Multi-task Transformation Learning for Robust Out-of-Distribution Detection

Sina Mohseni [1] **Arash Vahdat** [1] **JBS Yadawa** [1]

Abstract

Detecting out-of-distribution (OOD) samples plays a key role in open-world and safety-critical applications such as autonomous systems and healthcare. In this paper, we propose a simple framework that leverages *multi-task transformation learning* for training effective representation for OOD detection which outperforms state-of-the-art OOD detection performance and robustness on several image datasets. We empirically observe that the OOD performance depends on the choice of data transformations which can be optimized for different in-domain training sets. Finally, we demonstrate the performance and robustness of our proposed technique against a diverse range of the state-of-the-art OOD detection techniques.

1. Introduction

Despite advances in representation learning and their generalization to unseen samples, learning algorithms are bounded to perform well on source distribution and vulnerable to out-of-distribution (OOD) or outlier samples. Recent work on machine learning trustworthiness and safety have shown that model uncertainty estimation and OOD detection play a key role in open-world and safety-critical applications such as autonomous systems (Mohseni et al., 2019) and healthcare (Ren et al., 2019). However, uncertainty estimation in high dimensional domains like image data is a challenging task and often requires great computational resource (Gal & Ghahramani, 2016).

The recent surge in self-supervised learning techniques shows that learning pretext tasks can result in better semantic understanding of data by learning invariant representations (Dosovitskiy et al., 2014) and can improve model performance in different setups (Gidaris et al., 2018). However, the role of self-supervised learning in OOD detection is poorly explored as previous studies have been limited to one-class classification problems with rather simple geometric transformations tasks (Golan & El-Yaniv, 2018; Hendrycks et al., 2019b). Recently, (Tack et al., 2020) showed that contrastive learning can leverage shifting data transformations to improve OOD detection. However, (Tack et al., 2020) relied on a set of hand-crafted positive and negative transformations.

Inspired by the recent work, we investigate and observe that the OOD detection performance highly depends on the choice of self-supervised data transformations which itself depends on the in-domain training set. To address this problem, we propose a simple approach to select effective transformations and modulate their impact on OOD detection while learning representation, and we do so without requiring any OOD training samples. Additionally, we propose an ensemble score of OOD detection to harness the power in learning multiple transformations trained on a shared encoder.

Contributions In this paper, we propose an OOD framework that learns to identify multiple transformed views of the in-domain training set in self-supervised and fully-supervised manner (when data labels are available). The transformation learning tasks involve predicting a set of predefined shifting transformations applied to the training data such as rotation, noise, blur, etc. We empirically demonstrate that these transformation prediction tasks allow the model to learn better representation for both main classification and OOD detection compared to the augmentation-based approaches. Our technique achieves new state-of-the-art results on multi-class classification by improving averaged area under the receiver operating characteristics (AUROC) from 94.37% to 95.67% (+1.3%) for CIFAR-10 (Krizhevsky et al., 2009), from 79.98% to 84.35% (+4.37%) for CIFAR-100 (Krizhevsky et al., 2009), and from 95.67% to 96.69% (+1.02%) for ImageNet-30 (Hendrycks et al., 2019b) datasets. Finally, we analyze and compare our technique against a diverse set of OOD detection technique for their generalization and robustness.

2. Related Work

OOD detection methods can be generally divided into multiple categories; here we review techniques closely related to this work due to space limitations.

[1]NVIDIA, Santa Clara, CA. Correspondence to: Sina Mohseni <smohseni@nvidia.com>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

Distance-based Detection: Distance-based methods use different distance measures between the unknown test sample and source training set in the representation space. These techniques involve preprocessing or test-time sampling of the source domain distribution to measure their averaged distance to the novel input sample. The popular distance measures include Mahalanobis distance (Lee et al., 2018; Sehwag et al., 2021), cosine similarity (Techapanurak et al., 2020; Tack et al., 2020) and others semantic similarity metrics (Rafiee et al., 2020). These techniques usually work well with unlabeled data in unsupervised and one-class classification setups. Additionally, distance-based methods can benefit from ensemble measurements over input augmentations (Tack et al., 2020) or transformations (Bergman & Hoshen, 2020), network layers (Lee et al., 2018; Sastry & Oore, 2019), or source domain sub-distributions (Oberdiek et al., 2020) to improve detection results. For instance, (Tack et al., 2020) presents a detection score based on combining representation norm with cosine similarity between the outlier samples and their nearest training samples for one-class classification problem. They also show that OOD detection can be improved with ensembling over random augmentations, which carries a higher computational cost.

Classification-based Detection: These OOD detection techniques avoid costly distance-based and uncertainty estimation techniques (e.g., (Gal & Ghahramani, 2016)) by seeking effective representation learning to encode normality together with the main classification task. Various detection scores have been proposed including maximum softmax probability (Hendrycks & Gimpel, 2016), maximum logit scores (Hendrycks et al., 2019a), prediction entropy (Mohseni et al., 2020), and KL-divergence score (Hendrycks et al., 2019b). To improve the detection performance, (Lee et al., 2017; Hsu et al., 2020) proposed a combination of temperature scaling and adversarial perturbation of input samples and Another line of research proposed using auxiliary unlabeled and disjoint OOD training set (Hendrycks et al., 2018; Mohseni et al., 2020) for improved OOD detection. Recent work on self-supervised learning shows that adopting pretext tasks results in learning more invariant representations (Dosovitskiy et al., 2014) and which significantly improves OOD detection (Golan & El-Yaniv, 2018). Hendrycks et al. (Hendrycks et al., 2019b) extended self-supervised techniques with a combination of geometric transformation prediction tasks. Self-supervised contrastive training (Chen et al., 2020) is also shown to be effective to learn in-domain invariances, resulting in better OOD detection (Winkens et al., 2020; Sehwag et al., 2021; Tack et al., 2020). For example, (Tack et al., 2020) uses geometric transformations like rotation to shift different samples further away to improve OOD detection performance.

3. Method

Learning invariances with identity-preserving transformations is a technique to incorporate prior knowledge to improve representation learning and hence OOD detection. However, training an invariant network is challenging, and common data augmentation techniques do not guarantee for that (Lyle et al., 2020) and may even degrade training performance (Chen et al., 2020). To address this issue, Tack et al. (Tack et al., 2020) suggested to use some augmentations as negative samples in contrastive training, Xiao et al. (Xiao et al., 2021) proposed using multiple embedding sub-spaces for each augmentation, and Lee et al. (Lee et al., 2020) use label augmentation to train for image invariances. In contrast to these work, we suggest using multi-task learning for explicit and efficient training of a wide range of invariances to in-domain representations.

