How Much is an Augmented Sample Worth?

Hamid Eghbal-zadeh **Gerhard Widmer**
LIT AI Lab & Institute of Computational Perception
Johannes Kepler University, Linz, Austria
hamid.eghbal-zadeh@jku.at

Abstract

Data Augmentation (DA) methods are widely-used in various areas of machine learning, and have been associated with the generalization capabilities of deep neural networks. Data Augmentation incorporates certain invariances and Inductive Biases (IBs) into models by applying transformations that are aligned with the task at hand, and extends the training samples beyond the training set. Models trained on augmented data are then equipped with the priors incorporated by these IBs, allowing them to better generalize onto unseen examples. In addition to inductive bias, data augmentation methods introduce randomness, to increase the variety of augmented data, and prevent overfitting. However, in the literature the success of DA has been mostly associated with the choice of IBs, and the role of randomness has been mostly ignored. In this work, we investigate the role of randomness on the regularization effects of DA, by taking the number of augmented samples required to achieve a certain performance improvement into account. We provide a hypothesis that regularization effects of DA are not only due to IBs used, but that randomness has a causal effect in regularizing models incorporating DA. Further, we provide an experimental protocol to test and validate our hypothesis, comparing different popular DA algorithms. Finally, using our proposed protocol we evaluate different DAs under limited randomness, measuring the alignment of their IBs w.r.t the data and the task at hand.

1 Introduction

Data Augmentation (DA) is one of the main building blocks of Deep Learning which has been used from the early stages [19] in many areas such as Machine Vision [32], Medical Imaging [3], Audio Processing [30], Speech Recognition [26], Natural Language Processing [11], and many more. DA introduces certain "Inductive Biases" (IBs) into models as prior, that are often aligned with the distribution of the data, and the task at hand; extending the training data by applying stochastic transformations on samples. Using DA, certain invariances can be built into models during learning, mainly without the need for altering the model architecture (e.g, randomly rotating training samples could result in a rotation-invariant model).

The main difference between DA algorithms has been reported in the literature to be in the different IBs they introduce to models. For example, some DA methods create new samples by linearly combining existing data and labels with a randomly chosen weight [40], while some others [9] erase some randomly-chosen parts of input to make the models robust towards missing information in data.

While many of these IBs are based on heuristics [39, 15, 40, 36, 9], some DA methods learn invariances and IBs that are suitable for the task from data [8, 2, 22, 13, 21], or learn a proxy distribution of the training data using generative models [1, 41], and further generate additional data by randomly sampling the generative model. In [4] DA has been studied more fundamentally, under a group-theoretic formulation. They explain that in empirical risk minimization (ERM), using DA leads to minimizing an augmented loss, which is the average of the original loss under a group action

35th Pre-registration workshop (NeurIPS 2021), Remote.

Figure 1: a) The causal graph of the relationship between IB and Data Augmentation Regularization (DAR), which has been mainly investigated in the literature. b) The extended causal graph including the causal effect of randomness (RND).

that is acting on data. They show that due to the averaging over the group orbit (e.g., all random rotations of a sample) DA results in variance reduction, with a cost of introducing a bias that is related to the IB used in the DA.

As discussed above, it can be seen that in addition to the IBs used in DA algorithms, they all have one thing in common: *Randomness*. However, in all the previous work in the literature, this factor has been mostly ignored; often a fixed amount of randomness has been kept throughout training, and the relationship between randomness and the regularization the DA introduces to models have not been studied (see Figure 1.a). In this work, we investigate the role of randomness on the regularization effects of DA by intervening on the randomness in the data augmentation process, through limiting the amount of randomness in DA, and studying the causal effects of randomness on the regularization effects of data augmentation (see Figure 1.b). More specifically, we would like to answer the following questions:

1. Does randomness have a causal effect in the regularization effects of data augmentation?
2. Can limiting the randomness effect in DA enable us to better evaluate inductive biases in data augmentation without the interference of randomness?
3. Can such evaluation under limited randomness be useful for measuring the alignment of an IB of a DA w.r.t the data and the task at hand?

