Tailoring Self-Supervision for Supervised Learning

WonJun Moon, Ji-Hwan Kim, and Jae-Pil Heo*

Sungkyunkwan University
{wjun0830,damien,jaepilheo}@skku.edu

Abstract. Recently, it is shown that deploying a proper self-supervision is a prospective way to enhance the performance of supervised learning. Yet, the benefits of self-supervision are not fully exploited as previous pretext tasks are specialized for unsupervised representation learning. To this end, we begin by presenting three desirable properties for such auxiliary tasks to assist the supervised objective. First, the tasks need to guide the model to learn rich features. Second, the transformations involved in the self-supervision should not significantly alter the training distribution. Third, the tasks are preferred to be light and generic for high applicability to prior arts. Subsequently, to show how existing pretext tasks can fulfill these and be tailored for supervised learning, we propose a simple auxiliary self-supervision task, predicting localizable rotation (**LoRot**). Our exhaustive experiments validate the merits of LoRot as a pretext task tailored for supervised learning in terms of robustness and generalization capability. Our code is available at https://github.com/wjun0830/Localizable-Rotation.

Keywords: Pretext task, Auxiliary self-supervision, Supervised learning

1 Introduction

Beyond the success in visual recognition without human supervision [5, 8, 19, 56, 57, 61], there have been attempts to adopt self-supervision to supervised learning. Pioneering methods demonstrated that self-supervision indeed improves robustness along with human guidance [4, 29, 50]. They utilized self-supervision as an auxiliary task to support the feature learning. However, as these methods utilize the existing self-supervisions specialized for unsupervised representation learning, the benefits of self-supervision are restrained. Specifically, self-supervision is generally designed for representation learning which is performed better by the primary objective of supervised learning. Furthermore, employed transformations for pretext tasks often trigger significant data distribution shifts. For instance, rotation task, the most popular self-supervision in supervised domain, is not an exception since learning rotation-invariant features barely help the primary task [22, 35]. In fact, rotation task only achieves insignificant gains or sometimes even degrades the performance when applied in the form of multi-task learning or as an augmentation technique [10, 35]. This motivates us to develop a complementary pretext task to supervised objectives.

* Corresponding author

Therefore, in this paper, we first introduce three desirable properties of auxiliary self-supervision to maximize the benefits in supervised learning: 1) learning rich representations, 2) maintaining data distribution, 3) providing high applicability. First, the pretext task should guide the model to learn complementary features with the original ones from supervised learning. Models trained only with a primary task such as the classification often focuses on most discriminative parts of objects where such features provide shortcuts to solve the problem [20,46]. Thus, the primary goal of an auxiliary task is to help the model to capture additional detailed features, which are known to improve the robustness of the model such as detecting out-of-distribution samples as well as its accuracy [29]. Second, the transformation itself should not bring significant data distribution shifts. Although an ideal convolutional neural network (CNN) should be invariant to the transformations such as translation or rotation [42], in realistic circumstances, the shift of global views of images is often known to be harmful to classification tasks [10, 53]. Third, the pretext task is preferred to be highly applicable to existing model architectures and strategies of supervised learning in terms of the computational overhead and the amount of modification.

To validate the importance of these properties, we propose an auxiliary self-supervision task tailored for supervised learning, Localizable Rotation (LoRot), which forms the localization quizzes by rotating only a part of an image. Note that we choose rotation task to be modified into a tailored version for supervised domain on behalf of other pretext tasks due to its effectiveness of localizing the salient objects following previous works [29, 35]. LoRot provides complementary benefits to supervised learning since the model should first localize the patch to solve the rotation task. Specifically, it encourages the model to learn rich features for rotational clues within a part of the image even if they are less discriminative for the supervised objectives. Furthermore, we found that rotating a small patch does not incur a significant data distribution shifts. Finally, the LoRot requires small extra computational costs and implementation efforts, since it is designed for multi-task learning that only requires one additional classifier. In our extensive experiments, we validate that LoRot is effective at boosting the robustness and generalization capability of supervised models and even provides state-of-the-art results. Specifically, we evaluate LoRot on various tasks including out-of-distribution (OOD) detection, imbalanced classification, adversarial attack, image classification, localization, and transfer learning in Sec. 4.

2 Related Work

Self-Supervised Learning. Self-supervised learning has received considerable attention in past years. Its typical objective is to learn general features through solving pretext tasks. According to the number of instances to define pretext tasks, we can categorize the self-supervision into two groups, relation- and transformation-based ones. Relation-based approaches learn features to increase the similarity among a sample [6, 10, 11, 25, 54] and its transformed positive instances while some also treat other training samples as negative instances. The

memory bank [25,43] and in-batch [10,59] samplings are notable negative instance selection techniques. In contrast, there have been approaches to only use the positive pairs with siamese networks [6,24] or adding a relation module [49]. Moreover, transform-based self-supervision is another main stream of representation learning that substantial efforts are made. Remarkable methods are generating surrogate classes with data augmentation [18], predicting the relative position of patches [16], solving jigsaw puzzle [9,32,44,47,48], and predicting the degree of rotation [22]. LoRot also belongs to transform-based self-supervision but is devised for a different objective: to assist supervised learning.

Meanwhile, there have been new attempts to transfer the benefits of self-supervisions to supervised learning. SupCLR [31] modified the relation-based self-supervision framework to directly take advantage of labeled data since the class labels clearly define both positive and negative instances. Moreover, self-label augmentation (SLA) augmented the label space based on the Cartesian product of the supervised class label set and the data transformations label set because learning auxiliary pretext task degrades the performance [35]. In contrast, LoRot is an adequate self-supervision for supervised learning that can be directly applied to existing methods. Further discussion is in Sec. 3.1.