3.1. Multi-task Transformation Learning

We consider a set of geometric (translation, rotation) and non-geometric (blurring, sharpening, color jittering, Gaussian noise, cutout) transformations and we train the network with dedicated self-supervised and fully-supervised (when labels available) loss functions for each transformation. For the self-supervised transformation learning, given an unlabeled training set of $\mathcal{S} = \{(x_i)\}_{i=1}^{M}$, we denote the set of domain invariant transformations T_n by $\mathcal{T} = \{T_n\}_{n=1}^{N}$. We generate a self-labeled training set $\mathcal{S}_{T_n} = \{(T_n(x_i), \hat{y}_i)\}_{i=1}^{M}$ for each self-supervised transformation T_n where $\hat{y}_i$ are the transformation labels. For example, we consider the image rotation task with four levels of $\{0°, 90°, 180°, 270°\}$ self-labeled rotations and $\hat{y}_i \in \{0, 1, 2, 3\}$ in this case. The self-supervised loss $\mathcal{L}_{\text{ssl}}$ is the weighted average loss for all transformations in $\mathcal{T}$:

$$\mathcal{L}_{\text{ssl}}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \lambda_n \sum_{(T_n(x_i), \hat{y}_i) \in S_{T_n}} \ell(f_{\boldsymbol{\theta}}^{(n)}(T_n(x_i)), \hat{y}_i) \quad (1)$$

where $f_{\boldsymbol{\theta}}^{(n)}$ is a classification network with parameters $\boldsymbol{\theta}$ for the n^{th} task and $\boldsymbol{\lambda} = \{\lambda_n\}_{n=1}^{N}$ are transformation weights. When labels are available, given the labeled set $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{M}$, we generate transformed copies of the training sets $\mathcal{S}_{T_n} = \{(T_n(x_i), y_i)\}_{i=1}^{M}$ where the samples retain class labels. The supervised loss $\mathcal{L}_{\text{sup}}$ is defined by:

$$\mathcal{L}_{\text{sup}}(\boldsymbol{\lambda}', \boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \lambda'_n \sum_{(T_n(x_i), y_i) \in S_{T_n}} \ell(f_{\boldsymbol{\theta}}^{\prime(n)}(T_n(x_i)), y_i) \quad (2)$$

which measures the classification loss for transformed copies with $\boldsymbol{\lambda}' = \{\lambda'_n\}_{n=1}^{N}$ as transformation weights. In labeled setup, we define the auxiliary transformation loss as $\mathcal{L}_{\text{aux.}} := \mathcal{L}_{\text{ssl}} + \mathcal{L}_{\text{sup}}$ to combine both self-supervised and fully-supervised transformation losses. $\mathcal{L}_{\text{aux.}}$ can be combined with the main supervised learning loss $\mathcal{L}_{\text{main}}$:

$$\mathcal{L}_{\text{total}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}', \boldsymbol{\theta}) = \lambda_0 \mathcal{L}_{\text{main}}(\boldsymbol{\theta}) + \mathcal{L}_{\text{aux.}}(\boldsymbol{\lambda}, \boldsymbol{\lambda}', \boldsymbol{\theta}) \quad (3)$$

where λ_0 is a scalar coefficient. In one-class and unlabeled detection setups, we define $\mathcal{L}_{\text{total.}} := \mathcal{L}_{\text{main.}} := \mathcal{L}_{\text{ssl}}$ using the self-supervised loss. In the rest of the paper, for the ease of notation, we use $\boldsymbol{\lambda}$ to refer to all the coefficients $\{\lambda_0, \boldsymbol{\lambda}, \boldsymbol{\lambda}'\}$. Instead of training a separate network $f_{\boldsymbol{\theta}}^{(n)}$ or $f_{\boldsymbol{\theta}}^{\prime(n)}$ for each task, all the auxiliary tasks and the main classification task share a feature extraction network and each only introduces an additional 2-layer fully-connected head for each task. Training is done in a multi-task fashion in which the network is simultaneously trained for the main classification (if applicable) and all weighted auxiliary tasks using the cross-entropy loss.

3.2. Learning to Select Transformations

Previous work on self-supervised learning used ad-hoc heuristics for choosing data transformations for the training set (Hendrycks et al., 2019b; Golan & El-Yaniv, 2018; Tack et al., 2020). However, the choice of effective transformation(s) depends on the source distribution and heuristic approaches cannot scale up to diverse training distributions when there are many potential transformations to consider. Appendix A presents an example ablation study that illustrates the importance of choice of transformations. Given these observations, we propose a simple framework that relies on Bayesian optimization to first select effective transformation set $\mathcal{T}$, and then it uses meta-learning to learn optimum $\boldsymbol{\lambda}$ weights for OOD robustness.

Optimizing Transformations Set $\mathcal{T}$: We use Bayesian optimization to identify effective transformations for each in-domain training set as shown in Alg. 1 in Appendix A. Here, we assume that transformation weights $\boldsymbol{\lambda}$ are equal to one and we only search for effective transformations set from a pool of available transformations. We use a low-cost off-the-shelf Bayesian optimization (Akiba et al., 2019) library to find the optimum self-supervised task set due to the small $\mathcal{T}$ search space (i.e., 2^n for the $\mathcal{T}$ set of n). The Bayesian optimization objective seeks to minimize the representation learning loss $\mathcal{L}_{\text{main}}$ on the available D_{val}^{in}.

Optimizing Transformations Weights $\boldsymbol{\lambda}$: Next, we optimize $\boldsymbol{\lambda}$ coefficients for the selected transformation from the previous step to further improve representation learning. This is important because the $\boldsymbol{\lambda}$ coefficients modulate the impact of different transformations in the training objective in Eq. 1 and Eq. 3. Here, we assume that $\boldsymbol{\lambda}$ is a "meta-parameter" and we use a differentiable hyperparameter optimization algorithm (Maclaurin et al., 2015) for optimizing it as shown in Alg. 2 with details in Appendix A.

3.3. OOD Detection Scores

We consider two ways for computing the detection score: (i) We can aggregate the softmax prediction from the main task and all auxiliary transformation heads to compute an *ensemble score*. (ii) Alternatively, if the computational budget is limited for OOD detection, a faster detection score can be computed using the main classification head while discarding auxiliary transformation heads after training (see Table 3 for comparison). Given softmax scores obtained from either (i) or (ii), we use KL divergence between the softmax prediction and uniform distribution (Hendrycks et al., 2019b) as the OOD detection score. Unless mentioned otherwise, our main evaluation results are based on the ensembled score from available auxiliary heads.

4. Experiments and Results

Experiment Setup: We run our main experiments on ResNet-18 (He et al., 2016) network for direct and fair comparison with state-of-the-arts. We used 7 different self-labeled transformations including rotation, translation, Gaussian noise, Gaussian blur, cutout, sharpening, and color distortion with details explained in Appendix B. In all experiments, both transformations set $\mathcal{T}$ and their training weights $\boldsymbol{\lambda}$ are optimized using the framework with final ($\mathcal{T}$, $\boldsymbol{\lambda}$) sets presented in Appendix B.

