TSynD: Targeted Synthetic Data Generation for Enhanced Medical Image Classification

Leveraging Epistemic Uncertainty to Improve Model Performance

Joshua Niemeijer[3], Jan Ehrhardt[1], Hristina Uzunova[2], and Heinz Handels[1,2]

[1]Institute of Medical Informatics, University of Lübeck, Germany
[2]German Research Center for Artificial Intelligence, Lübeck, Germany
[3]German Aerospace Center, Braunschweig, Germany
Joshua.Niemeijer@dlr.de

Abstract. The usage of medical image data for the training of large-scale machine learning approaches is particularly challenging due to its scarce availability and the costly generation of data annotations, typically requiring the engagement of medical professionals. The rapid development of generative models enables us to tackle this problem by generating large amounts of realistic synthetic data for the training process. However, randomly choosing synthetic samples, might not be an optimal strategy.
In this work, we investigate the targeted generation of synthetic training data, in order to improve the accuracy and robustness of image classification. Therefore, our approach aims to guide the generative model to synthesize data with high epistemic uncertainty, since large measures of epistemic uncertainty indicate underrepresented data points in the training set. During the image generation we feed images reconstructed by an auto encoder into the classifier and compute the mutual information over the class-probability distribution as a measure for uncertainty. We alter the feature space of the autoencoder through an optimization process with the objective of maximizing the classifier uncertainty on the decoded image. By training on such data we improve the performance and robustness against test time data augmentations and adversarial attacks on several classifications tasks.

Keywords: synthetic data generation · generalization · robustness

1 Introduction

Creating imaging datasets for training deep neural networks consists of three major steps: data acquisition, data selection, and data labeling. These steps are especially challenging in the domain of medical image processing. Data acquisition is often limited and data delivery is impaired by privacy regulations. Also, relevant image data might further be bound by the frequency of certain

medical scenarios (e.g. rare diseases). Another main obstacle is the costly and time-intensive data labeling, which often requires medical professionals.

In this work, we address these problems by utilizing generative models to extend the distribution of the given training data. More specifically, we aim to create data points that represent missing parts of the relevant distribution. Such data points are marked by a high epistemic uncertainty when processed by a discriminative model (i.e. a classifier network). In this work, we present a novel approach called TSynD (***T****argeted* ***Syn****thetic* ***D****ata generation*): a method specifically designed to steer the generation process in order to synthesize data points from the missing parts of the training distribution and utilize them during the training of downstream tasks (here: classifier). For data generation, TSynD employs an autoencoder model that is able to reconstruct existing images of the training distribution. The autoencoder consists of an encoder that transforms the image into the latent space and a decoder that reconstructs the input image from the latent space. TSynD aims to optimize the latent space representations of the autoencoder in a way that the decoded images maximize the epistemic uncertainty in a given classifier. By further training the classifier on these images, we receive classification models that generalize well to unseen data, a feature that is especially important in medical image processing. We, therefore, show the performance of the TSynD method on several medical classification datasets. In order to simulate the smaller training datasets, typical for the medical image community, as well as recreate cases of out-of-distribution samples, this work primarily considers a low-data training setting. We provide experiments to investigate the out-of-distribution performance through random test time augmentations and investigate the robustness to adversarial attacks. Further, the robustness of our approach is investigated visually by applying class activation explanation approaches and we are able to show that a classifier trained with TSynD utilizes more meaningful image information.

2 Related Work

In our work, we present a novel method for training networks that generalize to out-of-distribution samples. We employ an adaptive data generation process that is based on generative models.

Data augmentation is a commonly used way of extending the given training distributions mostly in an untargeted way. As stated by Zhou et al. [21], there are four different types of data augmentation: Firstly, there are image transformations, which consist of e.g. random flipping, rotation, or color augmentations. Secondly, model-based augmentations, which, e.g., consist of random convolutions [19] or other augmentation networks like style transfer networks [1] or learnable image generators [22, 11]. Thirdly, latent space augmentations directly augment the latent space distributions of the tasks (e.g. classification) model as in Zhou et al. [23]. Finally, some approaches utilize adversarial gradients.

