A Experiments Details

Learning data invariances with semantic-preserving transformations is a common technique to incorporate inductive biases into representation learning and OOD detection. However, training an invariant network is challenging, and common data augmentation techniques do not guarantee that [15] and may even degrade training performance [2]. To address this issue, Tack et al. [20] suggested to use data augmentations as negative samples in contrastive training, Xiao et al. [23] proposed using multiple embedding sub-spaces for each augmentation, and Lee et al. [12] used label augmentation to train for image invariances. In contrast to these work, we propose to train semantic-preserving transformations in a multi-task training manner.

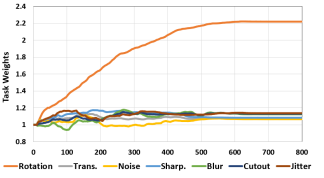
Dataset details: Our experiments are focused on image domain and we use CIFAR-10 [11], CIFAR-100 [11], and ImageNet-30 [9] in multi-class and unlabeled detection. CIFAR-10 and CIFAR-100 consist of 50,000 training and 10,000 test samples, respectively. ImageNet-30 is a selection of 30 classes from ImageNet [6] 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^{in}_{train}) and the other classes as test set (D^{out}_{test}). All D^{out}_{test} test images are resized to the D^{in}_{train} image size which is 32x32 in all CIFAR-10/100 experiments. In ImageNet-30 experiments, we first resize the images to 256x256 and then center crop to 224x224.

Training Details: All experiments are based on ResNet-18 network with 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 shifting 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 targets (e.g., all four rotations for the rotation transformation) from $\mathcal{T}$ on every mini-batch during the training, and therefore, the final mini-batch is the base mini-batch size multiplied by the total number of shifting transformations. We observe that learning multiple shifting transformations 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. Ablation studies are trained for only 200 epochs.

Evaluation Setup and Metrics: We evaluate the OOD detection performance using multiple diverse D^{out}_{test} sets to determine how well the detector can generalize on these unseen distributions, including test sets of SVHN [16], TinyImageNet [19]), Places365 [25], LSUN [24], and CIFAR-10 (or CIFAR-100 when CIFAR-10 is the source training set) for CIFAR-10 and CIFAR-100 experiments and Pets [18], Flowers-101 [17]), CUB-200 [22], Dogs [10], Food [1] for ImageNet-30 experiments. We choose the area under the receiver operating characteristic curve (AUROC) [5] 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^{out}_{test} (test set of the outlier dataset) as positive OOD samples and the D^{in}_{test} (test set of source training dataset) as negative samples for detection.

	Rot.	Trans.	Noise	Jitter	Sharp.	Cutout	Blur
Rot.	92.18	92.55	92.25	90.65	91.47	91.43	92.48
Trans.	92.55	90.16	90.65	90.62	90.54	89.95	90.57
Noise	92.25	90.65	90.08	90.24	90.41	89.99	90.35
Jitter	90.65	90.62	90.24	89.63	89.84	90.07	90.23
Sharp.	91.47	90.54	90.41	89.84	89.58	89.40	90.14
Cutout	91.43	89.95	89.99	90.07	89.40	89.93	89.92
Blur	92.48	90.57	90.35	90.23	90.14	89.92	90.50

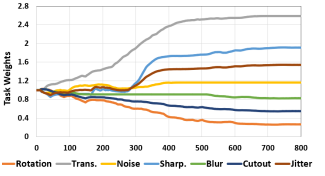
(a)



(b)

	Rot.	Trans.	Noise	Jitter	Sharp.	Cutout	Blur
Rot.	76.04	75.78	75.49	74.84	74.69	76.00	75.51
Trans.	75.78	78.81	76.58	78.96	78.20	76.77	77.47
Noise	75.49	76.58	78.96	76.70	76.40	75.26	76.70
Jitter	74.84	78.96	76.70	77.77	77.09	77.20	76.09
Sharp.	74.69	78.20	76.40	77.09	78.69	76.59	75.95
Cutout	76.00	76.77	75.26	77.20	76.59	77.99	76.08
Blur	75.51	77.47	76.70	76.09	75.95	76.08	77.86