Regional Data Transformation. Data augmentation is one of the most popular ways to improve classification accuracy [12,37,63]. Among them, we introduce methods that modify local regions of an image. Cutout [15] and random erasing [65] randomly mask out square regions of input, while Cutmix [60] cuts and pastes rectangular regions from other samples. LoRot shares the property of editing the local patch with them, however, the main difference is that our transformation retains all information within the image and is a solvable task.

3 Methodology

We first discuss three desired properties for the auxiliary self-supervision for supervised learning in Sec. 3.1. Based on those, we introduce and discuss two forms of LoRot (Localizable Rotation): having explicit and implicit localization tasks, both tailored for supervised learning in Sec. 3.2.

3.1 Desired Properties in Supervised Learning

In this section, we discuss three preferred properties of auxiliary self-supervision for supervised learning: (i) extracting rich representation, (ii) maintaining distribution, and (iii) high applicability, and point out the limitations of previous self-supervised methods in the manner of supervised learning.

In typical training of CNN for the classification task, the model tends to focus on identifying class-specific highly discriminative features to reach a high training accuracy. However, these features usually cover a limited portion of the objects since other parts can be unnecessary to achieve high training accuracy. This phenomenon is often called shortcut learning [20]. It can be problematic when the model faces samples that do not belong to known classes but have

Table 1: CIFAR10 classification accuracy (%) with rotation using different strategies.

	Accuracy
Baseline	95.01
+Rot (DA)	92.76
+Rot (MT)	93.38

Table 2: Distribution shift measured by Affinity score (%). Lower the score, the transformation function triggers a larger distribution shift.

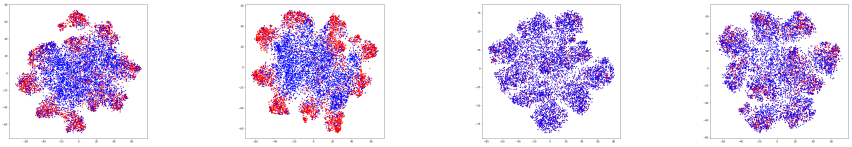
	Affinity
Rotation [22]	58.06
LoRot-I	93.78
LoRot-E	90.15

the learned discriminative features [46]. In such cases, discovering rich features including detailed parts of objects can enhance the robustness of the model [50]. Auxiliary tasks for supervised learning can encourage the model to learn such less discriminative features with the complementary objectives [2,51]. For instance, the popular rotation prediction task [22] can spread the attention of the model toward object parts for predicting the degree of rotation. However, discriminative clues for predicting rotation degrees also exist, e.g., location and orientation. Indeed, the rotation prediction task requires the model to focus on high-level object parts, which are roughly the same image regions as the supervised classification task [22]. So thus, the supervised learning with the auxiliary rotation task is still limited to identifying the most discriminative parts for both tasks.

Multi-task learning is an efficient and effective strategy when there exist multi objectives. It only employs a single shared feature extractor and improves generalization by utilizing the domain-specific information contained in the training signals of related tasks [7, 51]. To employ such a strategy, the transformation function should not incur the degree of the data distribution shift or should smooth the target label as the modified data distribution can impede the primary objective [60, 63]. However, previous self-supervision [22, 47] and following works in supervised domains [4, 21, 29, 50] do not satisfy above which leads them to use inefficient ways to adopt in supervised domains (Discussed in next paragraph). Particularly, in Tab. 1, we conduct a simple experiment with the rotation task [22] to show how the classification result is affected when the transformed input is utilized to learn the primary task. In the table, using the global rotation [22] only as an data augmentation technique (DA) shows the degraded performance. Moreover, sharing the input features for multi-classifiers as the multi-task learning (MT) also provides worse performance than the supervised learning (Baseline). To support our claims, we measure the distribution shift. To quantitatively measure the distribution shift, we use the affinity score [23] on CIFAR-10 in Tab. 2. Affinity score is a metric to evaluate the distribution shift measured as follows:

$$\text{Affinity} = \frac{A(m, D'_{val})}{A(m, D_{val})}, \tag{1}$$

where m is a model trained on training set, and $A(m, D)$ is the accuracy of m on a dataset D. D_{val} and D'_{val} are the original and augmented validation sets, respectively. Furthermore, we also qualitatively show in Fig. 1 (a) and (b) that transformations from previous self-supervision lead distribution gap from its original one. Therefore, we can derive that the data distribution shift triggers



(a) Rotation (b) Jigsaw puzzle (c) LoRot-I (d) LoRot-E

Fig. 1: t-SNE [40] visualization of feature distributions of original (Red) and transformed (Blue) test samples to see the data distribution shifts induced by each data transformation deployed in different self-supervision tasks. Embedding features are extracted from the last convolution layer by forwarding original and transformed test samples to ResNet18 [26] trained on the original train images of CIFAR-10.

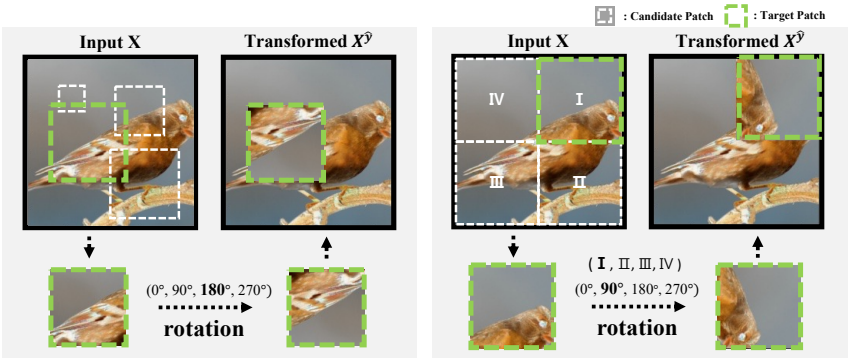
unstable training and a new transformation that maintains the semantics of the image is needed to employ multi-task learning.

