This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Improving Shift Invariance in Convolutional Neural Networks with Translation Invariant Polyphase Sampling

Sourajit Saha
University of Maryland, Baltimore County
ssaha2@umbc.edu

Tejas Gokhale
University of Maryland, Baltimore County
gokhale@umbc.edu

Abstract

Convolutional neural networks (CNNs), widely deployed in several applications, contain downsampling operators in their pooling layers which have been observed to be sensitive to pixel-level shift, affecting the robustness of CNNs. We study shift invariance through the lens of maximum-sampling bias (MSB) and find MSB to be negatively correlated with shift invariance. Based on this insight, we propose a learnable pooling operator called Translation Invariant Polyphase Sampling (TIPS) to reduce MSB and learn translation-invariant representations. TIPS results in consistent performance gains on multiple benchmarks for image classification, object detection, and semantic segmentation in terms of accuracy, shift consistency, shift fidelity, as well as improvements in adversarial and distributional robustness. TIPS results in the lowest MSB compared to all previous methods, thus explaining the strong empirical results. TIPS can be integrated into any CNN and can be trained end-to-end with marginal computational overhead. Code: https://github.com/sourajitcs/tips/

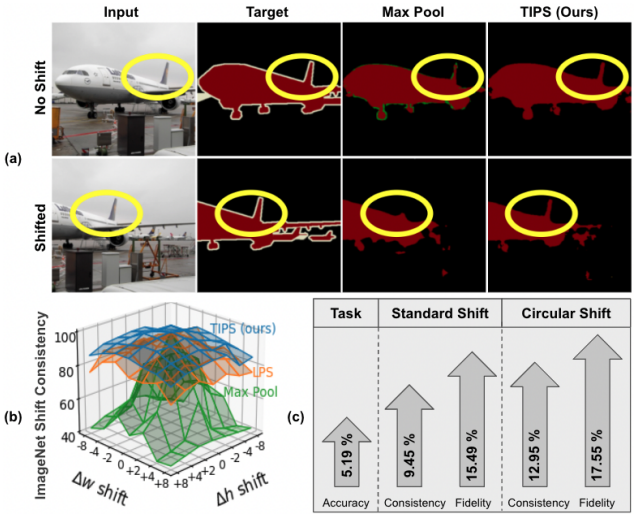

Figure 1. Translation-Invariant Polyphase Sampling (TIPS) is a pooling operator that improves shift invariance of CNNs. **(a)** An illustration of the improvements in semantic segmentation with TIPS; **(b)** Greater shift consistency of TIPS at higher degrees of pixel shift; **(c)** TIPS results in consistent and architecture-agnostic improvements in accuracy and four measures of shift invariance for image classification and semantic segmentation.

1. Introduction

Shift invariance is an ideal property for visual recognition models to ensure that outputs remain invariant to small pixel-level shifts in input images. Shift-invariance is desirable for image classification to ensure that categorical outputs are not affected by small horizontal and/or vertical pixel shift, and shift-equivariance is desirable for object detection and semantic segmentation to ensure that pixel-shift in the image results in equivalent shift in outputs. Recent studies have also found shift-invariant vision models to be more robust on *out-of-distribution* testing and adversarial attacks, making shift invariance an important facet of trustworthiness and risk assessment for visual recognition.

Although convolution layers are shift-equivariant [27], recent studies [1, 15] reveal that pooling operators such as max pooling, average pooling, and strided convolution break shift invariance by violating the Nyquist sampling theorem [36]. Since these pooling operators sample the highest feature activation in a pooling window, we study if this design choice can explain model's performance under pixel shift, by introducing the concept of Maximum-Sampling Bias (MSB) which denotes the tendency to pool larger values within pooling windows. We observe a strong negative correlation between MSB and shift invariance, i.e. models with higher MSB are the least shift invariant.

Informed by this finding, we design a pooling operator, Translation Invariant Polyphase Sampling (TIPS), to discourage MSB and improve invariance under shift transformations. We introduce two loss functions: $\mathcal{L}_{FM}$ – to discourage known *failure modes* of shift invariance and $\mathcal{L}_{undo}$ – to learn to *undo* standard shift. We demonstrate

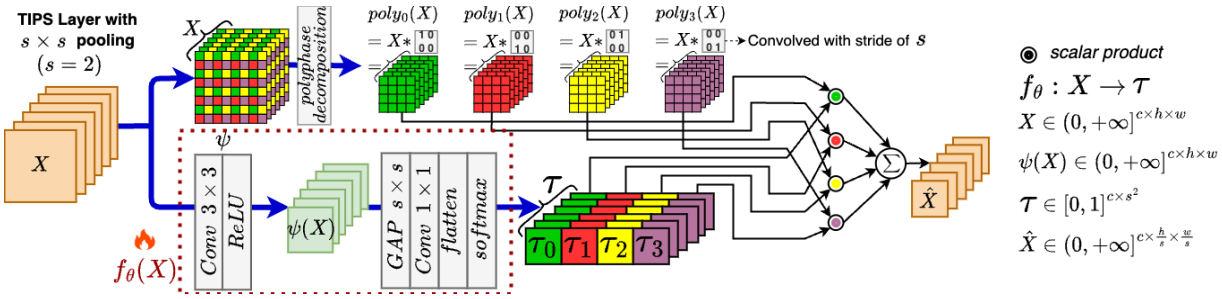

Figure 2. TIPS downsamples ReLU-activated intermediate feature map X into $\hat{X}$ with stride s and learns polyphase mixing coefficients τ (using a small fully convolutional function f_θ) which results in the output feature map as the weighted linear combination $\hat{X}$. The polyphase decomposition on input feature map X results in poly_i which are then mixed as a weighted linear combination with τ to compute $\hat{X}$.

that this approach consistently improves robustness under shift transformation on multiple image classification, object detection, and semantic segmentation benchmarks, outperforming data augmentation and contrastive learning strategies, and resulting in state-of-the-art performance in terms of accuracy, shift consistency, and shift fidelity under standard and circular shift transformations, while operating at a small computational overhead. In real world scenarios, standard shifts are more likely to occur than circular shifts; however, current literature largely focuses on circular shift invariance. We show that our method improves shift invariance on both circular shift and standard shift (shown in Figure 1). When tested on adversarial attacks, patch attacks, and natural corruptions, models trained with TIPS exhibit greater robustness than previous pooling operators.