2 Related Work

The availability of large amounts of data has been shown to be instrumental to the success of machine learning models. In [29] it was shown that the performance degradation in a model's prediction on data with corruption and perturbations (robust performance) disappears with the use of more data. Additionally, [31] demonstrate that training robust models requires *polynomially* more data than standard training. And in [23] the authors empirically demonstrate that by using larger models and more data, the performance of adversarial training of models can be improved.

The aforementioned work highlights the importance of additional data for training models. Where sufficient data is not available, data augmentation is a popular technique for creating additional training data, becoming an important factor in the success of data-hungry models such as deep neural networks. However, in a study [10] shows that the robustness of data augmentation methods may differ based on the inductive bias they infuse into models. They analyse data augmentation techniques such as heuristic-based augmentations and generative models, and show that data augmentations have a significant impact on the models not only in terms of the performance on a downstream task, but also in terms of the formation of their decision boundaries, as well as their adversarial robustness. [12] analyse Mixup augmentation [40], and show that it is primarily established empirically, and its effectiveness have not been explained in depth. They show that MixUp results in a form of "out-of-manifold regularization", which imposes certain "local linearity" constraints on the model's input space beyond the data manifold. They identify a limitation of MixUp, called "manifold intrusion" which is a form of under-fitting that results from conflicts between the synthetic labels of the mixed-up examples and the labels of original training data. [28] shows that heuristics-driven data

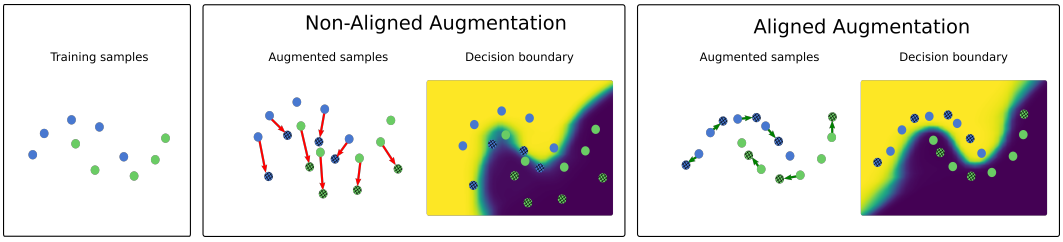


Figure 2: A simplified example of a non-aligned vs. aligned inductive bias (IB) in DA. The augmented examples are denoted with a hatched pattern.

augmentations are limited: these techniques tend to produce samples that are not complementary to the training set. In addition to the use of DA to incorporate invariance into models, there has also been a parallel body of work that focuses on designing architectures that are invariant/equivariant by design [5, 7, 6, 20, 24, 34, 35, 37, 38]. Such models can also achieve invariance, without the need of DA.

3 Definitions

Following [4], we define DA as a group action acting on the data space. Consider observations $X_1, \ldots, X_n \in \mathcal{X}$ (e.g., images) sampled i.i.d. from a probability distribution $\mathbb{P}$ on the sample space $\mathcal{X}$. Consider a group G of transforms (e.g., the set of all rotations of images), which *acts* on the sample space: there is a function $\phi : G \times \mathcal{X} \to \mathcal{X}, (g, x) \mapsto \phi(g, x)$, such that $\phi(e, x) = x$ for the identity element $e \in G$, and $\phi(gh, x) = \phi(g, \phi(h, x))$ for any $g, h \in G$ (we write $\phi(g, x) \equiv gx$ for notational simplicity).

Assuming that our data is invariant to certain transformations such as those in G, for any group element $g \in G$ and almost any $X \sim \mathbb{P}$, we have an "approximate equality" in distribution:

$$X \approx gX. \tag{1}$$

In supervised learning, approximate equality means that the probability of a sample being from a specific class is approximately the same as the probability of its augmented sample.

