Un-mix: Rethinking Image Mixtures for Unsupervised Visual Representation Learning

Zhiqiang Shen[1,4], **Zechun Liu**[2], **Zhuang Liu**[3], **Marios Savvides**[1], **Trevor Darrell**[3] **and Eric Xing**[1,4]

[1] Carnegie Mellon University [2] Reality Labs, Meta Inc. [3] University of California, Berkeley
[4] Mohamed bin Zayed University of Artificial Intelligence
{zhiqians,zechunl,marioss}@andrew.cmu.edu, zhuangl@berkeley.edu, trevor@eecs.berkeley.edu, epxing@cs.cmu.edu

Abstract

The recently advanced unsupervised learning approaches use the siamese-like framework to compare two "views" from the same image for learning representations. Making the two views distinctive is a core to guarantee that unsupervised methods can learn meaningful information. However, such frameworks are sometimes fragile on overfitting if the augmentations used for generating two views are not strong enough, causing the over-confident issue on the training data. This drawback hinders the model from learning subtle variance and fine-grained information. To address this, in this work we aim to involve the *soft distance concept* on label space in the contrastive-based unsupervised learning task and let the model be aware of the soft degree of similarity between positive or negative pairs through mixing the input data space, to further work collaboratively for the input and loss spaces. Despite its conceptual simplicity, we show empirically that with the solution – **Un**supervised image **mix**tures (Un-Mix), we can learn subtler, more robust and generalized representations from the transformed input and corresponding new label space. Extensive experiments are conducted on CIFAR-10, CIFAR-100, STL-10, Tiny ImageNet and standard ImageNet-1K with popular unsupervised methods SimCLR, BYOL, MoCo V1&V2, SwAV, *etc*. Our proposed image mixture and label assignment strategy can obtain consistent improvement by 1$\sim$3% following exactly the same hyperparameters and training procedures of the base methods. Code is publicly available at https://github.com/szq0214/Un-Mix.

1. Introduction

Unsupervised visual representation learning has attracted increasing attention [Noroozi and Favaro 2016, Zhang, Isola, and Efros 2016, Oord, Li, and Vinyals 2018, Hjelm et al. 2018, Gidaris, Singh, and Komodakis 2018, He et al. 2019, Chen et al. 2020a, Kim et al. 2020, Grill et al. 2020, Caron et al. 2020, Kalantidis et al. 2020] due to its enormous potential of being free from human-annotated supervision, *i.e.*, its extraordinary capability of leveraging the boundless unlabeled data. Previous studies in this field address this problem mainly in two directions: one is realized via a heuristic *pretext* task design that applies a transformation to the input image, such as colorization [Zhang, Isola, and Efros 2016], rotation [Gidaris, Singh, and Komodakis 2018], jigsaw [Noroozi and Favaro 2016], *etc.*, and the corresponding labels are derived from the properties of the transformation on the unlabeled data. Another direction is contrastive learning based approaches [He et al. 2019, Chen et al. 2020a] in the latent feature space, such as maximizing mutual information between different views [Bachman, Hjelm, and Buchwalter 2019, Tian, Krishnan, and Isola 2019], momentum contrast learning [He et al. 2019, Chen et al. 2020b] with instance discrimination task [Wu et al. 2018, Ye et al. 2019], larger batch sizes and nonlinear transformation [Chen et al. 2020a], symmetrized distance loss without negative pairs [Grill et al. 2020], contrasting cluster assignment [Caron et al. 2020]. SimSiam [Chen and He 2020] further found stop-gradient is critical to prevent from collapsing. These methods have shown great promise on this task, achieving state-of-the-art accuracy. However, these methods focus more on designing the training frameworks and loss formulations, ignoring crucial correlations between the input and loss spaces to enable fine-grained degrees of soft similarities between positive or negative pairs in the siamese-like unsupervised frameworks.

The motivation of our work stems from some simple observations of *label smoothing* in supervised learning [Szegedy et al. 2016]. Interestingly, it can be observed from visualizations of previous literature [Müller, Kornblith, and Hinton 2019, Shen et al. 2021] that *label smoothing* tends to force the output prediction of networks being less confident (*i.e.*, lower maximum probability of predictions), but the model representation and overall accuracy still increase significantly. The explanation for this seemingly contradictory phenomenon is that with *label smoothing*, the learner is encouraged to treat each incorrect instance/class as equally probable. Thus, more patterns are enforced to be explored in latent representations, enabling less variation across predicted instances and/or across semantically similar samples. This further prevents the network from overfitting on the training data. Otherwise, the network will be biased to produce over-confident predictions when evaluated on slightly different test samples. Considering that contrastive learning with InfoNCE loss is essentially classifying positive congruent and negative incongruent pairs with cross-entropy loss, such an observation reveals that a typical contrastive-based method can also encounter the over-confidence issue as in supervised learning.

Perspective of input and label spaces on un-/self-supervised learning. Contrastive learning methods adopt

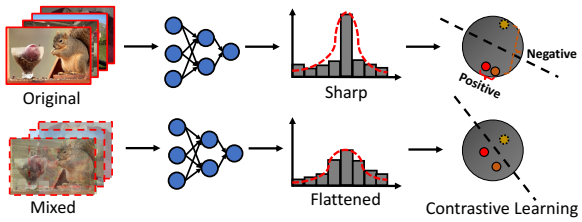

Figure 1: Illustration of the motivation in this work. We take the contrastive-based unsupervised learning approaches as an example. Contrastive learning measures the similarity of sample pairs in the latent representation space. With *flattened prediction/label*, the model is encouraged to treat semantically similar/dissimilar instances as equally probable, which will smooth decision boundaries and prevent the learner from becoming over-confident.