2. Related Work

Robustness of CNNs has been examined under geometric transformations [8, 12], domain shift [46], attribute shift [16], adversarial attacks [32], and natural corruptions [20]. Distributional robustness of CNNs has been explored through random or learned data augmentation [17, 47], contrastive learning [23], and Bayesian approaches [7].

Dense Sampling and Anti-aliasing. Conventional sliding window downsampling in computer vision algorithms is typically applied with stride that is bigger than 1 which breaks shift equivariance [42]. Shift invariance can be improved through dense sampling [28] with dilated convolutions [48] with susceptibility to griding artifacts. Zhang *et al*. [49] suggest BlurPool to enhance shift invariance through anti-aliasing before downsampling, whereas Zou *et al*. [50] propose DDAC, to learn low pass anti-aliasing filter.

Polyphase Sampling. APS [4] and LPS [40] use polyphase sampling to satisfy the Nyquist sampling theorem [36] and permutation invariance which provides robustness against circular shifts. APS enhances shift invariance by sampling the highest energy polyphase index (ℓ_p norm) while LPS learns the sampling. LPS being sensitive to gumble

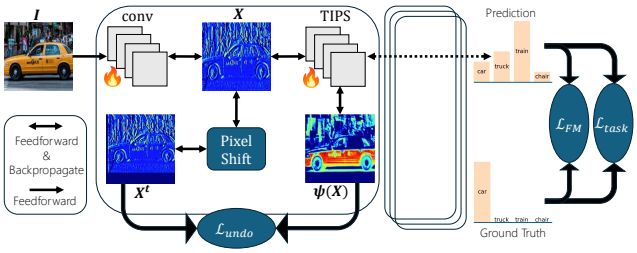

Figure 3. The *end-to-end* training pipeline with TIPS, regularization to undo shift $\mathcal{L}_{undo}$, regularization to discourage known failure modes $\mathcal{L}_{FM}$, and downstream task loss $\mathcal{L}_{task}$.

softmax temperature can sample polyphases to maximize downstream objective which does not consider shift invariance unless training data is shift-augmented. Evidence [4, 39, 49, 50] suggests that although polyphase sampling methods can improve shift invariance for circular shift, they still struggle to deal with standard shift; we show that TIPS can benefit both. The focus of this study is to improve robustness of CNNs against both standard and circular shifts, which constitute significant aspects of model evaluation.

3. Translation Invariant Polyphase Sampling

In this section, we discuss the design of our proposed TIPS pooling layer and the workflow to train CNNs with TIPS. Let $X \in \mathbb{R}^{c \times h \times w}$ be a ReLU-activated feature map, where c, h, w are the number of channels, height, and width. A pooling layer with sampling rate or stride s (also called $s \times s$ pooling) downsamples X into $\hat{X}$.

3.1. TIPS: A Learnable Pooling Layer

TIPS *learns* to sample polyphase decompositions of input feature maps X using two branches, as shown in Figure 2. In the first branch, polyphase components of X with stride s are computed similar to Chaman *et al*. [4]:

$$\text{poly}_{is+j}(X) = X[k, sn_1 + i, sn_2 + j], \quad (1)$$
$$\forall\, i, j \in \mathbb{Z}_0^{s-1}; k \in \mathbb{Z}_0^{c-1}; n_1 \in \mathbb{Z}_0^{\lceil \frac{h}{s} \rceil}; n_2 \in \mathbb{Z}_0^{\lceil \frac{w}{s} \rceil}.$$

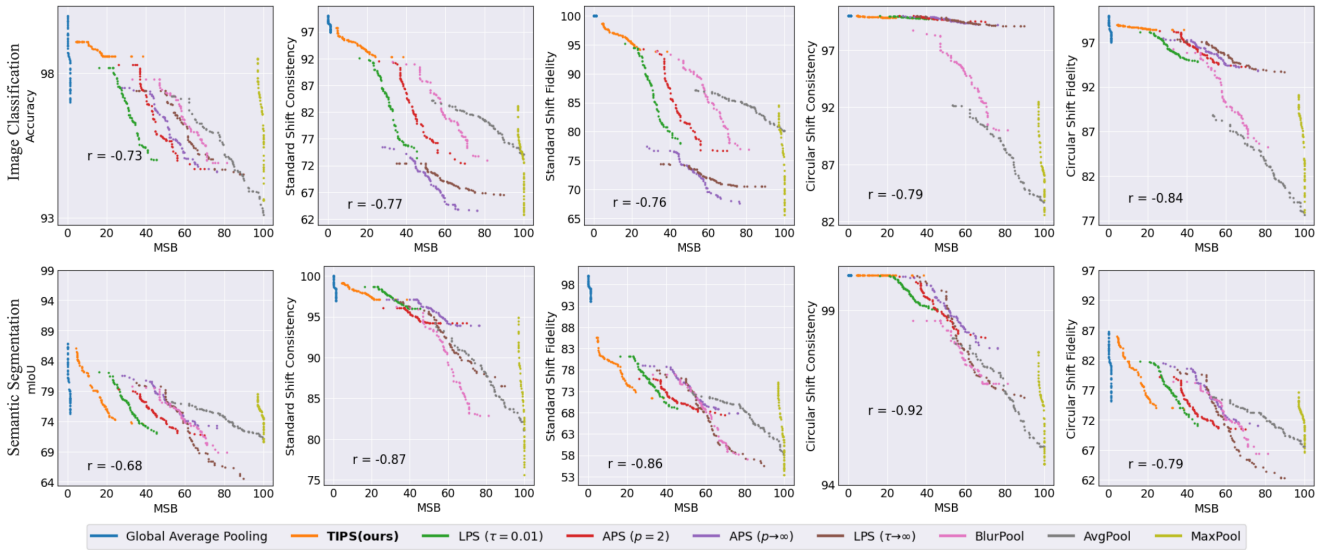

Figure 4. Our large-scale correlation study shows a strong negative correlation of performance with MSB (%) as indicated by Pearson's r. Linear clusters with negative correlation are also observed for points belonging to each pooling method.

