

# Supplementary Material for "Anchor-Based Masked Generative Distillation for Pixel-Level Prediction Tasks"

Xie Yu\*

yuxie\_scse@buaa.edu.cn

Wentao Zhang

zhangwt97@buaa.edu.cn

School of Computer Science,  
Beihang University,  
Beijing, China

\* Corresponding Author

## 1 Implementation

### 1.1 The generation process of Anchor-Based Masks

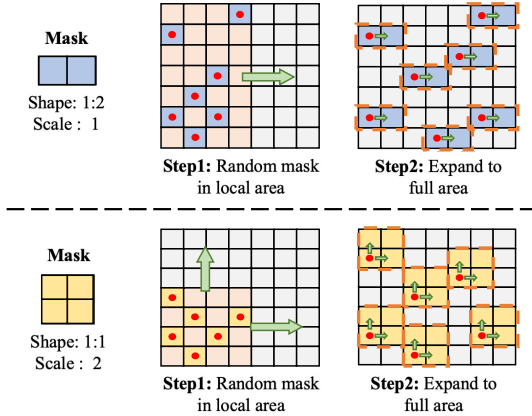


Figure 1: Visualization Results on Semantic Segmentation

The generation process of Anchor-Based Masks (ABM) is illustrated in Figure 1, where each mask map contains only one type of anchor-based mask with a certain shape and scale. The generation of mask maps should satisfy the following two conditions:

- To transfer more structured knowledge while reducing disturbance between masks, it is necessary to ensure that masks do not overlap with each other.
- To achieve sufficient generalization ability, the positions of anchor-based masks must have a certain degree of randomness.

Therefore, we initially generate a mask map within a localized region of the feature map, corresponding to the shape of masks. Each mask within this map covers only one  $1 \times 1$  pixel. Subsequently, we expand this mask map to match the size of the feature map. For instance, to generate a mask with a  $1 : 2$  aspect ratio on a feature map of dimensions  $H \times W \times C$ , we first create a random mask denoted as  $M$  on a scale of  $\frac{H}{2} \times W \times C$ . Then, we expand this random mask by a factor of two along the  $H$  dimension to obtain  $\hat{M}$ . If we need to increase the mask ratio, we proportionally reduce the size of  $M$ .

## 1.2 Implementation Details

**Settings in object detection.** We selected mean Average Precision ( $mAP$ ) as the main evaluation metric, and additionally report AP at object sizes  $AP_S$ ,  $AP_M$ ,  $AP_L$ . We conduct our experiments on MMDetection2 framework. Each model is trained using SGD optimization with momentum 0.9, weight decay  $1e^{-4}$  and batch size 8. The learning rate is set at 0.02.

**Settings in semantic segmentation.** We implement our method on MMSegmentation codebase. In training and evaluation, we use mean Intersection-over-Union (mIoU) to measure the performance of all methods. In training phase, all models are optimized by SGD with the momentum of 0.9, the initial learning rate of 0.02, and the batch size of 16. The input size is  $512 \times 512$ . In evaluation phase, we follow general settings in, which evaluate the performance with the original image size.

**Settings in super-resolution.** For evaluation, we select the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) on the  $Y$  channel of the  $YC_bC_r$  color space conventionally. For data, the low resolution images used for training were obtained by down-sampling the high-resolution images with the bicubic degradation. The  $\times 4$  scale super-resolution models are initialized with the corresponding  $\times 2$  ones. During training, each low resolution image is randomly cropped into  $48 \times 48$  patches and augmented by random horizontal and vertical flips and rotations. All the models are trained with ADAM optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.99$  and  $\epsilon = 1e^{-8}$ , with a batch size of 16. The initial learning rate is set to  $1e^{-4}$  and is decayed by a factor of 10 at every 105 iterations.

## 2 Datasets and Baselines

### 2.1 Datasets

We evaluate AMGD on three different tasks including object detection, semantic segmentation and super-resolution. For object detection, we train and evaluate AMGD on MS COCO dataset. For semantic segmentation, we train and evaluate AMGD on Cityscapes dataset. For super-resolution, we train AMGD on DIV2K and evaluated on four benchmark datasets: Set5, Set14, BSD100, and Urban100.

**MS COCO** is a large-scale dataset commonly used for object detection. It has 118,000 training samples and 5,000 validation samples, with totally 80 categories for object detection. In MS COCO dataset, each image contains 3.5 categories and 7.7 instance targets on average. The images in the COCO dataset contain more intricate backgrounds, a greater number of targets, and smaller target sizes. Less than 20% of the images contain only one category, and only 10% contain only a single instance.

