.7	42.0	55.6
	352	38.6	56.8	41.4	19.9	41.7	55.5
	336	37.9	55.8	40.4	18.8	40.6	55.6
	320	37.4	55.1	39.9	18.3	39.7	55.9
EL	200	31.8	47.8	33.4	12.2	31.4	52.2
	192	31.1	46.7	32.7	11.4	30.8	51.3
	184	30.4	45.8	31.9	11.3	29.8	51.2
	176	29.6	44.8	30.9	10.3	28.9	49.9
	168	28.7	43.6	30.2	9.4	27.7	49.1
	160	27.9	42.5	28.8	8.9	26.7	48.1

Table 2: Performance evaluation on the high(H)-, low(L)-, and extremely low(EL)-resolution student (with ResNet-50 backbone and FCOS) models guided by our final multi-scale fusion teacher trained on three base resolutions (800px, 400px, 200px). $1\times$ training schedule is used here.

Width (H)	Width (L)	AP^S	AP^S_{50}	AP^S_{75}	$AP^S_{\mathbb{S}}$	$AP^S_{\mathbb{M}}$	$AP^S_{\mathbb{L}}$
0.50×	1.00×	41.4	59.9	44.5	24.6	44.5	54.6
0.50×	0.75×	41.1	59.4	44.0	24.5	43.6	54.5
0.50×	0.50×	40.1	58.5	43.2	24.5	42.7	51.8
0.25×	0.50×	37.6	55.5	40.6	20.8	40.0	50.9

Table 3: Performance evaluation on using dual-resolution (high/H and low/L input resolutions) slimmed backbones within multi-scale fusion student models.

In the $3\times$ training schedule experiments, Mask R-CNN uses feature maps from P_2 to P_6 for high-resolution (800px) model. In our setting, we shift the pyramid feature level for the low-resolution (400px) model by setting m=1. In other words, we use the feature maps from P_1 to P_5 for the low-resolution model. For the low-resolution Mask R-CNN, we define P_1 as the combination of P_2 and C_1 features, following the standard FPN structure [3]. C_1 corresponds to the features that come after the first 7×7 convolution block of ResNet architecture.

F. Visualization results

We show the visualization results of the low-resolution student models (all with ResNet-50 backbones) trained with our framework on object detection, instance segmentation, and keypoint detection. For all tasks, we apply a threshold score of 0.7 to filter out unconfident detections. For each row, the left, middle and right sub-figures correspond to the the ground truth, the result from the multi-scale fusion (800px&400px) teacher model, and the result from the low-resolution (400px) student trained with our approach, respectively. It is notable that the low-resolution model performs very well even on small-sized objects.