Adversarial gradient augmentation and, more specifically, task adversarial augmentations are the most similar category to our approach. The approaches

of Sinha et al. [15], Volpi et al. [17], and Qiao et al. [13] utilize adversarial attacks by computing adversarial gradients w.r.t. to the task network in order to alter the training images. The alternation is hereby done by optimizing the pixel values of the image as parameters themselves. Such methods are often accused of introducing noise perturbations instead of larger image alternations e.g. representing domain shifts (Zhou et al. [21]).

In contrast, we optimize the latent space of a generative model as parameters of the image generation. The intuition behind this is that the latent space is a more abstract image representation, thus, its altering would lead to more complex and meaningful image changes. The work of Stutz et al. [16] is, therefore, the most related to our approach. They employ a VAE-GAN model [14] to represent the manifold, and similar to us, compute perturbations on the latent space to create adversarial examples. In contrast to us, they do not maximize the uncertainty, but rather maximize the cross-entropy loss. The optimization effectively changes the predicted label and thus introduces the need for constraints, to maintain the true image class. Our approach is inspired by active learning [9] and puts the main focus on generating images that maximize the epistemic uncertainty of the given classifier. This brings the advantage of not requiring any additional constraints. The work of Li et al. [7] also utilizes an autoencoder. Similar to us they compute perturbations on the latent space by e.g. using random noise. In our work, we utilize the randomly perturbed latent space as a starting point for our optimization to increase data diversity.

3 Methods

Given a labeled data set $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$ $(x_n \in \mathcal{X}, y_n \in \mathcal{Y})$, our approach aims to use the generative model in a targeted way to make a classification network $\mathcal{C} : \mathcal{X} \to \mathcal{Y}$ more robust to missing parts of the data distribution that are not included in the labeled set $\mathcal{D}$. The generative model, e.g. an autoencoder, consists of an encoding function $f_{\text{enc}} : \mathcal{X} \to \mathcal{Z}$ and a decoder $f_{\text{dec}} : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{Z}$ is the latent space. It can be trained in an unsupervised way using a larger amount of unlabeled data from the domain $\mathcal{X}$. Inspired by active learning strategies, we utilize the generative model to create images that maximize the epistemic uncertainty of our classification network $\mathcal{C}$. Samples yielding a high epistemic uncertainty represent missing parts of the learned distribution, and training on such samples can make the classification network more robust. Figure 1 shows an overview of our approach: starting by the encoded labeled images, the latent code z is optimized to reconstruct new images that locally maximize the epistemic uncertainty of the classifier $\mathcal{C}$. The newly generated samples are now used together with the labeled images for the training of the classifier.

3.1 Estimation of the epistemic uncertainty

Given the classifier $\mathcal{C}$ with model parameters θ, the predictive class probability distribution for a decoded image $\hat{x} = f_{\text{dec}}(z)$ with latent code z is computed by

$$p(y|\hat{x}, \theta) = p(y|z, \theta) = \sigma(\mathcal{C}(f_{\text{dec}}(z); \theta)),$$

