Supplementary Materials for the Paper: Self-Supervision Can Be a Good Few-Shot Learner

Yuning Lu[1⋆], Liangjian Wen[2], Jianzhuang Liu[2],
Yajing Liu[1], and Xinmei Tian[1,3]

[1]University of Science and Technology of China [2]Huawei Noah's Ark Lab
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{lyn0, lyj123}@mail.ustc.edu.cn, xinmei@ustc.edu.cn,
{wenliangjian1, liu.jianzhuang}@huawei.com

A Implementation Details

A.1 Training Details

Self-supervised pre-training (UniSiam). We use the SGD optimizer with a weight decay of 10^{-4}, a momentum of 0.9, and a cosine decay schedule of learning rate. Note that our method does not require early stopping with the accuracy in the validation set (unlike many previous FSL works). The validation set is only used for model selection. The model of the last epoch is used for subsequent fine-tuning. For *tiered*-ImageNet, we follow SimSiam by setting the learning rate to 0.1 and the batch size to 512. For the smaller dataset *mini*-ImageNet, we use a larger learning rate of 0.3 with a smaller batch size of 256 to guarantee the convergence of pre-training. The numbers of epochs are 200 and 400 for *tiered*-ImageNet and *mini*-ImageNet, respectively. For our loss $\mathcal{L}_{AMINE}$ (Eq. 6), we set $\lambda = 0.1$. The temperature scalar τ is 2.0. All models are trained on 4 or 8 V100 GPUs.

Self-supervised knowledge distillation (UniSiam+dist). The optimization details and hyper-parameters of self-supervised knowledge distillation are the same as in the pre-training, except that we set $\lambda = 0.2$ for *tiered*-ImageNet.

A.2 Data Augmentation

The **default data augmentation** (in Section 4.2) follows the practice in existing works. It includes *RandomResizedCrop* with scale in $[0.2, 1.0]$, *RandomHorizontalFlip* with probability 0.5, *ColorJitter* [6] of {brightness, contrast, saturation, hue} with probability 0.8 and strength {0.4,0.4,0.4,0.1}, *grayscale* with probability 0.2, and *GaussianBlur* with probability 0.5 and the std of Gaussian kernel in $[0.1, 2.0]$. The **strong data augmentation** (in Section 4.2) adds *RandomVerticalFlip* with probability 0.5 and *RandAugment* [2] to the default data augmentation. The image size is 224×224 unless specified.

⋆ This work was done during an internship in Huawei Noah's Ark Lab.

In the paragraph about the effect of data augmentation (Section 4.2), the **simple data augmentation** is a common data augmentation strategy in supervised pre-training, which includes *RandomResizedCrop* with scale in $[0.2, 1.0]$, *RandomHorizontalFlip* with probability 0.5, and *ColorJitter* of {brightness, contrast, saturation} with strength {0.4,0.4,0.4}.

A.3 Linear Classifier

The logistic regression is the default linear classifier in our experiments. Similar to the implementation of [7], we transform features with the power transformation in all our experiments. The value of power is 0.5.

A.4 Compared Methods

The projection head of SimCLR is a 2-layer MLP following the original paper. The hidden dimensions of the projection head are the same as our model. Our variant method with the symmetric alignment (Table 1 in the manuscript) uses the same network architecture as SimCLR. For the unsupervised FSL methods (UMTRA and ProtoCLR), we use the same data augmentation strategy and backbone as ours.

A.5 Mutual Information Estimation

We compare the mutual information (MI) estimators I_{MINE} and I_{NCE} in the correlated Gaussian experiment [1]. The two random variables $\mathbf{x} \in \mathbb{R}^{16}$ and $\mathbf{y} \in \mathbb{R}^{16}$ come from a multivariate Gaussian distribution with component-wise correlation $corr(\mathbf{x}_i, \mathbf{y}_j) = \delta_{i,j}\rho$, where $\rho \in (-1, 1)$ and $\delta_{i,j}$ is Kronecker's delta. We consider the standardized Gaussian for marginal distributions $p(\mathbf{x})$ and $p(\mathbf{y})$ following [1]. We employ I_{MINE} and I_{NCE} to estimate the MI $I(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$.