Applicability is an another important aspect of an auxiliary task in practical view. Devising lightweight architectures and methods is one of the current research trend [30]. However, since current self-supervised methods are not studied thoroughly in supervised learning, they often trigger high extra cost. Specifically, previous works [4, 29, 50] adopted self-supervision into the supervised domains in rather inefficient ways: parallel-task learning strategy or label augmentation strategy. We define parallel-task learning as each separate input sets being forwarded to handle each tasks in contrast to multi-task learning. For more details, See Fig. 3 (c). We further note that label augmentation requires all possible transformations to be applied per sample at both the training and inference times. At this point, when the usage of pretext tasks in the supervised domain is increasing on the sacrifice of expensive costs, applicability is the next mission for self-supervision to be more widely applied in supervised domains.

3.2 Localizable Rotation (LoRot)

Localizable rotation is designed to rotate a local region. To solve the localizable rotation task, the model should first localize the patch and then identify the high-level clues to predict the rotation degree, e.g., object parts such as eyes, tails, and heads within the patch [22]. Therefore, an explicit localization task to predict the position of the patch may not be necessary to guide the model to learn the localization capability. In this context, we introduce two versions of LoRot with explicit and implicit localization tasks as shown in Fig. 2. For the rest of this paper, we define LoRot-E and LoRot-I as each LoRot having the localization task explicitly and implicitly.

Let $X \in \mathbb{R}^{H \times W \times C}$ be a training image with the width W, height H, and channels C, and y be its class label of supervised learning. Unless mentioned, LoRot is used in the form of multi-task learning with two classifiers each for the primary task and localizable rotation. Let also the feature extractor and two softmax classifiers be F_θ, σ_u, and σ_v parameterized by θ, u and v, respectively.



(a) LoRot-I (b) LoRot-E

Fig. 2: Illustration of LoRot-I and LoRot-E. (a) LoRot-I draws and rotates a random patch from the image, while (b) LoRot-E chooses and rotates a cell from the predefined grid layout. For both methods, the degree of rotation is randomly chosen from $\{0°, 90°, 180°, 270°\}$. Note that, white and green boxes indicate possible and selected patches in this example, respectively.

We also define the transformation function T and the pretext label $\hat{y}$ in which T generates the transformed sample $X^{\hat{y}}$ as follows:

$$X^{\hat{y}} = T(X|\hat{y}, S), \tag{2}$$

where S stands for patch selection strategy. We define possible rotation degrees to $(0, 90, 180, 270°)$ following the rotation task [22]. Then, as shown in Fig. 2, the number of classes for LoRot-I would be 4 and 16 for LoRot-E with the position in the 2x2 grid layout. Note that, for LoRot-E, we keep redundant cases with 0° at every cell as we pursue to place more weights on the original image.

Moreover, we define $P_u(X^{\hat{y}}) = \sigma_u(F_\theta(X^{\hat{y}}))$ and $P_v(X^{\hat{y}}) = \sigma_v(F_\theta(X^{\hat{y}}))$ as the probability distributions over the labels of the primary and pretext tasks, respectively. When $P^j(.)$ is the probability of the j-th class with a batch of N training images $\{X_i\}_{i=1}^N$, the overall objective is:

$$\min_\theta -\frac{1}{N}\sum_{i=1}^{N}(\log(P_u^y(X_i^{\hat{y}})) + \lambda \log(P_v^{\hat{y}}(X_i^{\hat{y}}))), \tag{3}$$

where λ is a hyperparameter to control the weight of learning LoRot.

Patch Selection. We use different strategies to generate patches for LoRot-E and LoRot-I. Simply put, we pre-define the 2x2 grid layout for LoRot-E to easily design the localization task. Then, the sampling method S does not need any parameters. Specifically, we divide each image into a $K \times K$ uniform grid and rotate a single cell of the grid with the dimension of $\mathbb{R}^{\frac{H}{K} \times \frac{W}{K} \times C}$. In this paper, we set K to 2 which each quadrant of an image can be the target.

Meanwhile, we choose random sampling method for LoRot-I. We randomly sample a length l and the position of the top-left corner (p_x, p_y) from the uniform

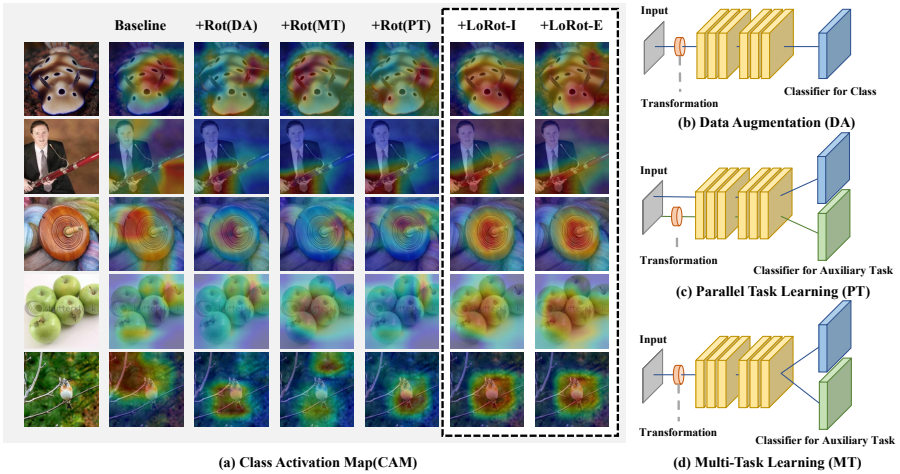


Fig. 3: (a) Comparison of class activation maps (CAMs) [66] of differently learned models. Rot indicates the global rotation task. DA, PT and MT stand for each strategy of utilizing the rotation task: Data Augmentation, Parallel-Task learning, and Multi-Task learning, as illustrated in (b), (c) and (d). DA and MT take augmented input to predict single- or multi-tasks. On the other hand, PT requires separate input batches to predict primary and auxiliary tasks, respectively. LoRot-I and LoRot-E are applied to the baseline by MT as designed. The CAMs show that our LoRot provide the activation of higher coverage to the object compared to global rotations. In other words, LoRot auxiliary task encourages the model to learn rich features. Best viewed in color.