4.1. Classification Accuracy

Since our technique enjoys from learning domain knowledge through multiple data transformations, we first compare the supervised classification accuracy of our technique with state-of-the-art self-supervised and contrastive learning techniques. Figure 2 in Appendix C presents classification accuracy of CIFAR-10 and CIFAR-100 datasets trained on WideResNet-40-2, ResNet-18, and ResNet-50 networks. Our technique outperforms other techniques across both datasets in WideResNet-40-2 (3.7% gain over Baseline in CIFAR-10) and ResNet-18 (8.64% gain over Baseline in CIFAR-100) networks which indicates the effectiveness of transformation learning for better generalization.

4.2. Detection Performance

We review OOD detection results in multi-class classification setup, one-class classification, and unlabeled multi-class detection.

Multi-class Classification Table 1 presents main evaluation results with AUROC metric on ImageNet-30 (Hendrycks et al., 2019b), CIFAR-10, CIFAR-100 datasets each with 6 disjoint D_{test}^{out} sets (see detail results in Table 2). We compare our technique with the full supervised Baseline (Hendrycks & Gimpel, 2016) and current state-of-the-art methods including self-supervised learning (Geometric) (Hendrycks et al., 2019b), supervised contrastive learning (SupSimCLR) (Khosla et al., 2020) and SSD (Sehwag et al., 2021), and contrasting shifted instances (CSI) (Tack et al., 2020) and with its ensembled version (CSI-ens). All techniques are trained on ResNet-18 network with equal training budget, and all except SSD+ use their softmax prediction as OOD detection score in multi-class classification. Our results are reported with both ($\mathcal{L}_{\text{ssl}}$) and ($\mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{ssl}}$) as

Table 1. Comparison of OOD detection results (average detection AUROC% on five D_{test}^{out} sets) between the supervised Baseline and state-of-the-art self-supervised learning, contrastive learning, and our multi-task transformation learning trained on CIFAR-10, CIFAR-100, and ImageNet-30 for multi-class classification.

Method	CIFAR-10	CIFAR-100	ImageNet-30
Baseline	88.98	76.84	87.01
Geometric (Hendrycks et al., 2019b)	94.05	77.46	91.42
SupSimCLR (Khosla et al., 2020)	92.52	76.88	92.81
SSD+ (Sehwag et al., 2021)	94.14	76.35	94.14
CSI-ens (Tack et al., 2020)	94.37	79.98	95.67
Ours ($\mathcal{L}_{\text{ssl}}$)	95.52	80.78	84.35
Ours ($\mathcal{L}_{\text{ssl}} + \mathcal{L}_{\text{sup}}$)	**95.67**	**84.35**	**96.69**

the $\mathcal{L}_{\text{aux}}$ training loss to compare the impact of both loss functions. The averaged standard deviation (STD) of detection AUROC over six test sets from 5 runs of our techniques shows 0.18% STD for ImageNet-30, 0.13% STD for CIFAR-10, and 0.33% STD for CIFAR-100.

The Choice of Transformations: Next, we investigate the importance of selecting a suitable set of transformations as opposed to prior work which focused on geometric transformations in image domain. Table 4 in Appendix C presents an analysis of the effects of our transformation set $\mathcal{T}$ and $\boldsymbol{\lambda}$ weights optimizations on OOD detection performance. Results show both transformation selection ($\mathcal{T}$ optimization) and their training weights ($\boldsymbol{\lambda}$ optimization) have significant impact on OOD performance.

Advantage of Multi-task Transformation Learning: Appendix C presents a comparison between our ensemble detection score and model classification prediction score. We observe that the ensemble of auxiliary heads outperforms the main head by a large margin indicating the effectiveness of using multiple transformation head in the network.

Comparison to Data Augmentation Techniques: Lastly, we compare our results with RandAugment (Cubuk et al., 2020) and AutoAugment (Cubuk et al., 2019) techniques in Appendix C to observe how combinations of multiple light data augmentations compare with our heavy transformation learning tasks. Results show that despite achieving competitive classification accuracy on ResNet18, however, they perform significantly lower in OOD detection compared to our proposed transformation learning.

Unlabeled Detection Next, we test our technique for multi-class unlabeled and one-class OOD detection trained with the $\mathcal{L}_{\text{ssl}}$ loss (Eq. 1) based on our transformation optimization framework. Looking at detailed results in Appendix C, Table 6(a) presents AUROC results on unlabeled multi-class datasets in which our technique outperforms state-of-the-art methods with a large margin in CIFAR-100 and ImageNet-30 experiments. Table 6(b) shows detailed one-class classification results for each of the CIFAR-10 classes as D_{train}^{in} and the remaining classes as D_{test}^{out}. Our technique with 90.9% averaged AUROC on CIFAR-10 one-class detection outperforms previous works except for the CSI with a far more computationally expensive distance-based detection score.

4.3. OOD Detection Generalizability and Robustness

We introduce four criteria required for an ideal OOD detection technique, including i) zero-shot OOD training, ii) no hyperparameter dependency, iii) generalization to diverse unseen OOD distributions, and iv) robustness against test-time perturbations. We analyze and situate our proposed technique against a diverse range of state-of-the-art OOD detection techniques trained on the same network with the same training budget. Appendix D discussed evaluation details in Table 7 and Figure 3 for detectors trained on CIFAR-10 and being tested on six D_{test}^{out} sets. Note that this is not intended to be a ranking of different OOD detection techniques; instead, we are aiming to review trade-offs and limitations among different detection approaches. Albeit the simplicity, we show that our proposed approach based on transformation learning outperforms the state-of-the-art techniques on robustness and generalizability criteria, in addition to the commonly used OOD detection performance that we discussed in the previous section.

5. Conclusion

Given the importance of reliable and trustworthy machine learning in open-world, we emphasize on the value of representation learning research for OOD detection in addition to the in-domain prediction performance. In this work, we empirically demonstrated a data-free technique for learning domain-specific representation for OOD detection which outperforms the state-of-the-art OOD detection techniques on a range of diverse unseen OOD sets. Thorough evaluation of OOD frameworks is still an underdeveloped subject in this research area. We sketched out a set of criteria for ideal OOD detection and examined our method along with several state-of-the-art techniques under these criteria. Future research is needed to propose faster and more efficient ways for selecting and optimizing identity-preserving transformations for OOD robustness in open-world applications such as object detection and image segmentation.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Bergman, L. and Hoshen, Y. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Cubuk, E. D., Zoph, B., Shlens, J., and 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*, pp. 702–703, 2020.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *ICML*, pp. 233–240. ACM, 2006.

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

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pp. 766–774, 2014.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pp. 1050–1059, 2016.

Gidaris, S., Singh, P., and Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Golan, I. and El-Yaniv, R. Deep anomaly detection using geometric transformations. In *NIPS*, pp. 9758–9769, 2018.

