FerKD: Surgical Label Adaptation for Efficient Distillation
Supplementary Material

Zhiqiang Shen
Mohamed bin Zayed University of AI
Zhiqiang.Shen@mbzuai.ac.ae

Appendix

In the appendix, we provide more details omitted in the main paper, including:

- Section A: Implementation details.
- Section B: More visualization of identified crops.

Backbone	ResNet-50	ViT-S/16
Epoch	300	300
Batch size	1,024	1,024
Optimizer	AdamW	AdamW
Init. *lr*	0.002	0.002
lr scheduler	cosine	cosine
Weight decay	0.05	0.05
Warmup epochs	5	5
Num crops	4	4
Label smoothing	✗	✗
Dropout	✗	✗
Stoch. Depth	✗	0.1
Repeated Aug	✗	✗
Gradient Clip.	✗	✗
Rand Augment	✗	✗
Mixup prob.	✗	0.8
Cutmix prob.	✗	1.0
SelfMix prob.	1.0	✗
Random erasing	✗	✗

Table 1: Pre-training setting for ImageNet-1K.

A. Implementation Details

Training details for ResNet-50 and ViT-S/16 in the main text. We elaborate the detailed training settings and hyper-parameters of `FerKD` for pre-training from scratch on ImageNet-1K with ResNet-50 and ViT-S/16 backbones, as provided in Table 1. Generally, the training protocol follows FKD [2]'s training strategy on ViT, DeiT and SReT. We employ `SelfMix` for ResNet-50, Mixup and CutMix for ViT-S/16 separately. We also use 4 as the number of crops in each image, batch size $= 1,024$ during training.

Training details for finetuning ViT-G/14 and RegY-128GF in the main text. The finetuning settings and hyper-parameters of `FerKD` with ViT-G/14 [1] and RegY-128GF [3] backbones are provided in Table 2, which are similar to the training protocol in EVA [1]. We employ `SelfMix` for both of the two pretrained backbones.

Backbone	ViT-G/14 [1] | RegY-128GF [3]
Peak learning rate	3e-5
Optimizer	AdamW
Optimizer hyper-parameters	$\beta_1, \beta_2, \epsilon$ = 0.9, 0.999, 1e-8
Layer-wise lr decay	0.95
Learning rate schedule	cosine decay
Weight decay	0.05
Input resolution	336
Batch size	512
Warmup epochs	2
Training epochs	15
Num crops	2
Drop path	0.4 | 0.0
Augmentation	RandAug (9, 0.5)
Label smoothing	✗
Cutmix	✗
Mixup	✗
Random erasing	✗
SelfMix prob.	1.0
Random resized crop	(0.08, 1)
Ema	0.9999
Test crop ratio	1.0

Table 2: Fine-tuning setting for ImageNet-1K.

Data augmentation details for Mixup, Cutmix and `SelfMix`. The data augmentation configurations adopted in training are: for Mixup, we use probability 0.8 to generate the *Beta distribution*, and 1.0 for CutMix and `SelfMix`.

B. More Visualization

Fig. 1 illustrates the identified crops by teacher for hard and easy samples. We do not involve any localization information, but the teacher's probability can reflect object and background areas automatically based on their magnitudes.

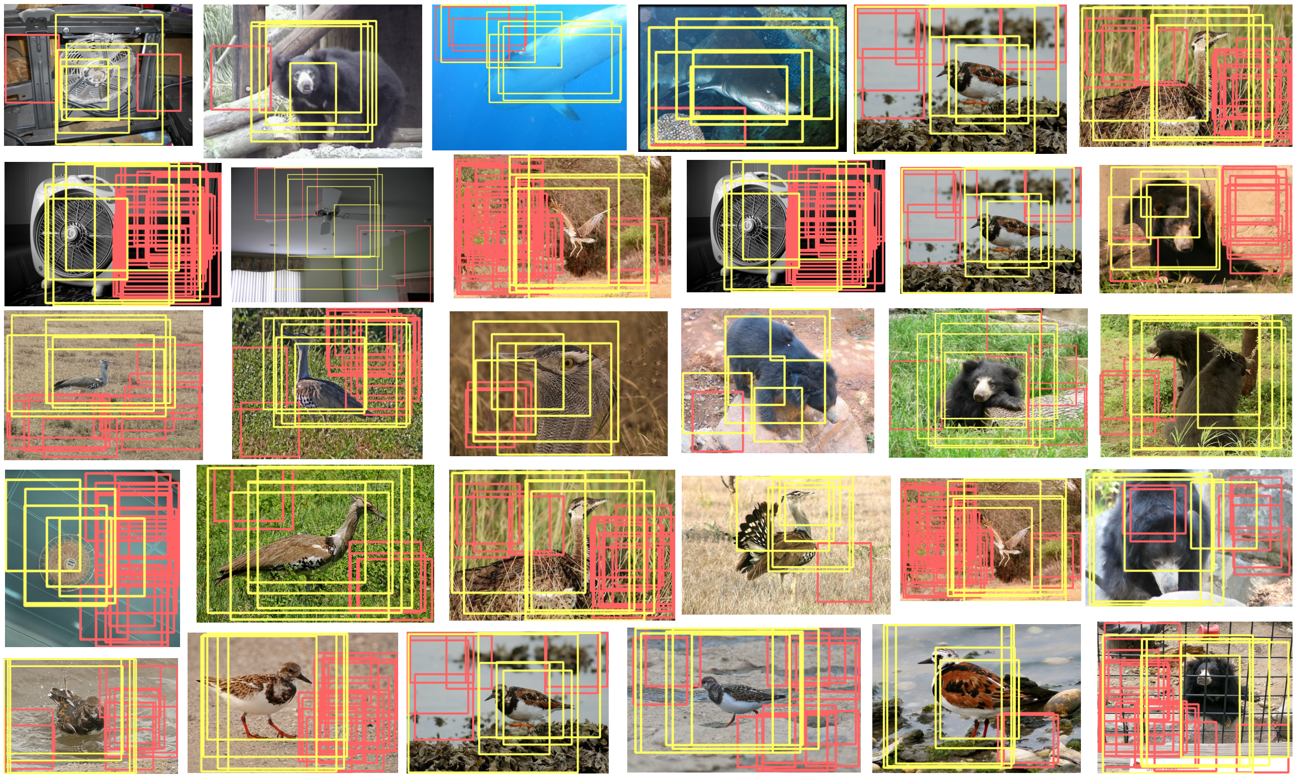

Figure 1: Illustration of the identified crops by teacher model for hard (background) and easy (foreground) samples. The teacher's probability can reflect object and background areas visually based on their magnitudes.

References

[1] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 1

[2] Zhiqiang Shen and Eric Xing. A fast knowledge distillation framework for visual recognition. In *ECCV*, 2022. 1

[3] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *CVPR*, 2022. 1