distribution U to form sampling strategy S as follows:

$$S(l, p_x, p_y) \begin{cases} l \sim \mathrm{U}(2, \min(\lfloor W/2 \rfloor, \lfloor H/2 \rfloor)), \\ p_x \sim \mathrm{U}(0, W - l), \\ p_y \sim \mathrm{U}(0, H - l) \end{cases} \tag{4}$$

Note that only a square-shaped patch is used in our work for simplicity. Also, we limit the length of the patch up to half of $\min(H, W)$ to prevent rotating an overly large region. Next, we detail how LoRot satisfies the desired properties.

Rich Representations. LoRot encourages the model to consider even the less-discriminative features by setting rotation prediction quizzes on different locations within an image. Particularly, the model should learn rich features to solve rotation tasks for patches of various sizes at different locations. One may ask that the LoRot can produce many useless features. For instance, the rotation problem with patches totally outside object regions can disturb learning good representations. However, this is alleviated by joint optimization with the supervised objective. Since the primary loss is rather dominant compared to LoRot's, such certainly unnecessary features are dropped throughout the training iterations. To support our argument, we investigate the class activation maps [66] of various models trained with only the primary task (Baseline), and the primary task with auxiliary tasks of global rotation (Rot(*)), and LoRot.

As shown in Fig. 3 (a), LoRot provides larger coverage of activations on the object compared to others. We also quantitatively validate the aforementioned in terms of model robustness and localization performance in Sec. 4. As discussed, LoRot discovers features and asks their necessity to the primary task. In Fig. 4, we show how LoRot discovers sub-discriminative features. Whenever the position of the patch moves, the model adjusts its focus to the other parts of the image to solve the auxiliary task. This also promotes the model to neglect unnecessary parts to solve specific LoRot tasks since the salient clues for predicting LoRot is random.

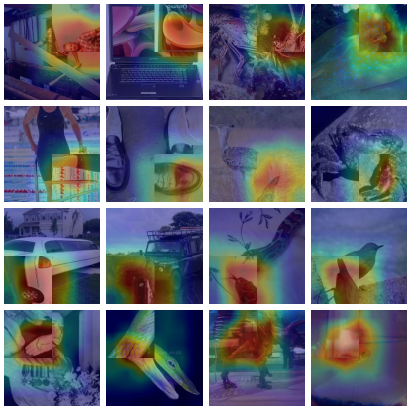


Fig. 4: Class activation mapping visualizations for predicting LoRot. From top to bottom, we show that the model focuses on the high-level clues in the rotated patch at each quadrant, i.e., in the first column, the model spotlights the head, leg, wheel, and hand.

Maintaining Data Distribution. Unlike the existing transformations of pretext tasks, LoRot is less likely to incur data distribution shift since it only carries out geometric transformations locally. Specifically, most parts are kept intact in LoRot so that data distribution shift is restrained. Not just the smaller number of the transformed pixels but also the preserved high-level semantics contribute to keep the distribution close to the original one. In Fig. 1 (c) and (d), we observe that LoRot maintains the data distribution as the blue obscures the red in the embedding space. Affinity score also validates it as shown in Tab. 2.

Applicability. To apply self-supervision of LoRot, we adopt multi-task learning so that a single transformed input is shared by both the primary and pretext tasks as shown in Fig. 3 (d). This provides the advantages in the computational cost and easy deployment to existing models since LoRot only requires one extra classifier without requiring multi-batches. In fact, previous pretext tasks usually require several times more samples to be forwarded and backpropagated to achieve their performances as they apply all possible transformations per sample [29, 35]. Throughout Sec. 4, we validate LoRot's high applicability, a lightweight task boosting baselines' performances. Particularly, we observe that LoRot is not only complementary to standard baselines, but also to contrastive approaches in Tab. 3 and Tab. 8. This spotlights LoRot as an easily attachable self-supervised module for many supervised methods.

4 Experiments

We first examine the robustness of LoRot in Sec. 4.1 and validate the generalization capability in Sec. 4.2. Unless otherwise mentioned, LoRot is applied to supervised baseline with cross-entropy loss. For baselines, we compare rotation [22], the most

Table 3: AUROC scores for distinguishing in- and out-distribution data for image classification. The model is trained with CIFAR-10 dataset and evaluated on both CIFAR-10 and each OOD dataset. '$*$' indicates our reproduced version based on official implementations to unify the backbone network and training protocols. All experiments are averaged over five runs and '$\pm$' denotes the standard deviation. 'FS' and 'IN' stand for the number of forwarded samples to train each model and ImageNet, respectively.