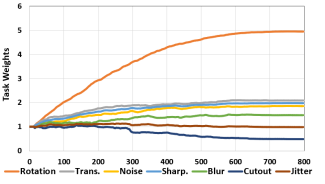
(c)



(d)

	Rot.	Trans.	Noise	Jitter	Sharp.	Cutout	Blur
Rot.	95.28	95.10	94.77	94.76	94.46	94.69	94.95
Trans.	95.10	94.20	94.10	93.90	93.56	93.33	93.80
Noise	94.77	94.10	93.47	93.48	93.32	93.44	93.44
Jitter	94.76	93.90	93.48	93.77	93.57	92.90	93.51
Sharp.	94.46	93.56	93.32	93.57	94.38	93.83	93.98
Cutout	94.69	93.33	93.44	92.90	93.83	93.10	92.92
Blur	94.95	93.80	93.22	93.51	93.98	92.92	93.47

(e)



(f)

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

B Additional Results

In this section, we provide additional ablation experiments to quantify contributions made by different parts of our model.

B.1 Ablation Studies

Figure 1 present additional ablation study results with CIFAR-10, CIFAR-100 and ImageNet-30 [9] datasets as in-distribution and CIFAR-100, CIFAR-10, and Pets [18] as example OOD test sets, respectively.

Transformation Optimization: Table 1-a presents an analysis of the effects of our transformation set $\mathcal{T}$ and λ weights optimizations on OOD detection when training on CIFAR-10 and testing on CIFAR-100 as D_{test}^{out}. When training with all available transformations with equal $\lambda = \mathbf{1}$ weights (first row), the detection AUROC drops 2.76% compared to training with both $\mathcal{T}$ and λ optimized (forth row). We also observe that when only enabling λ or $\mathcal{T}$ optimizations (second and third rows, respectively), the OOD detection performance is sub-optimal. We hypothesize that the gradient-based λ optimization only has access to training signal from a few gradient updates on θ and it does not capture the full effect of mini-batch transformations on OOD detection for long training. In contrast, Bayesian optimization in $\mathcal{T}$ optimization can capture the effect of each transformation on the full training and improves

access to D^{out}	λ_n opt.	$\mathcal{T}$ opt.	AUROC
–	–	–	90.36
–	✓	–	91.72
–	–	✓	92.90
–	✓	✓	**93.12**
✓	✓	✓	93.24

(a)

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**

(b)

Table 1: Ablations studies. **(a)** effectiveness of $\mathcal{T}$ and λ_n optimizations and optimization loss. **(b)** Detection using only the main classification head vs. ensemble of auxiliary heads.

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**

Table 2: Detection AUROC with our technique vs. data augmentation.

OOD detection by a large margin. Nevertheless, we obtain the best OOD detection result (fourth row) with both λ and $\mathcal{T}$ optimizations.

Our method avoids making any assumption on the availability of any D^{out} for model training or hyperparameter optimization unlike prior work such as ODIN [13], Mahalanobis [14], Outlier Exposure [8] detectors. In the absence of OOD samples for training, we utilize the in-domain validation loss as a proxy and we examine the generalization capability of our model to unseen examples to guide the optimization of shifting transformations.

In an ablation experiment in Table 1 last row, we use a disjoint subset of 80 Million Tiny Images dataset [21] as the source of unlabeled D^{out} and the KL divergence between D^{out} predictions and uniform distribution as the optimization objective for both $\mathcal{T}$ and λ optimizations. Interestingly, we only observe a slight performance improvement when assuming access to an auxiliary D^{out} for transformation optimization.

Advantage of Ensemble Detection Score: Table 1-b presents OOD detection performance when using only clean samples for the main classification head compared to using transformed samples for generating an ensemble of detection scores from all auxiliary heads. Results are based on averaged OOD detection AUROC over six D^{out}_{test} sets. We observe a clear performance increase when using the ensemble of auxiliary heads compared to the main classification head.

