

TRIBYOL: TRIPLET BYOL FOR SELF-SUPERVISED REPRESENTATION LEARNING

Guang Li[†] Ren Togo^{††} Takahiro Ogawa^{†††} Miki Haseyama^{†††}

[†] Graduate School of Information Science and Technology, Hokkaido University, Japan

^{††} Education and Research Center for Mathematical and Data Science, Hokkaido University, Japan

^{†††} Faculty of Information Science and Technology, Hokkaido University, Japan

E-mail: {guang, togo, ogawa}@lmd.ist.hokudai.ac.jp, miki@ist.hokudai.ac.jp

ABSTRACT

This paper proposes a novel self-supervised learning method for learning better representations with small batch sizes. Many self-supervised learning methods based on certain forms of the siamese network have emerged and received significant attention. However, these methods need to use large batch sizes to learn good representations and require heavy computational resources. We present a new triplet network combined with a triple-view loss to improve the performance of self-supervised representation learning with small batch sizes. Experimental results show that our method can drastically outperform state-of-the-art self-supervised learning methods on several datasets in small-batch cases. Our method provides a feasible solution for self-supervised learning with real-world high-resolution images that uses small batch sizes.

Index Terms— Self-supervised learning, representation learning, triplet network.

1. INTRODUCTION

Deep supervised learning has shown outstanding performance in many areas, especially on various computer vision tasks such as image classification, object detection and semantic segmentation [1]. However, supervised learning methods have over-reliance on manually designed labels and suffer from generalization problems, and hence are meeting their bottlenecks [2]. As an alternative, self-supervised learning is a learning framework that conforms to human cognition, which can learn information from the data itself without the need for manually designed labels [3]. Self-supervised learning has shown performance comparable to supervised learning methods on multiple tasks and has received widespread attention.

Early self-supervised learning methods often learn representations via a pretext task that applies a transformation to the input image and requires the learner to predict the properties of transformation (*e.g.*, rotation [4] and jigsaw [5]) from the transformed image [3]. Although such transformations are beneficial for predicting 3D correlations, it is undesirable for most semantic recognition tasks. PIRL [6] proposes to learn invariant representations rather than covariant ones. The method constructs image representations similar to the representations of transformed versions of the same image and different from the representations of other images. By combining the jigsaw or rotation pretext task with PIRL, the performance surpassed supervised learning in some computer vision tasks.

This work was partly supported by AMED Grant Number JP21zf0127004. This study was conducted on the Data Science Computing System of Education and Research Center for Mathematical and Data Science, Hokkaido University.

Recently, self-supervised learning methods based on the siamese network [7] achieved high representation learning performance on natural images. These methods define the inputs as two augmented views from one image, then input to the siamese network and maximize the similarity between the representations of views [8]. When the batch size decreases, these methods face accuracy degradation problems due to the reduced mutual information and transformation-invariant representation [9]. Therefore, these methods typically need to increase the number of samples in a batch and need heavy computational resources (*e.g.*, at least 4 GPUs or 32 TPU cores) for learning better representations from images [10]. For example, the state-of-the-art methods, SimCLR [11], SimSiam [12], and BYOL [13] all use batch sizes over 128. Requiring such a huge batch size is expensive and impractical in the real-world (*e.g.*, high-resolution medical images [14–16] and remote sensing images [17]). Hence, it is crucial to explore self-supervised learning in small-batch cases.

In this paper, we propose a novel self-supervised learning method called TriBYOL for learning better representations with small batch sizes. Considering the appropriate addition of augmented views can increase mutual information and encourage a more transformation-invariant representation in small-batch cases [18], we present a new triplet network combined with a triple-view loss based on the state-of-the-art method BYOL [13]. The conventional triplet network [19] is a variant of the siamese network and typically has three inputs, including anchor, positive and negative samples. Different from the conventional triplet network, our method only has three augmented views from one image. Experimental results show that our method can drastically outperform the state-of-the-art self-supervised learning methods on several datasets in small-batch cases.

Our main contributions can be summarized as follows.