instance classification pretext, the features from different transformations (data augmentation) of the same images are compared directly to each other. The label of each image pair is binary (positive or negative) or continuous distance metrics. Augmentation is used as a transformation to make the distance of the same image to be larger. Different from data augmentation that enlarges the dissimilar distance but the label for calculating loss is still unchanged, our proposed mixtures will manipulate the semantic distance between two images, while adjusting the label for unsupervised loss accordingly. In other words, *data augmentation only changes the distance of input space*, *i.e.*, heavier data augmentation makes two images look more different, *but remains unchanged in label space* in training. However, mixture will modulate both input and label spaces simultaneously and the degree of change is controllable, which can further help capture the fine-grained representations from the unlabeled images and force models to learn more precise and smoother decision boundaries on the latent features. As a result, neural networks trained with new spaces learn flattering class-agnostic representations, that is, with fewer directions of variance on semantically similar classes. The mechanism of image mixtures in unsupervised learning is generally different from the data augmentation. Whereas, from the perspective of enlarging the training data space, mixtures can be considered as a broader concept of augmentation scheme in unsupervised learning.

We verify our method on five recently proposed unsupervised learning methods: SimCLR [Chen et al. 2020a], MoCo V1&V2 [He et al. 2019, Chen et al. 2020b], BYOL [Grill et al. 2020], SwAV [Caron et al. 2020] and Whitening [Ermolov et al. 2020b] as our baseline approaches. We conduct extensive experiments on CIFAR-10, CIFAR-100, STL-10, Tiny ImageNet, ImageNet-1K classification, as well as downstream object detection task on PASCAL VOC and COCO to demonstrate the effectiveness of our proposed approach. We observe that our mixture learned representations are extraordinarily effective for the downstream detection task which empirically proves that our method can improve the model's generalizability. For instance, our 200-epoch trained model outperforms the baseline MoCo V2 by 0.6% (AP_{50}), and is even better than the MoCo V2 800-epoch model.

Our contributions are summarized as follows:

- We provide empirical analysis to reveal that *mixing input images* and *smoothing labels* could improve performance favorably for a variety of unsupervised learning methods. We applied two simple image mixture methods based on previous literature [Zhang et al. 2018, Yun et al. 2019] to encourage neural networks to predict less confidently.
- We show that input and label spaces matter. We provide empirical evidence on how flattening happens under ideal conditions of latent space, validate it empirically on practical situations of contrastive learning, connect it to previous works on analyzing the discipline inside the unsupervised learning behavior. We explain the difficulties raised with original image space when visualizing distributions of predictions. Thus, we conclude that *good input and label spaces* are crucial for unsupervised optimization.
- Our proposed method is simple, flexible and universal. It can be utilized in nearly all mainstream unsupervised representation learning methods and only requires *a few lines of PyTorch codes* to incorporate in an existing framework. We demonstrate with a variety of base approaches and datasets, including SimCLR, BYOL, MoCo V1&V2, SwAV, *etc.*, on CIFAR-10, CIFAR-100, STL-10, Tiny ImageNet and ImageNet-1K. Our method obtains consistent accuracy improvement by 1$\sim$3% across them.

2. Related Work

(i) Un/Self-supervised Visual Feature Learning. Unsupervised learning aims to exploit the internal distributions of data and learn a representation without human-annotated labels. To achieve this purpose, early works mainly focused on reconstructing images from a latent representation, such as autoencoders [Vincent et al. 2008, 2010, Masci et al. 2011], sparse coding [Olshausen and Field 1996], adversarial learning [Goodfellow et al. 2014, Donahue, Krähenbühl, and Darrell 2016, Donahue and Simonyan 2019]. After that, more and more studies tried to design handcrafted pretext tasks such as image colorization [Zhang, Isola, and Efros 2016, 2017], solving jigsaw puzzles [Noroozi and Favaro 2016], counting visual primitives [Noroozi, Pirsiavash, and Favaro 2017], rotation prediction [Gidaris, Singh, and Komodakis 2018]. Recently, contrastive-based visual representation learning [Hadsell, Chopra, and LeCun 2006] has attracted much attention and achieved promising results. For example, Oord et al. [Oord, Li, and Vinyals 2018] proposed to use autoregressive models to predict the future samples in latent space with probabilistic contrastive loss. Hjelm et al. [Hjelm et al. 2018] proposed to maximize mutual information from the encoder between inputs and outputs of a deep network. Bachman et al. [Bachman, Hjelm, and Buchwalter 2019] further extended this idea to multiple views of a shared context. Moreover, He et al. [He et al. 2019] proposed to adopt momentum contrast to update the models and Misra&Maaten [Misra and van der Maaten 2019] developed the pretext-invariant representation learning strategy that learns invariant representations from the pre-designed pretext tasks. The clustering-based methods [Caron et al. 2018, 2020] are also a family for the unsupervised visual feature learning. **(ii) Smoothing Label/Prediction in Super-**

vised Learning. Explicit label smoothing has been adopted successfully to improve the performance of deep neural models across a wide range of tasks, including image classification [Szegedy et al. 2016], object detection [Krothapalli and Abbott 2020], machine translation [Vaswani et al. 2017], and speech recognition [Chorowski and Jaitly 2016]. Moreover, motivated by mixup, Verma et al. [Verma et al. 2019] proposed to implicitly interpolate hidden states as a regularizer that encourages neural networks to predict less confidently (softer prediction) on interpolations of hidden representations. They found that neural networks trained with this kind of operation can learn flatter class representations that possess better generalization, as well as better robustness to novel deformations and even adversarial examples in testing data. Some recent work [Müller, Kornblith, and Hinton 2019, Shen et al. 2021] further demonstrated that label smoothing implicitly calibrates the prediction of learned networks, so that the confidence of their outputs is more aligned with the true labels of the trained dataset. However, all of these studies lie in supervised learning. **(iii) Differences to *i*-Mix** [Lee et al. 2021] and **MixCo** [Kim et al. 2020]. These two concurrent works also employ the idea of image mixtures on unsupervised learning but the similarity to our Un-Mix is more in the spirit than the concrete solution. We achieve the mixture operation by using a self-mixture strategy within a *mini*-batch of samples during training, which is simpler and more manageable for incorporating the proposed method into the existing unsupervised frameworks for mixture purpose.

