X-Learner: Learning Cross Sources and Tasks for Universal Visual Representation(Supplementary Material)

Yinan He[1⋆], Gengshi Huang[2]1, Siyu Chen[3]1, Jianing Teng[4]1,
Kun Wang[4], Zhenfei Yin[4], Lu Sheng[5], Ziwei Liu[6], Yu Qiao[1], and Jing Shao[4]

[1]Shanghai AI Laboratory [2]Sun Yat-sen University [3] Carnegie Mellon University
[4]SenseTime Research [5]College of Software, Beihang University
[6]S-Lab, Nanyang Technological University
heyinan@pjlab.org.cn huanggsh3@mail2.sysu.edu.cn siyuche3@cs.cmu.edu
{tengjianing, wangkun, yinzhenfei, shaojing}@senseauto.com
lsheng@buaa.edu.cn ziwei.liu@ntu.edu.sg qiaoyu@pjlab.org.cn

1 HalfResNet-50 Used in X-Learner$_r$

Model	ResNet-50	HalfResNet-50
Channels	(256, 512, 1024, 2048)	(180, 360, 724, 1448)
Parameters	23,508,032	11,761,825

Table 1. The number of parameters and channel configuration of ResNet-50 and HalfResNet-50.

For implementing X-Learner$_r$, we use a HalfResNet-50 as sub-backbone with only $1/\sqrt{2}$ of the original ResNet-50 channels (see Tab. 1 for details).

2 Comparison with MuST

Table 2. Comparison with MuST. We use the same pre-training datasets as MuST without MiDaS [14]. Our setting can take advantage of most existing vision datasets. We replace the sub-backbone of X-Learner with ResNet-152. * represents that depth estimation (NYU-Depth V2) is an unseen task for X-Learner, but it is included in the training process of MuST.

Method	Backbone	Pre-training Settings	CIFAR-100 [8]	PASCAL Det [4]	PASCAL Seg [4]	NYU-Depth V2 [17]
MuST [6]	ResNet-152	ImageNet + OBJ365 + COCO + MiDaS	86.3	85.1	80.6	87.8
MuST [6]	ResNet-152	JFT300M + OBJ365 + COCO + MiDaS	88.3	87.9	**82.9**	89.5
X-Learner$_{R152}$	ResNet-152	ImageNet + OBJ365 + COCO	**88.7** (+2.4)	**88.5** (+3.4)	**81.4** (+0.8)	**91.2***(+3.4)

⋆ equal contribution

For fair comparison with MuST [6], a self-training approach using ResNet-152 as the backbone, we also conduct the experiment with the same backbone and dataset (except for MiDaS [14]). Table 2 provides the performance comparison on four different downstream tasks. We observe that our framework outperforms MuST by significant margins over all evaluated downstream tasks. When compared with MuST pre-trained with JFT-300M, X-Learner$_{\mathrm{R152}}$ shows the high data efficiency. Moreover, it is worth mentioning that our X-Learner surpasses MuST on NYU-Depth V2 without any depth estimation pre-training. However, MuST takes MiDaS, for pre-trainning which is large depth estimation dataset including 1.9M images. This zero-shot performance further demonstrates the strong generalization capability of X-Learner.

3 Dataset Details

In this section, we list 10 data sources used in the pre-training and 13 downstream datasets. All data are publicly available for non-commercial use.

3.1 Pre-Training Datasets

ImageNet [15] is a general classification dataset with 1.28M training data. Each image is labeled with one of the 1,000 classes.
Places365 [22] is a scene recognition dataset sampled from the Places database. We use the challenge version composed of 8 million training images comprising 365 scene classes.
iNat2021 [18] is a large-scale image dataset collected and annotated by community scientists and contains over 2.7M images from 10k different natural species.
CompCars [20] contains data from two scenarios, including images from web-nature and surveillance-nature. We only use the former part with 163 car makes with 1,716 car models. We include all 136,726 images capturing entire cars, and predict their car make labels.
Tsinghua Dogs [24] is a fine-grained classification dataset for dogs, over 65% of whose images are collected from real life. Each dog breed in the dataset contains at least 200 images and a maximum of 7,449 images.
COCO [10] has 118k training images labeled with 80 object detection categories.
Objects365 [16] is a large-scale object detection dataset with 609k training data and 365 classes.
WIDER FACE dataset [21] contains 32,203 images and 393,703 face labels with a high degree of variety in scale, pose and occlusion. For simplicity, we use the same mAP metric as COCO to report the pre-training performance on its validation set.
ADE20K [23] has 20k images with 150 non-background classes covering scene categories from the SUN and Places databases.
COCO-Stuff [2] is a dataset for scene understanding tasks like semantic segmentation and image captioning. It is constructed by annotating the original COCO images with additional stuff classes. There are 164k images spanning 172 categories including 80 things, 91 stuffs, and 1 unlabeled class.