4 Problem Statement

In DA, usually training samples are augmented using some transformation g to extend the training set. But this transformation is often applied with randomness. For example, if the DA adds a small amount of noise N to a sample X, then this noise is randomly drawn from a distribution. Every time X is augmented, a new noise vector is observed, hence creating a different augmented data point $gX = X + N$. In this instance, the inductive bias introduced by g is that changing the sample by a small amount will not negatively affect the characteristics of the datapoint X w.r.t the task at hand (e.g, $X + N$ won't cross the class boundary). Hence adding this small amount of noise is **aligned** with the task and the data at hand. However, if the amount of additive noise increases so much that it changes the class of the augmented datapoint, this IB is **not aligned** with the task anymore.

We provide a simplistic example in Figure 2 to visualise the case explained above. The task in this example is to separate the two classes denoted by blue and green dots (left figure), and we follow the simplistic DA defined above (adding noise). In the middle, we show an example where the DA is *not aligned* with the task and data. As can be seen, augmented samples are often crossing the boundary between the two classes. On the right, we visualise applying a DA that is *aligned* with the task and data. Hence, the small additive noise never causes the augmented samples to cross the boundary. For both cases, we also show a decision boundary of a model trained on the original training plus the augmented samples. We can observe that the decision boundary of the model trained with *aligned DA* (right) better separates the two classes than the one trained with the *non-aligned DA* (middle).

5 Proposed Experimental Protocol

During training, every time a new transformation configuration is sampled, and consequently, a new augmented data is being created. Hence, in model updates, the original training samples are used several times, while each unique augmented sample is being used only once, due to the differences in the specific augmentation configuration chosen.

To evaluate the hypotheses detailed in Section 1, we propose a setup where an augmented sample is used to update the model several times. In the other words, we limit the number of unique augmented variants gX created from a training sample X. This way, we can control the amount of randomness in data augmentation, and consequently, in the underlying models using them. To this end, we evaluate models under the following regimes:
1) **Maximum Randomness Regime (MRR)**: each augmented sample is used in the model update only once. In the other words, every augmented version of a training sample is almost always slightly different from other versions, due to the randomness of the DA. This regime is the default of DAs currently used in the literature.
2) **Limited Randomness Regime (LRR)**: all augmented samples are kept in a pool, and are reused during training. The number of times a sample is reused is determined by the *percentage of randomness* p, and the number of times a training sample is augmented and added to the pool is determined by *number of augmented variants* k, for a model trained for e epochs, where $p = \frac{k}{e} \times 100$. This regime allows for control over the number of times a specific augmented sample has been seen by the model.

Under each regime, we will evaluate models for the task of image classification, and will report: 1) training and test classification accuracy, 2) training loss curves on original and augmented data during training, 3) Approximate Invariance erroR (AIR) on training and test sets as defined below:

$$\text{AIR}_{gX} = \frac{1}{s}\sum_{i=1}^{s} \mathbf{1}[f(X_i) \neq f(gX_i)] \tag{2}$$

where f is a supervised model (e.g, a neural network) trained to classify X, and $\mathbf{1}[a \neq b] = 1$ if $a \neq b$ and 0 otherwise. and 4) Augmentation Worth (AW) defined as:

$$\text{AW}_{g^k} = \frac{Acc_{g^k} - Acc_{g^0}}{k^{sgn(Acc_{g^k} - Acc_{g^0})}} \tag{3}$$

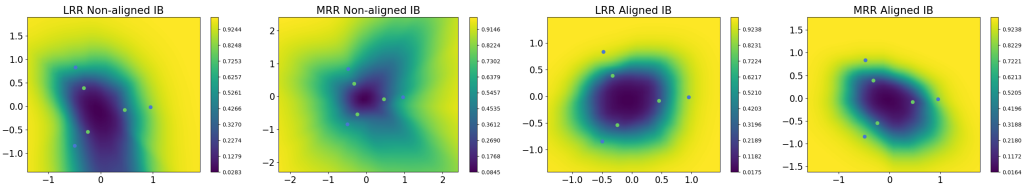
where Acc_{g^k} is the test classification accuracy, incorporating data augmentation introduced by g, k is the number of augmented variants used during training[1], and sgn is the sign function.