Goyal, S., Raghunathan, A., Jain, M., Simhadri, H. V., and Jain, P. DROCC: Deep robust one-class classification. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3711–3721. PMLR, 13–18 Jul 2020.

Grefenstette, E., Amos, B., Yarats, D., Htut, P. M., Molchanov, A., Meier, F., Kiela, D., Cho, K., and Chintala, S. Generalized inner loop meta-learning. *arXiv preprint arXiv:1910.01727*, 2019.

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

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., and Song, D. A benchmark for anomaly segmentation. *arXiv preprint arXiv:1911.11132*, 2019a.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pp. 15663–15674, 2019b.

Hsu, Y.-C., Shen, Y., Jin, H., and Kira, Z. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Khosla, A., Jayadevaprakash, N., Yao, B., and Li, F.-F. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, 2009.

Lee, H., Hwang, S. J., and Shin, J. Self-supervised label augmentation via input transformations. In *International Conference on Machine Learning*, pp. 5714–5724. PMLR, 2020.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 7167–7177, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Lyle, C., van der Wilk, M., Kwiatkowska, M., Gal, Y., and Bloem-Reddy, B. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

Maclaurin, D., Duvenaud, D., and Adams, R. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pp. 2113–2122. PMLR, 2015.

Mohseni, S., Pitale, M., Singh, V., and Wang, Z. Practical solutions for machine learning safety in autonomous vehicles. *arXiv preprint arXiv:1912.09630*, 2019.

Mohseni, S., Pitale, M., Yadawa, J., and Wang, Z. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI Conference on Artificial Intelligence*, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nilsback, M.-E. and Zisserman, A. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1447–1454. IEEE, 2006.

Oberdiek, P., Rottmann, M., and Fink, G. A. Detection and retrieval of out-of-distribution objects in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 328–329, 2020.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Rafiee, N., Gholamipoor, R., and Kollmann, M. Unsupervised anomaly detection from semantic similarity scores. *arXiv preprint arXiv:2012.00461*, 2020.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pp. 14707–14718, 2019.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. Deep one-class classification. In *ICML*, pp. 4393–4402, 2018.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., and Kloft, M. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Sastry, C. S. and Oore, S. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *arXiv preprint arXiv:1912.12510*, 2019.

Sehwag, V., Chiang, M., and Mittal, P. {SSD}: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=v5gjXpmR8J.

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J. V., Ren, J., and Nado, Z. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

Tack, J., Mo, S., Jeong, J., and Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *34th Conference on Neural Information Processing Systems (NeurIPS) 2020*. Neural Information Processing Systems, 2020.

Techapanurak, E., Suganuma, M., and Okatani, T. Hyperparameter-free out-of-distribution detection using cosine similarity. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Winkens, J., Bunel, R., Roy, A. G., Stanforth, R., Natarajan, V., Ledsam, J. R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2017.

A. Framework Details

Previous work on self-supervised learning used ad-hoc heuristics for choosing data transformations for the training set (Hendrycks et al., 2019b; Golan & El-Yaniv, 2018; Tack et al., 2020). However, the choice of effective transformation(s) depends on the source distribution and heuristic approaches cannot scale up to diverse training distributions when there are many potential transformations to consider. To illustrates this, we train a ResNet-18 (He et al., 2016) with single and paired self-supervised transformations that are selected from a pool of seven transformations. The OOD detection results are reported in Figure 1-top with CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-30 (Hendrycks et al., 2019b) datasets as in-distribution and CIFAR-100, CIFAR-10, and Pets (Parkhi et al., 2012) as OOD test sets, respectively. The heatmap visualization presents a clear view of how different transformations (and the combinations of two) have a different impact on the OOD detection performance depending on the source distribution. For example, although rotation is the most effective transformation on CIFAR-10 and ImageNet-30, it is among the least effective ones for CIFAR-100. On the other hand, sharpening and color jittering are from the most effective transformations for CIFAR-100, but they perform worse on CIFAR-10.

After searching for transformation set $\mathcal{T}$, optimizing $\boldsymbol{\lambda}$ coefficients for the selected transformations can further improve representation learning. We assume that $\boldsymbol{\lambda}$ is a "meta-parameter" and we use a differentiable hyperparameter optimization algorithm (Maclaurin et al., 2015) for optimizing it as shown in Alg. 2 with details in Appendix A. Our optimization algorithm consists of an inner training updates that trains network parameters $\boldsymbol{\theta}$ using $\mathcal{L}_{\text{total}}$ on D^{in}_{train} for K steps. Given the current state of parameters $\boldsymbol{\theta}$, we update $\boldsymbol{\lambda}$ in the outer loop such that $\mathcal{L}_{\text{main}}(\boldsymbol{\theta})$ is minimized on D^{in}_{val}. Note that the gradient of $\mathcal{L}_{\text{main}}(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\lambda}$ is defined only through the gradient updates in the inner loop. Thus, the $\boldsymbol{\lambda}$ update in the outer loop requires backpropagating through the gradients updates in the inner loop which can be done using differentiable optimizers (Grefenstette et al., 2019).

Algorithm 1 Transformations $\mathcal{T}$ Optimization

Input: Transformation set $\mathcal{T}$
Output: Optimal transformation set $\mathcal{T}_{opt}$
while *not converged* **do**
  Sample a new $\mathcal{T}$ set.
  Train a classifier with $\mathcal{L}_{\text{total}}$ loss given the $\mathcal{T}$ transformations with $\boldsymbol{\lambda} = 1$.
  Calculate $\mathcal{L}_{\text{main}}$ on D^{in}_{val} as fitness measure.
  Update the acquisition function.

Algorithm 2 $\boldsymbol{\lambda}$ Weights Optimization

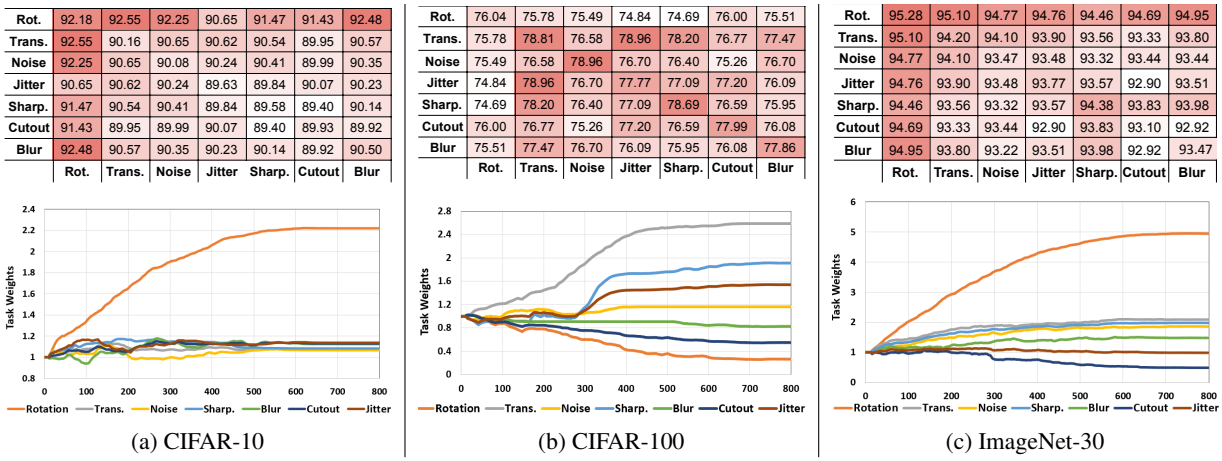
Input: Optimal set $\mathcal{T}_{opt}$, learning rate α, β, inner steps K
Output: Optimal $\boldsymbol{\lambda}$ coeficients
Initialize with $\boldsymbol{\lambda} = 1$.
while *not converged* **do**
  for K *steps* **do**
    $\boldsymbol{\theta} = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{total}}(\boldsymbol{\lambda}, \boldsymbol{\theta})$ on D^{in}_{train}
  $\boldsymbol{\lambda} = \boldsymbol{\lambda} - \beta \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{\text{main}}(\boldsymbol{\theta})$ on D^{in}_{val}