Note that polyphase sampling can also be achieved by a strided convolution with a $s \times s$ kernel equal to 1 at index (i, j) and 0 elsewhere. A visualization of polyphase sampling with $s = 2$ is shown in Figure 2. The second branch is a function $f_\theta : X \to \tau$ that learns mixing coefficients $\tau \in [0, 1]^{c \times s^2}$. $\psi()$ is a 3×3 convolutional layer followed by ReLU activation. All operations in f_θ are shift invariant. The output of TIPS layer is computed as a weighted linear combination of the polyphase components of it's input:

$$\hat{X} = \sum_{i,j} \tau_{is+j} \, \text{poly}_{is+j}(X) \tag{2}$$

Regularizing TIPS to Discourage Known Failure Modes Chaman *et al.* [4] have shown that skewed mixing coefficients (*e.g.* $\tau = \{0, 1, 0, 0\}$ for $s=2$) are not robust against standard shift. TIPS with uniform mixing coefficients and LPS with higher softmax temperature (*e.g.* $\tau = \{0.25, 0.25, 0.25, 0.25\}$ for $s=2$) is identical to average pooling, which hurts shift invariance [49]. Based on these known failure modes, we introduce a regularization on the mixing coefficients in TIPS – the first term discourages skewed τ and the second term discourages uniform τ.

$$\mathcal{L}_{FM} = (\|\tau\|_2 - 1) + (1 - s^2 \|\tau\|_2) = (1 - s^2) \|\tau\|_2 . \tag{3}$$

Learning to Undo Standard Shift Although prior work have shown the benefits of using polyphase sampling to counter circular shift, performance still degrades with standard shifts due to information loss beyond the pooling boundary. To address this, we shift the input feature map with a random amount of vertical and horizontal standard shift sampled from uniform distribution $U(0, \frac{h}{10})$ and $U(0, \frac{w}{10})$ respectively to obtain a shifted X^t and design a regularization with the objective of undoing this shift:

$$\mathcal{L}_{undo} = \mathbb{E}_{h' \in h} \mathbb{E}_{w' \in w} [X^t_{h',w'} - \psi(X_{h',w'})]^2 \tag{4}$$

3.2. Training CNNs with TIPS

Let N be the number of training epochs. For the first ϵN epochs, we train only with the task loss $\mathcal{L}_{task}$ and the regularization to discourage failure modes $\mathcal{L}_{FM}$. For subsequent epochs, the undo regularization $\mathcal{L}_{undo}$ is introduced. The final training loss is the Lagrangian (with $\alpha \in [0, 1]$):

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{task} + \mathbb{1}(epoch < \epsilon N)\alpha\mathcal{L}_{undo} + \mathcal{L}_{FM} \tag{5}$$

We gradually introduce $\mathcal{L}_{undo}$ to stabilize training so that the downstream performance is not compromised. $\psi(X)$ contains a 3×3 convolution layer, ReLU, followed by global average pooling, 1×1 convolution, flattening, and a softmax operation to obtain τ as shown in Figure 2. Weights θ are Kaiming-normal initialized [19].

4. Maximum-Sampling Bias

In this section, we set up a framework to study shift invariance in CNNs, by defining maximum-sampling bias (MSB). Through a large-scale analysis, we show that MSB is negatively correlated with shift invariance.

Definition of MSB Existing pooling operators exhibit a common tendency to propagate signals based on activation strength. We denote this phenomenon as maximum-sampling bias (MSB), defined as the fraction of window locations for which the maximum signal value is sampled; a higher MSB indicates a higher probability of maximum signal values being selected during pooling.

Let $p()$ be a pooling operator in a CNN and s be the downsampling factor; *e.g.* $s=2$ for max-pooling with a 2×2 window. Let $X \in \mathbb{R}^{h \times w}$ be the input to a pooling layer. Applying a pooling operator $p()$ on X with downsampling fac-

	Method	Unshifted Acc. ↑	Standard Shift Consistency ↑	Standard Shift Fidelity ↑	Circular Shift Consistency ↑	Circular Shift Fidelity ↑
CNN (ResNet-101)	MaxPool	78.54±0.22	88.45±0.15	69.47±0.14	92.82±.14	79.20±0.15
	APS	83.01±0.08	91.37±0.06	75.85±0.04	**100.00±0.00**	83.01±0.08
	LPS	85.67±0.18	92.95±0.04	79.63±0.05	**100.00±0.00**	85.67±0.18
	TIPS	**86.78±0.19**	**94.27±0.14**	**81.80±0.15**	**100.00±0.00**	**86.78±0.19**
	BlurPool (LPF-5)	82.83±0.13	90.81±0.17	75.22±0.12	95.87±0.19	79.41±1.12
	APS (LPF-5)	83.52±0.03	92.00±0.20	76.84±0.11	**100.00±0.00**	83.52±0.03
	LPS (LPF-5)	86.74±0.09	93.38±0.17	80.99±0.04	**100.00±0.00**	86.74±0.09
	TIPS (LPF-5)	**86.91±0.13**	**94.55±0.06**	**88.20±0.07**	**100.00±0.00**	**86.91±0.13**
ViT	ViT-B/16 (I21k)	89.34±0.06	73.47±0.03	65.64±0.11	72.94±0.19	65.16±0.08
	ViT-L/16 (I21k)	90.75±0.15	**74.39±0.16**	**67.51±0.14**	73.85±0.06	67.02±0.19
	Swin-B (I21k)	**91.19±0.04**	72.14±0.15	65.78±0.17	**75.49±0.05**	**68.84±0.10**

(a) TinyImageNet

	Method	Unshifted Acc. ↑	Standard Shift Consistency ↑	Standard Shift Fidelity ↑	Circular Shift Consistency ↑	Circular Shift Fidelity ↑
CNN (ResNet-101)	MaxPool	76.31±0.18	89.05±0.19	67.05±0.06	87.56±0.13	66.82±0.17
	APS	76.07±0.15	90.95±0.13	69.19±0.13	**100.00±0.00**	76.07±0.15
	LPS	78.29±0.14	91.74±0.03	71.82±0.13	**100.00±0.00**	78.29±0.14
	TIPS	**80.24±0.09**	**92.87±0.08**	**74.52±0.18**	**100.00±0.00**	**80.24±0.09**
	BlurPool (LPF-5)	76.33±0.08	90.70±0.14	69.23±0.15	90.55±0.17	69.12±0.19
	APS (LPF-5)	76.49±0.08	91.23±0.17	69.78±0.05	99.98±0.00	76.41±0.06
	LPS (LPF-5)	78.31±0.05	92.49±0.15	72.43±0.08	**100.00±0.00**	78.31±0.05
	TIPS (LPF-5)	**81.36±0.10**	**93.11±0.03**	**75.75±0.14**	**100.00±0.00**	**81.36±0.10**
ViT	ViT-B/16 (I21k)	83.89±0.07	84.38±0.05	70.79±0.27	81.03±0.11	67.98±0.19
	ViT-L/16 (I21k)	85.06±0.02	83.19±0.12	70.76±0.17	81.64±0.15	69.44±0.14
	Swin-B (I21k)	**85.16±0.05**	**85.24±0.19**	**72.59±0.05**	**82.79±0.08**	**70.50±0.18**