We report experiments on image and audio classification tasks. For images, we use ImageNet [18] and CIFAR100 [17], datasets, with two different architectures, VGG [33] and ResNet [14]. As for data augmentations, we study Mixup [12], Cutout [9], Cutmix [39], and classic augmentation (flipping, rotation, color jitter) [17, 14]. For Audio, we use the DCASE2020 dataset [25], with an Audio Resnet architecture [16], trained on spectrogram features, and for augmentations we use Mixup, SpecAugment [26], and time-rolling on the audio spectrograms. In Section 6, we provide some preliminary result using two toy datasets.

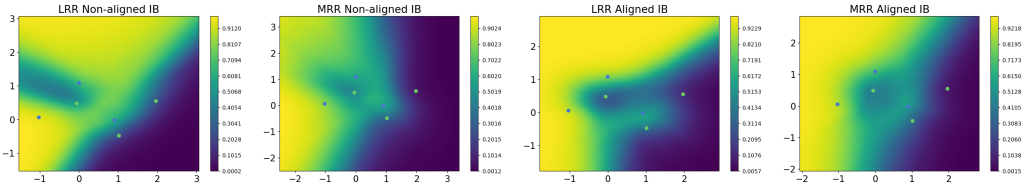
6 Preliminary Experiments

In this section, we use two toy datasets [27] to demonstrate how DA with a non-aligned IB will affect models, and how can we identify a non-aligned IB using *DA under LRR*. To this end, we train models[2] on each dataset, while applying DAs with either an Aligned IB (DA-AIB) or a Non-Aligned IB (DA-NAIB). For the DA with an Aligned IB, we add noise from a zero-mean Gaussian distribution with a small std, with careful consideration that augmented samples won't cross the boundary between classes. For the DA with a Non-Aligned IB, we add noise to samples without considering the distance between classes, intentionally forcing some augmented samples to cross the class boundary. We expect that models trained with DA-NAIB perform significantly worse than models trained with DA-AIB. The noise for DA-AIB is drawn from $\mathcal{N}([0,0],[0.1,0.1])$ and $\mathcal{N}([0,0],[0.2,0.4])$ on circles and

[1] e.g, Acc_{g^2} shows the test accuracy of a model where each training sample has been augmented twice.
[2] A neural network with 3 fully-connected layers and ReLU non-linearity.



(a) Cicles Dataset.



(b) Moons Dataset.

Figure 3: Decision boundaries of models for Circles and Moons datasets. The colums from left to right: 1) Non-Aligned DA under LRR, 2) Non-Alighted DA under MRR, 3) Aligned DA under LRR, 4) Alighted DA under MRR

moons datasets, respectively. For DA-NAIB, $\mathcal{U}([-1,-1],[1,1])$ was used in both datasets. $\mathcal{N}(\mu,\sigma)$ and $\mathcal{U}(a,b)$ denote multivariate normal and uniform distributions, respectively.

To test the effect of randomness, we reduce the amount of randomness by augmenting the 6 training samples once, and then train a model on these 12 original and augmented examples. In this setting, each training example has been augmented only once, but has been seen by the model 100 times during training (our reference is training with 100 epochs under MRR, as we will explain next). Similarly, the model has seen the original training data 100 times. We will refer to this setting as *DA under LRR*. As a control experiment, we also apply DA on the fly, that is, augment the training data with the probability of 0.5 in each update. We refer to this as *DA under MRR*. Each model is trained for 100 epochs on a training set of 64 samples, and is then evaluated on a test set of 1500 unseen samples. Each experiment is repeated 5 times, and the mean and standard deviation of classification accuracies, as well as the mean Augmentation Worths are reported in Table 1.

Preliminary Results: Looking at the results, we observe that under the MRR, models trained using DA with the intentionally non-aligned IBs, achieve similar results to those trained using DA with an aligned IB. However, under the LRR, models trained using DA with non-aligned IB perform significantly worse than those trained with an aligned-IB DA. We can also observe that the Augmentation Worth measure can reflect the value of the IB to the task at hand, showing a *larger negative worth* for the non-aligned DA. These results suggest that in order to determine whether an IB is aligned with the task and data, limiting the amount of stochasticity is crucial, and it can help us to better evaluate the IBs in DA and their effect on performance. In Figure 3 we additionally provide visualisation of the original training data, as well as the decision boundary[3] in models trained on our toy datasets. In these figures, we can observe that the decision boundaries of models trained with DA-NAIB under LRR (1st column) are different from MRR (2nd column); and samples of one class are much closer to the boundary in the model trained with DA-NAIB under LRR.

