

HEAVY LABELS OUT! DATASET DISTILLATION WITH LABEL SPACE LIGHTENING

-SUPPLEMENTARY MATERIALS-

Anonymous authors

Paper under double-blind review

A IMPLEMENTATION DETAILS

A.1 DATASETS

We primarily conduct experiments on **high-resolution datasets** ImageNet-100, Places365-Standard Zhou et al. (2017), and ImageNet-1K Deng et al. (2009). Among them, ImageNet-100 is a subset of ImageNet-1K. As there are multiple versions of ImageNet-100, we follow the previous work IDC Kim et al. (2022) to build the dataset. In addition, all images are resized to 224×224 .

A.2 NETWORKS

Our proposed method utilizes ResNet-18 He et al. (2016) networks in different training stages as guidance to perform LoRA-like low-rank knowledge transfer to the surrogate projection model. Here, we adopt the official torchvision code to obtain the training trajectories of the teacher models for ImageNet-100, Places365-Standard, and ImageNet-1K. For detailed parameter settings, please refer to Table 1. To better align with downstream tasks, we propose a multi-weak-teacher strategy. The stage of the model is closely related to the complexity of the dataset. For simple dataset or small IPC, we adopt teachers from the very early stage, while for complex or large IPC, we adopt teachers from the later stage.

For baseline performance evaluation, we adopt ResNet-18 He et al. (2016) as the evaluation architecture. For cross-architecture performance evaluation, we adopt ShuffleNet-V2 (X0_5) Ma et al. (2018), MobileNet-V2 Sandler et al. (2018), EfficientNet-B0 Tan & Le (2019), Swin-V2-Tiny Liu et al. (2022), and VGG-11 Simonyan (2014) as the evaluation architectures.

A.3 DETAILS OF SURROGATE PROJECTION

For the architecture of the surrogate projection models, we adopt CLIP (ResNet-50) Radford et al. (2021) from the Open-AI as the base model. Here, we use the image encoder part from CLIP, followed by a 1024×1000 linear transformation. To improve the initial performance of the whole projection model and save storage space, we employ the text embedding of the text encoder to initialize the linear transformation. We adopt prompts from the official prompt engineering to ensemble generate text embedding, which is as follows: “itap of a {}.”; “a bad photo of the {}.”; “a origami {}.”; “a photo of the large {}.”; “a {} in a video game.”; “art of the {}.”; “a photo of the small {}.”. Also, the computational process of the projection model is equivalent with the original CLIP model. Subsequent updates of the whole projection model will be carried out through LoRA-like low-rank knowledge transfer, which significantly reduce the required storage.

A.4 DETAILS OF LoRA-LIKE LOW-RANK KNOWLEDGE TRANSFER

Here, for the ImageNet-100 dataset, we adopt rank 8 and 64 for the image encoder part and linear transformation part, respectively. As for both Places365-Standard and ImageNet-1K, we adopt rank 8 and 128 for the image encoder part and linear transformation part, respectively. At the same time, we adopt multiple weak teachers to guide the projector learning. The hyper-parameters of the knowledge transfer process are listed in Table 2.

Table 1: The hyper-parameters of teacher model generation for all three datasets.

Hypeparameter	Value
Optimizer	SGD
Base Learning Rate	0.1
Learning Rate Scheduler	Step
Weight Decay	1e-4
Learning Rate Step Size	30
Momentum	0.9
Batch Size	256
Model Pool Size	9
Training Epochs	90

Table 2: The hyper-parameters of LoRA-like knowledge transfer for all three datasets.

Hypeparameter	Value
Optimizer	AdamW
Learning Rate	5e-4
Training Epochs	20
Loss Type	MSE + CE
CE Weight	0.1
Batch Size	256

A.5 DETAILS OF IMAGE INITIALIZATION AND UPDATE

For image initialization, we adopt the state-of-the-art method RDED Sun et al. (2024b) to select and concatenate the important patches from the original dataset by well-pretrained ResNet-18. Then, to narrow the performance gap between the observer and the projector, we follow the LIC Sun et al. (2024a), and adapt it to our cases. We match the information loss of the features encoded by the image encoder of the projector. Here, the learning rate of the image update is 0.01, the optimizer is Adam, and the number of epochs is 300.

A.6 DETAILS OF EVALUATION

For all three datasets, the number of training epochs for evaluation is 300. Following the previous works Yin et al. (2024); Sun et al. (2024b); Shao et al. (2024), the augmentation strategy used here is CutMix Yun et al. (2019). For more details, please refer to Table 3.