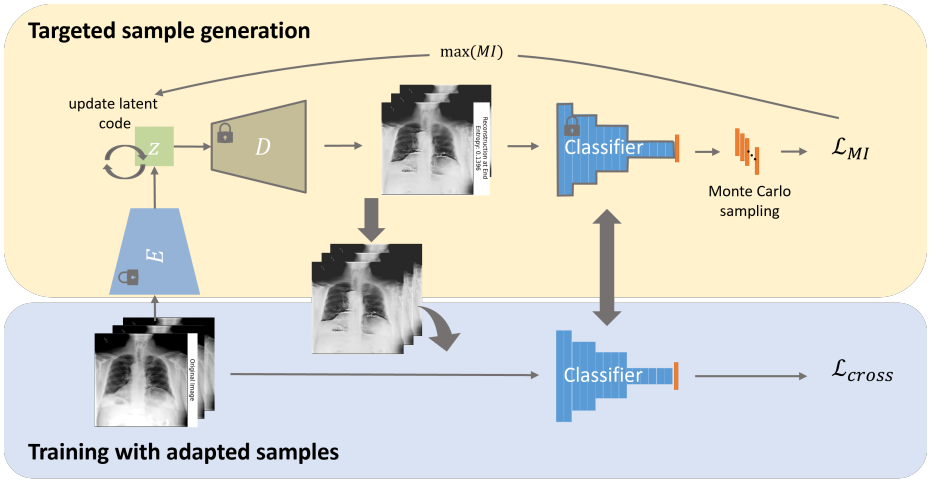

Fig. 1. The overall framework of TSynD (***T**argeted **Syn**thetic **D**ata generation*) for the robust training of a classifier: The autoencoder is pre-trained unsupervised, then its weights are frozen. During classifier training, the latent spatial representations of original images are optimized to maximize the classifier's epistemic uncertainty in the decoded images. The new images then serve as additional training data.

where the function $\sigma(\cdot)$ transfers the classifier logit outputs into probabilities. Here, $\sigma(\cdot)$ is the softmax function in the case of a multilabel classification or a sigmoid function in the case of a binary classification. The primary objective is to guide the reconstruction process $\hat{x} = f_{\text{dec}}(z)$ in a manner that the resulting sample $\hat{x}$ contributes meaningfully to the training of the classifier $\mathcal{C}$. This guidance involves modifying a latent variable $z \in \mathcal{Z}$ of the autoencoder with the aim of generating samples with a high epistemic uncertainty in the classifier $\mathcal{C}$.

The uncertainty of the predictive distribution is defined by the entropy

$$\mathbf{U}_H(z) = H(p(y|z,\theta)) = -\sum_{y\in\mathcal{Y}} \hat{p}(y|z,\theta)\log(\hat{p}(y|z,\theta)),$$

however, the epistemic uncertainty associated with a data sample $\hat{x} = f_{\text{dec}}(z)$ stems from uncertainty in model parameters. This can be quantified by the expected change in entropy of the model parameter posterior distribution, expressed by the conditional mutual information [8]:

$$\mathbf{U}_{MI}(z) = MI(z;\theta) = H(\mathbb{E}_\theta(p(y|z,\theta))) - \mathbb{E}_\theta(H(p(y|z,\theta))),$$

where the expectation is computed over Monte Carlo Dropouts [6]. Mutual information is considered to be a better measure of epistemic uncertainty [6]. To keep the additional computational effort low we only iterate over the last layers of $\mathcal{C}$ with K dropout masks to compute samples of $p(y|z,\theta^k)$, $k = 1 \dots K$.

3.2 Targeted Synthetic Data Generation

An optimization-based approach is used to find latent codes z that locally maximize the given measure for uncertainty $\mathbf{U}(z)$. As shown in Fig. 1, starting with the latent code $z_n = f_{\text{enc}}(x_n)$ of a random image of the training distribution, we search a local maximum

$$z^* = \arg\max_z \ \mathbf{U}(z). \tag{1}$$

Since $\mathcal{C}$ and f_{dec} both are differentiable, $\mathbf{U}(z)$ can be maximized by standard backpropagation. The resulting sample $\hat{x} = f_{\text{dec}}(z^*)$ is added to the training set, assuming that it belongs to the same class as x_n, but lies in missing parts of the learned data distribution, as indicated by the high uncertainty.

Latent space noise. Apart from the uncertainty that a sample yields, the diversity w.r.t. the training distribution is crucial. To introduce further varieties into the reconstructed image we generate samples by adding uniform noise to latent codes $z_n = f_{\text{enc}}(x_n)$:

$$\hat{x} = f_{\text{dec}}(z_n + \epsilon), \quad \epsilon \sim N(0, \sigma\mathbf{I}).$$

The resulting image $\hat{x}$ is an alternative representation of x_n and therefore increases the diversity in the dataset. Latent space noise can be used as a standalone augmentation or as an initial augmentation before optimization according to Eq. 1 to generate even more diverse samples.