Method	SVHN	LSUN	IN	LSUN (FIX)	IN (FIX)	CIFAR-100	FS
Cross Entropy$*$	$84.6_{\pm 5.2}$	$90.9_{\pm 0.7}$	$87.8_{\pm 1.4}$	$84.3_{\pm 1.0}$	$85.3_{\pm 0.6}$	$83.5_{\pm 0.5}$	5M
Cutmix [60]$*$	$75.5_{\pm 9.5}$	$92.5_{\pm 3.3}$	$92.1_{\pm 2.0}$	$86.2_{\pm 1.0}$	$84.3_{\pm 1.0}$	$80.9_{\pm 1.1}$	5M
SLA+SD [35]$*$	$89.1_{\pm 4.4}$	$90.7_{\pm 1.3}$	$89.8_{\pm 0.8}$	$82.9_{\pm 1.6}$	$86.0_{\pm 0.7}$	$83.6_{\pm 0.4}$	20M
Rotations [29]$*$	$96.1_{\pm 1.8}$	$97.3_{\pm 0.5}$	$96.9_{\pm 0.9}$	$91.0_{\pm 0.4}$	$91.8_{\pm 0.2}$	$89.1_{\pm 0.4}$	25M
SupCLR [31]	$97.3_{\pm 0.1}$	$92.8_{\pm 0.5}$	$91.4_{\pm 1.2}$	$91.6_{\pm 1.5}$	$90.5_{\pm 0.5}$	$88.6_{\pm 0.2}$	70M
CSI [53]	$96.5_{\pm 0.2}$	$96.3_{\pm 0.5}$	$96.2_{\pm 0.4}$	$92.1_{\pm 0.5}$	$92.4_{\pm 0.0}$	$90.5_{\pm 0.1}$	280M
LoRot-I	$92.6_{\pm 2.1}$	$\underline{98.6}_{\pm 0.7}$	$\underline{98.0}_{\pm 0.8}$	$94.4_{\pm 0.9}$	$93.6_{\pm 1.0}$	$90.1_{\pm 0.7}$	5M
LoRot-E	$94.4_{\pm 0.9}$	$\mathbf{98.7}_{\pm 0.6}$	$\mathbf{98.1}_{\pm 0.5}$	$94.1_{\pm 0.3}$	$93.1_{\pm 0.4}$	$\underline{90.6}_{\pm 0.3}$	5M
CSI+LoRot-I	$\mathbf{97.7}_{\pm 0.6}$	$98.3_{\pm 0.1}$	$\underline{98.0}_{\pm 0.3}$	$\mathbf{95.7}_{\pm 0.1}$	$\mathbf{95.6}_{\pm 0.1}$	$\mathbf{93.8}_{\pm 0.0}$	280M
CSI+LoRot-E	$\underline{97.5}_{\pm 0.4}$	$98.0_{\pm 0.2}$	$97.8_{\pm 0.1}$	$\underline{95.5}_{\pm 0.2}$	$\underline{95.4}_{\pm 0.2}$	$\mathbf{93.8}_{\pm 0.1}$	280M

popular pretext task in supervise domain, and previous works that adopted self-supervision in supervised domains (Rotations [29], SLA+SD [35], and SSP [58]). We also compare SOTA methods between benchmarks and show that contrastive learning is a complementary method, not our baseline. For other pretext tasks, we claim these are neither our baseline nor better than our baselines since our baselines are modified versions for supervised domain on top of existing pretext tasks. Throughout this section, we use bolds and underlines to represent the best and the second best scores. Furthermore, as LoRot is robust to λ, we set λ to 0.1 for all experiments and further explore the effects of λ in the supplementary.

4.1 Robustness

Out-Of-Distribution Detection is to assess the model's uncertainty against unknown data. It is essential when deploying the model in real-world systems since DNNs are vulnerable to shortcut learning [20,27,46]. For the experiment, we train the model with CIFAR-10 [34] which we call it in-distribution dataset. Then, we use SVHN [45], resized ImageNet and LSUN [36], fixed versions of ImageNet and LSUN [53], and CIFAR-100 [34] as out-of-distribution datasets. We compare our method against the previous SOTA works on OOD detection [29,53] as well as approaches that utilize a rotation technique [35] and regional modification [60] to enhance the robustness. Unlike the previous SOTA works that require huge costs either at the training [53] or inference time [29] to yield their best performance, LoRot does not require huge costs neither at the training nor inference time. Indeed, we only take 3.6% of training time compared to [53] and 50% of inference time compared to [29] (Measured with 2 Quadro RTX 8000). Still, we acquired significant gains on detecting OOD samples. For fair comparison, we use the ResNet18 [26] following the previous SOTA work [53] to unify the benchmarks.

As we can see in Tab. 3, our proposed LoRot outperforms the state-of-the-art methods on five benchmarks. To measure the performance of LoRot, we utilize

Table 4: Imbalanced classification accuracy (%) on CIFAR-10/100. We add LoRot and other self-supervised approaches on LDAM-DRW and compare the gains.

Imbalance Ratio	0.01	0.02	0.05	0.01	0.02	0.05
LDAM-DRW [3]	77.03	80.94	85.46	42.04	46.15	53.25
+ Rot (DA)	71.91	74.50	77.94	40.32	43.70	46.97
+ Rot (MT)	71.63	74.26	78.02	39.22	43.43	46.76
+ Rot (PT)	75.86	81.13	84.90	43.08	47.67	52.81
+ SSP [58]	77.83	82.13	-	43.43	47.11	-
+ SLA+SD [35]	80.24	-	-	45.53	-	-
+ LoRot-I	81.13	83.69	86.52	45.82	49.33	**54.69**
+ LoRot-E	**81.82**	**84.41**	**86.67**	**46.48**	**50.05**	54.66

the KL-divergence between the softmax predictions and the uniform distribution as in [28, 29]. However, we use the softmax predictions for SLA+SD [35] and CutMix [60] as the softmax results fit better with their methods. The results for SupCLR [31] and CSI [53] are from its paper and we further report the performances of LoRot when applied to contrastive approach, CSI. Interestingly, we observe that the AUROC score of CutMix [60] degrades on harder benchmarks in OOD detection. We conjecture that the label smoothing effect of CutMix could degrade their robustness to unseen samples in harder benchmarks. In contrast, LoRot consistently improves the baselines by large margin (including 11%p and 4%p improvement on LSUN(FIX) dataset to cross-entropy and CSI, respectively). For the slightly low performance on SVHN, we think that it is because there is no difference when a small patch is rotated against a plain background of the SVHN dataset. Thus, we believe LoRot-E is better when images are composed of a simple background and, otherwise both approaches would work fine.