- We propose a novel self-supervised learning method called TriBYOL for learning better representations with small batch sizes.
- We confirm that our method can drastically outperform state-of-the-art self-supervised learning methods on several datasets in small-batch cases.

2. METHODOLOGY

2.1. Description of TriBYOL

The overview of our method is shown in Fig. 1. Motivated by the fact that appropriate addition of augmented views can increase mutual information and encourage a more transformation-invariant representation in small-batch cases [18], different from BYOL [13] which uses the siamese network, we propose the triplet network combined

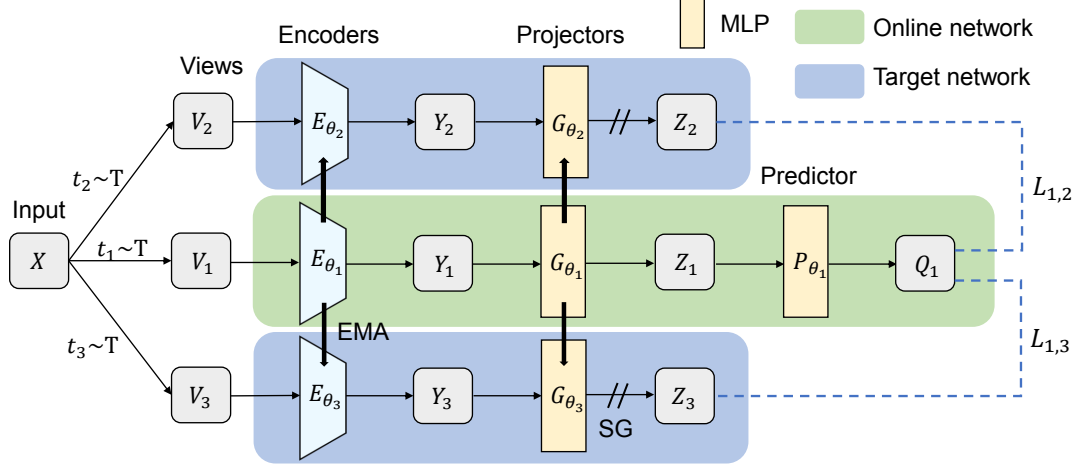


Fig. 1. Overview of the proposed method. Our method minimizes a triple-view loss between representations of three views from the triplet network. The weights of the target networks (θ_2 and θ_3) are an exponential moving average (EMA) of the weights of the online network (θ_1). MLP denotes multilayer perceptron. SG denotes stop-gradient.

with a triple-view loss for learning better representations with small batch sizes. Our method uses a triplet network to learn discriminative representations from images. The triplet network contains one online network and two target networks with the same structure, where the weights of the target networks are an exponential moving average (EMA) of the weights of the online network [20]. The conventional triplet network is a variant of the siamese network and typically has three inputs, including anchor, positive and negative samples. Different from the conventional triplet network, our method only has three augmented views from one image.

Encoder E_{θ_1} and predictor G_{θ_1} belong to the online network. Encoders E_{θ_2} and E_{θ_3} belong to the target networks. Given an input image X from a dataset D without label information, three transformations t_1 , t_2 , and t_3 are randomly sampled from a distribution T to generate three views $V_1 = t_1(X)$, $V_2 = t_2(X)$, and $V_3 = t_3(X)$. Similar to BYOL, our method exchanges views for learning better representations. Y_1 , Y_2 , and Y_3 are representations processed by the encoders. Z_1 , Z_2 , and Z_3 are representations processed by the projector. Q_1 is representations processed by the predictor. Note that we only use the predictor in the online network to make the triplet network asymmetric, which can prevent learning from collapsing, thereby improving the representation learning performance [13]. Finally, we define the loss L to compare the normalized representations from different views of the same image. The triple-view loss L comparing representations of V_1 , V_2 , and V_3 is defined as follows:

$$L_{i,j} = \|\hat{Q}_i - \hat{Z}_j\|_2^2, \quad (1)$$

$$= 2 - 2 \cdot \frac{\langle Q_i, Z_j \rangle}{\|Q_i\|_2 \cdot \|Z_j\|_2},$$

$$L = L_{1,2} + L_{1,3}, \quad (2)$$

where $\hat{Q}_i = Q_i / \|Q_i\|_2$ and $\hat{Z}_i = Z_i / \|Z_i\|_2$ denote the normalized representations of V_i ($i = 1, 2, 3$). The consistency among views from the triplet network helps to learn discriminative representations from images. The weights of the online network (θ_1) are updated as follows:

$$\theta_1 \leftarrow \text{Opt}(\theta_1, \nabla_{\theta_1} L, \alpha), \quad (3)$$