(b) ImageNet

Table 1. Image classification performance on TinyImageNet and ImageNet averaged over five trials.

tor s results in an output $\hat{X} = p(X) \in \mathbb{R}^{\frac{h}{s} \times \frac{w}{s}}$. It is trivial to see that $MSB = 1$ for max-pool, as max pool selects $\hat{X}[i,j] = \max_{m,n} X[is+m, js+n]$. Average pooling produces $X[i,j] = \mathbb{E}_{m,n} x[is+m, js+n]$ and is equivalent to max-pooling if all values within the window are identical. For all other cases, the average value (less than the maximum) is sampled and thus $MSB \leq 1$. APS pooling [4] samples the polyphase component of X with maximum ℓ_p norm; $\hat{X}[i,j] = \max_{is+j} \{\|poly_j(x)\|_p\}_{i,j=0}^{s-1}$. As the polyphase function $poly()$ in APS and LPS is a monotonic function, it also exhibits a preference for sampling larger signals in the pooling window.

Negative Correlation between MSB and Shift Invariance. We investigate whether MSB is linked to shift invariance, by conducting a large-scale analysis of correlation between MSB and shift invariance on a number of visual recognition benchmarks with multiple CNN architectures and pooling methods [1]. We evaluated 576 models across different architectures, datasets, and pooling methods and conducted a correlation study as shown in Figure 4 with MSB on x-axis and performance metrics on the y-axis for both image classification and semantic segmentation. We observe a strong negative correlation between MSB and shift invariance metrics and also with downstream task performance (accuracy, mIoU). Circular consistency is more negatively correlated with MSB than standard consistency. Linear relationships are observed for points corresponding to each pooling method across architectures and datasets. Using Global Average Pooling (GAP) [19] with no spatial downsampling of the intermediate feature maps helps shift invariance, but with significant computational costs. TIPS achieves high shift invariance with marginal computational overhead (see supplementary material for comparison).

5. Experiments

We perform experiments on multiple benchmarks for both image classification, object detection, and semantic segmentation. For image classification, we evaluate shift invariance while for object detection and semantic segmentation we evaluate shift equivariance. However, following conventions used in prior work [40, 50], this is also referred to as "shift invariance" in the semantic segmentation results. For both classification, detection, and segmentation, we avoid using pre-trained CNNs since the pre-training step use non-shift invariant strided convolution and maxpool.

5.1. Image Classification Experiments

Datasets and Baselines We benchmark the performance of TIPS and prior work on five image classification datasets: CIFAR-10 [24], Food-101 [3], Oxford-102 [35], Tiny ImageNet [26], and ImageNet [25]. Results for Food-101 and Oxford-102 are in the Supplementary Material. Our baselines include MaxPool, APS (p=2), and LPS (τ=0.23), as well as BlurPool, APS, and LPS with anti-aliasing using $n \times n$ Gaussian low-pass filter (LPF-5). We also compare with three Vision Transformer (ViT) architectures: ViT-B/16, ViT-L-16 [11], and Swin-B [31] which are pre-trained on the larger ImageNet-21k dataset [10].

Hyperparameters For TIPS, we choose $\epsilon = 0.4$ and $\alpha = 0.35$ in Equation 5. All models are trained using an SGD optimizer with initial learning rate 0.05, momentum 0.9, and weight decay 1e-4 with early stopping. No models in our experiments were trained on shifted images. For each dataset-backbone pair, for fair comparison, TIPS and all baselines are trained with identical hyperparameters.

Evaluation Metrics In addition to reporting accuracy on the unshifted test set, we use the *consistency* definition from [50] which compares the predictions for two shifted images. However, as *consistency* does not consider the ground truth label (y) for evaluation, we introduce *fidelity* as a new metric. Note: x^{h_1,w_1} denotes image x shifted by $h \sim U(0, \frac{h}{8})$ vertically and $w \sim U(0, \frac{w}{8})$ horizontally.

$$\text{Consistency} = \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{(h_1,w_1),(h_2,w_2)} \mathbb{1}[f(x^{h_1,w_1}) = f(x^{h_2,w_2})]$$

$$\text{Fidelity} = \mathop{\mathbb{E}}_{x} \mathop{\mathbb{E}}_{(h_1,w_1),(h_2,w_2)} \mathbb{1}[y = f(x^{h_1,w_1}) = f(x^{h_2,w_2})]$$

[1] see supp.mat. for details on experiment setup for correlation analysis.

	Method	Unshifted Acc. ↑	Standard Shift Consistency ↑	Standard Shift Fidelity ↑	Circular Shift Consistency ↑	Circular Shift Fidelity ↑
CNN (ResNet-18)	MaxPool	91.43±0.04	87.43±0.05	79.94±0.05	90.18±0.03	82.45±0.08
	APS	94.02±0.07	92.89±0.08	87.33±0.05	**100.00**±0.00	94.02±0.07
	LPS	94.45±0.05	93.11±0.07	87.94±0.03	**100.00**±0.00	94.45±0.05
	TIPS	**95.75**±0.11	**98.38**±0.37	**94.20**±0.08	**100.00**±0.00	**95.75**±0.11
	BlurPool (LPF-5)	94.29±0.11	91.04±0.09	85.84±0.12	98.27±0.11	92.66±0.07
	APS (LPF-5)	94.44±0.09	93.25±0.13	88.06±0.17	**100.00**±0.00	94.44±0.09
	LPS (LPF-5)	95.17±0.12	94.87±0.08	90.09±0.15	**100.00**±0.00	95.17±0.12
	TIPS (LPF-5)	**96.05**±0.13	**98.65**±0.11	**94.75**±0.10	**100.00**±0.00	**96.05**±0.13
ViT	ViT-B/16 (I21k)	98.89±0.04	82.34±0.07	81.43±0.05	83.79±0.15	82.86±0.12
	ViT-L/16 (I21k)	99.15±0.02	82.72±0.09	82.01±0.08	**84.41**±0.11	**83.69**±0.06
	Swin-B (I21k)	**99.22**±0.03	**83.19**±0.07	**82.54**±0.05	84.05±0.04	83.40±0.04

Table 2. Image classification performance on CIFAR-10.