We use $K = 1$ step for the inner-loop optimization with SGD when updating $\boldsymbol{\theta}$ and we use Adam (Kingma & Ba, 2014) to update $\boldsymbol{\lambda}$. Figure 1-bottom presents $\boldsymbol{\lambda}$ values during optimization from an ablation study on three training sets and reveals the dependency and importance of optimum $\boldsymbol{\lambda}$ weights.

B. Experiments Details

Transformations Details In contrast to common data augmentations, we followed (Dosovitskiy et al., 2014) to apply transformations to the extreme degree. We used 7 different self-labeled transformations including rotation ($\{0°, 90°, 180°, 270°\}$) , translation (combinations of $\pm$ 30 % horizontal and vertical translations) , Gaussian noise (with standard deviations of {0, 0.3, 0.5, 0.8}), Gaussian blur (with sigmas of {0, 0.3, 0.5, 0.8}), cutout (rectangular cut with sizes of {0, 0.3, 0.5, 0.8}), sharpening (image blended with its convolution-based edges with alphas of {0, 0.3, 0.6, 1.0}), and color distortion (jittered brightness, contrast, saturation, and hue by rates of {0, 0.4, 0.6, 0.8}).

Training Details: All experiments use mini-batch size of 64, SGD optimizer with momentum of 0.9, initial learning rate of 0.1 (decayed using a cosine annealing schedule). Despite the set of self-supervised transformations, we still use a few data invariant "native augmentations" including random horizontal flip and small random crop and padding for the main supervised head. We use cross-entropy loss for all supervised and self-supervised branches with labels generated for self-supervised tasks. We apply all transformation tar-



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet-30

Figure 1. Ablation studies show that the optimum transformations set $\mathcal{T}$ and their training weights $\boldsymbol{\lambda}$ depend on the in-domain training set. **(Top)** Ablation study to measure effects of individual and paired transformations on OOD detection performance. **(Bottom)** Optimizing training weights ($\boldsymbol{\lambda}$) for auxiliary self-supervised tasks for each training set. Experiments are done in multi-class classification setup on different training sets.

gets (e.g., all four rotations for the rotation transformation) from $\mathcal{T}$ on every mini-batch during the training using the cross-entropy loss with labels generated for self-supervised tasks. Therefore, the final mini-batch is the base mini-batch size multiplied by the total number of supervised and self-supervised branches. We observe that multi-transformation learning benefits from longer training time similar to contrastive learning setups. So we train all multi-class classification models for 800 epochs and unsupervised models for 200 epochs.

$\mathcal{T}$ and $\boldsymbol{\lambda}$ Optimization: We first run the Bayesian optimization in Algorithm 1 to find the optimum $\mathcal{T}$ set, followed by the Algorithm 2 to optimize all training weights. Our framework finds the following transformation and weight pairs for the CIFAR-10 dataset {(Jitter, 3.2791), (Rotation, 2.7547), (Sharpening, 2.6906)} with $\lambda_0 = 4.0760$. In CIFAR-100 dataset: {(blur, 4.3051), (Jitter, 2.2612), (Translate, 2.9636), (Sharpening, 3.9634)} with $\lambda_0 = 8.6546$. In ImageNet-30: {(Noise, 5.3806), (Rotation, 3.3754), (Sharpening, 4.7626)} with $\lambda_0 = 9.3599$.

Dataset details: Our experiments are focused on image domain and we use CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-30 (Hendrycks et al., 2019b) in multi-class and unlabeled detection. CIFAR-10 and CIFAR-100 consist of 50,000 training and 10,000 test samples, respectively. ImageNet-30 is a selection of 30 classes from ImageNet (Deng et al., 2009) dataset that contains 39,000 training and 3,000 test samples. In one-class classification, we only used single classes of CIFAR-10 as training source (D_{train}^{in}) and the other classes as test set (D_{test}^{out}).

Evaluation Setup and Metrics: We evaluate the OOD detection performance using multiple diverse D_{test}^{out} sets to determine how well the detector can generalize on these unseen distributions, including test sets of SVHN (Netzer et al., 2011), TinyImageNet (Russakovsky et al., 2015)), Places365 (Zhou et al., 2017), LSUN (Yu et al., 2015), and CIFAR-10 (or CIFAR-100 when CIFAR-10 is the source training set) for CIFAR-10 and CIFAR-100 experiments and Pets (Parkhi et al., 2012), Flowers-101 (Nilsback & Zisserman, 2006)), CUB-200 (Wah et al., 2011), Dogs (Khosla et al., 2011), Food (Bossard et al., 2014) for ImageNet-30 experiments. We choose the area under the receiver operating characteristic curve (AUROC) (Davis & Goadrich, 2006) as a threshold agnostic metric in all evaluations. The AUROC will be 100% for the perfect detector and 50% for a random detector. In all evaluations, we use D_{test}^{out} (test set of the outlier dataset) as positive OOD samples and the D_{test}^{in} (test set of source training dataset) as negative samples for detection.