Acknowledgments and Disclosure of Funding

The LIT AI Lab is financed by the Federal State of Upper Austria. This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904. We thank Werner Zellinger from the Software Competence Center Hagenberg GmbH (SCCH), Rosanne Liu, and the ML Collective for the valuable discussions and their feedback on this work.

[3]We visualise the probability landscape of models for the data space, which represent their decision boundary.

Table 1: Classification Accuracy of models trained with different DAs on the toy datasets. No-augmentation model achieves 93.51 ± 5.83 and 92.02 ± 5.01 on circles & moons datasets, respectively. N.A.: Non-Aligned IB. A.: Aligned IB.

	LRR		MRR	
	N.A.	**A.**	**N.A.**	**A.**
	ACC			
Circles	76.97 ± 9.66	94.39 ± 6.48	81.13 ± 7.77	93.59 ± 3.32
Moons	77.74 ± 11.21	94.64 ± 1.69	87.77 ± 4.97	93.34 ± 1.78
	AIR			
Circles	0.37 ± 0.05	0.08 ± 0.03	0.40 ± 0.06	0.08 ± 0.02
Moons	0.27 ± 0.03	0.15 ± 0.04	0.25 ± 0.07	0.17 ± 0.02
	$\boldsymbol{AW_{g^1}}$		$\boldsymbol{AW_{g^{100}}}$	
Circles	-16.54	0.88	-1238.03	0.00
Moons	-14.28	2.62	-425.53	0.01

References

[1] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Augmenting image classifiers using data augmentation generative adversarial networks. In *International Conference on Artificial Neural Networks*, pages 594–603. Springer, 2018.

[2] Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning invariances in neural networks. *arXiv preprint arXiv:2010.11882*, 2020.

[3] Marcus D Bloice, Peter M Roth, and Andreas Holzinger. Biomedical image augmentation using augmentor. *Bioinformatics*, 35(21):4522–4524, 2019.

[4] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

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

[6] Taco S Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. *Advances in neural information processing systems*, 32:9145–9156, 2019.

[7] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016.

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[10] Hamid Eghbal-zadeh, Khaled Koutini, Paul Primus, Verena Haunschmid, Michal Lewandowski, Werner Zellinger, Bernhard A Moser, and Gerhard Widmer. On data augmentation and adversarial risk: An empirical analysis. *arXiv preprint arXiv:2007.02650*, 2020.

[11] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*, 2017.

[12] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019.

[13] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. *arXiv preprint arXiv:1911.06987*, 2019.

[14] 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.

[15] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[16] Khaled Koutini, Hamid Eghbal-zadeh, and Gerhard Widmer. Receptive field regularization techniques for audio classification and tagging with deep convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

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

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[20] Dmitry Laptev, Nikolay Savinov, Joachim M Buhmann, and Marc Pollefeys. Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 289–297, 2016.

[21] Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxing Yang. Dada: Differentiable automatic data augmentation. *arXiv preprint arXiv:2003.03780*, 2020.

[22] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *Advances in Neural Information Processing Systems*, pages 6665–6675, 2019.

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[24] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5048–5057, 2017.

[25] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018.

[26] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

[29] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[30] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[31] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

[32] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*, 2019.

[35] Mark van der Wilk, Matthias Bauer, ST John, and James Hensman. Learning invariances using the marginal likelihood. In *Advances in Neural Information Processing Systems*, pages 9938–9948, 2018.

[36] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[37] Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *Advances in Neural Information Processing Systems*, pages 7366–7378, 2019.

[38] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

[39] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[40] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[41] Xiaofeng Zhang, Zhangyang Wang, Dong Liu, and Qing Ling. Dada: Deep adversarial data augmentation for extremely low data regime classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2807–2811. IEEE, 2019.