3. Our Approach

In this section, we begin by presenting different paradigms using mixtures in the unsupervised learning framework, including mixing both two branches and a single one. Then, we discuss image mixture strategies and the circumstances that contain a memory bank or not. Lastly, we elaborate the loss functions for our approach and provide the analysis for explaining the information gain of our proposed method.

Conventional siamese-like framework for unsupervised learning. Given an image I, we first augment it to two transformed views I_A and $\hat{I}_A$ by applying a pre-defined random transformation. Then, we feed into a two-branch framework with a projection head to produce latent representations. Finally, we define metric loss, such as InfoNCE, distance losses for optimization, as shown in Fig. 2 (1).

3.1. Paradigms of Mixtures

The proposed mixtures follow the image transformations of input samples. We define I_A^M and $\hat{I}_A^M$ as the mixed images which can be $\{I_{\mathbf{g\_m}}, I_{\mathbf{r\_m}}\}$ (global and region-level image mixtures, respectively) according to the type of mixture operation we choose in the current training iteration. The mixed images are forwarded through the target network f_θ, then a non-linear projection head p_θ is adopted to obtain the representations of the input sample for the unsupervised distance loss. Image mixture with relabeling can provide additional subtle information to force two branches unequally distant, instead of solely learning *positive* or *negative* pairs for the representations. In the following, we discuss two circumstances in such a framework, as shown in Fig. 2 (2) (3).

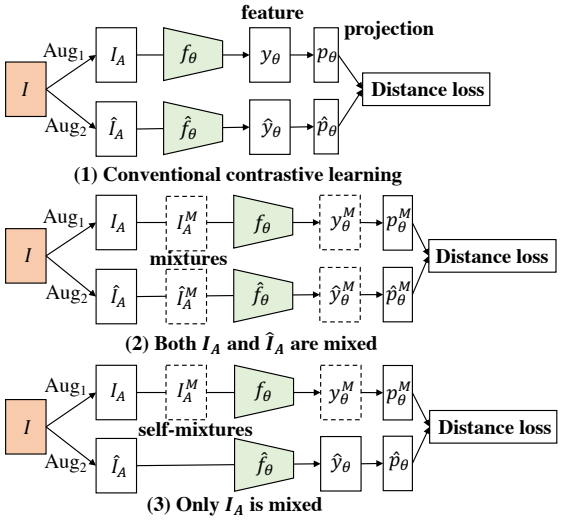

Figure 2: Comparison of different paradigms of utilizing mixtures in unsupervised learning. (1) is the conventional instance classification based framework. (2) and (3) are the strategies of applying the proposed image mixtures. "self-mixtures" denotes that the images of mixture operations only happen in current batch samples. The dashed bounding box represents the mixed image and its representation.

Both I_A and $\hat{I}_A$ are mixed (Fig. 2 (2)). This solution is to mix both the two views of an input image. Thus, the similarity between mixtures will remain unchanged. While this mixture strategy on two branches will suffer from undesirable equilibria as the mixture ratio of images is not used on loss space, namely, the additional information of mixture degree is not fully utilized. We found this strategy is effective on relatively small-scale datasets like CIFAR, STL-10, *etc.*, but is barely helpful on the large-scale ImageNet-1K.

Only I_A is mixed (Fig. 2 (3)). This is the main strategy that we use in this work. Compared to the one above, this solution is more efficient since it only needs one additional forward pass. Also, reverse order outputs can be obtained by permutation from normal order outputs. From our experimental results, it is also more effective for obtaining accuracy gain.

3.2. Image Mixture Strategies

We introduce two widely-used mixture methods in supervised learning: **(i) Mixup** [Zhang et al. 2018] and **(ii) Cutmix** [Yun et al. 2019]. Since they are designed for supervised learning with available ground-truth for calculating mixed labels, in this work, we focus on exploring the way to sample training data in a *mini*-batch and assign new softened distance loss formulations in the unsupervised learning frameworks.

Mixup can be written as:

$$I_{g\_m} \leftarrow \alpha I_1 + (1-\alpha) I_2 \tag{1}$$

where $\{I_1, I_2\}$ denote the images that we want to mix. I_{g_m} is the output mixture, $\alpha \in [0, 1]$ is the mixture coefficient.

Cutmix replaces within particular locations of a region:

$$I_{r\_m} \leftarrow \mathbf{M}_b \odot I_1 + (\mathbf{1} - \mathbf{M}_b) \odot I_2 \tag{2}$$

where $\mathbf{M}_b \in \{0, 1\}^I$ denotes a binary mask as defined in [Yun et al. 2019]. $\mathbf{1}$ is a binary mask with all values equaling one. $\odot$ denotes element-wise multiplication.

Both Mixup and Cutmix can be regarded as the regularization techniques to prevent the models from overfitting and make the predictions less confident.

Dealing with Memory Banks (MB). In this part, we describe different scenarios regarding how to design the framework using the proposed mixture training strategy if the base model contains a memory bank or not. Our goal is to enhance visual feature representations by leveraging additional mixture information and the different mixing ratios between two images in the unsupervised scheme. To this end, we propose a way to re-measure the distance of one pair of samples for the MB-based or non-MB-based unsupervised frameworks.