C. Detailed Performance Results

Classification Performance: Since our technique enjoys from learning domain knowledge through multiple data transformations, we first compare the supervised classification accuracy of our technique with state-of-the-art self-supervised and contrastive learning techniques. Figure 2 presents classification accuracy of CIFAR-10 and CIFAR-100 datasets trained on WideResNet-40-2, ResNet-18, and ResNet-50 networks. Our technique outperforms other techniques across both datasets in WideResNet and ResNet-18 networks and achieves competitive performance in ResNet-50, which indicates the effectiveness of transformation learning for better generalization. The improvement is more visible in smaller network sizes like WideResNet-40-2 (e.g.,

Table 2. Comparison of OOD detection results (AUROC %) with the supervised Baseline (Hendrycks & Gimpel, 2016) and state-of-the-art self-supervised learning (Geometric (Hendrycks et al., 2019b)), contrastive learning (SupSimCLR (Khosla et al., 2020), SSD (Sehwag et al., 2021), and CSI (Tack et al., 2020)), and our multi-task transformation learning

D^{in}_{train}	D^{out}_{test}	Detection AUROC						
		Baseline	Geometric	SupSimCLR	SSD+	CSI (ens)	Ours ($\mathcal{L}_{\text{ssl}}$)	Ours ($\mathcal{L}_{\text{ssl}} + \mathcal{L}_{\text{sup}}$)
ImageNet-30	Flowers 101	87.70	92.13	93.81	96.47	95.43 (96.18)	94.19	**97.18**
	CUB-200	85.26	90.58	89.19	**96.57**	93.32 (94.15)	93.34	96.44
	Dogs	90.30	93.25	95.16	95.23	96.43 **(97.64)**	93.63	97.07
	Food	78.93	85.09	83.61	85.48	88.48 (89.04)	82.51	**96.49**
	Pets	92.88	95.28	96.38	96.24	97.35 **(98.49)**	94.82	96.37
	Texture	86.98	92.16	98.70	94.86	97.63 **(98.54)**	93.99	96.56
	Average	87.01	91.42	92.81	94.14	94.77 (95.67)	92.08	**96.69**
CIFAR-10	SVHN	92.89	97.96	97.22	93.80	96.11 (97.38)	**99.92**	96.60
	Texture	87.69	96.25	94.21	94.05	95.92 (97.18)	**97.61**	96.91
	Places365	88.34	92.57	91.11	91.77	92.21 (93.11)	93.72	**98.73**
	TinyImageNet	87.44	92.06	92.10	90.28	91.33 (92.49)	92.99	**93.57**
	LSUN	89.87	93.57	92.13	94.40	92.91 (94.02)	**95.03**	94.12
	CIFAR-100	87.62	91.91	88.36	90.40	90.60 (92.06)	93.24	**94.07**
	Average	88.98	94.05	92.52	92.45	93.18 (94.37)	95.42	**95.67**
CIFAR-100	SVHN	79.18	83.62	81.55	83.60	79.22 (87.38)	87.11	**90.64**
	Texture	75.28	82.39	76.83	81.35	78.33 (78.31)	**85.47**	77.99
	Places365	76.07	74.57	75.37	79.16	77.15 (78.1)	77.87	**92.62**
	TinyImageNet	78.53	77.56	80.77	76.29	80.07 **(82.41)**	80.66	79.25
	LSUN	73.73	71.86	73.50	63.77	74.89 **(75.22)**	74.32	74.01
	CIFAR-10	78.26	74.73	73.28	73.94	75.98 (78.44)	79.25	**91.56**
	Average	76.84	77.46	76.88	76.35	77.61 (79.98)	80.78	**84.35**

3.7% gain over Baseline in CIFAR-10) and the smaller training sets (e.g., 8.64% gain over Baseline in CIFAR-100).

OOD Detection Performance: Table 2 presents main evaluation results with AUROC metric on ImageNet-30 (Hendrycks et al., 2019b), CIFAR-10, CIFAR-100 datasets each with 6 disjoint D^{out}_{test} sets. We compare our technique with the full supervised Baseline (Hendrycks & Gimpel, 2016) and current state-of-the-art methods including self-supervised learning (Geometric) (Hendrycks et al., 2019b), supervised contrastive learning (SupSimCLR) (Khosla et al., 2020) and SSD (Sehwag et al., 2021), and contrasting shifted instances (CSI) (Tack et al., 2020) and with its ensembled version (CSI-ens). All techniques are trained on ResNet-18 network with equal training budget, and all except SSD+ use their softmax prediction as OOD detection score in multi-class classification. Our final results in Table 2 are reported with both ($\mathcal{L}_{\text{ssl}}$) and ($\mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{ssl}}$) as the $\mathcal{L}_{\text{aux}}$ training loss to compare the impact of both loss functions. The averaged standard deviation (STD) for detection AUROC over six test sets from 5 runs of our techniques shows 0.18% STD for ImageNet-30, 0.13% STD for CIFAR-10, and 0.33% STD for CIFAR-100.

Table 3. Detection using only the main classification head vs. ensemble of auxiliary heads.

D^{in}	only main head	ensemble aux. heads
CIFAR-10	93.83	95.67
CIFAR-100	79.30	84.35
ImageNet-30	92.59	96.69

Advantage of Multi-task Transformation Learning: Experiment results in Table 2 show that both self-supervised and contrastive learning can learn in-domain normality to a comparable extend when using similar transformations. Yet, our technique outperforms other techniques by learning a diverse set of transformations optimized for the source D^{in}_{train}. Notably, unlike contrastive learning, our framework modulates the impact of different transformations via trainable $\boldsymbol{\lambda}$ instead of explicitly dividing transformations into positive and negative sets as done in CSI. Table 3 presents averaged OOD detection performance results when using only the main head compared to using an ensemble of detection scores from all auxiliary heads.

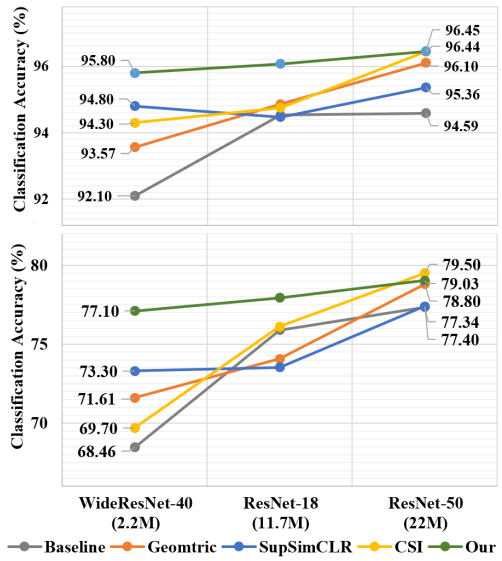


Figure 2. Classification accuracy on CIFAR-10 **(Top)** and CIFAR-100 **(Bottom)** datasets trained on fully supervised Baseline and state-of-the-art self-supervised and contrastive techniques trained on different networks, WideResNet-40-2 (2.2M parameters) and ResNet-18 (11.7M parameters)

The Choice of Transformations: Our results highlight the importance of selecting a suitable set of transformations as opposed to prior work which focused on geometric transformations in image domain. For example, experiments on the CIFAR-100 dataset show that the rotation transformation is not the optimum choice for invariance (see Figure 1) on this dataset, and our multi-tasked technique with $\mathcal{L}_{\text{ssl}}$ loss improved OOD detection on this dataset by a large margin of 3.32% AUROC compared to Geometric based on rotation and translation transformations. In fact, Table 2 shows that all prior work based on rotation transformation performs worse than the Baseline on the CIFAR-100 experiment when testing with CIFAR-10 as the D^{out}_{test} with the exception of CSI-ens. In some regards, the popularity of the rotation transformation in the OOD detection literature may indicate a case of unintentional hyperparameter tuning for CIFAR-10, that does not apply to other datasets. Table 4 presents an analysis of the effects of our transformation set $\mathcal{T}$ and $\boldsymbol{\lambda}$ weights optimizations on OOD detection performance. When training with all available transformations with equal $\boldsymbol{\lambda} = \mathbf{1}$ weights (first row), the OOD performance drops 2.76% compared to training with both $\mathcal{T}$ and $\boldsymbol{\lambda}$ optimized for the in-domain training set (last row).