Algorithm 1 TriBYOL

Require: D : dataset; B : batch size; T : distribution; E : training epochs

Ensure: E_{θ_1} : the trained encoder

- 1: **for** each training epoch $t = 1$ to E **do**
- 2: Load a minibatch X from D with B samples
- 3: Sample three transformations from T as t_1 , t_2 , and t_3
- 4: Generate views with random augmentation:
 $V_1 = t_1(X)$, $V_2 = t_2(X)$, and $V_3 = t_3(X)$
- 5: Calculate the triple-view loss L with Eqs. (1) and (2)
- 6: Update the weights of the online network (θ_1) with Eq. (3)
- 7: Update the weights of the target networks (θ_2 and θ_3) with Eqs. (4) and (5)
- 8: **end for**

where Opt and α denote the optimizer and the learning rate, respectively. Since the stop-gradient (SG) is critical for preventing the collapse of self-supervised learning [12], the target networks (E_{θ_2} and E_{θ_3}) are not updated using backpropagation. The weights of target networks (θ_2 and θ_3) are an exponential moving average of the weights of θ_1 and are updated as follows:

$$\theta_2 \leftarrow \tau \theta_2 + (1 - \tau) \theta_1, \quad (4)$$

$$\theta_3 \leftarrow \tau \theta_3 + (1 - \tau) \theta_1, \quad (5)$$

where τ denotes a momentum coefficient, and we alternately update the weights θ_2 and θ_3 after every iteration. After the self-supervised learning, we can use the trained encoder (E_{θ_1}) for downstream tasks such as linear evaluation and fine-tuning. The self-supervised learning process of our method is summarized in Algorithm 1.

2.2. Self-supervised learning in small-batch cases

In this subsection, we propose a new problem setting and discuss its feasibility. Self-supervised learning methods based on the siamese network have achieved high representation learning performance on

Table 1. Linear evaluation results (%) with different batch sizes. “b32”, “b64”, and “b128” denote the batch size of 32, 64, and 128, respectively.

Method	CIFAR-10			CIFAR-100			STL-10		
	b32	b64	b128	b32	b64	b128	b32	b64	b128
TriBYOL	79.09	85.35	87.31	49.07	59.90	63.05	75.41	83.16	88.19
Cross [9]	76.01	82.06	83.50	48.04	54.65	58.79	69.66	78.38	83.79
BYOL [13]	68.67	81.47	83.79	41.21	49.68	58.34	49.60	80.09	84.88
SimSiam [12]	58.42	71.25	75.58	1.00	37.06	49.21	10.00	65.20	71.78
PIRL-rotation [6]	-	-	55.78	-	-	31.55	-	-	50.26
PIRL-jigsaw [6]	-	-	49.94	-	-	27.36	-	-	48.55
SimCLR [11]	-	-	52.58	-	-	21.26	-	-	44.50
ImageNet	82.37			60.72			91.76		

natural images. However, these methods typically need to increase the number of samples in a batch for learning better representations from images. Requiring such a huge batch size is expensive and impractical in the real world. For example, in practical applications, high-resolution medical images and remote sensing images are difficult to perform self-supervised learning with large batch sizes due to the limitation of GPU memory. To solve the above problem, we propose the triplet network combined with a triple-view loss based on the state-of-the-art method BYOL [13]. The appropriate addition of augmented views can increase mutual information and encourage a more transformation-invariant representation in small-batch cases [18]. Our method can provide a feasible solution for subsequent practical applications (*e.g.*, high-resolution medical images and remote sensing images) of self-supervised learning in small-batch cases.