Table 3: The hyper-parameters of evaluation for all three datasets.

Hypeparameter	Value
Optimizer	AdamW
Learning Rate	1e-3
Training Epochs	300
Loss Type	MSE + CE
CE Weight	0.025
Batch Size	100
Augmentation	CutMix
Alpha	1.0

B MORE EXPERIMENTS RESULTS

B.1 RESULTS ON HIGHER IPCS

To further show the effectiveness of our proposed method, here, we conduct experiments under the settings of ImageNet-100 with IPC 100. The performance results and the required extra storage costs are shown in Table 4 (here, we adopt online soft label generation and regard the teacher model as the extra storage costs). The results indicate that our proposed method can maintain high performance at higher IPCs.

Table 4: The results on higher IPCs. The experiments are conducted under the ImageNet-100 with IPC 100. Here, the required extra storage costs for RDED is the storage costs for teacher model.

	Downstream Accuracy	Required Extra Storage Costs
RDED	75.9 ± 0.1	42.83MB
Ours	76.2 ± 0.2	10.20MB (0.24× of the original extra storage)

B.2 RESULTS ON MORE TRANSFORMER-BASED ARCHITECTURES

To demonstrate the generalizability of our proposed method on transformer-based architectures, we conduct cross-architecture evaluation experiments on Swin-V2-Tiny under the ImageNet-1K setting with IPC 10, as presented in the paper. In addition, we perform supplementary experiments on ViT-B-16 under the same settings, with the results shown in Table 5. The evaluation results indicate that our proposed method exhibits stronger generalization ability on transformer-based models.

Table 5: The cross-architecture results on transformer-based architectures. The experiments are conducted under the setting of ImageNet-1K with IPC 10.

	ViT-B-16	Swin-V2-Tiny
RDED	19.1 ± 0.4	17.8 ± 0.1
Ours	$23.3 \pm 0.3 (+ 4.2)$	$29.5 \pm 0.1 (+11.7)$

B.3 EFFICIENCY EVALUATION FOR DOWNSTREAM TASKS

Here, we conduct experiments under the ImageNet-1K with IPC 10 setting. All methods are evaluated based on the official code, and all experimental configurations, including hyperparameters, are set according to the official default values provided by the authors. Specifically, we measure the time required for each single downstream training iteration as well as the overall memory cost. All experiments are conducted on one single NVIDIA RTX A5000 GPU.

Table 6: The comparison for the efficiency of the downstream training. Here, we evaluate the required runtime for single iteration and over all peak memory.

	Ours	SRe ² L	G_VBSM	RDED
Runtime	0.21s	0.75s	0.77s	0.12s
Peak Memory	4004MiB	20850MiB	23310MiB	4538MiB

B.4 RESULTS ON OTHER TASKS

We also conduct experiments on semantic segmentation tasks, here we follow the definition and setting of dataset distillation, utilizing a very small subset of 10,000 images from the SA-1B

dataset (around 0.09% of the original dataset) for downstream model training. Here, the baseline performance is evaluated on the 89.85M SAM-B model (storage costs 358.32MB) to generate soft labels for the downstream tasks. Our proposed method adopts the SAM-B model as the base model and applies our proposed label space lightening strategies, finally with only 5.64M (storage costs 22.56MB, 6.3% of the original storage costs) learnable and required to store parameters. The baseline performance is mIoU 49.46%, while our proposed method, uses only 6.3% storage costs, and achieves better performance with mIoU of 50.75%.

Table 7: Performance on semantic segmentation tasks. Here, SAM-B Directly Guided refers to directly adopting the SAM-B model to generate the soft labels for downstream tasks. Our method utilizes SAM-B as the base model with only 6.3% of the original costs to achieve better performance.

	Downstream mIoU	Required Extra Storage Costs
SAM-B Directly Guided	49.46%	358.32MB
Ours	50.75%	22.56MB (0.063× of the original extra storage)

REFERENCES

- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoon Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pp. 11102–11118. PMLR, 2022.
- Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022.
- Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.
- Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16709–16718, 2024.
- Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Peng Sun, Bei Shi, Xinyi Shang, and Tao Lin. Information compensation: A fix for any-scale dataset distillation. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop*, 2024a.

216 Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An
217 efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer
218 Vision and Pattern Recognition*, pp. 9390–9399, 2024b.

219
220 Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural net-
221 works. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

222 Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at
223 imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36,
224 2024.

225 Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo.
226 Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceed-
227 ings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

228
229 Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10
230 million image database for scene recognition. *IEEE transactions on pattern analysis and machine
231 intelligence*, 40(6):1452–1464, 2017.

232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269