Appendix

A. Ablations

We provide more detailed ablation results of MKD in Tab. 1. Baseline denotes the student network which is trained from scratch, and the experiment is conducted under a 3× schedule. To be more specific, we train the model for 36 epochs and reduce the learning rate by 0.1× at the 27th and 33rd epochs.

method	AP	AP50	AP75
baseline	37.1	56.0	39.5
MKD mask=0.1	41.4	61.0	44.2
MKD mask=0.2	41.6	61.1	44.5
MKD mask=0.3	**41.7**	**61.3**	**44.6**
cutout 0.1	39.6	58.9	42.4
FitNet	39.9	59.2	42.7
FGD	41.0	60.4	43.6
MGD	41.2	60.8	44.0

Table 1: Detailed results on COCO dataset under 3x schedule with different distillation methods and MKD with different mask ratios.

Compared with the commonly used cutout augmentation, our masking strategy with feature generation operation can achieve a much higher improvement. This indicates that the effectiveness of our method is not mainly come from the data augmentation brought about by the masked input image, but from fully learning the teacher's corresponding information in the adjacent regions. We observe that as the training schedule gets longer, conventional distillation methods reach saturation, while our methods continue to rise and get a larger performance gap. This shows that our method can enhance the distillation process and more fully explore the potential of the student model. As is shown in Tab. 1, MKD with a masking ratio of 0.3 appears to perform the best in longer training schedules such as 3×. This indicates the effectiveness of the mask autoencoding scheme in the distillation and shows its difference with common data augmentation.

We further examine different combinations of mask patch sizes and SAM resolutions. As is shown in Tab. 4, SAM resolution $[H/32, W/32]$ with the patch size of 32 outperforms the others. This suggest that larger mask patch size such as 64 may distort the input image too much while patch size of 32 is more moderate in our proposed method.

The current MKD pipeline consists of a decoder and an encoder. To justify the necessity of adding a decoder for generating a complete teacher feature maps, we try to add an additional L2 loss L_{MSE} to the output of the backbone of the student network (i.e. the encoder), making the student's feature directly mimic the teacher's feature in the unmasked areas. The result is shown in Tab. 5. The result shows that the encoder output directly mimicking the teacher's feature did not result in any further improvement, suggesting that using the decoder to force the student to generate full feature maps can better enhance the distillation.

Ability compared with MAE pretraining. Table 2 shows that MIM or vanilla KD alone can give a boost, while the gain brought by KD is more significant and our MKD performs the best. In Table 3, MKD continues to improve on the basis of MIM pre-training. These results indicate that MKD indeed improves over standard MIM and KD.

B. Error analysis on COCO

Fig. 1 illustrates the error analysis on four ran-

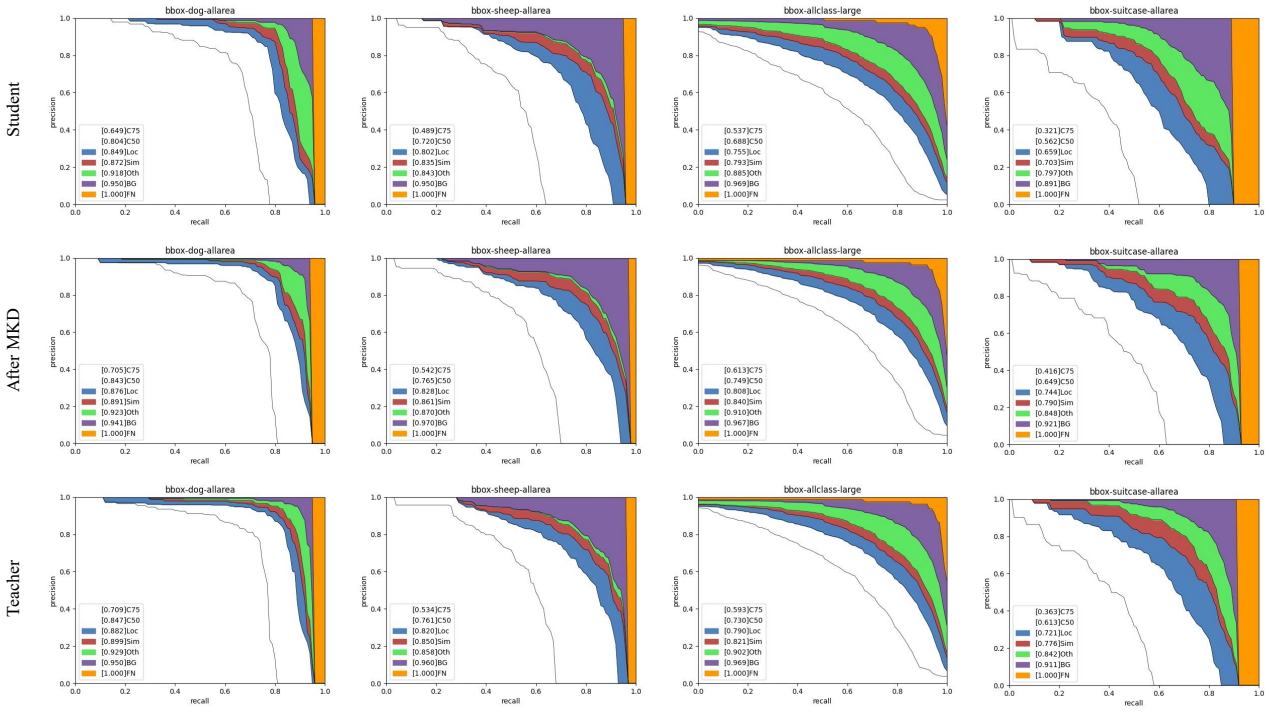

Figure 1: Precision-Recall curves and error analyses of different distillation methods on COCO dataset. The top curves correspond to the original student model (RetinaNet-Res50, 37.4 mAP), the middle curves correspond to the student model after MKD (RetinaNet-Res50, 41.5 mAP), and the bottom curves correspond to the teacher model(RetinaNet-ResX101, 41.0 mAP).

Method	mAP
RetinaNet-Res101(T)	38.9
RetinaNet-Res18(S)	33.4
MIM* (self-KD)	34.1(+0.7)
Vanilla KD	35.9(+2.5)
MKD	**37.3(+3.9)**

Table 2: Comparison of MIM, vanilla KD and our MKD. MIM* denotes MKD with pretrained student as the teacher.

Method	mAP
ViTDet ViT-Base(T)	51.1
ViTDet ViT-Small-MAE(S)	47.2
Vanilla KD	48.4
MGD	49.2
MKD	**50.3**

Table 3: Results on the transformer-based backbone with MIM pretraining.

domly selected classes. Our proposed MKD significantly improves the overall performance and the background error, compared with the original student model. The Precision-Recall curves further indicate that our MKD helps the smaller student model make better predictions and surpass the teacher model.

patch size	resolution	mAP	AP50	AP75
32	$[H/32, W/32]$	**39.9**	**59.3**	42.5
32	$[H/64, W/64]$	39.8	59.0	**42.7**
64	$[H/64, W/64]$	39.7	58.9	42.6

Table 4: Ablation on the relation between mask patch size and resolution in SAM under 1x schedule.

L_{MSE}	schedule	mAP	AP50	AP75
baseline	2x	37.4	56.7	39.6
w/o	2x	**41.1**	**60.6**	**44.0**
w/	2x	40.8	60.3	43.7

Table 5: Ablation on the effectiveness of decoder.