(i) Without a memory bank. Under this circumstance, the unsupervised frameworks will use positive pairs only for training (*e.g.*, BYOL [Grill et al. 2020]) or contrastive-based pipelines (*e.g.*, SimCLR [Chen et al. 2020a]). Therefore, we only need to design the new distance of the positive pairs, as shown in Fig. 4. In our proposed self-mixture strategy, the new distance scale $\mathcal{D}_{\text{dis}}$ of a positive pair will be:

$$I_A^M = \lambda I_1 + (1-\lambda) I_2,$$
$$\mathcal{D}_{\text{dis}}\left(I_A^M, \hat{I}_A\right) = \begin{cases} \lambda & \text{if } \hat{I}_A = \hat{I}_1, \\ 1-\lambda & \text{if } \hat{I}_A = \hat{I}_2. \end{cases} \quad (3)$$

where $\hat{I}_1, \hat{I}_2$ are another views of I_1, I_2 from the same images. In traditional unsupervised scheme, they are a positive pair. λ is the mix ratio controlled by the degree of mixture we use in the current iteration of training, when employing global mixture, $\lambda = \alpha$ as in Eq. 1, otherwise, $\lambda = \frac{\mathbf{M}_b}{\mathbf{1}}$ as in Eq. 2.

(ii) With a memory bank. Using a memory bank with mixtures will solely affect the constitution of negative pairs as the distance/label between them is always "zero" in instance classification based contrastive learning. We keep the distance of negative pairs as original values, whatever they are the combination of one original and one mixed images. In particular, negative pairs (samples) can be {*original, original*}, {*original, mixed*}, {*mixed, mixed*} images. We found in experiments that maintaining one MB with the representations from original/unmixed images is enough to obtain good performance, we explain this through the enlarged training data space. However, this is inapplicable in the multi-scale training scheme, as we will discuss later in our Appendix.

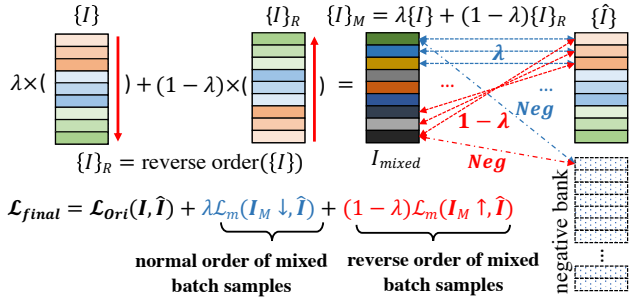

Figure 3: Illustration of self-mixture within a *mini*-batch. In each iteration, we randomly choose one mixture operation for all the current samples with a pre-defined probability P, thus the formulation of λ depends on the chosen mixture type.

3.3. Loss Functions

Self-Mixtures within Per *Mini*-Batch of Training. In our method, the mixing ratio of two images is the core gifted extra

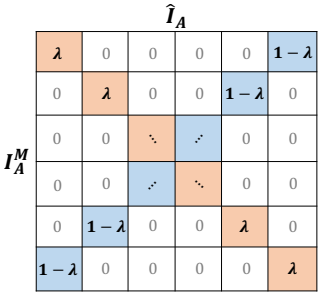

Figure 4: The distance matrix of proposed mixture strategy between the mixed I_A (*i.e.*, I_A^M) and $\hat{I}_A$ for calculating the softened distance loss. Here we take six images in the *mini*-batch as an example.

information that can be utilized in the unsupervised methods. Also, properly proposing a strategy to reflect the image mixture information in the loss space is crucial for leveraging image mixture in the self-supervised domain. Here we introduce the strategy of how to retain such information for loss calculation. We propose to directly mix the first image with the last one in each *mini*-batch of training, the second one is mixed with the penultimate, and so on. Our strategy is visualized in Fig. 3 and Fig. 4, the advantages of such a strategy are: **(i)** Different from employing individual ratio for each image in one *mini*-batch, the proposed scheme can be realized through calculating the batch loss with a weighted coefficient, which is well-regulated, manageable, more efficient for implementing and can facilitate the design of label assignment in unsupervised frameworks. **(ii)** The proposed strategy will make the soft distances between the mixtures and original samples to be consistent across all pairs within a *mini*-batch. Hence, the calculation rule of loss function will be simplified and independent from the different frameworks that are employed, for instance, contrastive learning frameworks that use both positive and negative pairs or positive only, memory bank or without it, *etc.*, as scaling similarity distance is equivalent to weighting these loss values.

We now elaborate the loss functions. We compute an extra loss from a mixed pair of images. Given two mixed images I_A^M and $\hat{I}_A^M$ from two different augmentations of the same image (the case that both branches are mixed), we compute their loss together with the original one as the following:

$$\mathcal{L}_{both} = \mathcal{L}_{ori}(I_A, \hat{I}_A) + \underbrace{\mathcal{L}_m(I_A^M, \hat{I}_A^M)}_{\text{extra term of mixtures}} \quad (4)$$

where $\mathcal{L}_{ori}$ is the original loss function corresponding to the base method we use, like InfoNCE, ℓ_2 distance, *etc.*, and $\mathcal{L}_m$ measures the fit between samples I_A^M and $\hat{I}_A^M$.