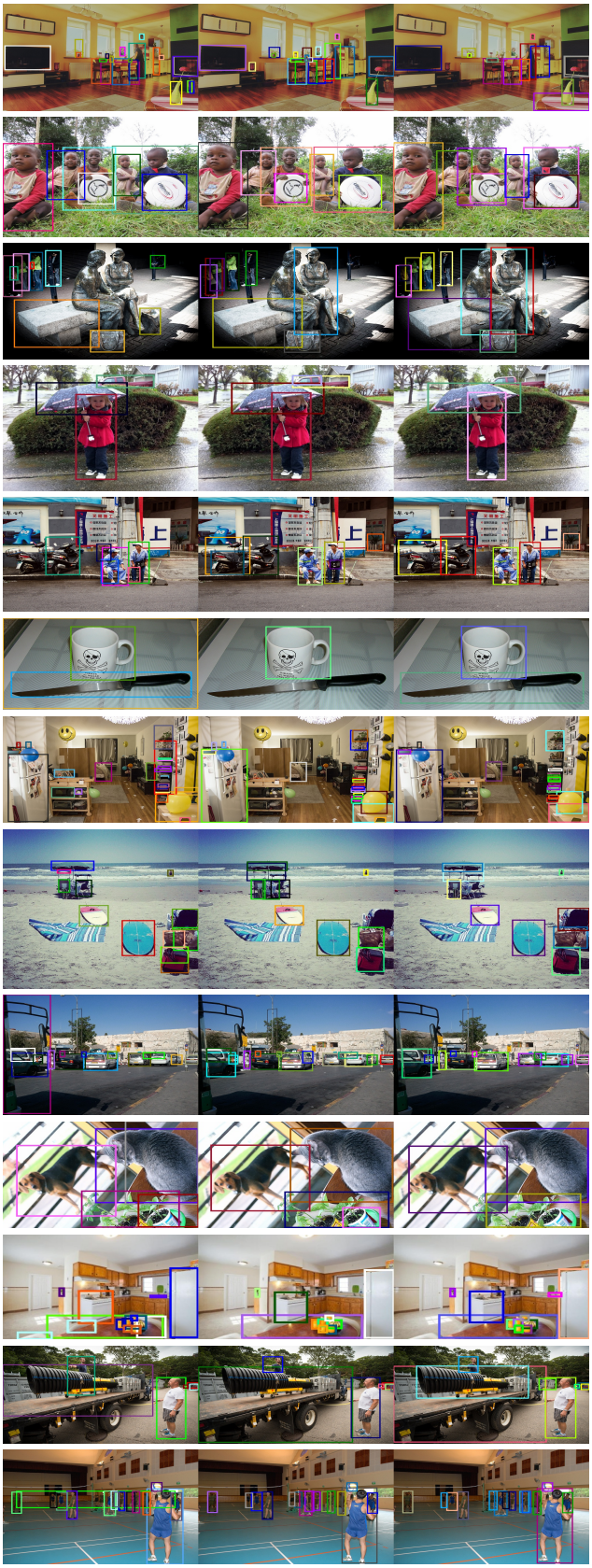

Figure 1: Object detection with FCOS.

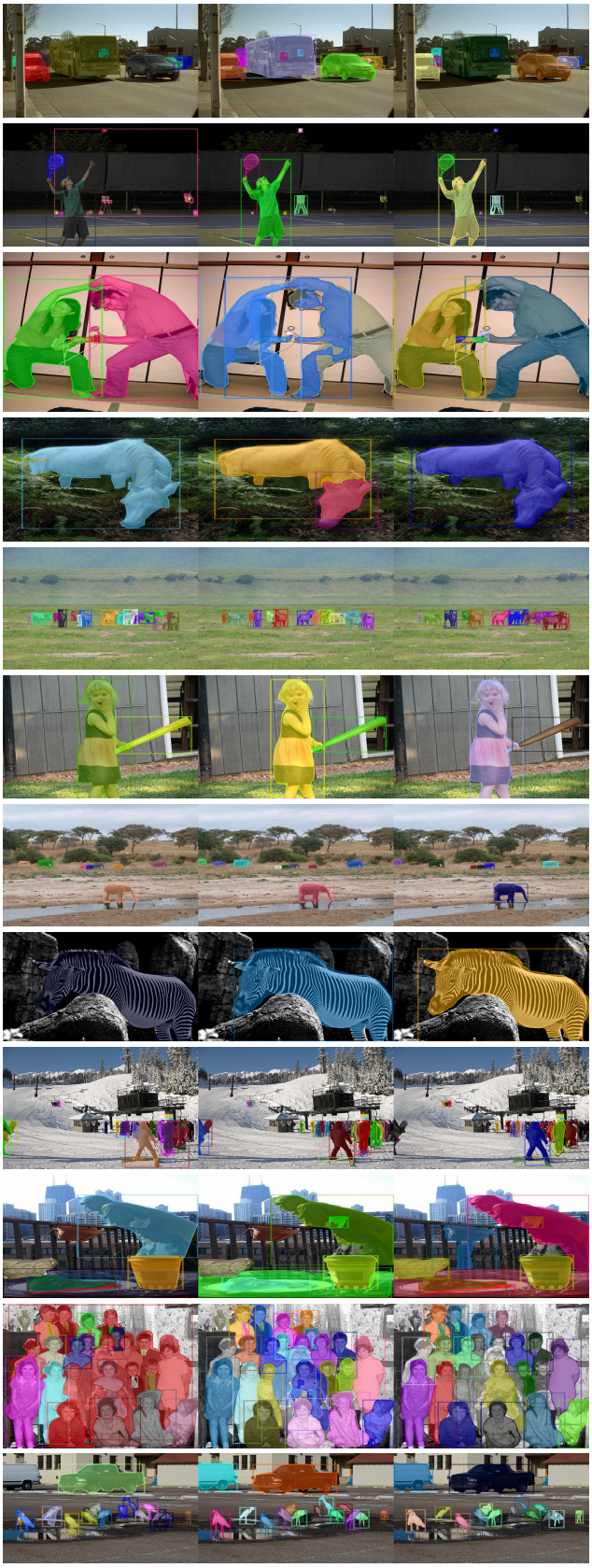

Figure 2: Instance segmentation with Mask R-CNN.

Figure 3: Keypoint Detection with Mask R-CNN.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[2] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *NeurIPS*, 2019. 2

[3] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[5] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 2

[6] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In *ICLR*, 2019. 1, 2

[7] Rufeng Zhang, Zhi Tian, Chunhua Shen, Mingyu You, and Youliang Yan. Mask encoding for single shot instance segmentation. In *CVPR*, 2020. 2