Comparison to Data Augmentation Techniques: A natural question to ask is whether the improvements in OOD detection could also be obtained with the supervised baseline that is trained with data augmentations. In Table 2, we compare our method with the supervised baseline trained with RandAugment [4] and AutoAugment [3] techniques as two popular augmentation techniques. We observe that both data augmentation techniques achieve competitive OOD detection performance on ResNet-18, improving upon the supervised baseline. However, they perform significantly lower than our proposed shifting transformation learning. Note that we could not use our shifting transformations as the data augmentation in supervised training as our transformations heavily change the input distribution and do not let training converge.

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**

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

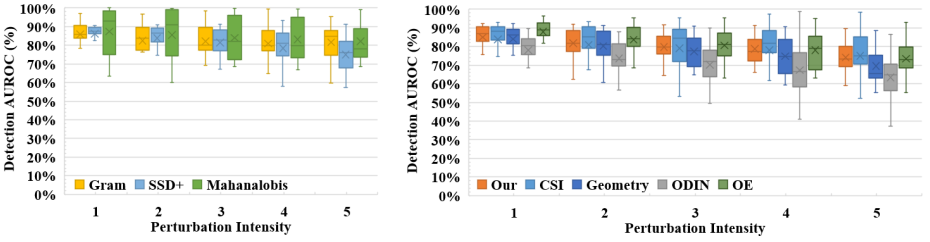
D^{in}	DROCC	GOAD	Geom.	SSD	CSI -ens	Ours ($\mathcal{L}_{\text{ssl}}$)
Airplane	81.7	75.5	80.2	82.7	90.0	84.3
Automobile	76.7	94.2	96.6	98.5	99.1	96.0
Bird	66.6	82.4	85.9	84.2	93.3	87.7
Cat	67.2	72.1	81.7	84.5	86.4	82.3
Deer	73.6	83.7	91.6	84.8	94.8	91.0
Dog	74.4	84.8	89.8	90.9	94.4	91.5
Frog	74.4	82.8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

(b) One-class Detection on CIFAR-10

Table 3: Comparison of OOD detection results (AUROC %) with different one-class classification and unlabeled multi-class datasets on CIFAR-10, CIFAR-100, and ImageNet-30.

B.2 Unlabeled OOD Detection

Table 3-a presents results for unlabeled multi-class detection in which averaged detection AUROC over the six D^{out}_{test} sets is outperforming state-of-the-art methods with a large margin except for the CIFAR-10 experiment. Table 3-b shows detailed one-class classification results for each of the CIFAR-10 classes as D^{in}_{train} and the remaining classes as D^{out}_{test}.



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

Figure 2: 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.

C Additional Robustness Results

Figure 2 presents a robustness evaluation for different distance-based and classification-based techniques under test-time perturbations. In this experiment, both in-domain and OOD samples are perturbed with 15 natural perturbations on 5 levels of intensity proposed by [7]. Our results indicate that all techniques show more performance drop at the higher levels of perturbation intensity. Specifically, distance-based detectors show significantly less performance drop compared to classification-based detectors like Outlier Exposure and Geometric with over 14% AUROC drop. Noticeably, the Mahalanobis detector with access to perturbed D_{test}^{out} samples for tuning maintains fairly high detection performance under all perturbation types. Looking into individual corruption types, all techniques perform best under image contrast distortion and the worst on glass blur corruption.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[5] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *ICML*, pages 233–240. ACM, 2006.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[8] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

[9] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15663–15674, 2019.

[10] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2. Citeseer, 2011.

[11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

[12] Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Self-supervised label augmentation via input transformations. In *International Conference on Machine Learning*, pages 5714–5724. PMLR, 2020.

[13] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[14] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[15] Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. *arXiv preprint arXiv:2005.00178*, 2020.

[16] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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

[18] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[20] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. 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.

[21] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 30(11):1958–1970, 2008.

[22] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[23] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021.

[24] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[25] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6):1452–1464, 2017.