Finally, we define the following sum of three loss terms from the original and mixed predictions as the ultimate objective:

$$\mathcal{L}_{final} = \mathcal{L}_{ori} + \underbrace{\lambda \mathcal{L}_m(I_A^M(\downarrow), \hat{I}_A)}_{\text{normal order of mixtures}} + \underbrace{(1-\lambda)\mathcal{L}_m(I_A^M(\uparrow), \hat{I}_A)}_{\text{reverse order of mixtures}} \quad (5)$$

where the last two loss terms measure the fit between samples I_A^M and $\hat{I}_A$, as detailed above. We take contrastive loss with InfoNCE as an example (notations refer to MoCo):

$$\begin{aligned}\underbrace{\mathcal{L}_m(I_A^M(\downarrow),\hat{I}_A)}_{\text{normal order of mixtures}} &= -\log\frac{\exp\left(q_m\cdot k^*/\tau\right)}{\sum_{i=0}^{K}\exp\left(q_m\cdot k_i/\tau\right)}\\ \underbrace{\mathcal{L}_m(I_A^M(\uparrow),\hat{I}_A)}_{\text{reverse order of mixtures}} &= -\log\frac{\exp\left(q_{rm}\cdot k^*/\tau\right)}{\sum_{i=0}^{K}\exp\left(q_{rm}\cdot k_i/\tau\right)}\end{aligned} \tag{6}$$

where q_m, q_{rm} are normal/reverse orders of mixed queries in a *mini*-batch, k^* is the unmixed single key, τ is temperature.

Justifications from the Mutual Information (MI) Theory. Suppose the latent representations of inputs are calculated as $\boldsymbol{z}_{ori} = f_{\theta_1}(I_{ori}), \boldsymbol{z}_{mix} = f_{\theta_2}(I_{mix})$. According to [Oord, Li, and Vinyals 2018], the mutual information $\boldsymbol{I}(\boldsymbol{z}_{ori}, \boldsymbol{z}_{mix})$ of InfoNCE loss can be formulated as:

$$\boldsymbol{I}(\boldsymbol{z}_{ori}, \boldsymbol{z}_{mix}) \geq \log(\boldsymbol{N}) - \mathcal{L}_{\boldsymbol{N}} \tag{7}$$

where $\boldsymbol{N}$ is the number of training samples (one positive and $\boldsymbol{N}$-1 negative samples). To maximize the lower bound of MI, one way is to minimize the InfoNCE objective $\mathcal{L}_{\boldsymbol{N}}$, while, $\boldsymbol{I}$ can also increase when $\boldsymbol{N}$ becomes larger which equivalently maximizes a lower bound on $\boldsymbol{I}(\boldsymbol{z}_{ori}, \boldsymbol{z}_{mix})$. Considering contrastive pairs without mixture, we build $\binom{n}{1}$ relationships (n is the number of images) in the dataset, only (n-1) images are negative pairs to the original one. After adding mixtures of two images, the MI we utilized is $\binom{n}{2}$ relationships. In general, using additional mixtures (equivalent to enlarge values of $\boldsymbol{N}$) does increase the tightness of mutual information $\boldsymbol{I}$.

4. Experiments

We demonstrate the effectiveness and superiority of our UnMix learned models with unsupervised pretraining on a variety of datasets. We first evaluate the representation ability in linear evaluation protocol. We then measure its transferability using object detection task on PASCAL VOC and COCO.

4.1. Datasets

CIFAR-10/100 [Krizhevsky and Hinton 2009] consist of tiny colored natural images with a size of 32×32. In each dataset, the train and test sets contain 50K and 10K images.

STL-10 [Coates, Ng, and Lee 2011] is inspired by CIFAR-10 with 10 classes, while each class has fewer labeled training examples (500 training images and 800 test images per class, and 100K unlabeled images). The size of images is 96×96.

Tiny ImageNet is a lite version of ImageNet which contains 200 classes with images resized down to 64×64. The train and test sets contain 100K and 10K images, respectively.

ImageNet-1K [Deng et al. 2009], aka ILSVRC 2012 classification dataset consists of 1000 classes, with a number of 1.28 million training images and 50K validation images.

4.2. Baseline Approaches

We perform our evaluation of image mixtures and label assignment strategy on the following five recently proposed unsupervised methods with state-of-the-art performance:

MoCo V1&V2 [He et al. 2019, Chen et al. 2020b]. MoCo is a contrastive learning method using momentum updating for unsupervised visual feature learning. MoCo V2 further improves momentum contrastive learning by adopting an MLP projection head and more/heavier data augmentation from the following SimCLR [Chen et al. 2020a].

SimCLR [Chen et al. 2020a]. SimCLR is a simple framework for contrastive learning without requiring specialized architectures or a memory bank. It introduces a learnable nonlinear transformation that substantially improves the quality of the learned representations.

BYOL [Grill et al. 2020]. BYOL adopts online and target networks that learn from each other. It trains the online network to predict the target network representation of the same image under a different augmented view. At the same time, it updates the target network with a slow-moving average of the online network without the negative pairs.

SwAV [Caron et al. 2020]. SwAV is a clustering-based method for unsupervised learning. Unlike contrastive learning that compares features directly, it clusters the data while enforcing consistency between cluster assignments produced for different augmentations of the same image.

Whitening [Ermolov et al. 2020b]. Whitening is a loss function proposed for unsupervised representation learning which is based on the whitening of the latent space features. The whitening operation has a scattering effect to avoid degenerate solutions of collapsing to a simple status.

Our baseline approach implementations follow their official codebases which are all publicly available [He et al. 2020b,a, Ermolov et al. 2020a, Caron et al. 2020].

4.3. Implementation Details in Pre-training

The goal of our experiments is to demonstrate the effectiveness of our proposed image mixture and label assignment upon various unsupervised learning frameworks, isolating the effects of other settings, such as the architectural choices, data augmentations, hyper-parameters. As this, we use the same encoder ResNet-18 for all non-ImageNet experiments and ResNet-50 for ImageNet-1K. We use the same training settings, hyper-parameters, *etc.*, as our comparisons. Therefore, all gains in this paper are "minima", and further tuning the hyper-parameters in the baseline approaches to fit our mixture strategies might achieve more considerable improvement, while it is not the focus of this work.

Non-ImageNet Datasets. Following [Ermolov et al. 2020b], on CIFAR-10 and CIFAR-100, we train for 1,000 epochs with learning rate 3×10^{-3}; on Tiny ImageNet, 1,000 epochs with learning rate 2×10^{-3}; on STL-10, 2,000 epochs with learning rate 2×10^{-3}. We also apply warm-up for the first 500 iterations, and a 0.2 learning rate drop at 50 and 25 epochs before the end.