2.3. Implementation details

Image Augmentations Three random views are generated by a combination of standard data augmentation methods, including cropping, resizing, flipping, color jittering, grayscaling, and Gaussian blurring based on BYOL [13] data augmentation pipeline. The following operations are performed sequentially to produce each view.

- Random cropping with an area uniformly sampled with a scale between 0.2 to 1.0, followed by resizing to a resolution of 96×96 pixels.
- Random horizontal flipping with an applying probability of 0.5.
- Color jittering of brightness, contrast, saturation, and hue with the strength of 0.8, 0.8, 0.8, and 0.2 with an applying probability of 0.8.
- Grayscale with an applying probability of 0.2.
- Gaussian blurring with kernel size 9 and std between 0.1 to 2.0.

Architecture The encoders (E_{θ_1} , E_{θ_2} , and E_{θ_3}) in our method are ResNet50 [21] backbone. The projectors (G_{θ_1} , G_{θ_2} , and G_{θ_3}) are multilayer perceptron (MLP) whose architecture contains a linear layer with an output size of 512, a batch normalization layer, a ReLU activation function, and a linear layer with an output size of 128. The predictor (G_{θ_1}) is also an MLP with the same architecture.

Optimization The optimizer (Opt) used in our method is an SGD optimizer, whose learning rate (α), momentum, and weight

decay are set to 0.03, 0.9, 0.0004, respectively. We search for learning rate, momentum, and weight decay based on 40-epoch results and then apply it for longer learning such as 80 epochs or 200 epochs. We use different small batch sizes such as 32, 64, and 128, which are friendly to typical single GPU implementations.

3. EXPERIMENTS

In this section, we conduct experiments to verify the effectiveness of TriBYOL. First, we show the experimental settings in section 3.1. Next, we show the linear evaluation results, fine-tuning results, and transfer learning results in sections 3.2, 3.3, and 3.4, respectively. All of our experiments were conducted by using the PyTorch framework with an NVIDIA Tesla P100 GPU that has 16G memory.

3.1. Experimental settings

In the experiments, we used 9 benchmark datasets to verify the effectiveness of TriBYOL including MNIST [22], FashionMNIST [23], KMNIST [24], USPS [25], SVHN [26], CIFAR-10 [27], CIFAR-100 [27], STL-10 [28], and Tiny ImageNet [29]. We used training images without label information for self-supervised learning on well-known MNIST, FashionMNIST, KMNIST, USPS, SVHN, CIFAR-10, and CIFAR100. The STL-10 dataset contains 100,000 unlabeled images and 13,000 labeled images from 10 classes, among which 5,000 images are for training while the remaining 8,000 images are for testing. For STL-10, we used 100,000 unlabeled images and 5,000 training images without label information for self-supervised learning. The training set of the Tiny ImageNet dataset contains 100,000 images of 200 classes (500 images for each class). We only used the training set of Tiny ImageNet without label information for self-supervised learning and transfer learning performance test in subsection 3.4.

To verify the effectiveness of TriBYOL, we use the following methods as comparative methods. Cross [9], BYOL [13], and SimSiam [12] are state-of-the-art self-supervised learning methods without negative sample pairs. PIRL-rotation [6], PIRL-jigsaw [6], and SimCLR [11] are state-of-the-art self-supervised learning methods with negative sample pairs. ImageNet and From Scratch denote pre-training on ImageNet [30] and training from scratch (*i.e.*, random initial weights), respectively. We conducted experiments of different downstream tasks containing linear evaluation, fine-tuning, and transfer learning.

Table 2. Fine-tuning results (%) with different numbers of labels. “1%”, “10%”, and “100%” denote the percentage of used labels.

Method	CIFAR-10			CIFAR-100			STL-10		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
TriBYOL	56.60	71.73	87.07	9.50	23.57	58.92	56.66	67.72	97.34
Cross [9]	50.88	67.34	86.03	6.81	20.96	57.23	42.80	59.22	93.28
BYOL [13]	56.28	70.94	86.87	9.38	22.51	58.17	53.96	65.98	97.26
SimSiam [12]	43.16	63.08	84.76	4.86	14.76	54.70	40.38	49.96	88.38
From Scratch	32.29	57.24	83.87	5.95	17.47	56.70	20.38	39.10	83.96
ImageNet	69.99	84.27	91.29	27.48	52.41	70.80	81.24	86.34	98.00

Table 3. Transfer learning results (%) on different datasets.