The training process. In each training iteration, the batch is divided into two halves: The first half consists of original image-label pairs $\{(x_n, y_n)\}_{n=1}^{\frac{B}{2}}$, and the second half consists of the optimized reconstructions $\hat{x}_n = f_{\text{dec}}(z_n^*)$ resulting from Eq. (1), along with their corresponding labels $\{(\hat{x}_n, y_n)\}_{n=1}^{\frac{B}{2}}$. Since the maximization of the epistemic uncertainty depends on the current state of the classifier $\mathcal{C}$ we need to redo the generation process of $\hat{x}_n$ after each training iteration. This also prevents the so-called mode collapse problem that would occur if we ran the image generation only once. The resulting images would be similar since similar images are likely to maximize the uncertainty of the given classifier. However, since we retrain and generate in an alternating way, the network is updated and the generation process yields new alternations.

Optimizing latent codes vs. pixel values as parameters. Optimizing the pixel values like in [15, 17, 13] likely results in salt and pepper noise. Altering abstract representations gives us more substantial alternations, since each element of $z \in \mathcal{Z}$ represents larger receptive fields in the image. Additionally, the autoencoder is learned on the distribution of relevant images. The reconstruction process is therefore already constrained w.r.t. this distribution. Constraints that need to be introduced when optimizing the image pixels directly like in

	BreastMNIST		DermaMNIST		OCTMNIST		OrganaMNIST		OrgansMNIST		PathMNIST	
	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%	1%	10%
Baseline	70.7	75.2	66.8	65.2	59.9	67.5	71.9	89.3	49.6	67.8	67.4	77.2
Noise	72.0	**78.2**	66.7	65.8	61.0	**71.7**	73.8	87.8	52.9	68.6	64.7	**83.5**
TSynD	**73.3**	77.8	**66.9**	**66.7**	**61.4**	66.7	**77.2**	**89.4**	**54.2**	**71.4**	**73.1**	78.5
Gaussian Noise Augmentation during Test												
Baseline	62.6	**73.1**	66.8	65.1	24.9	29.9	44.5	78.9	37.1	52.6	12.6	10.6
Noise	**73.5**	**73.1**	66.7	65.7	24.5	34.9	44.1	65.3	37.3	52.4	13.5	11.5
TSynD	73.3	**73.1**	**66.9**	**66.7**	**28.7**	**36.4**	**63.5**	**85.1**	**45.6**	**66.4**	**28.2**	**12.8**
Adversarial Attacks during Test												
Baseline	65.6	7.1	66.4	48.8	5.3	3.3	34.4	68.4	13.8	25.6	28.5	21.1
Noise	68.6	21.6	**66.7**	53.0	8.1	4.1	39.2	71.6	25.1	25.6	31.7	26.1
TSynD	**71.4**	**28.2**	**66.7**	**64.1**	**12.8**	**42.8**	**53.5**	**83.9**	**27.1**	**51.3**	**43.5**	**47.8**

Table 1. Accuracy results of different MedMNIST datasets with a subsampling of the training dataset to 1% and 10%. The results are reported for the respective test set of the datasets and two augmented versions of the tests sets (Gaussian Noise and adversarial attacks).

the approaches of [15, 17, 13] are not needed. However, it is important to maximize epistemic uncertainty (model uncertainty) rather than aleatoric uncertainty (data uncertainty). Maximizing aleatoric uncertainty would result in ambiguous data, such as altering an image so that it can no longer be classified (e.g. to a noise image). By solely optimizing the epistemic uncertainty, the optimization process is implicitly constrained to generate meaningful, unambiguous data.

4 Experiments

Our experiments aim to show the effect of TSynD on the generalization performance and robustness of classification networks. Since the test and validation sets of available datasets are often drawn from similar distributions as the training distribution, the generalization of networks is hard to measure. For that reason, we introduce a sub-sampling of the training dataset to 1% and 10% of the respective datasets. This introduces a sampling bias and makes it more likely that the test and validation distributions contain out-of-distribution data. This also mirrors the common scenario in medical data where training datasets are often small. Our experiments concentrate on two main questions: 1) Does the proposed TSynD improve classification results when training in a low-data setting? 2) Is the training using the proposed approach more robust, e.g., against random test data augmentations and test time adversarial attacks? To investigate 1), we train and evaluate using three different settings: baseline classifier without any additional training time augmentations; augmentation through random latent space noise during the training (see section 3); and training using TSynD. For research question 2), the three previously trained settings are used and tested in three scenarios: no test data augmentation; Gaussian noise with

	OrganSMNIST		**Chest-XRay**		**OCTMNIST**	
	1%	10%	1%	10%	1%	10%
Entropy	73.5	86.0	60.8	67.9	79.5	82.9
MI	68.1	84.1	61.5	68.0	81.6	89.6

Table 2. Comparison between TSynD maximizing mutual information (MI) and Entropy on the validation sets of the respective datasets.