Standard ImageNet-1K. Unless otherwise stated, all the hyperparameter configurations strictly follow the baseline MoCo V2 on ImageNet-1K. For example, we use a *mini*-batch size of 256 with 8 NVIDIA V100 GPUs on ImageNet-1K, considering our primary objective is to verify the effectiveness of proposed method instead of suppressing state-of-the-art results. For image mixtures and label assignment, we use $\gamma = 1.0$ in beta sampling for all experiments, and $P = 0.5$ for non-ImageNet and 0 for ImageNet-1K based on our ablation study.

4.4. Linear Classification

Our linear classification experiments consist of two parts: **(i)** ablation studies on small datasets including CIFAR-10,

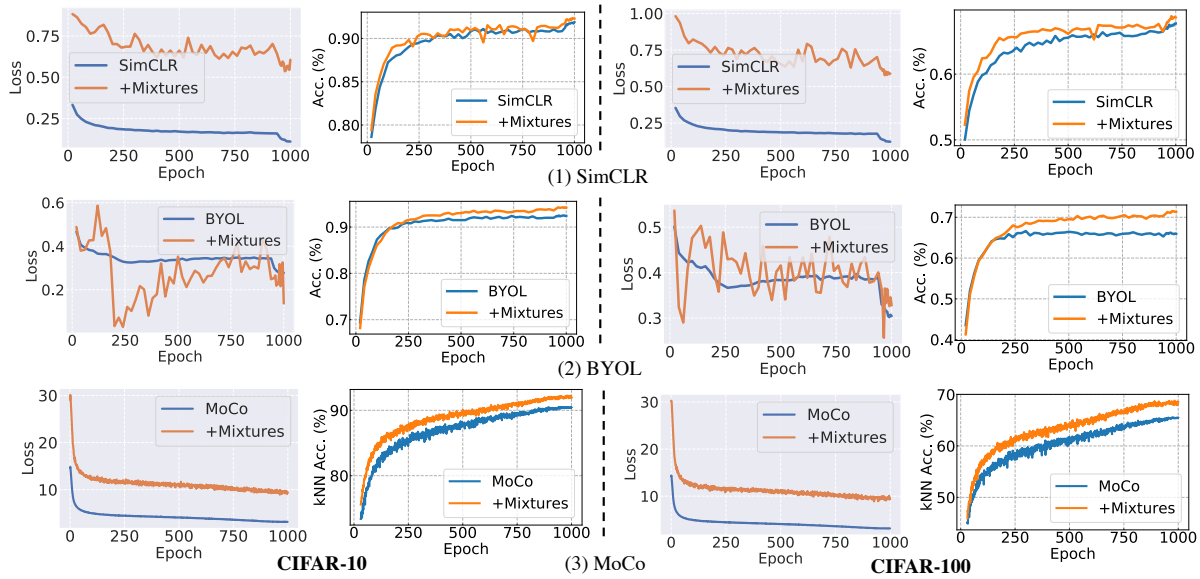

Figure 5: Training loss and testing accuracy of SimCLR, BYOL and MoCo on CIFAR-10/100.

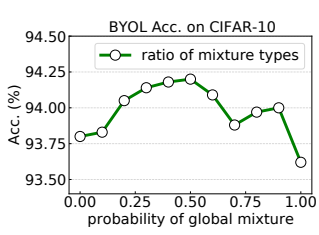

Figure 6: Acc. with various P.

CIFAR-10	γ in **Beta** sampling		
	1.0	0.8	0.5
Acc. (%)	**94.20**	94.12	93.93

Table 1: Sensitivity for γ.

ImageNet-1K	P of global mixture		
	1.0	0.5	0.0
Acc. (%)	67.6	68.3	**68.6**

Table 2: Sensitivity for P.

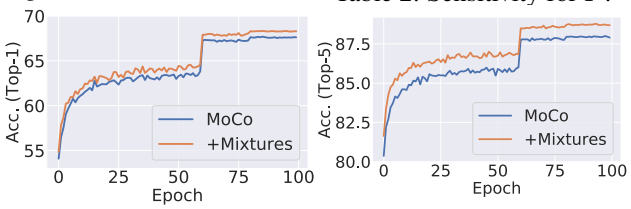

Figure 7: Linear classification accuracy of Top-1 (left) and Top-5 (right) with MoCo V2 and ours on ImageNet-1K dataset.

CIFAR-100, STL-10 and Tiny ImageNet with various base approaches to explore the optimal mixture hyperparameters and demonstrate the effectiveness of our strategy; **(ii)** the final results on the standard ImageNet-1K using MoCo V2.

Ablation Study. We investigate the following aspects in our methods: **(i)** the probability P between global and region mixtures; **(ii)** sensibility of γ in beta distribution sampling.

(1) Probability P for choosing global or region-level mixtures in each iteration. The results are shown in Fig. 6 (non-ImageNet) and Tab. 2 (ImageNet-1K). They show that P=0.5 is optimal for small datasets and choosing region-level only (*i.e.*, $P = 0$) is best for the large-scale ImageNet-1K.

(2) Beta distribution hyperparameter γ. The combination ratio λ between two sample points is sampled from the beta distribution **Beta**(γ, γ). Our results on different γs are presented in Tab. 1, $\gamma = 1.0$ is the best and we use it for all our experiments, which means that λ is sampled from a uniform distribution $[0, 1]$.