Results Tables 1 and 2 show strong dataset- and backbone-agnostic evidence for the efficacy of TIPS in terms of accuracy and shift invariance for both standard shift and circular shift. TIPS results in large gains in consistency and fidelity on standard shift, which was a challenge for prior work. It is important to note that TIPS with LPF-5 also improves upon prior work that uses LPF-5 anti-aliasing. For ViTs, shift invariance performance is inferior to CNNS, even though they consistently achieve higher accuracy. ViT architectures - despite being pre-trained on a very large scale dataset ImageNet21k (I21k) does not improve shift invariance which indicates that large-scale pre-training has no implications on shift invariance. While CNNs in general perform better on circular shift than standard shift, there is no such clear trend for ViT – for example, ViTs are more robust on standard shift for Oxford-102 and Tiny ImageNet and more robust on circular shift for the remaining datasets.

Comparison with other strategies for improving shift invariance. Besides pooling, we compare with methods that use data augmentation while training: standard shift, circular shift, and their combination, and two contrastive learning approaches: SimCLR [6] (self-supervised) and SupCon [23] (supervised) with the objective of aligning shifted samples closer and are used for downstream image classification. Table 3 shows that while data augmentation does not significantly improve performance, both contrastive learning methods outperform MaxPool on all metrics. TIPS, without any contrastive learning or data augmentation, results in greater shift invariance.

Applicability of TIPS to Vision Transformers We focus on improving shift invariance of CNNs – models that are widely used in real-world applications. Our method is based on polyphase downsampling and cannot be directly extended to vision transformers (ViT). However, in Tables 1 and 2, we demonstrate that ViTs are not shift invariant and our simple plug-in solution for CNNs (TIPS) achieves higher shift invariance than ViTs. We note that in vision transformers, three modules break shift invariance:

- *Patch embeddings* convert image patches into vectors using strided convolution (not shift invariant).

Strategy	Method	Unshifted Acc. ↑	Standard Shift Consistency ↑	Standard Shift Fidelity ↑	Circular Shift Consistency ↑	Circular Shift Fidelity ↑
Pooling (no anti-aliasing)	MaxPool	64.88	82.41	53.14	80.39	50.71
	DDAC	67.59	85.43	57.74	80.90	54.68
	APS	67.05	86.39	57.92	**100.00**	67.05
	LPS	67.39	86.17	58.07	**100.00**	67.39
	TIPS	**69.02**	**87.42**	**60.34**	**100.00**	**69.02**
Pooling (with LPF-5) (anti-aliasing)	BlurPool	66.85	87.43	58.54	87.88	58.75
	DDAC	66.98	86.92	58.22	80.35	53.82
	APS	67.52	87.02	58.76	99.98	67.51
	LPS	69.11	86.58	59.84	**100.00**	69.11
	TIPS	**70.01**	**87.51**	**61.27**	**100.00**	**70.01**
Data Aug.	circular	64.25	83.58	53.71	84.27	54.14
	standard	63.91	84.45	53.97	81.27	51.94
	both	**64.87**	**84.99**	**55.13**	**85.64**	**55.55**
Contrastive Learning	SimCLR	71.15	85.63	60.93	78.26	55.68
	SupCon	**72.49**	**86.17**	**62.46**	**81.75**	**59.26**

Table 3. Comparison of ResNet18 trained on ImageNet using pooling, data augmentation, and contrastive learning. Best per section: **bold**; lower than MaxPool: red ; overall best: cyan .

- *Positional encodings* for both shifted and non-shifted inputs are identical (amount of shift is not encoded).
- *Window-based self-attention* is computationally cheap, but applying local attention on window sizes larger than the amount of input shift causes token values to vary.

5.2. Semantic Segmentation Experiments

Datasets and Baselines We use the following datasets: PASCAL VOC 2012 [13], Cityscapes [9], Kvasir [22], and CVC-ClinicDB [2]. Results for Kvasir and CVC-ClinicDB are in the Supplementary Material. Our baselines include MaxPool, APS, LPS, BlurPool (LPF-3), and DDAC (groups g=8, LPF-3). BlurPool and DDAC [50] perform antialiasing by using a fixed low-pass filter and a learnable low pass group-wise convolution filter, respectively.

Hyperparameters We use SGD optimizer with an initial learning rate 0.01, momentum 0.9, weight decay 5e-4 with early stopping. We use DeepLabV3+ [5] with ResNet-18 backbone for the Pascal-VOC dataset and with ResNet-101 backbone for the Cityscapes dataset. For Kvasir and CVC-ClinicDB, we use a UNet [41] model with "Kaiming Normal" initialization. For TIPS, we use $\epsilon = 0.4, \alpha = 0.35$.

Evaluation Metrics For semantic segmentation, we compute consistency and fidelity by comparing the overlapping window for images with and without shift. Within the common crop, we compute the percentage of pixels for which the predicted category matches the ground truth.

Results Comparison of mIOU, consistency and fidelity in Table 4 shows that TIPS improves mIOU in comparison to all baselines on all four benchmarks. Consistent with our finding in image classification, we observe a sharper increase in shift consistency under standard shift than with circular. Models trained with TIPS pooling have higher fidelity on both standard and circular shifts, depicting the efficacy of TIPS in learning both shift invariant and high quality segmentation. In Figure 5, we compare the quality

Method	Anti-Alias	PASCAL VOC 2012 - DeepLabV3+ (ResNet-18)					Cityscapes - DeepLabV3+ (ResNet-101)				
		Unshifted	Standard Shift		Circular Shift		Unshifted	Standard Shift		Circular Shift	
		mIOU ↑	Consistency ↑	Fidelity ↑	Consistency ↑	Fidelity ↑	mIOU ↑	Consistency ↑	Fidelity ↑	Consistency ↑	Fidelity ↑
MaxPool	-	70.03	95.17	66.65	95.42	66.82	78.50	96.03	75.38	97.07	76.20
Blurpool	LPF-3	71.02	95.52	67.84	96.03	68.20	78.90	96.09	75.82	97.94	77.27
DDAC	LPF-3	72.28	96.77	69.95	95.98	69.37	79.52	96.28	76.54	98.21	78.09
APS	LPF-3	72.37	97.05	70.24	96.70	69.98	79.84	97.53	77.87	98.32	78.50
LPS	LPF-3	72.37	97.98	70.92	**100.00**	72.37	80.15	98.60	79.03	**100.00**	80.15
TIPS	LPF-3	**73.84**	**98.65**	**72.84**	**100.00**	**73.84**	**81.37**	**99.02**	**80.57**	**100.00**	**81.37**

Table 4. Semantic segmentation performance on Pascal VOC and Cityscapes datasets.