Imbalanced Classification. Following [3], we use CIFAR to design imbalanced scenarios. To make imbalanced set, $\upsilon \in (\mu, 1)^K$ is multiplied to define the sample numbers for each class as $n_i = n_i \upsilon_i$ where i and n are the class index and the number of the original train set. μ and K denote imbalance ratio and number of classes, respectively. Then, we measure the accuracy using the original test set. As the baseline, we deploy LDAM-DRW [3] and follow experimental configurations from them. Meanwhile, to compare ours with other self-supervision techniques, we also report the results of Rotation [22], SLA+SD [35], and SSP [58]. To be specific, we apply rotation in the form of DA, MT, and PT as described in Fig. 3. SSP [58] is the method of pre-training the network with self-supervised learning.

In Tab. 4, we show LoRot has clear complementary effects and consistently improves the SOTA model by a large gain of up to +4.44%p (10.56%) in the highly imbalanced scenario in CIFAR100. As an analysis, the classifier might not learn the discriminative parts for specific classes only with a few examples in an imbalanced setting since the classifier has a bias towards a small number of samples for such categories. However, LoRot alleviates this issue since LoRot complements the classifier by discovering sub-discriminative features. More results with the fully supervised baseline are in the supplementary report.

Table 5: Classification accuracy (%) against the adversarial attack on CIFAR10. The results show that our model outperforms the baselines in 20-step PGD and 100-step PGD with less degradation of the accuracy for the clean dataset.

Method	Clean	20-step	100-step
Baseline	95.3	0.0	0.0
Adv. Training	83.4	46.5	46.5
+ Rotations [29]	82.8	49.3	49.2
+ LoRot-I (**Ours**)	82.1	52.7	52.6
+ LoRot-E (**Ours**)	82.6	**52.8**	**52.8**

Table 6: Top-1 and Top-5 Classification accuracy (%) on ImageNet. Numbers in the parenthesis are the baseline accuracy.

Method	Backbone	Top-1	Top-5
Baseline	ResNet50	76.32	92.95
+ Rot(DA)	ResNet50	76.42	93.06
+ Rot(MT)	ResNet50	76.68	93.10
+ Rot(SS)	ResNet50	76.79	93.16
SLA+SD [35]	ResNet50	$76.17_{(75.17)}$	-
LoRot-I (**Ours**)	ResNet50	77.71	93.60
LoRot-E (**Ours**)	ResNet50	77.72	93.65
Cutout [15]	ResNet50	77.07	93.34
CutMix [60]	ResNet50	**78.60**	**94.08**

Adversarial Perturbations. Substantial efforts were put into improving DNN's robustness [1, 17, 41] to compensate for the vulnerability against adversarial noise [52]. For the evaluation, we adopt the PGD training [41] as the baseline following the settings from Rotations [29]. We conduct experiments on CIFAR10 against ℓ_∞ perturbations with ϵ set to $8/255$. We adversarially train the network with 10-step adversaries and use 20-step and 100-step adversaries. We set the α to $2/255$ for 10, 20-step and $0.3/255$ for 100-step as in [29, 41]. Tab. 5 shows the results of LoRot along with the rotation task under the same codebase. Using LoRot led the network to be robust with the increase in PGD attacks by large improvement compared to the baselines. Note that the tradeoff between accuracy and robustness against adversarial noise is very natural [62].

4.2 Generalization Capability

Image Classification. To validate LoRot's benefits in terms of the generalization capability, we evaluate on ImageNet [14] and CIFAR datasets. We compare ours with rotation [22] in multiple forms, SLA+SD [35], and patch-based augmentations [15, 60]. SLA+SD augments the class label by applying rotation and utilizes self-distillation to yield a similar output to ensemble results at inference time. Tab. 6 shows that LoRot clearly achieves the best performance among the methods utilizing rotation. Furthermore, we newly spotlight the potential of self-supervision in the perspective of generalization capability in that the gap between LoRot and popular patch-based augmentation, CutMix, is less than 1% on ImageNet while achieving robustness multifariously.

Table 7: Additive benefits of LoRot with augmentation methods on ImageNet classification (%). LoRot shows a consistent trend of performance gains.

Method	Backbone	-		+LoRot-I		+LoRot-E	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Mixup [63]	ResNet50	77.58	93.60	**78.36**	**94.15**	78.18	94.05
AutoAug [12]	ResNet50	77.60	93.80	78.09	93.76	**78.22**	**93.86**
RandAug [13]	ResNet50	77.52	93.47	78.12	93.84	**78.24**	**93.95**

Table 8: Additive benefits of LoRot with contrastive learning on CIFAR-10/100 classification and OOD detection. '†' indicates the number taken from the paper [31] using batch size of 1024. The rest of the results were reproduced with batch size of 512 due to lack of GPU memory. OOD scores are measured with the trained model on CIFAR-10 and averaged over the datasets in Table. 3. All results are averaged on three trials.

