Data-Agnostic Augmentations for Unknown Variations: Out-of-Distribution Generalisation in MRI Segmentation

Puru Vaish[*1] P.VAISH@UTWENTE.NL
Felix Meister[2] FELIX.MEISTER@SIEMENS-HEALTHINEERS.COM
Tobias Heimann[2] TOBIAS.HEIMANN@SIEMENS-HEALTHINEERS.COM
Christoph Brune[1] C.BRUNE@UTWENTE.NL
Jelmer M. Wolterink[1] J.M.WOLTERINK@UTWENTE.NL
[1] *Department of Applied Mathematics, Technical Medical Centre, University of Twente*
[2] *Digital Technology and Innovation, Siemens Healthineers, Erlangen, Germany*

Editors: Under Review for MIDL 2024

Abstract

Medical image segmentation models are often trained on curated datasets, leading to performance degradation when deployed in real-world clinical settings due to mismatches between training and test distributions. While data augmentation techniques are widely used to address these challenges, traditional visually consistent augmentation strategies lack the robustness needed for diverse real-world scenarios. In this work, we systematically evaluate alternative augmentation strategies, focusing on MixUp and Auxiliary Fourier Augmentation. These methods mitigate the effects of multiple variations without explicitly targeting specific sources of distribution shifts. We demonstrate how these techniques significantly improve out-of-distribution generalization and robustness to imaging variations across a wide range of transformations in cardiac cine MRI and prostate MRI segmentation. We quantitatively find that these augmentation methods enhance learned feature representations by promoting separability and compactness. Additionally, we highlight how their integration into nnU-Net training pipelines provides an easy-to-implement, effective solution for enhancing the reliability of medical segmentation models in real-world applications.

Keywords: MRI, segmentation, data augmentation, generalisation, robustness

1. Introduction

Medical image analysis requires deep learning models that are accurate, robust, and generalize well to new and unseen data. However, when deployed in real-world scenarios, deep neural networks often suffer performance degradation (Hendrycks and Dietterich, 2019; Kamann and Rother, 2021). This generalization gap can be attributed to a range of factors, including variations in patient populations, differences in image acquisition, and imaging artefacts. Among strategies to improve generalization are data augmentation techniques like intensity shifts, affine transforms, and noise addition (Garcea et al., 2023; Goceri, 2023). As they can demonstrably improve out-of-distribution generalization (Boone et al., 2023), they are among the standard set of augmentations used in many deep learning segmentation models, such as nnU-Net (Isensee et al., 2021).

However, standard augmentation strategies cannot cover more complex underlying image formation mechanisms, which in MRI could include bias fields due to coil miscalibration,

* Corresponding author

Rician noise when MRI is taken at higher resolutions, ghosting artefacts, or random RF spikes during acquisition (see Fig. 1). Artifact-specific augmentation policies might mitigate this problem (Boone et al., 2023), but such policies are not guaranteed to exist. For instance, variations may only happen at inference due to data mishandling (Shimron et al., 2022), and explicitly anticipating all possible variations is often infeasible. Hence, as an alternative, augmentation strategies that are not specifically designed for any variation and yet manages to mitigate the effect of multiple variations would greatly benefit medical imaging with deep learning.

In this work, we systematically investigate general, *data-agnostic* augmentation strategies, namely MixUp (Zhang et al., 2018) and Auxiliary Fourier Augmentation (AFA) (Vaish et al., 2024). By data-agnostic, we mean augmentations that do not seek to maintain the visual consistency of the data being augmented. We demonstrate the effect of these techniques in nnU-Net models for segmentation of cardiac cine MRI and prostate MRI. Neither MixUp nor AFA explicitly addresses specific sources of variation in these data, yet we show how they improve segmentation performance in various out-of-distribution generalization settings. Moreover, we include an analysis of the learned feature representations, showing improved structure and interoperability when MixUp and AFA are used. Our findings provide new insights into the effectiveness and limitations of these augmentation methods in medical image analysis scenarios and show that MixUp and AFA can improve the performance of deep neural networks in multiple tasks and generalization settings.

2. Materials and Methods

2.1. Data

We investigate to what extent data augmentation strategies can mitigate the effects of distribution shifts in MRI. We perform experiments on cardiac cine MRI and bi-parametric prostate MRI. For each of these, we include two separate datasets, allowing us to consider differences within and between datasets.

Cardiac cine MR We use the Automated Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018), which contains 150 cardiac cine MRI scans (100 training, 50 test) acquired at Hospital of Dijon, France (in-plane resolution 1.37-1.67mm, slice thickness 5-10mm). As an external test set used to measure generalization performance, we include the M&Ms (Campello et al., 2021) test set of 268 scans with different pathologies from multiple centers (in-plane resolution 0.85-1.45mm, slice thickness 10mm). In both ACDC and M&Ms, manual annotations of the left ventricle (LV), myocardium (MYO), and right ventricle (RV) are provided in end-diastolic (ED) and end-systolic (ES) frames.

Prostate MRI We include prostate bi-parametric MRI (bpMRI) scans from the Prostate 158 (P158) dataset (Adams et al., 2022), which has 139 scans for training and 19 scans for testing. This data set contains patients with prostate cancer (PCa) lesions. In addition, as an external test set to measure generalization, we use the ProstateX (PX) dataset (Armato et al., 2018), which includes 141 test scans. Both datasets provide T2w scans at an in-plane resolution of 1.45-1.5 mm, ADC maps at 0.45-0.5 mm, and slice thicknesses of 3-4mm. Segmentation masks of the transitional zone (TZ) and the prostate peripheral zone (PZ) are available in all images. The test masks for PX are taken from (Xu et al., 2023).