Comparison to Data Augmentation Techniques: Lastly, we compare our results with RandAugment (Cubuk et al., 2020) and AutoAugment (Cubuk et al., 2019) techniques to observe how combinations of multiple light data augmentations compare with our heavy transformation learning tasks. Both data augmentation techniques achieve competitive classification accuracy on ResNet18 with 95.81% and 96.65% on CIFAR-10 (and 74.68 and 78.98% on CIFAR-100) for RandAugment and AutoAugment, respectively. Table 5 presents averaged AUROC results. Both augmentation techniques improved OOD detection over Baseline; however, they perform significantly lower than our proposed transformation learning.

Table 4. Ablation study on $\mathcal{T}$ and λ_n Optimizations

λ_n opt.	$\mathcal{T}$ opt.	AUROC
–	–	90.36
✓	–	90.93
–	✓	92.90
✓	✓	93.12

Table 5. Detection performance with our technique vs. data augmentation.

D^{in}	Baseline	AutoAug.	RandAug.	Ours
CIFAR-10	88.98	92.46	92.72	95.67
CIFAR-100	76.84	78.68	78.62	84.35

D. Detailed Robustness Results

We now analyze our model against four main criteria required from an ideal OOD detection technique, including i) zero-shot OOD training, ii) no hyperparameter dependency, iii) generalization to various unseen OOD distributions, and iv) robustness against test-time perturbations. We situate our proposed technique against a diverse range of state-of-the-art OOD detection techniques trained on the same network with the same training budget. Table 7 and Figure 3 present results for training on CIFAR-10 and testing on six D^{out}_{test} sets stated in Section 4. Note that this is not intended to be a ranking of different OOD detection techniques; instead, we are aiming to review trade-offs and limitations among different detection approaches. Albeit the simplicity, we show that our proposed approach outperforms the state-of-the-art techniques on robustness and generalizability criteria, in addition to the commonly used OOD detection performance that we discussed in the previous section.

Hyperparameters Dependency: While hyperparameter tuning for normal training of in-domain samples is done using a held-out validation set, the hyperparameter disentanglement is a crucial property for OOD detection. Specifically, an ideal detector should not be sensitive to hyperparameters tied to the target outlier distribution. Table 7 compares different techniques w.r.t their dependency on detection hyperparameters into three levels of high, low, and no dependency.

Techniques with high dependency like ODIN (Liang et al., 2017) and Mahalanobis (Lee et al., 2018) optimize their detection threshold using a validation set of D^{out} and hence

Table 6. Comparison of OOD detection results (AUROC %) with **(a)** unlabeled multi-class datasets on CIFAR-10, CIFAR-100, and ImageNet-30 and **(b)** Deep-SVDD (Ruff et al., 2018), DROCC (Goyal et al., 2020), GOAD (Bergman & Hoshen, 2020), Geometric (Hendrycks et al., 2019b), SSD (Sehwag et al., 2021), and CSI-ens (Tack et al., 2020) one-class classification techniques on CIFAR-10.

(a) Unlabeled CIFAR-10, CIFAR-100, and ImageNet-30

D^{in}	Geometric	SimCLR	SSD	CSI-ens	Ours ($\mathcal{L}_{\text{ssl}}$)
CIFAR-10	86.04	77.84	84.54	**91.99**	89.8
CIFAR-100	75.28	48.81	66.41	71.91	**83.95**
ImageNet-30	85.11	65.27	87.42	92.13	**96.57**

(b) One-class Detection on CIFAR-10

D^{in}	Deep-SVDD	DROCC	GOAD	Geometric	SSD	CSI-ens	Ours ($\mathcal{L}_{\text{ssl}}$)
Airplane	61.7	81.7	75.5	80.2	82.7	90.0	84.3
Automobile	65.9	76.7	94.2	96.6	98.5	99.1	96.0
Bird	50.8	66.6	82.4	85.9	84.2	93.3	87.7
Cat	59.1	67.2	72.1	81.7	84.5	86.4	82.3
Deer	60.9	73.6	83.7	91.6	84.8	94.8	91.0
Dog	65.7	74.4	84.8	89.8	90.9	94.4	91.5
Frog	67.7	74.4	82.8	90.2	91.7	94.4	91.1
Horse	67.3	71.3	93.4	96.1	95.2	95.2	96.3
Ship	75.9	80.0	92.6	95.1	92.9	98.2	96.3
Truck	73.1	76.2	85.1	92.8	94.4	97.9	92.3
Average	64.8	74.2	84.7	90.0	90.0	**94.3**	90.9

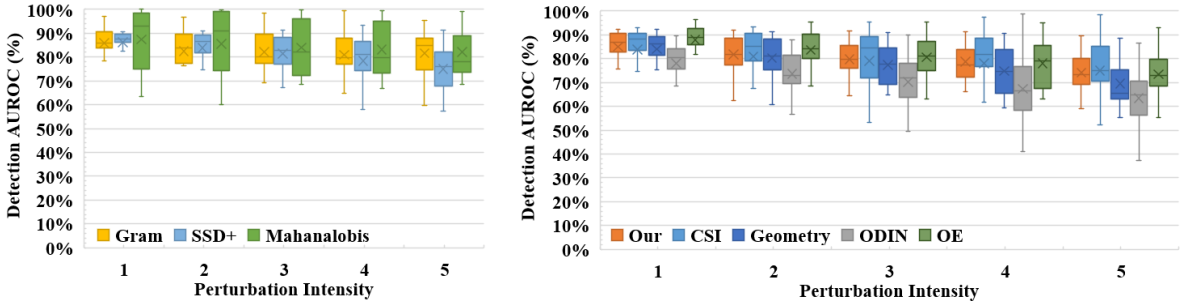
do not perform well under unseen or diverse mixture of outlier distribution. Over 3% performance drop is a clear indicator of strong D^{out} hyperparameter dependency as seen in Table 7 from the gap between the averaged detection performance on six D^{out}_{test} sets (Column 5) and detection performance under an equal mixture of the same test sets (Column 6). Techniques with low dependency do not use a subset of D^{out}_{test}, however, they depend on hyperparameters such as the choice of D^{out}_{train} set (e.g., Outlier Exposure (Hendrycks et al., 2018)), or hand-crafted self-supervised tasks (e.g., (Golan & El-Yaniv, 2018), (Hendrycks et al., 2019b)), or data augmentation (e.g., (Tack et al., 2020)) that requires post training D^{out}_{test} for validation. These techniques would suffer significantly in settings in which the new source training set is invariant to previous hand-crafted self-supervised tasks and augmentations as seen in Figure 1. On the other hand, techniques with no hyperparameters like Gram Matrices (Sastry & Oore, 2019), SSD (Sehwag et al., 2021), and our proposed framework bear no hyperparameter dependency on the choices of in-domain or outlier distribution. Note that many techniques, like ours, use a λ training hyperparameter to balance training between in-domain classification and auxiliary tasks, which is chosen based on the normal classification loss.