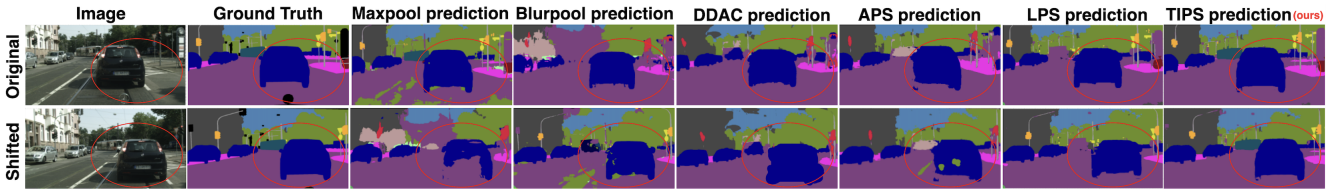

Figure 5. Qualitative comparison of segmentation masks predicted on original and shifted images. Regions where TIPS achieve improvements (i.e. consistent segmentation quality) under linear shifts are highlighted with circles.

Method	RetinaNet [29]								FasterRCNN [38]							
	Standard Shift				Circular Shift				Standard Shift				Circular Shift			
	AP ↑	ΔAP ↓	AP_{50} ↑	ΔAP_{50} ↓	AP ↑	ΔAP ↓	AP_{50} ↑	ΔAP_{50} ↓	AP ↑	ΔAP ↓	AP_{50} ↑	ΔAP_{50} ↓	AP ↑	ΔAP ↓	AP_{50} ↑	ΔAP_{50} ↓
MaxPool	36.5	2.2	56.7	5.1	36.4	2.3	56.7	5.4	37.6	2.9	59.0	6.4	37.8	3.4	59.3	4.3
BlurPool	35.2	1.4	55.1	3.4	36.3	1.6	56.3	5.4	37.8	2.0	58.7	4.6	38.3	3.7	59.5	1.3
APS	36.8	1.2	56.7	4.6	37.4	2.1	56.9	4.5	38.5	1.8	59.4	2.2	38.3	3.6	59.6	1.1
LPS	36.9	1.1	**56.8**	3.8	37.5	**1.9**	57.2	4.2	38.5	1.7	60.3	1.1	38.7	**2.4**	59.9	1.1
TIPS	**37.0**	**0.9**	**56.8**	**2.7**	**38.0**	2.0	**57.4**	**3.7**	**38.6**	**1.3**	**60.5**	**1.0**	**38.9**	2.5	**60.1**	**0.8**

Table 5. Standard and Circular Shift Equivariance for Object Detection on MS-COCO [30] validation set with RetinaNet, FasterRCNN.

of the masks predicted on shifted images when using prior pooling operators and TIPS. The areas highlighted with red circles demonstrate that TIPS segments objects with higher consistency than other pooling operators under pixel shifts.

5.3. Object Detection Experiments

Datasets, Baselines, and Hyperparameters Although there is limited work studying shift equivariance in object detection, we evaluated the efficacy of TIPS by extending the experimental settings from Manfredi *et al*. [33] to evaluate both standard and circular shift equivariance analogous to our semantic segmentation experiments. We train RetinaNet [29] and FasterRCNN [38] on the MS-COCO dataset [30] and present our findings in Table 5.

Evaluation Metrics and Results. In Table 5 we report AP, $\Delta AP (=$ bestAP$-$worstAP$)$, AP_{50}, $\Delta AP_{50} (=$ best$AP_{50}-$ worst$AP_{50})$ for both standard and circular shift invariance. Where higher value of AP and AP_{50} denotes superior object detection performance and lower value of ΔAP and AP_{50} denotes superior shift equivariant object detection performance. Table 5 demonstrates that TIPS consistently achieves lower ΔAP_{50} and higher AP_{50} which demonstrates that TIPS is also effective for shift equivariant object detection in comparison to existing pooling methods.

6. Analysis

We analyze the impact of our regularization through ablation studies, the impact of the choice of hyperparameters, and the effect of TIPS on various measures of robustness.

6.1. Effect of $\mathcal{L}_{undo}$ and $\mathcal{L}_{FM}$ Regularization

Figure 7 shows the impact of $\mathcal{L}_{undo}$ by visualizing $|\psi(X) - X^t|$ at different training stages. As training progresses, $\mathcal{L}_{undo}$ pushes $|\psi(X) - X^t|$ closer to 0, guiding TIPS to offset standard shift on intermediate features. Figure 8 quantifies this observation with accuracy and shift invariance measurements with different ϵ (optimal ϵ=0.4).

We perform an ablation study to examine the efficacy of each term in the loss function from Equation 5. Our results in Figure 6 reveal a clear trend: ○> ⋆ >△>□. Across all datasets and tasks, using both regularizations together (denoted by ○) leads to the highest accuracy, highest shift invariance, and lowest MSB. Moreover, only using one of $\mathcal{L}_{undo}$ or $\mathcal{L}_{FM}$ also improves performance compared to $\mathcal{L}_{task}$. These results demonstrate the impact of each component of our loss function on shift invariance and the inverse relationship between MSB and shift invariance.

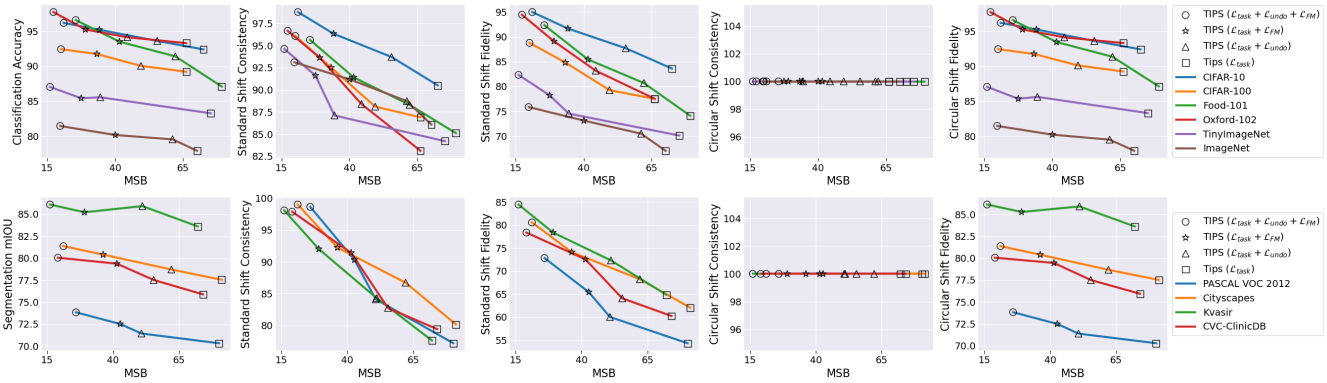

Figure 6. The top row shows results for six image classification datasets and the bottom row shows results for four semantic segmentation datasets. Regularization using both $\mathcal{L}_{undo}$ and $\mathcal{L}_{FM}$ results in the best performance.