	MNIST	FashionMNIST	KMNIST	USPS	SVHN	CIFAR-10	CIFAR-100	STL-10
TriBYOL	98.74	91.76	92.00	96.61	75.23	80.09	55.88	79.11
Cross [9]	98.54	91.28	90.33	96.21	71.29	77.55	51.53	76.26
BYOL [13]	98.41	90.77	89.88	96.06	68.75	75.31	48.51	74.04
SimSiam [12]	97.58	89.25	83.31	94.02	58.70	64.51	35.63	63.44
ImageNet	98.58	90.85	90.77	96.56	77.34	82.37	60.72	91.76

3.2. Linear evaluation results with different small batch sizes

In this subsection, we evaluate the performance of TriBYOL by performing linear evaluation on CIFAR-10, CIFAR-100, and STL-10 with different small batch sizes (32, 64, and 128). The training epochs of self-supervised learning are set to 80. Following the linear evaluation protocol from [11], we trained a linear classifier on top of the frozen ResNet50 backbone pretrained with TriBYOL and other comparative methods for 200 epochs and test for every 10 epochs. Since self-supervised learning methods with negative sample pairs performed poorly even with a batch size of 128, we did not conduct experiments with a batch size of 32 or 64.

Table 1 shows the linear evaluation results with different small batch sizes. “b32”, “b64”, and “b128” denote the batch size of 32, 64, and 128, respectively. From Table 1, we can see that TriBYOL drastically outperforms all comparative methods in small-batch cases. Especially, the performance of TriBYOL even exceeds that of the pretrained model on ImageNet on CIFAR-10 and CIFAR-100 with small batch sizes. When the batch size is 32, the training of SimSiam [12] collapses (*i.e.*, output was constant), and the test results are always a certain category, resulting in 1.00% and 10.00% on CIFAR-100 and STL-10, respectively. Linear evaluation results on CIFAR-10, CIFAR-100, and STL-10 show that TriBYOL can learn better representations with different small batch sizes than other state-of-the-art self-supervised learning methods.

3.3. Fine-tuning results with different numbers of labels

In this subsection, we evaluate the performance of TriBYOL by fine-tuning a linear classifier with different numbers of labels on CIFAR-10, CIFAR-100, and STL-10. The number of training epochs of self-supervised learning was set to 80. We used the ResNet50 backbone that was trained with a batch size of 128. We fine-tuned the linear classifier for 10 epochs and tested it for every epoch.

Table 2 shows the fine-tuning results with different numbers of labels. “1%”, “10%”, and “100%” denote the percentage of used labels. We can see that TriBYOL outperforms all comparative methods with different numbers of labels. Compared with training from scratch, TriBYOL shows better representation learning ability. Fine-tuning results on CIFAR-10, CIFAR-100, and STL-10 show that

TriBYOL can learn better representations with different numbers of labels than other state-of-the-art self-supervised learning methods.

3.4. Transfer learning results with different datasets

In this subsection, we evaluate the performance of TriBYOL by training a linear classifier on top of the frozen representations learned from Tiny ImageNet on a variety of datasets containing MNIST, FashionMNIST, KMNIST, USPS, SVHN, CIFAR-10, CIFAR-100, and STL-10. The number of training epochs of self-supervised learning was set to 80. We use the ResNet50 backbone that was trained with a batch size of 128. We trained the linear classifier for 200 epochs and tested it for every 10 epochs.

Table 3 shows the transfer learning results on different datasets. As shown in Table 3, TriBYOL outperforms all comparative methods on different datasets. Furthermore, when transferring to MNIST, FashionMNIST, KMNIST, and USPS, the performance of TriBYOL even exceeds that of the pretrained model on ImageNet. Transfer learning results on different datasets show that TriBYOL can learn more generalizable representations than other state-of-the-art self-supervised learning methods.