Zero-shot Training: A previous trend in OOD detection techniques considered using a subset of the target D^{out}_{test} for model tuning (e.g., ODIN (Liang et al., 2017) and Mahalanobis (Lee et al., 2018)) or using an auxiliary D^{out}_{train} set as a part of model training (e.g., Outlier Exposure (Hendrycks et al., 2018)). Although these techniques can achieve high detection performance with the right training set, having access to the specific D^{out}_{tune} for tuning or even any D^{out}_{train} for training the detector is not a realistic assumption in many setups. An efficient proposal to use these techniques is to integrate them into zero-shot techniques as presented by (Sastry & Oore, 2019; Hendrycks et al., 2019b) when D^{out}_{train} is available or to benefit from taking semi-supervised approaches to collect D^{out}_{train} samples as done by (Ruff et al., 2019; Sehwag et al., 2021).

Detection Generalizability: Recent work on OOD detection recognized the necessity of diverse D^{out}_{test} sets to evaluate the generalizability of OOD detection techniques (Mohseni et al., 2020; Winkens et al., 2020; Sastry & Oore, 2019) w.r.t. the distance between the D^{in}_{train} set and D^{out}_{test} set. Typically, near-OOD and far-OOD sets are chosen based on the semantic and imaging similarities between the in-domain and outlier distributions and in some cases measured by relevant similarity metrics (e.g., confusion log probability (Winkens et al., 2020)). Following the previous works, we chose CIFAR-100 as the near-OOD test distribution and SVHN as the far-OOD test distribution for detectors trained on CIFAR-10. While Table 7 shows high performance on far-OOD for all techniques, Gram Matrices, Mahalanobis, and ODIN show 20.5%, 10.9%, and

Table 7. Review of OOD detection criteria, averaged detection performance, and generalizability to unseen OOD test distributions (AUROC %) for a diverse set of OOD detection techniques.

Detection Technique	OOD Detection Criteria			Averaged Detection Performance	Generalizability Tests		
	Hyp.-Para. Dependency	Generalizable	Zero Shot		Mixed Distribution	Far-OOD	Near-OOD
ODIN (Lee et al., 2017)	High	–	–	91.15	88.10	96.70	85.80
Mahalanobis (Lee et al., 2018)	High	–	–	95.35	92.24	99.10	88.51
Outlier Exposure (Hendrycks et al., 2018)	Low	✓	–	96.24	96.88	98.76	93.41
Geometric (Hendrycks et al., 2019b)	Low	✓	✓	94.05	94.29	97.96	91.91
CSI-ens (Tack et al., 2020)	Low	✓	✓	94.37	94.10	97.38	92.06
SSD (Sehwag et al., 2021)	No	✓	✓	92.45	92.70	93.8	90.40
Gram Matrices (Sastry & Oore, 2019)	No	–	✓	94.17	95.08	99.50	79.01
Ours	No	✓	✓	95.67	95.55	96.60	94.07



(a) Distance-based techniques (b) Classification-based techniques

Figure 3. OOD detection robustness results for **(Left)** distance-based and **(Right)** Classification-based techniques with both D_{test}^{out} and D_{test}^{in} perturbed under 5 levels of intensity. $\times$ sign represents the mean AUROC.

10.6% detection performance drop for near-OOD distribution compared to the far-OOD test distribution, respectively. In comparison, our technique shows 2.53% performance between far-OOD and near-OOD test distributions.

Detection Robustness: Evaluating the effects of distribution shift on predictive uncertainty have been previously studied in (Snoek et al., 2019) for its real-world application. We investigate the effect of natural perturbations and corruptions proposed in (Hendrycks & Dietterich, 2019) on OOD detection performance. Figure 3 presents averaged OOD detection results for all 15 image distortions on 5 levels of intensity where both D_{test}^{in} and D_{test}^{out} are treated with the same distortion. Mahalanobis detector with access to perturbed D_{test}^{out} samples for tuning achieves the best robustness with 84.46% mean AUROC under all perturbations levels compared to SSD+, Outlier Exposure, Grams, Ours, CSI-ens, Geometric, and ODIN techniques with 81.02%, 80.70%, 80.19%, 79.92%, 79.45%, 77.22%, 70.58% AUROC, respectively. All techniques show more performance drop at higher levels of perturbation intensity. For example, Outlier Exposure, Geometric, and ODIN show the most mean detection AUROC drop over 14% by increasing perturbation intensity from level 1 to 5; compared to our technique with 10.79% AUROC drop. On the other hand, Gram and Mahalanobis distance-based detectors (Figure 3-a) show significantly less mean detection AUROC drop compared to classification-based detectors (Figure 3-b) with 4.23% and 5.24% AUROC drop, respectively. Appendix D reviews detection AUROC under different perturbation types for more insights.

Computation Complexity: Finally, we review the importance of test-time computation costs in real-world applications with resource and time constraints. Classification-based techniques can perform faster as they only use the model prediction probability from the main classification or an ensemble of multiple tasks. For example, the Baseline, Outlier Exposure, and ODIN technique only use the model softmax prediction as the OOD detection score. However, distance-based methods carry an overhead to measure the unknown D_{test}^{out} samples' distance from the seen D_{train}^{in} set. For instance, the CSI-ens (Tack et al., 2020) technique uses an ensemble of distances (cosine similarity distance) from multiple augmented copied of the D_{test}^{out} to the entire or a subset of D_{train}^{in}. On the other hand, SSD (Sehwag et al., 2021) measures the Mahalanobis distance between D_{test}^{out} and a pre-trained representation of D_{train}^{in} based on k-mean clustering which significantly improves the detection time. Table 8 presents a comparison between test-time detection time between under study classification-based and distance-based techniques running on the same system with

Table 8. Detection inference time averaged on five D_{test}^{out} sets.

Detector	Inference Time (s)
Baseline, OE	9.3s
Geometric	39.1s
CSI	18.7s
CSI-ens	163.2s
SSD	23.6
Gram	323.9
Ours	76.3s

a single RTX 1080ti GPU. We encourage future research to investigate diverse opportunities for distance-based and classification-based detectors from many applications of OOD detection.