B Additional Experiments

B.1 Cross-Domain Few-Shot Image Classification

The recent work [3] evaluates existing self-supervised learning methods on the benchmark of cross-domain few-shot learning (CDFSL) [4]. The goal of CDFSL is to evaluate the performance of FSL methods in real scenarios, where there are significant domain shifts between the unknown downstream tasks and the pre-training dataset. The BSCD-FSL benchmark [4] includes four different downstream datasets: CropDisease (crop disease images), EuroSAT (satellite images), ISIC (dermatology images), and ChestX (radiology images). We also evaluate our UniSiam model on these widely varying datasets, which is pre-trained on natural images.

	CropDiseases			EuroSAT		
	5-shot	20-shot	50-shot	5-shot	20-shot	50-shot
InsDis	88.01 ± 0.58	91.95 ± 0.44	92.70 ± 0.43	81.29 ± 0.63	86.52 ± 0.51	88.25 ± 0.47
MoCo-v1	87.87 ± 0.58	92.04 ± 0.43	92.87 ± 0.42	81.32 ± 0.61	86.55 ± 0.51	87.72 ± 0.46
PCL-v1	72.89 ± 0.69	80.74 ± 0.57	82.83 ± 0.55	66.56 ± 0.76	75.19 ± 0.67	76.41 ± 0.63
PIRL	86.22 ± 0.63	91.19 ± 0.49	92.18 ± 0.44	82.14 ± 0.63	87.06 ± 0.50	88.55 ± 0.44
PCL-v2	87.57 ± 0.60	92.58 ± 0.44	93.57 ± 0.40	81.10 ± 0.54	87.94 ± 0.40	89.23 ± 0.37
SimCLR-v1	90.29 ± 0.52	94.03 ± 0.37	94.49 ± 0.37	82.78 ± 0.56	89.38 ± 0.40	90.55 ± 0.36
MoCo-v2	87.62 ± 0.60	92.12 ± 0.46	93.61 ± 0.40	84.15 ± 0.52	88.92 ± 0.41	89.83 ± 0.37
SimCLR-v2	90.80 ± 0.52	94.92 ± 0.34	95.80 ± 0.29	86.45 ± 0.49	91.05 ± 0.36	92.07 ± 0.30
SeLa-v2	90.96 ± 0.54	94.75 ± 0.37	95.40 ± 0.33	84.56 ± 0.57	88.34 ± 0.57	88.51 ± 0.59
InfoMin	87.77 ± 0.61	92.34 ± 0.44	92.93 ± 0.40	81.68 ± 0.59	86.76 ± 0.47	87.61 ± 0.43
BYOL	92.71 ± 0.47	96.07 ± 0.33	96.69 ± 0.27	83.64 ± 0.54	89.62 ± 0.39	90.46 ± 0.35
DeepCluster-v2	**93.63 ± 0.44**	96.63 ± 0.29	97.04 ± 0.27	**88.39 ± 0.49**	92.02 ± 0.37	93.07 ± 0.31
SwAV	93.49 ± 0.46	96.15 ± 0.31	96.72 ± 0.28	87.29 ± 0.54	91.99 ± 0.36	93.36 ± 0.31
Supervised	89.37 ± 0.55	93.09 ± 0.43	94.32 ± 0.36	83.81 ± 0.55	88.36 ± 0.43	89.62 ± 0.37
UniSiam (Ours)	92.05 ± 0.50	**96.83 ± 0.27**	**98.14 ± 0.19**	86.53 ± 0.47	**93.24 ± 0.30**	**95.34 ± 0.23**