$\sigma = 0.2$ added to the test data; and the test data is altered using adversarial attacks as described in [3].

The datasets used in our experiments are MedMNIST v2 [20] datasets and the Chest-Xray [18] dataset for classification, since they are openly available and suitable for establishing a baseline. We utilized the commonly used ResNet-18 [4] and DenseNet [5] as classifiers, and a state-of-the-art autoencoder VQ-VAE [12, 2] trained unsupervised on the full training set as the generative model. In each experiment, the classifier was trained for 100 epochs and the model with the best validation performance was selected. The training was repeated three times, and the averaged values were reported. Our TSynD process is influenced by the learning rate (chosen as 0.1) and the number of iterations (either 100 or 50) for the optimizer to maximize the epistemic uncertainty. The noise factor that is added to the feature space is chosen empirically (either 0.1 or 1.0 in our experiments).

4.1 Classification and Robustness Results

Table 1 shows the classification results across different MedMNIST datasets using a ResNet-18 model to compare baseline training without augmentation, augmentation with latent space noise (Noise) and TSynD. The TSynD models improve over the baseline model and even over the Noise models on the standard test sets in almost all low data scenarios that were tested. This shows that TSynD is an effective method for training models that generalize well in such low-data settings. It further shows the advantage of the targeted optimization-based generation of new samples compared to random sampling. When we apply Gaussian noise to the test set or introduce adversarial attacks we can observe that the TSynD models are always better than the baseline models and even improve over the noise models, as well. This indicates that the samples that were generated by TSynD made the resulting model more robust against these out-of-distribution samples.

4.2 Uncertainty Maximization

Table 4.2 presents an ablation study w.r.t. to the uncertainty that is maximized during the image generation. In section 3 we introduce the entropy and the mutual information (MI) as measures for the uncertainty. We can see that the

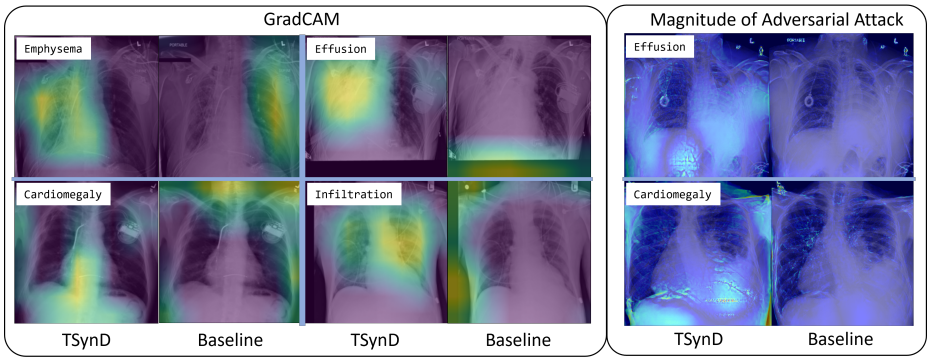

Fig. 2. Left: EigenGradCAM maps of the baseline classifier and classifier trained with TSynD. Right: Perturbation of images to minimize the probability of the given class. Depicted is the difference of the images at the start and end of the minimization.

MI performs better than the entropy. This also aligns with the theory, since entropy often is rather viewed as a measure of the aleatoric uncertainty, and the MI is viewed as a measure of the epistemic uncertainty. However, the difference between maximizing the MI and the entropy is not large, indicating that the entropy is not a strict measure for the aleatoric uncertainty and the MI is not a strict measure for the epistemic uncertainty. Improving the measure of epistemic uncertainty could further constrain the image generation process to produce more meaningful and unambiguous data (see section 3.2).

