MixMask: Revisiting Masking Strategy for Siamese ConvNets – *Supplementary Material*

Kirill Vishniakov[1]
kirill.vishniakov@mbzuai.ac.ae
Eric Xing[1, 2]
eric.xing@mbzuai.ac.ae
Zhiqiang Shen[1]
zhiqiang.shen@mbzuai.ac.ae

[1] Mohamed bin Zayed University of Artificial Intelligence
[2] Carnegie Mellon University

1 Base Models & Datasets

In this section, we provide a description of self-supervised learning frameworks and datasets that we used in the experiments. To test our method, we tried to select a diverse set of frameworks that incorporate different mechanisms to avoid model collapse and follow different design paradigms.

1.1 Base Models

MoCo V1&V2 [3, 5] is a self-supervised contrastive learning framework that employs a memory bank to store negative samples. MoCo V2 is an extension of the original MoCo, which introduces a projection head and stronger data augmentations.
Un-Mix [9] is an image mixture technique with state-of-the-art performance for unsupervised learning, which uses CutMix and Mixup at its core. It smooths decision boundary and reduces overconfidence in model predictions by introducing an additional mixture term to the original loss value, which is proportional to the degree of the mixture.
SimCLR [1] is a siamese framework with two branches that uses contrastive loss to attract positive and repel negative instances using various data augmentations.
BYOL [8] is a self-supervised learning technique that does not use negative pairs. It is composed of two networks, an online and a target. The task of an online network is to predict the representations produced by the target network. EMA from the online network is used to update the weights of the target.
SimSiam [2] The authors examined the effect of the different techniques which are commonly used to design siamese frameworks for representation learning. As a result, they proposed a simple framework with two branches that relies on the stop gradient operation on one branch and an extra prediction module on the other.

1.2 Datasets

CIFAR-100 [6] consists of 32×32 images with 100 classes. There are 50,000 train images and 10,000 test images, 500 and 100 per class, respectively.
Tiny-ImageNet [7] is a dataset containing 64 × 64 colored natural images with 200 classes. The test set is composed of 10,000 test images, whilst the train contains 500 images per category, totaling 100,000 images.
ImageNet-1K [4] has images with a size of 224×224. 1,281,167 images span the training set, with 1K different classes, and the validation set includes 50K images.

2 Training Configurations

In this section we provide hyperparameter settings for:

- Training on CIFAR-100 and Tiny-ImageNet in Table 1.
- Pretraining and linear probing on ImageNet-1K configurations are shown in Table 2.
- Configurations for semi-supervised and supervised fine-tuning on ImageNet-1K are given in Table 3.
- For object detection and segmentation we use the Detectron2[10] library and follow the 1× recipe on COCO and standard 24k training protocol on VOC07.

MoCo		SimCLR & BYOL		SimSiam	
hparam	value	hparam	value	hparam	value
backbone	resnet18	backbone	resnet18	backbone	resnet18
optimizer	SGD	optimizer	Adam	optimizer	SGD
lr	0.06	lr	0.003/0.002	lr	0.03
batch size	512	batch size	512	batch size	512
opt momentum	0.90	proj layers	2	opt momentum	0.90
epochs	1,000	epochs	1,000	epochs	1,000
weight decay	5e-4	weight decay	5e-4	weight decay	5e-4
embed-dim	128	embed-dim	64/128	embed-dim	128
moco-m	0.99	Adam l2	1e-6	warmup epochs	10
moco-k	4,096	proj dim	1,024	proj layers	2
unmix prob	0.50	unmix prob	0.50	unmix prob	0.50
moco-t	0.10	byol tau	0.99		

Table 1: Training settings on CIFAR-100 and Tiny-ImageNet. Slash separated values correspond to CIFAR-100 and Tiny-ImageNet, respectively.

3 Training Loss and Accuracy Curves

In Fig. 1, we present the training loss and *k*-NN accuracy curves for different base frameworks trained for 1,000 epochs on CIFAR-100 dataset. MixMask consistently outperforms baseline on all methods. MixMask has a higher (in case of SimSiam lower because it can attain the value of -1) training loss than baseline due to the presence of the additional asymmetric loss term.

