

Supplementary Material for FRTAlign

1. Experimental Details

1.1. STL10 Experiments

STL10 consists of 96×96 color images from 10 classes, including airplane, bird, car, and dog. It provides 5,000 labeled training images, 8,000 test images, and 100,000 unlabeled images. We perform pretraining on 10,500 images by combining the labeled and unlabeled subsets. Each image is augmented twice using semantic and geometric transformations. Semantic transformations include random resized cropping (scale $[0.2, 1.0]$), horizontal flipping (probability 0.5), color jittering (brightness, contrast, saturation, hue; prob. 0.8), grayscale conversion (prob. 0.2), and Gaussian blur (kernel size 9). The image is then normalized (mean = $[0.43, 0.42, 0.39]$, std = $[0.27, 0.26, 0.27]$). In addition, each view is randomly rotated by $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, with the identity rotation (0°) sampled more frequently, controlled by the `nr_ratio`. The rotation label is recorded for supervised rotation prediction. This augmentation scheme encourages invariance to both appearance and geometric transformations. For evaluation, we follow the linear probing protocol: the pretrained backbone is frozen and a linear classifier is trained on 5,000 labeled training images and evaluated on 8,000 test images. All experiments were conducted on a single NVIDIA RTX 3090 GPU.

SimCLR We use a rotation predictor with a hidden dimension of 512. Pretraining runs for 400 epochs with a batch size of 512, learning rate 0.6, cosine decay, and the LARS optimizer [11]. For linear probing, we train a classifier for 100 epochs with learning rate 1.0.

MoCo v2 We similarly use a 512-dim rotation predictor. Pretraining uses 400 epochs, batch size 256, learning rate 0.03 (cosine decay), and SGD. MoCo-specific settings include a queue size of 16,384, momentum 0.999, and temperature 0.2. For linear probing, the classifier is trained for 60 epochs with step-wise learning rate decay (divided by 10 at epochs 20 and 40).

1.2. ImageNet100 Experiments

ImageNet100 is a subset of ImageNet with 100 classes, as introduced by Tian et al. [10], containing approximately 130,000 images (about 1,300 training and 50 validation samples per class). It serves as a computationally efficient yet semantically diverse benchmark for self-supervised learning. We resize all images to 224×224 and apply two types of augmentation: semantic transformations and rotation. Semantic augmentations include random resized cropping (scale $[0.2, 1.0]$), horizontal flip (probability 0.5), color jitter (brightness, contrast, saturation, hue; prob. 0.8), grayscale conversion (prob. 0.2), and Gaussian blur (kernel size 23). Images are normalized with dataset-specific statistics (mean = $[0.485, 0.456, 0.406]$, std = $[0.229, 0.224, 0.225]$). Each view is further rotated by $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, sampled uniformly. The rotation label is recorded for rotation prediction, encouraging invariance to both photometric and geometric changes. For evaluation, we perform linear probing on 126,689 training and 5,000 validation images. All experiments were conducted on a single NVIDIA H100 SXM GPU.

SimCLR We use a rotation predictor with a hidden dimension of 512. Pretraining is conducted for 400 epochs with batch size 256, learning rate 0.3, cosine decay schedule, and the LARS optimizer. For linear evaluation, a classifier is trained for 100 epochs with a learning rate of 1.0 on the frozen backbone.

1.3. EMNIST Experiments

EMNIST [3] is an extension of the original MNIST dataset, built from NIST’s Special Database 19. It includes both handwritten digits and letters, offering a more diverse and challenging benchmark. In our experiments, we use the EMNIST

Balanced split, which consists of 47 classes—10 digits and 37 upper- and lower-case letters—balanced across approximately 131,600 samples. To reflect low-resource deployment scenarios, we adopt a lightweight 5-layer CNN backbone (*LightCNN*, see Table 1), a 2-layer Rotation Predictor (hidden dim 128), and a Linear FRT head. All images are resized to 28×28 , and pretraining is conducted using only the training split. Each image undergoes a combination of semantic and geometric augmentations: random resized cropping (scale $[0.6, 1.0]$), random affine transforms (translation up to ± 2 pixels, scaling $[0.9, 1.1]$, and in-plane rotation within $\pm 15^\circ$), Gaussian noise ($\sigma = 0.1$), normalization (mean = $[0.5]$, std = $[0.5]$), and random erasing (probability 0.15, area ratio $[0.02, 0.15]$). Each view is then randomly rotated by $\{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$ with uniform probability, and the corresponding label is used for supervised rotation prediction. Linear evaluation is performed using 112,800 training and 18,800 validation images. Pretraining is conducted on a single NVIDIA RTX 3090 GPU.

SimCLR We train for 400 epochs with a batch size of 512, learning rate 0.6 (cosine decay), and the LARS optimizer. A linear classifier is then trained for 100 epochs with learning rate 1.0 on the frozen backbone.

Table 1. Architecture of the LightCNN encoder. All convolution layers use a 3×3 kernel with padding 1.

Layer	Details
Conv1	Conv2d(1, 32), BatchNorm2d, ReLU
Conv2	Conv2d(32, 32), BatchNorm2d, ReLU, MaxPool2d(2)
Conv3	Conv2d(32, 64), BatchNorm2d, ReLU
Conv4	Conv2d(64, 64), BatchNorm2d, ReLU, MaxPool2d(2)
Conv5	Conv2d(64, feat_dim), BatchNorm2d, ReLU
GAP	AdaptiveAvgPool2d(1)

2. Method Comparison

Table 2. Overview of recent self-supervised methods addressing rotation. Standard methods lack equivariant modeling, and prior approaches only partially incorporate rotation-aware components. Our method, **FRTAlign**, introduces a novel human-inspired alignment strategy that integrates rotation prediction, feature-level equivariance, and alignment into a unified framework.

Method	Joint-Embedding	Rotation Prediction	Feature-Level Equiv.	Alignment
<i>Standard SSL Methods</i>				
SimCLR [2]	✓			
MoCo v2 [8]	✓			
RotNet [7]		✓		
<i>Implicit Equivariant Learning</i>				
E-SSL [4]	✓	✓		
AugSelf [9]	✓	✓		
RefosNet [1]	✓	✓		
<i>Explicit Equivariant Learning</i>				
EquiMod [5]	✓		✓	
SIE [6]	✓		✓	
STL [12]	✓		✓	
<i>Ours</i>				
FRTAlign	✓	✓	✓	✓

References

- [1] Gairui Bai, Wei Xi, Xiaopeng Hong, Xinhui Liu, Yang Yue, and Songwen Zhao. Robust and rotation-equivariant contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

- [2] 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, 2020. [2](#)
- [3] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre van Schaik. Emnist: an extension of mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017. [1](#)
- [4] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljacic. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022. [2](#)
- [5] Alexandre Devillers and Mathieu Lefort. Equimod: An equivariance module to improve self-supervised learning. In *arXiv preprint arXiv:2211.01244*, 2022. [2](#)
- [6] Quentin Garrido, Laurent Najman, and Yann LeCun. Self-supervised learning of split invariant equivariant representations. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10975–10996, 2023. [2](#)
- [7] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. [2](#)
- [8] 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. [2](#)
- [9] Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. *Advances in Neural Information Processing Systems*, 34:17710–17722, 2021. [2](#)
- [10] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. [1](#)
- [11] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. [1](#)
- [12] Jaemyung Yu, Jaehyun Choi, Dong-Jae Lee, HyeongGwon Hong, and Junmo Kim. Self-supervised transformation learning for equivariant representations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [2](#)