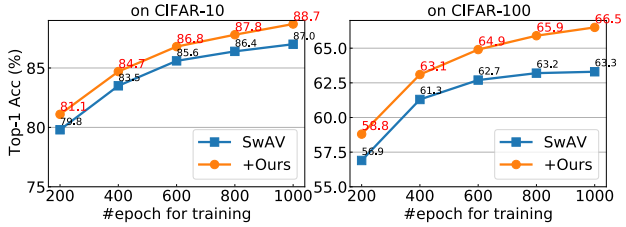

Figure 8: Comparison under different training budgets.

Results on CIFAR-10/100, STL-10 and Tiny ImageNet. Our results are shown in Tab. 3 and Fig. 5. All experiments are conducted on a single scale since the input sizes of these datasets are small, also for fair comparisons to baselines. Our method obtains consistency of $1 \sim 3\%$ gains. In particular, our loss values are usually larger than the baselines (except BYOL, which is unstable since it has no negative pairs), but our accuracy is still superior. We also verify with different training budgets on SwAV, are shown in Fig. 8. It can be seen our method still significantly benefits from longer training.

Results on ImageNet-1K with MoCo V2. As in Tab. 4, our method obtain 1.1% improvement than baseline MoCo V2. Employing multi-scale training as in Appendix further boosts accuracy by 2.3%. It is possible that tuning hyperparameters in MoCo V2, *e.g.*, temperature to fit our mixed training samples has potential to further improve performance.

4.5. Downstream Tasks

In this section, we evaluate the transferability of our learned representation on the object detection task. We use PASCAL VOC [Everingham et al. 2010] and COCO [Lin et al. 2014] as our benchmarks and we strictly follow the same setups and hyperparameters of the prior works [He et al. 2019, Chen et al. 2020b] on the transfer learning stage. We use Faster R-CNN [Ren et al. 2015] and Mask R-CNN [He et al. 2017] implemented in Detectron2 [Wu et al. 2019] with a ResNet-50 [He et al. 2016] backbone.

Method	CIFAR-10				CIFAR-100				STL-10				Tiny ImageNet			
	linear	**ours**	5-nn	**ours**	linear	**ours**	5-nn	**ours**	linear	**ours**	5-nn	**ours**	linear	**ours**	5-nn	**ours**
SimCLR	91.80	**92.35**	88.42	**89.74**	66.83	**68.83**	56.56	**58.82**	90.51	**90.86**	85.68	**86.16**	48.84	**49.58**	32.86	**34.46**
BYOL	91.73	**94.20**	89.45	**93.03**	66.60	**71.50**	56.82	**63.83**	91.99	**93.34**	88.64	**90.46**	51.00	**53.39**	36.24	**39.27**
Whitening (W = 2)	91.55	**93.04**	89.69	**91.33**	66.10	**70.12**	56.69	**61.28**	90.36	**92.21**	87.10	**88.88**	48.20	**51.33**	34.16	**36.78**
Whitening (W = 4)	91.99	**93.18**	89.87	**91.70**	67.64	**69.70**	56.45	**60.74**	91.75	**91.96**	88.59	**88.71**	49.22	**50.67**	35.44	**36.13**
MoCo (Sym. Loss)	–	–	90.49*	**92.25***	–	–	65.49*	**68.83***	–	–	–	–	–	–	–	–

Table 3: Linear and 5-nearest neighbors classification results for different loss functions and datasets with a ResNet-18 backbone. Table is adapted from [Ermolov et al. 2020b] and multi-scale training is not used for fair comparisons. Note that MoCo is trained with symmetric loss, 1000 epochs and evaluated with 200 in kNN monitor* following [He et al. 2020a].

Arch.	**Method**	**#Params**	**Budget (#ep)**	**Top-1 (%)**
R50	MoCo	24	200	60.6
R50	CMC	24	200	66.2
R50	SimCLR	24	200	66.6
R50	MoCo V2	24	200	67.5
R50	MoCo V2 + **Ours**	24	200	**68.6**$^{\uparrow 1.1}$
R50	MoCo V2 + **Ours**†	24	200	**69.8**$^{\uparrow 2.3}$
R50	PIRL	24	800	63.6
R50	SimCLR	24	1000	69.3
R50	MoCo V2	24	800	71.1
R50	MoCo V2 + **Ours**	24	800	**71.8**$^{\uparrow 0.7}$

Table 4: Comparison of linear classification on standard ImageNet-1K. †denotes the result using multi-scale training, more details can be referred to our Appendix. Note that all the hyperparameters follow the baseline MoCo V2 so they might not be optimal on our mixture training scheme, the gains are generally "minima".

.

PASCAL VOC. We fine-tune our models on the split of `trainval07+12` and evaluate on the VOC `test2007` following [Wu et al. 2018, He et al. 2019, Misra and van der Maaten 2019]. All models are fine-tuned for 24k iterations on VOC. It can be observed that significant improvements are consistently obtained by our proposed mixtures.

COCO. We fine-tune on the `train2017` and evaluate on the `val2017` split. The total training budget is 180K iterations. The whole schedule follows the Detectron2 (`coco_R_50_C4_2x`) default setting. Our results are shown in Tab. 5 (b), it can be observed that our results are consistently better than the baseline by a significant margin.

4.6. Visualization and Analysis

Learned representations. To further explore what our model indeed learned, we visualize the embedded features in Fig. 9 from baseline MoCo (left) and our mixture model (right) using t-SNE with the last conv-layer features (128-dimension) from ResNet-18. Our model has more separate embedding clusters, especially on classes 9, 8 and 1. We also visualize the histogram of weights in particular convolutional layers, as shown in our Appendix with discussions.

Limitation. The only limitation we observed in our method is that it will take one additional forward pass for the mixed images. Since in our strategy, calculating the normal and reverse order of images' representations can share the same forwarding operation and there is no extra back-propagation, so the total extra cost will be less than one-third. We emphasize that the information gain of our method is from the mixed representations and mixture ratios, rather than the "longer" training. The mechanism of our method is different from longer training and cannot be replaced by it.