Method	CIFAR10	CIFAR100	OOD
SupCLR [31]†	96.0	76.5	N/A
SupCLR [31]	95.75	76.52	96.98
+ Rotation (MT)	94.24	71.80	96.28
+ Rotation (PT)	96.07	76.73	96.90
+ LoRot-I	**96.79**	**78.78**	**97.95**
+ LoRot-E	96.73	78.77	97.92

We further apply our LoRot with data augmentation techniques and contrastive learning. Particularly, we test with AutoAugment [12], RandAugment [13], and Mixup [63] on ImageNet [14] and SupCLR [31] of contrastive learning on CIFAR datasets. The results in Tab. 7 show the consistent trend of the performance gain with three data-augmentation methods without a large number of additional parameters (+0.12%) and extra training time (+6%). Interestingly, we notice that Mixup [63] better fits to LoRot-I while Auto- and Rand-Augment are better with LoRot-E. In the viewpoint of LoRot-E, we speculate that this is because Auto- and Rand-Augment provide the randomness to the grid layout which results in more diverse inputs while Mixup causes a large modification to the image when used with LoRot-E. Note that LoRot's limitation is that it does not bring surplus benefits to CutMix [60] (±0%) since LoRot and patch-based augmentations may modify overlapped region and interrupt each other.

Contrastive Learning has achieved promising results for both unsupervised [6, 10] and supervised learning [31]. As it is shown that relation-based and transform-based methods are complementary in PIRL [43], we also examine it in terms of supervised domain. We report the performance of SupCLR [31] both from its paper and our reproduced version in Tab. 8. We first applied rotation [22] with two different strategies: MT and PT. However, applying rotation with MT provoked the decline in the performance as mentioned in contrastive learning [8] that rotation as augmentation degrades the discriminative performance. As is, using rotation only for self-supervised loss (PT) was not very efficient either, in that it requires twice more computational cost to yield insignificant increase. On the contrary, LoRot benefits additive effects to contrastive learning by enriching the representation vectors that are to be pushed or pulled between other samples.

Table 9: Weakly Supervised Object Localization accuracy (%) on ImageNet.

Threshold	0.5	0.6	0.7	0.8	0.9
Baseline	46.72	31.55	14.49	4.22	1.91
CutMix	47.39	30.24	13.86	4.57	2.03
LoRot-I	49.73	35.49	17.21	5.08	2.03
LoRot-E	**50.24**	**36.07**	**17.81**	**5.49**	**2.12**

Table 10: AP (%) of object detection and instance segmentation models initialized with each pretrained method.

Pretrained	RetinaNet	SOLOv2
Baseline	33.8	33.7
LoRot-I	**35.3**	**34.5**
LoRot-E	35.2	34.4

Localization and Transfer Learning are important criteria to evaluate the model's localization capability. For these experiments, we used our pretrained model yielded from Tab. 6. Briefly, for weakly supervised object localization, the model needs to localize the object when only given with class labels. Thus, the model is required to not only find the class-descriptive clues but also understand the image. Tab. 9 demonstrates that LoRot better guides the model to focus on salient regions. Particularly, we observe that LoRot-E, explicitly having the localization task, leads to better localization capability. For evaluation, we use CAM [66] following ACOL [64] and Co-mixup [33]. As ACOL searched for threshold for CAM results between 0.5 to 0.9, we report all these results.

Object detection and instance segmentation are another tasks that require precise localization capability of the model. Indeed, backbones are commonly initialized with ImageNet pretrained weights to deal with the lack of labeled train data. Thus, we examine whether pretrained models trained with LoRot yield any benefits. For evaluation, we employ Retinanet [38] and SOLOv2 [55] for each task and use COCO 2017 dataset [39] for experiments. Tab. 10 shows our findings: pretrained models with LoRot consistently outperform standard models.

4.3 Further Study

To understand why LoRot is effective in enhancing robustness, we conducted an in-depth analysis of OOD detection shown in Tab. 3. In Fig. 5, we compare the class-wise average confidence scores for OOD (SVHN) dataset between the baseline and the LoRot. We observe a clear tendency that both the LoRot-I and LoRot-E effectively lower all the confidence scores for OOD classes while retaining high confidence for in-distribution dataset. Therefore, LoRot can achieve higher AUROC scores.

Furthermore, we visualized the final embedding space with t-SNE to explore underneath reason for why a better separation has been achieved by the proposed method. In Fig. 6, ten red clusters and blue dots can be found for three methods. Red clusters represent classes in in-distribution dataset, CIFAR10, and blue dots are embeddings of OOD dataset, SVHN. Yet, we can observe that the red and blue dots are significantly mixed in the baseline's embedding space. Meanwhile, two colors overlie less on top of the other in LoRot's feature space and tighter boundaries are formed for red clusters. As discussed, this observation is because the model obtains rich features through learning LoRot which enables the model to understand the input even when the most discriminative hint for each class

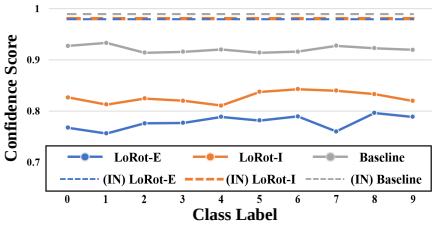


Fig. 5: Average confidence scores for in- and out-of-distribution data (CIFAR10 and SVHN) of the baseline and LoRot. Dotted lines are the averaged confidence scores of in-distribution (IN-) dataset for each method. Solid lines represent the confidence scores (y-axis) for each class in out-distribution dataset (x-axis). These results demonstrate that LoRot improves the capability of the models to detect unknown samples.