4.3 Qualitative Robustness Evaluation

We trained a classifier on the Chest-Xray [18] dataset with and without TSynD. On average, we obtained an AUC improvement of about 1% using TSynD on the validation set (both on the 1% and 10% subsampling of the training dataset). In this experiment, however, we do not concentrate on performance gain, moreover, we investigate the robustness of the proposed training mechanism. We explore the reasoning process of the classifier, by applying a commonly used explanation approach – EigenGradCAM [10]. The results can be seen on the left-hand side in Figure 2. It can be observed, that the classifier trained using TSynD utilizes more relevant regions of the image than the baseline classifier trained without TSynD. We, additionally, employed our synthetic data generation process to create adversarial examples by minimizing class probabilities instead of maximizing the classifier uncertainty. The magnitude of the difference between the original reconstruction and the optimized adversarial image can be seen on the right-hand side of Figure 2. We can observe that in order to minimize the probability for the classifier trained with TSynD, much larger and more relevant image regions must be altered, further indicating the increased robustness introduced by TSynD.

5 Conclusion

In this work, we have shown how to utilize generative models to create synthetic data that is exploring unknown and relevant parts of the training distribution. We thus take a first step toward replacing the acquisition of large real-world data distributions with a more targeted data generation process that creates important data points. We have shown that training on this synthetic data yields a model that generalizes better to out-of-distribution samples and is more robust against adversarial attacks.

In the current state our generation method only augments given samples. This is not ideal from a distribution diversity standpoint. As a future direction, we want to extend the method to generate new samples that yield a high epistemic uncertainty and are therefore relevant for the training process.

References

1. Borlino, F.C., D'Innocente, A., Tommasi, T.: Rethinking domain generalization baselines. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9227–9233. IEEE (2021)
2. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
3. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
5. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2018)
6. Kirsch, A., van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
7. Li, P., Li, D., Li, W., Gong, S., Fu, Y., Hospedales, T.M.: A simple feature augmentation for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8886–8895 (2021)
8. Linander, H., Balabanov, O., Yang, H., Mehlig, B.: Looking at the posterior: accuracy and uncertainty of neural-network predictions. Machine Learning: Science and Technology **4**(4), 045032 (2023)
9. Mittal, S., Niemeijer, J., Schäfer, J.P., Brox, T.: Best practices in active learning for semantic segmentation. In: Köthe, U., Rother, C. (eds.) Pattern Recognition. pp. 427–442. Springer Nature Switzerland, Cham (2024)
10. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 international joint conference on neural networks (IJCNN). pp. 1–7. IEEE (2020)
11. Niemeijer, J., Schwonberg, M., Termöhlen, J.A., Schmidt, N.M., Fingscheidt, T.: Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2830–2840 (January 2024)

12. van den Oord, A., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning (2018)
13. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12556–12565 (2020)
14. Rosca, M., Lakshminarayanan, B., Warde-Farley, D., Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. arXiv preprint arXiv:1706.04987 (2017)
15. Sinha, A., Namkoong, H., Volpi, R., Duchi, J.: Certifying some distributional robustness with principled adversarial training. arXiv preprint arXiv:1710.10571 (2017)
16. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
17. Volpi, R., Namkoong, H., Sener, O., Duchi, J.C., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
18. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. CoRR **abs/1705.02315** (2017), http://arxiv.org/abs/1705.02315
19. Xu, Z., Liu, D., Yang, J., Niethammer, M.: Robust and generalizable visual representation learning via random convolutions. CoRR **abs/2007.13003** (2020), https://arxiv.org/abs/2007.13003
20. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. CoRR **abs/2110.14795** (2021), https://arxiv.org/abs/2110.14795
21. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4396–4415 (2023). https://doi.org/10.1109/TPAMI.2022.3195549
22. Zhou, K., Yang, Y., Hospedales, T., Xiang, T.: Learning to generate novel domains for domain generalization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. pp. 561–578. Springer (2020)
23. Zhou, K., Yang, Y., Qiao, Y., Xiang, T.: Domain generalization with mixstyle. CoRR **abs/2104.02008** (2021), https://arxiv.org/abs/2104.02008