Pre-train	AP_{50}	AP	AP_{75}
Random init.	60.2	33.8	33.1
Supervised IN-1M	81.3	53.5	58.8
MoCo V2 (200ep)	82.4	57.0	63.6
Ours (200ep)	**83.0**$^{\uparrow 0.6}$	**57.7**$^{\uparrow 0.7}$	**64.3**$^{\uparrow 0.7}$
MoCo V2 (800ep)	82.5	57.4	64.0
Ours (800ep)	**83.2**$^{\uparrow 0.7}$	**58.1**$^{\uparrow 0.7}$	**65.2**$^{\uparrow 1.2}$
Ours (200ep), MS	83.2	57.8	64.5

(a) Faster R-CNN, **R50-C4** on **PASCAL VOC**

Pre-train	AP	AP_{50}	AP_{75}
Random init.	35.6	54.6	38.2
Supervised IN-1M	40.0	59.9	43.1
MoCo V2 (200ep)	40.9	60.7	44.4
Ours (200ep)	**41.2**$^{\uparrow 0.3}$	**60.9**$^{\uparrow 0.2}$	**44.7**$^{\uparrow 0.3}$

(b) Mask R-CNN, **R50-C4 2**$\times$ on **COCO**

Table 5: Object detection results fine-tuned on PASCAL VOC (a) and COCO (b) datasets. Models are fine-tuned with the same number of iterations as the baseline, *e.g.,* 24k on VOC. On the VOC dataset, we run three trials and report the means.

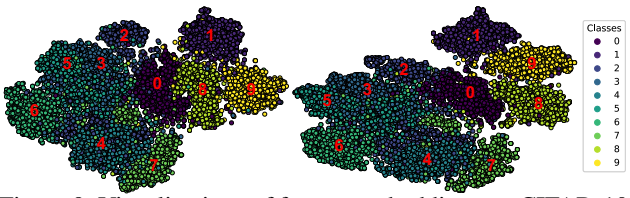

Figure 9: Visualizations of feature embeddings on CIFAR-10.

5. Conclusion

We have investigated the feasibility of mixture operations in an unsupervised scheme, and proposed the strategy of image mixtures and corresponding label re-assignment for flattening inputs and predictions in various architectures of unsupervised frameworks. Through extensive experiments on SimCLR, BYOL, MoCo V1&V2, *etc.*, and downstream tasks like object detection, we have shown that neural networks trained with our newly constructed input space have better representation capability in terms of generalization and transferability, as well as better robustness for different pretext tasks or frameworks (contrastive or non-contrastive learning, with or without memory banks, multi-scale training, *etc.*). Considering its simplicity to implement and it only incurs rational extra cost, we hope the proposed method can be a useful technique for the unsupervised learning problem.

Acknowledgements

We thank all reviewers for their constructive and helpful comments in reviewing our paper. Our full paper with Appendix is available on arXiv: https://arxiv.org/abs/2003.05438.

References

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, 15509–15519.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv preprint arXiv:2002.05709*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020b. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; and He, K. 2020. Exploring Simple Siamese Representation Learning. *arXiv preprint arXiv:2011.10566*.

Chorowski, J.; and Jaitly, N. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 215–223. JMLR Workshop and Conference Proceedings.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*.

Donahue, J.; and Simonyan, K. 2019. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, 10541–10551.

Ermolov, A.; Siarohin, A.; Sangineto, E.; and Sebe, N. 2020a. https://github.com/htdt/self-supervised.

Ermolov, A.; Siarohin, A.; Sangineto, E.; and Sebe, N. 2020b. Whitening for self-supervised representation learning. *arXiv preprint arXiv:2007.06346*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Gidaris, S.; Singh, P.; and Komodakis, N. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, 1735–1742. IEEE.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2019. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020a. https://colab.research.google.com/github/facebookresearch/moco/blob/colab-notebook/colab/moco_cifar10_demo.ipynb.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020b. https://github.com/facebookresearch/moco.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive learning. *arXiv preprint arXiv:2010.01028*.

Kim, S.; Lee, G.; Bae, S.; and Yun, S.-Y. 2020. MixCo: Mix-up Contrastive Learning for Visual Representation. *arXiv preprint arXiv:2010.06300*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario.

Krothapalli, U.; and Abbott, A. L. 2020. Adaptive Label Smoothing. arXiv:2009.06432.

Lee, K.; Zhu, Y.; Sohn, K.; Li, C.-L.; Shin, J.; and Lee, H. 2021. i-Mix: A Domain-Agnostic Strategy for Contrastive Representation Learning. *arXiv preprint arXiv:2010.08887*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Masci, J.; Meier, U.; Cireşan, D.; and Schmidhuber, J. 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks*, 52–59. Springer.

Misra, I.; and van der Maaten, L. 2019. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 4696–4705.

Noroozi, M.; and Favaro, P. 2016. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, 69–84. Springer.

Noroozi, M.; Pirsiavash, H.; and Favaro, P. 2017. Representation learning by learning to count. In *Proceedings of the IEEE International Conference on Computer Vision*, 5898–5906.

Olshausen, B. A.; and Field, D. J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583): 607–609.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Shen, Z.; Liu, Z.; Xu, D.; Chen, Z.; Cheng, K.-T.; and Savvides, M. 2021. Is Label Smoothing Truly Incompatible with Knowledge Distillation: An Empirical Study. In *International Conference on Learning Representations*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, 6438–6447. PMLR.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec): 3371–3408.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

Ye, M.; Zhang, X.; Yuen, P. C.; and Chang, S.-F. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6210–6219.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, 6023–6032.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision*, 649–666. Springer.

Zhang, R.; Isola, P.; and Efros, A. A. 2017. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1058–1067.