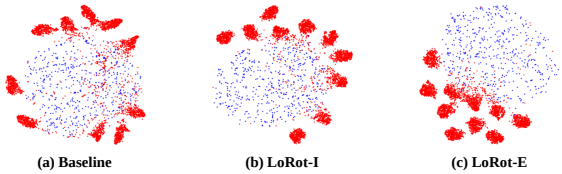


Fig. 6: From left to right, t-SNE visualization for the baseline, LoRot-I, and LoRot-E, respectively. We plot the feature distributions of in-distribution instances (Red) and out-distribution instances (Blue). Unlike the baseline where many red dots are scattered with the blue dots, it is evident that clusters appearing under LoRot is more compact.

is not available. In other words, the model is less vulnerable to mispredicting OOD samples with learned features of some classes because it considers a broader spectrum of class-descriptive features.

5 Conclusion

Although self-supervision has been proved to be powerful in supervised domain, its potential is still an untapped question since existing works are designed for unsupervised condition. Thus, we presented three desirable properties of self-supervision to be tailored for supervised learning: enriching representations, maintaining data distribution, and high applicability. To comply with them, we introduced LoRot, a self-supervised localization task that assists supervised learning to further improve robustness and generalization capability. Our extensive experiments demonstrated the merits of LoRot as well as the complementary benefits to prior arts. Furthermore, as we revisited the potential of self-supervision in a simple applicable way in supervised settings, we believe this line is worth further study to be a standard technique in supervised learning.

Acknowledgements. This work was supported in part by MSIT/IITP (No. 2022-0-00680, 2020-0-00973, 2020-0-01821, and 2019-0-00421), MCST/KOCCA (No. R2020070002), and MSIT&KNPA/KIPoT (Police Lab 2.0, No. 210121M06).

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International Conference on Machine Learning. PMLR (2018)
2. Baxter, J.: A bayesian/information theoretic model of learning to learn via multiple task sampling. Machine learning (1997)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: Advances in Neural Information Processing Systems (2019)
4. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020)
7. Caruana, R.: Multitask learning. Machine learning (1997)
8. Chen, G., Qiao, L., Shi, Y., Peng, P., Li, J., Huang, T., Pu, S., Tian, Y.: Learning open set network with discriminative reciprocal points (2020)
9. Chen, P., Liu, S., Jia, J.: Jigsaw clustering for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR (2020)
11. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
12. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
13. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee (2009)
15. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
16. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE international conference on computer vision (2015)
17. Dong, Y., Fu, Q.A., Yang, X., Pang, T., Su, H., Xiao, Z., Zhu, J.: Benchmarking adversarial robustness on image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
18. Dosovitskiy, A., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with convolutional neural networks. In: Advances in Neural Information Processing Systems. Citeseer (2014)

19. Feng, Z., Xu, C., Tao, D.: Self-supervised representation learning by rotation feature decoupling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
20. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence (2020)
21. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Boosting few-shot visual learning with self-supervision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8059–8068 (2019)
22. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018)
23. Gontijo-Lopes, R., Smullin, S., Cubuk, E.D., Dyer, E.: Tradeoffs in data augmentation: An empirical study. In: International Conference on Learning Representations (2020)
24. 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.: Bootstrap your own latent: A new approach to self-supervised learning. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020)
25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
27. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. International Conference on Learning Representations, ICLR 2017 (2016)
28. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations (2019)
29. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. Advances in Neural Information Processing Systems (NeurIPS) (2019)
30. Khan, A., Sohail, A., Zahoora, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. Artificial Intelligence Review (2020)
31. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning (2020)
32. Kim, D., Cho, D., Yoo, D., Kweon, I.S.: Learning image representations by completing damaged jigsaw puzzles. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2018)
33. Kim, J.H., Choo, W., Jeong, H., Song, H.O.: Co-mixup: Saliency guided joint mixup with supermodular diversity. International Conference on Learning Representations, ICLR 2021 (2021)
34. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
35. Lee, H., Hwang, S.J., Shin, J.: Self-supervised label augmentation via input transformations. In: International Conference on Machine Learning. PMLR (2020)
36. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. International Conference on Learning Representations, ICLR 2018 (2017)

37. Lim, S., Kim, I., Kim, T., Kim, C., Kim, S.: Fast autoaugment. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
38. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (2017)
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. Springer (2014)
40. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
41. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations, ICLR 2018 (2017)
42. Mallat, S.: Understanding deep convolutional networks. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (2016)
43. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
44. Mundhenk, T.N., Ho, D., Chen, B.Y.: Improvements to context based self-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
45. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011 (2011)
46. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2015)
47. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. Springer (2016)
48. Noroozi, M., Vinjimoor, A., Favaro, P., Pirsiavash, H.: Boosting self-supervised learning via knowledge transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
49. Patacchiola, M., Storkey, A.: Self-supervised relational reasoning for representation learning (2020)
50. Perera, P., Morariu, V.I., Jain, R., Manjunatha, V., Wigington, C., Ordonez, V., Patel, V.M.: Generative-discriminative feature representations for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
51. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
52. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. International Conference on Learning Representations, ICLR 2014 (2013)
53. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. In: Advances in Neural Information Processing Systems (2020)
54. Tian, Y., Chen, X., Ganguli, S.: Understanding self-supervised learning dynamics without contrastive pairs. Proceedings of the International Conference on Machine Learning, (ICML) (2021)

55. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
56. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
57. Yang, J., Parikh, D., Batra, D.: Joint unsupervised learning of deep representations and image clusters. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
58. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)
59. Ye, M., Zhang, X., Yuen, P.C., Chang, S.F.: Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
60. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
61. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., et al.: A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867 (2019)
62. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. PMLR (2019)
63. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018)
64. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
65. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2020)
66. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)