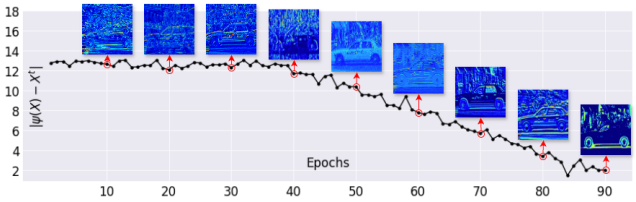

Figure 7. The effect of $\mathcal{L}_{undo}$ in terms of $|\psi(X) - X^t|$ and example feature maps (ResNet-101 with TIPS trained on ImageNet).

Figure 8. $\epsilon = 0.4$ (i.e. training without $\mathcal{L}_{undo}$ for the first 40% of epochs and with $\mathcal{L}_{undo}$ for the rest of the epochs) is optimal. ϵ>0.4 yields suboptimal shift invariance, but is better than low values of ϵ, demonstrating the impact of $\mathcal{L}_{undo}$.

	MaxPool		TIPS	
CNN	**Param.** ↓	**Consistency** ↑	**Param.** ↓	**Consistency** ↑
ResNet-18	**11.9**	87.43	**12.5**	**98.65**
DenseNet-BC (k=24) [21]	15.3	**90.02**	16.9	96.71
ResNet-34	21.3	88.93	22.3	98.60
EfficientNet-B7 [44]	64.0	89.14	67.3	93.67

Table 6. Simply increasing the number of parameters in CNN models does not improve shift invariance. However, using TIPS as the pooling operator leads to architecture-agnostic gains.

6.2. Size of Models and Number of TIPS Layers

Table 6 shows that with max pooling, even a > 400% increase in parameters (from 11.9M in ResNet18 to 64.0M in EfficientNet) only results in a 1.71% improvement in shift consistency. TIPS with ResNet18 (a 5% increase in number of parameters from 11.9M to 12.5M), leads to a 11.22% increase in shift consistency. This is observed across the board for different architectures. Figure 11 shows that as we train with more TIPS layers, shift invariance increases and MSB decreases, further indicating the efficacy of TIPS.

6.3. Performance at Different Levels of Shift

In Figure 9, we demonstrate shift invariance under all possible levels of shifts $d \in \{0, 1, ..., D\}$, and observe that circular shift consistency drops slower for TIPS than previous methods at higher degrees of shift. TIPS outperforms other pooling methods on average and at all degrees of shifts in terms of shift invariance as well as downstream accuracy.

6.4. Robustness Evaluation

Adversarial Attacks Recent studies reveal the vulnerability of CNNs against adversarial attacks if they are optimized for domain generalization [14] or shift invariance [43], and a trade-off between adversarial robustness and other forms of generalization [18, 34, 45]. We investigate ℓ_2 and ℓ_∞ robustness of TIPS (with ResNet-34 backbone) using PGD and FGSM attacks from Foolbox [37]. Figure 10 shows TIPS to have superior adversarial robustness compared to previous methods; better shift invariance is generally correlated with better adversarial robustness.

Patch Attacks We adopt the experiment setup from [4] where square patches are randomly erased from the input image and test models trained on the clean CIFAR-10 and ImageNet datasets using a ResNet-18 backbone. Figure 10 demonstrates that TIPS outperforms other methods (pooling and data augmentation) in robustness to such patch attacks. On ImageNet, shift consistency is more pronounced than other methods, especially when larger patches are erased.

Natural Corruptions We evaluated TIPS under an *out-of-distribution* setting, where models are trained on clean images, but tested on images with natural corruptions [20] due

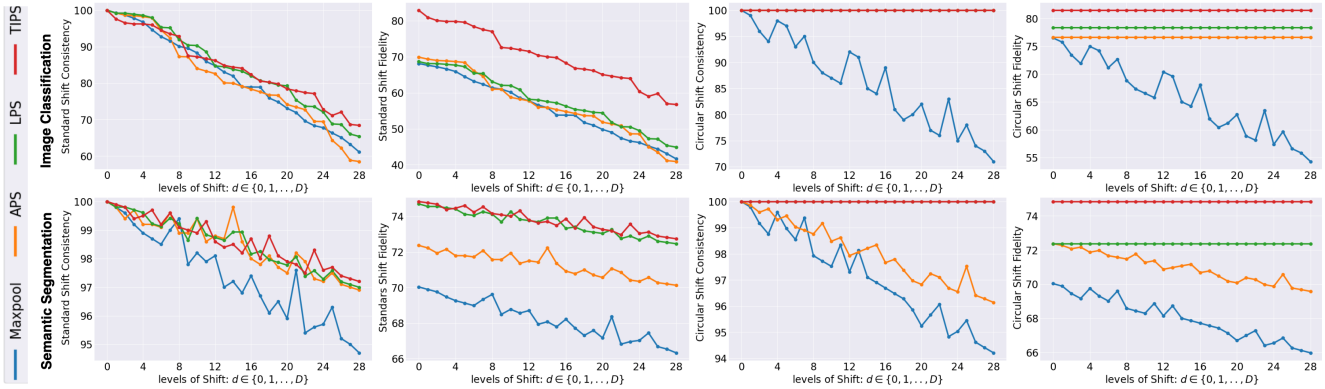

Figure 9. Shift invariance (consistency and fidelity) for standard and circular shifts on ImageNet classification and Cityscapes segmentation for varying degrees of shift. TIPS outperform previous methods in all evaluation metrics for shift invariance.