	ISIC			ChestX		
	5-shot	20-shot	50-shot	5-shot	20-shot	50-shot
InsDis	43.90 ± 0.55	52.19 ± 0.53	55.76 ± 0.50	25.67 ± 0.42	29.13 ± 0.44	31.77 ± 0.44
MoCo-v1	44.42 ± 0.55	53.79 ± 0.54	56.81 ± 0.52	25.92 ± 0.45	30.00 ± 0.43	32.74 ± 0.43
PCL-v1	33.21 ± 0.48	38.01 ± 0.44	39.77 ± 0.45	23.33 ± 0.40	25.54 ± 0.43	27.40 ± 0.42
PIRL	43.89 ± 0.54	53.24 ± 0.56	56.89 ± 0.52	25.60 ± 0.41	29.48 ± 0.45	31.44 ± 0.47
PCL-v2	37.47 ± 0.52	44.40 ± 0.52	46.82 ± 0.46	24.87 ± 0.42	28.28 ± 0.42	30.56 ± 0.43
SimCLR-v1	43.99 ± 0.55	53.00 ± 0.54	56.16 ± 0.53	26.36 ± 0.44	30.82 ± 0.43	33.16 ± 0.47
MoCo-v2	42.60 ± 0.55	52.39 ± 0.49	55.68 ± 0.53	25.26 ± 0.44	29.43 ± 0.45	32.20 ± 0.43
SimCLR-v2	43.66 ± 0.58	53.15 ± 0.53	56.83 ± 0.54	26.34 ± 0.44	30.90 ± 0.44	33.23 ± 0.47
SeLa-v2	39.97 ± 0.55	48.43 ± 0.54	51.31 ± 0.52	25.60 ± 0.44	30.43 ± 0.46	32.81 ± 0.44
InfoMin	39.03 ± 0.55	48.21 ± 0.54	51.58 ± 0.51	25.78 ± 0.44	29.48 ± 0.44	31.58 ± 0.44
BYOL	43.09 ± 0.56	53.76 ± 0.55	58.03 ± 0.52	26.39 ± 0.43	30.71 ± 0.47	34.17 ± 0.45
DeepCluster-v2	40.73 ± 0.59	49.91 ± 0.53	53.65 ± 0.54	26.51 ± 0.45	31.51 ± 0.45	34.17 ± 0.48
SwAV	39.66 ± 0.54	47.08 ± 0.50	51.10 ± 0.50	26.54 ± 0.48	30.91 ± 0.45	33.86 ± 0.46
Supervised	39.38 ± 0.58	48.79 ± 0.53	52.54 ± 0.56	25.22 ± 0.41	29.26 ± 0.44	32.34 ± 0.45
UniSiam (Ours)	**45.65 ± 0.58**	**56.54 ± 0.55**	**62.27 ± 0.54**	**28.18 ± 0.45**	**34.58 ± 0.46**	**39.48 ± 0.50**

Table A1: Average accuracy (%) of 5-way few-shot classification and 95% confidence interval on the BSCD-FSL dataset. The compared results are taken from [3].

We compare our results with those reported in [3]. All methods use the same backbone of ResNet-50. In contrast to the compared models in [3], which use the ImageNet [5] dataset for pre-training, our model is pre-trained on a small subset of ImageNet (i.e., the training classes of *mini*-ImageNet). As shown in Table A1, though pre-trained on a smaller dataset, our UniSiam overall outperforms the previous self-supervised methods and the supervised baseline by a large margin.

References

1. Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D.: Mutual information neural estimation. In: ICML (2018)
2. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: NeurIPS (2020)
3. Ericsson, L., Gouk, H., Hospedales, T.M.: How well do self-supervised models transfer? arXiv:2011.13377 (2020)
4. Guo, Y., Codella, N.C., Karlinsky, L., Codella, J.V., Smith, J.R., Saenko, K., Rosing, T., Feris, R.: A broader study of cross-domain few-shot learning. In: ECCV (2020)
5. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV (2015)
6. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018)
7. Yang, S., Liu, L., Xu, M.: Free lunch for few-shot learning: Distribution calibration. In: ICLR (2021)