3.2 Downstream Datasets

CIFAR-10 [8] is composed of 50k low-resolution training data labeled with 10 classes.
CIFAR-100 [8] is similar to CIFAR-10, but has 100 classes.
Food-101 [1] consists of 101 food categories with 750 training and 250 test images per category, making a total of 101k images. The labels for the test images have been manually cleaned.
Oxford-IIIT Pets [13] Dataset has 37 categories with roughly 200 images for each class. The images have large variations in scale, pose and lighting.
Oxford 102 Flower [12] is an image classification dataset consisting of 102 flower categories. The flowers are chosen to be commonly occurring in the United Kingdom. Each class consists of between 40 and 258 images.
SUN397 [19] contains 899 classes and 130,519 images. There are 397 well-sampled categories to evaluate algorithms for scene recognition.
Stanford Cars [7] dataset consists of 196 classes of cars with a total of 16,185 images, taken from the rear. The data is divided into almost a 50-50 train/test split with 8,144 training images and 8,041 testing images.
Describable Textures Dataset (DTD) [3] contains 5,640 texture images in the wild. They are annotated with human-centric attributes inspired by the perceptual properties of textures.
Caltech-101 [5] dataset is composed of images from 101 object categories and one additional background class. Most classes have about 50 images.
FGVC-Aircraft [11] contains 10,200 images of aircraft, with 100 images for each of the 102 different aircraft model variants, most of which are airplanes.
PASCAL Detection [4] refers to the PASCAL VOC object detection dataset with 20 object classes. We use the VOC07+12 set with 16.5k data for training, and evaluate on the PASCAL VOC 2007 test set.
PASCAL Segmentation [4] denotes the PASCAL VOC 2012 segmentation dataset with 1.5k training images and 20 classes.
NYU-Depth V2 [17] is a depth estimation dataset. It contains 120K RGB and depth pairs acquired as video sequences using a Microsoft Kinect from 464 indoor scenes. We follow the BTS [9] split, using 249 scenes (24,231 images) for training and 215 scenes (654 images) for testing.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: ECCV (2014) 3
2. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018) 2
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3606–3613 (2014) 3
4. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010) 1, 3
5. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: CVPR workshop. pp. 178–178. IEEE (2004) 3
6. Ghiasi, G., Zoph, B., Cubuk, E.D., Le, Q.V., Lin, T.Y.: Multi-task self-training for learning general representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8856–8865 (2021) 1, 2
7. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013) 3
8. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 1, 3
9. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019) 3
10. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 2
11. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013) 3
12. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR. vol. 2, pp. 1447–1454. IEEE (2006) 3
13. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: CVPR. pp. 3498–3505. IEEE (2012) 3
14. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2020) 1, 2
15. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 2
16. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8430–8439 (2019) 2
17. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012) 1, 3

18. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: CVPR. pp. 12884–12893 (2021) 2
19. Xiao, J., Ehinger, K.A., Hays, J., Torralba, A., Oliva, A.: Sun database: Exploring a large collection of scene categories. IJCV **119**(1), 3–22 (2016) 3
20. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3973–3981 (2015) 2
21. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: A face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5525–5533 (2016) 2
22. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) 2
23. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019) 2
24. Zou, D.N., Zhang, S.H., Mu, T.J., Zhang, M.: A new dataset of dog breed images and a benchmark for finegrained classification. Computational Visual Media **6**(4), 477–487 (2020) 2