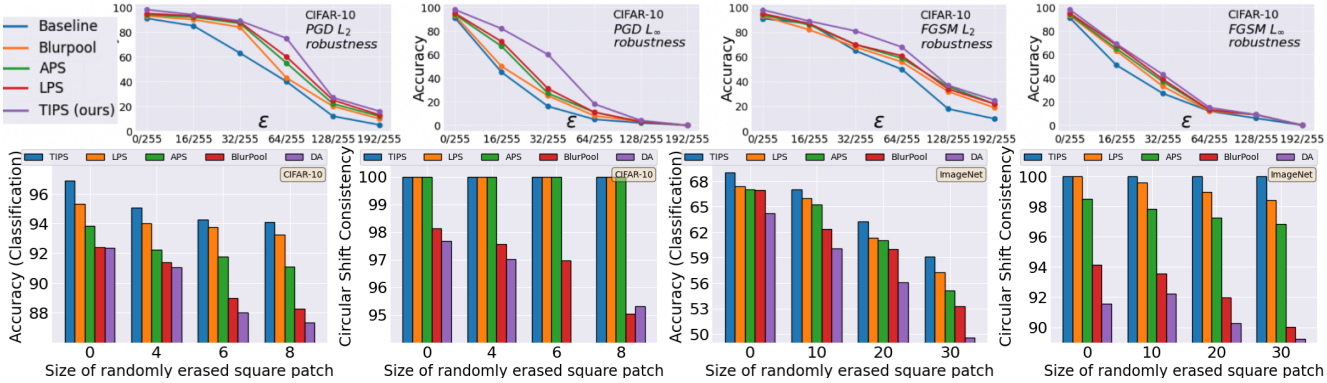

Figure 10. Evaluation of robustness of image classifiers. **Top:** Adversarial robustness under different levels (ε) of input perturbations. **Bottom:** Shift invariance under *patch attacks* (randomly erasing image patches). TIPS improves robustness of CNNs on both metrics.

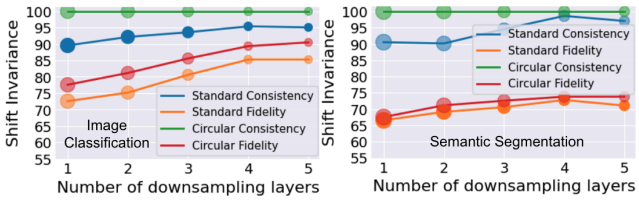

Figure 11. With more downsampling layers, MSB (bubble diameter in the plot) decreases and shift invariance increases.

Method	Clean ↓	mCE ↓	Noise	Blur	Weather	Digital
VGG-19	25.8	81.6	84.33	86.50	72.00	84.00
VGG-19 +**TIPS**	**25.1**	**81.1**	84.17	86.13	71.40	83.53
ResNet-18	30.2	84.7	88.67	87.75	79.25	83.25
ResNet-18 +**TIPS**	**28.7**	**83.9**	88.27	87.20	78.40	83.00

Table 7. Errors on clean (ImageNet) and corrupted (ImageNet-C) test sets. Models are trained only on clean data.

to noise, blur, weather, or digital artifacts. Table 7 shows that with TIPS, both the clean error and the mean corruption error (mCE) for VGG-19 and ResNet-18 improves.

7. Conclusion

Through a large-scale correlation analysis, we identify a strong inverse relationship between shift invariance of CNNs and the maximum sampling bias (MSB) of pooling operators. We find that optimizing model weights to reduce MSB is a good strategy for improving shift invariance. We introduce the Translation Invariant Polyphase Sampling (TIPS) pooling layer and regularization to promote low MSB, resulting in state-of-the-art results for shift invariance on several image recognition benchmarks, outperforming data augmentation and contrastive learning strategies. Our analysis reveals additional benefits of TIPS such as improved robustness to adversarial attacks and corruptions. Our work lays the foundation for further investigations on factors causing sensitivity to shifts of input signals in CNNs.

Acknowledgments. The authors acknowledge support from UMBC ORCA, UMBC HPCF GPU cluster, and Kowshik Thopalli for valuable feedback. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

[1] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019. 1

[2] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 5

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 4

[4] Anadi Chaman and Ivan Dokmanic. Truly shift-invariant convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3773–3783, 2021. 2, 3, 4, 7

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[7] Sheng Cheng, Tejas Gokhale, and Yezhou Yang. Adversarial bayesian augmentation for single-source domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11400–11410, 2023. 2

[8] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5

[10] 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. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

[12] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pages 1802–1811. PMLR, 2019. 2

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5

[14] Spencer Frei, Gal Vardi, Peter L Bartlett, and Nathan Srebro. The double-edged sword of implicit bias: Generalization vs. robustness in relu networks. *arXiv preprint arXiv:2303.01456*, 2023. 7

[15] Ziteng Gao, Limin Wang, and Gangshan Wu. Lip: Local importance-based pooling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3355–3364, 2019. 1

[16] Tejas Gokhale, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J Thiagarajan, Chitta Baral, and Yezhou Yang. Attribute-guided adversarial training for robustness to natural perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7574–7582, 2021. 2

[17] Tejas Gokhale, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 434–443, 2023. 2

[18] Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2705–2718, 2022. 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[20] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 2, 7

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 7

[22] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462, 2020. 5

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2, 5

[24] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009. 4

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 4

[26] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 4

[27] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989. 1

[28] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43:29–44, 2001. 2

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2

[33] Marco Manfredi and Yu Wang. Shift equivariance in object detection. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 32–45. Springer, 2020. 6

[34] Mazda Moayeri, Kiarash Banihashem, and Soheil Feizi. Explicit tradeoffs between adversarial and natural distributional robustness. *Advances in Neural Information Processing Systems*, 35:38761–38774, 2022. 7

[35] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 4

[36] Harry Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928. 1, 2

[37] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. 7

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 6

[39] Renan A Rojas-Gomez, Teck-Yian Lim, Minh N Do, and Raymond A Yeh. Making vision transformers truly shift-equivariant. *arXiv preprint arXiv:2305.16316*, 2023. 2

[40] Renan A Rojas-Gomez, Teck-Yian Lim, Alex Schwing, Minh Do, and Raymond A Yeh. Learnable polyphase sampling for shift invariant and equivariant convolutional networks. *Advances in Neural Information Processing Systems*, 35:35755–35768, 2022. 2, 4

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5

[42] Eero P Simoncelli, William T Freeman, Edward H Adelson, and David J Heeger. Shiftable multiscale transforms. *IEEE transactions on Information Theory*, 38(2):587–607, 1992. 2

[43] Vasu Singla, Songwei Ge, Basri Ronen, and David Jacobs. Shift invariance can reduce adversarial robustness. *Advances in Neural Information Processing Systems*, 34:1858–1871, 2021. 7

[44] Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 7

[45] Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[46] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 2

[47] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*, 2020. 2

[48] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 2

[49] Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019. 2, 3

[50] Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee. Delving deeper into anti-aliasing in convnets. In *BMVC*, 2020. 2, 4, 5