Pretraining		Linear probing	
hparam	value	hparam	value
backbone	resnet50	backbone	resnet50
optimizer	SGD	optimizer	SGD
lr	0.03	lr	30
batch size	256	batch size	256
opt momentum	0.90	opt momentum	0.90
lr schedule	cosine	lr schedule	[60, 80]
epochs	200/800	epochs	100
weight decay	0	weight decay	0
moco-dim	128		
moco-m	0.999		
moco-k	65,536		
moco-t	0.2		
unmix probability	0.5		
mask type	block		
grid size	8		

Table 2: Hyperparameter values for pretraining and linear probing on ImageNet-1K. This configuration achieves the highest score. All experiments are conducted on 4 × NVIDIA A100 SXM4 40GB GPU.

hparam	value
backbone	resnet50
optimizer	SGD
lr stem	0.002/0.002/0.001
lr classifier	0.5/0.5/0.05
batch size	256
opt momentum	0.90
lr schedule	[12, 16]
epochs	20
weight decay	0

Table 3: Hyperparameter values for semi-supervised and supervised finetuning on ImageNet-1K. Slash separated values correspond to 1%, 10% and 100% percent data regimes, respectively.

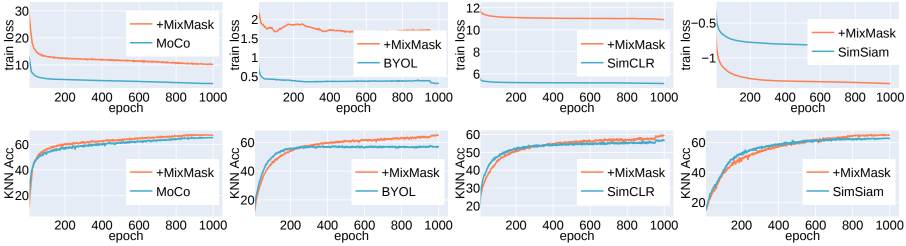

Figure 1: Training losses (top row) and *k*-NN evaluation accuracies (bottom row) on CIFAR-100 for experiments with 1,000 epochs for different self-supervised frameworks. MixMask (red) outperforms vanilla baseline (blue) on all frameworks by a significant margin.

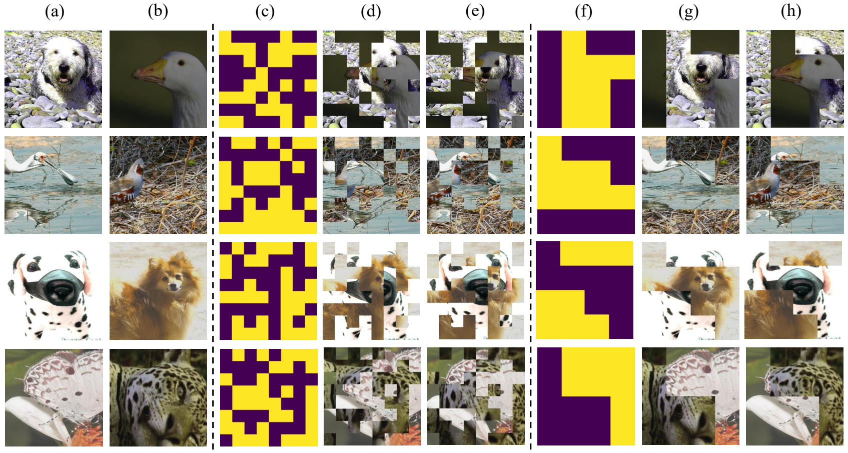

Figure 2: Illustration of the different mask patterns with a mask grid size of 8. (a) and (b) are input images. (c) is the discrete/random mask pattern, and (d) and (e) are mixed images using this mask. (f) is the blocked mask pattern, and (g) and (h) are mixed images with a blocked mask. Discrete masking breaks (c) – (e) the completeness of an object which is important for the contrastive loss because it operates on the global object level. On the other hand, blocked masking (f) – (h) preserves important global features leading to superior performance.

4 Illustrations of Different Mask Patterns

We provide additional illustrations for the different mask patterns and images generated by them. In Fig. 2 illustrations we use mask with grid size 8. All original images are sampled from ImageNet-1K.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[3] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[4] 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*, pages 248–255. Ieee, 2009.

[5] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[7] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[8] Pierre H Richemond, Jean-Bastien Grill, Florent Altché, Corentin Tallec, Florian Strub, Andrew Brock, Samuel Smith, Soham De, Razvan Pascanu, Bilal Piot, et al. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

[9] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[10] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.