**Cityscapes** is a dataset for real-world semantic urban scene understanding. It has 5,000 image samples with high quality pixel-level annotations and 20,000 image samples with

coarse annotations collected from 50 different cities. In semantic segmentation, only samples with pixel-level annotations are used, which contain 2,975 training samples, 500 validation samples and 1,525 testing samples, with totally 19 classes.

**DIV2K** is a single-image super-resolution dataset which contains 1,000 images with different scenes and is splitted to 800 for training, 100 for validation and 100 for testing. Each high-resolution image in DIV2K has a 2K resolution. Low resolution images are generated from its corresponding high resolution images with magnification factors of  $\times 2$ ,  $\times 3$  and  $\times 4$ .

**Set5**, **Set14**, **BSD100** and **Urban100** are all commonly used datasets for super-resolution evaluation. Specifically, the Set5 dataset comprises five images categorised as "baby", "bird", "butterfly", "head", and "woman". The Set14 dataset consists of fourteen images categorised as "baboon", "barbara", "bridge", "coastguard", "comic", "face", "flowers", "foreman", "lenna", "man", "monarch", "pepper", "ppt3", and "zebra". The Urban100 dataset contains 100 images of urban scenes, featuring architectural structures. The BSDS100 dataset includes 100 images representing a diverse range of categories, from natural scenes to specific objects, such as plants, people, and food.

## 2.2 Baselines

In this part, we conducted comparative experiments on three different tasks. For object detection, we selected FKD[[10](#)], CWD[[5](#)], FGD[[9](#)] and MGD[[11](#)] as baselines; For semantic segmentation, we selected SKD[[8](#)], IFVD[[7](#)], CWD[[5](#)], CIRKD[[6](#)], MGD[[11](#)] and MaskD[[9](#)] as baselines; For super-resolution, we selected RKD[[4](#)], FAKD[[12](#)], CSD[[13](#)] and DUKD[[14](#)] as baselines.

**Baselines in object detection.** FKD is designed to be a fast knowledge distillation (FKD) framework that achieves the same high level of performance as vanilla KD. CWD firstly extended the feature distillation methods into channel-wise dimension by normalizing the activation map of each channel. FGD took the relationship among pixels into consideration. It proposed focal and global distillation, which enables the student not only to focus on the teacher's critical pixels and channels, but also to learn the relation between pixels. MGD is the first methods to involve masked generative paradigm into knowledge distillation tasks. And it has a great potential in building unified architecture for various distillation tasks. MasKD represents the current state-of-the-art (SOTA) methodology within the masked generative distillation paradigm. It introduced a learnable embedding, termed the receptive token, to localize vital features within the feature map, thereby enhancing the performance of traditional masked generative distillation methods.

**Baselines in semantic segmentation.** Beyond CWD, MGD, and MasKD, we further introduce two distillation methods specialised for semantic segmentation, namely IFVD and CIRKD. IFVD proposed an intra-class feature variation distillation method for semantic segmentation. It forced the student model to mimic the set of similarity between the feature on each pixel and its corresponding class-wise prototype, aiming to alleviate the difference of feature distributions between the student model and the teacher model. While CIRKD focused more on transferring cross image knowledge such as structured pixel-to-pixel and pixel-to-region relations among teacher's feature maps and student's feature maps.

**Baselines in super-resolution.** RKD and FAKD are originally proposed for high-level CV tasks, but they are compatible with SR task and applicable to CNN models. CSD is a self-distillation method through channel-wise contrastive learning. DUKD is the current SOTA method in super-resolution distillation. It aims to leverage the label consistency regularization into knowledge distillation for super-resolution tasks.

## References

- [1] Zibin He, Tao Dai, Jian Lu, Yong Jiang, and Shu-Tao Xia. Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 518–522. IEEE, 2020.
- [2] Tao Huang, Yuan Zhang, Shan You, Fei Wang, Chen Qian, Jian Cao, and Chang Xu. Masked distillation with receptive tokens. *arXiv preprint arXiv:2205.14589*, 2022.
- [3] Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2604–2613, 2019.
- [4] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.
- [5] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021.
- [6] Yanbo Wang, Shaohui Lin, Yanyun Qu, Haiyan Wu, Zhizhong Zhang, Yuan Xie, and Angela Yao. Towards compact single image super-resolution via contrastive self-distillation. *arXiv preprint arXiv:2105.11683*, 2021.
- [7] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 346–362. Springer, 2020.
- [8] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022.
- [9] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022.
- [10] Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *European Conference on Computer Vision*, pages 53–69. Springer, 2022.
- [11] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020.
- [12] Yun Zhang, Wei Li, Simiao Li, Jie Hu, Hanting Chen, Hailing Wang, Zhijun Tu, Wenjia Wang, Bingyi Jing, and Yunhe Wang. Data upcycling knowledge distillation for image super-resolution. *arXiv preprint arXiv:2309.14162*, 2023.