4. CONCLUSION

This paper has proposed a novel self-supervised learning method called TriBYOL. The proposed method presents a new triplet network combined with a triple-view loss based on the state-of-the-art method BYOL for better representation learning performance with small batch sizes. We conduct extensive experiments of different downstream tasks containing linear evaluation, fine-tuning, and transfer learning to verify the effectiveness of TriBYOL. Experimental results show that our method can drastically outperform state-of-the-art self-supervised learning methods on several datasets in small-batch cases. Although we only take experiments in small-batch cases due to the limitations of computing resources, we believe that our method can also have good performance even with large batch sizes. Evaluation of TriBYOL on other high-resolution image datasets in the real world will be one of our future works.

5. REFERENCES

- [1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.
- [2] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Zhaoyu Wang, Li Mian, Jing Zhang, and Jie Tang, “Self-supervised learning: Generative or contrastive,” *arXiv preprint arXiv:2006.08218*, 2020.
- [3] Longlong Jing and Yingli Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [4] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, “Unsupervised representation learning by predicting image rotations,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.
- [5] Mehdi Noroozi and Paolo Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 69–84.
- [6] Ishan Misra and Laurens van der Maaten, “Self-supervised learning of pretext-invariant representations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6707–6717.
- [7] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, “Signature verification using a “siamese” time delay neural network,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 1993.
- [8] Yuandong Tian, Lantao Yu, Xinlei Chen, and Surya Ganguli, “Understanding self-supervised learning with dual deep networks,” *arXiv preprint arXiv:2010.00578*, 2020.
- [9] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “Cross-view self-supervised learning via momentum statistics in batch normalization,” in *Proceedings of the IEEE International Conference on Consumer Electronics – Taiwan (ICCE-TW)*, 2021.
- [10] Yuandong Tian, Xinlei Chen, and Surya Ganguli, “Understanding self-supervised learning dynamics without contrastive pairs,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.
- [11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.
- [12] Xinlei Chen and Kaiming He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., “Bootstrap your own latent—a new approach to self-supervised learning,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 21271–21284.
- [14] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “Self-supervised learning for gastritis detection with gastric x-ray images,” *arXiv preprint arXiv:2104.02864*, 2021.
- [15] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “Triplet self-supervised learning for gastritis detection with scarce annotations,” in *Proceedings of the IEEE Global Conference on Consumer Electronics (GCCE)*, 2021.
- [16] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama, “Self-knowledge distillation based self-supervised learning for covid-19 detection from chest x-ray images,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [17] Chao Tao, Ji Qi, Weipeng Lu, Hao Wang, and Haifeng Li, “Remote sensing image scene classification with self-supervised paradigm under limited labeled samples,” *IEEE Geoscience and Remote Sensing Letters*, 2020.
- [18] Yonglong Tian, Dilip Krishnan, and Phillip Isola, “Contrastive multiview coding,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
- [19] Elad Hoffer and Nir Ailon, “Deep metric learning using triplet network,” in *Proceedings of the International Workshop on Similarity-based Pattern Recognition*, 2015, pp. 84–92.
- [20] Antti Tarvainen and Harri Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 1195–1204.
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [22] Yann LeCun, Corinna Cortes, and CJ Burges, “Mnist handwritten digit database,” 2010.
- [23] Han Xiao, Kashif Rasul, and Roland Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” *arXiv preprint arXiv:1708.07747*, 2017.
- [24] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha, “Deep learning for classical japanese literature,” *arXiv preprint arXiv:1812.01718*, 2018.
- [25] Jonathan J. Hull, “A database for handwritten text recognition research,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550–554, 1994.
- [26] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, “Reading digits in natural images with unsupervised feature learning,” in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2011.
- [27] Alex Krizhevsky, Geoffrey Hinton, et al., “Learning multiple layers of features from tiny images,” 2009.
- [28] Adam Coates, Andrew Ng, and Honglak Lee, “An analysis of single-layer networks in unsupervised feature learning,” in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 215–223.
- [29] Ya Le and Xuan Yang, “Tiny imagenet visual recognition challenge,” 2015.
- [30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.