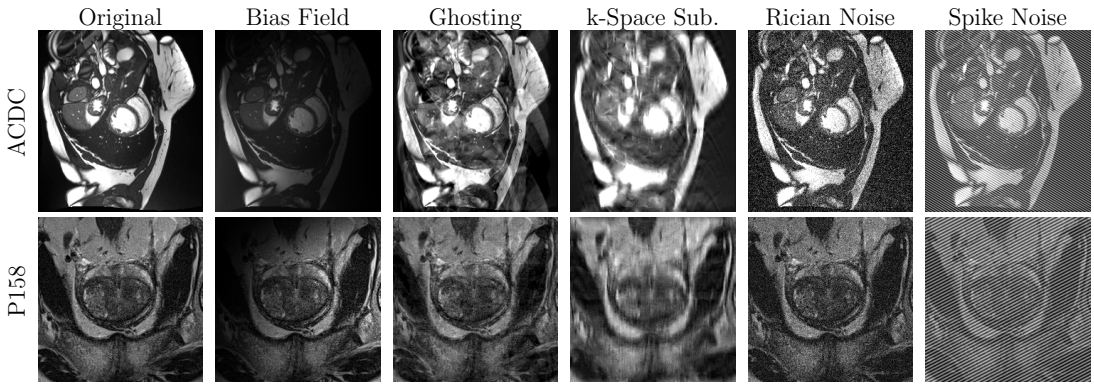

Figure 1: Corruptions (severity level 3) in cardiac cine MRI images and the T2w channel of prostate bpMRI images as a result of our image variation model.

2.2. Modelling Image Variation

We simulate image variation in a controlled manner by applying a range of transformations to images. This allows a thorough evaluation of model robustness under different augmentation strategies. Following the methodology outlined in ROOD-MRI (Boone et al., 2023), we apply elastic deformation, isotropic downsampling, anisotropic downsampling, bias field amplification, contrast compression, contrast expansion, ghosting, random motion, Rician noise addition, and smoothing, at five different levels (1: mild, 5: severe). In addition, we include spike noise artifacts (radio frequency noise) and k-space subsampling, again at five severity levels. In total, we apply 14 corruptions to each image (see Fig. 1 for examples).

2.3. Augmentation Strategies

We conduct experiments in which we train models with three different kinds of augmentation strategies within the nnU-Net framework (Isensee et al., 2021). In all experiments, we keep the pre-processing and post-processing fixed to the default nnU-Net options.

Base augmentations By default, nnU-Net employs eight augmentation strategies, namely rotation, scaling, Gaussian noise injection, Gaussian blurring, brightness and contrast adjustments, simulation of low-resolution imaging, gamma correction, and mirroring.

MixUp In this strategy, new samples are generated through linear interpolation of pairs of training samples. Here a sample consists of input x_i and label y_i. Formally, given two samples (x_i, y_i) and (x_j, y_j), MixUp creates a new synthetic sample as:

$$x_{\text{mix}} = \lambda x_i + (1-\lambda)x_j, \quad y_{\text{mix}} = \lambda y_i + (1-\lambda)y_j$$

where $\lambda \in [0, 1]$ is drawn from a Beta distribution. We adapt the original MixUp setup, in which y_i is a one-hot encoded sample label, to segmentation by linearly interpolating one-hot encoded segmentation masks.

Table 1: DSC and HD95 on the original and transformed test set of ACDC and P158 using either using no augmentations or a combination of base, MixUp, and AFA augmentations. Blue and red colors indicate statistically significant ($p < 0.05$) differences between models without and with MixUp or AFA. Bold-faced numbers indicate the best result for each column.

Augmentation			ACDC				P158			
			Original		Transformed		Original		Transformed	
Base	MixUp	AFA	DSC	HD95 (mm)	DSC	HD95 (mm)	DSC	HD95 (mm)	DSC	HD95 (mm)
			0.891	**5.81**	0.755	12.68	**0.789**	4.63	0.705	8.23
	✓		0.889	6.08	0.760	12.86	0.786	4.99	0.696	8.29
		✓	0.866	6.48	0.801	**9.67**	0.786	4.77	**0.724**	**6.84**
	✓	✓	0.876	6.35	**0.804**	10.67	0.780	5.15	0.711	7.85
✓			**0.925**	**3.37**	0.801	9.40	0.825	4.60	0.730	7.39
✓	✓		0.924	3.49	0.842	7.90	**0.832**	4.35	0.758	6.78
✓		✓	0.920	3.72	0.850	7.61	0.826	4.79	0.760	6.67
✓	✓	✓	0.921	3.60	**0.862**	**7.02**	0.829	**4.29**	**0.770**	**6.33**

Auxiliary Fourier Augmentation (AFA) augments images in the frequency domain under the hypothesis that visual augmentation techniques are unable to cover the vulnerability of neural networks to perturbations in the frequency domain (Vaish et al., 2024). AFA samples frequency basis functions and adds them to the training samples, leaving the label unchanged. Formally, let $\mathcal{F}$ denote the Fourier transform operator. For a training sample (x_i, y_i), the n-dimensional Fourier transform of x_i is given by $X_i = \mathcal{F}(x_i)$. To augment the data, a perturbation α is added at a randomly chosen frequency coordinate $(k_1, k_2, \ldots)$ in the Fourier domain, modifying X_i as:

$$X_i^{\text{aug}}(k_1, k_2, \ldots) = X_i(k_1, k_2, \ldots) + \alpha.$$

The augmented image in the spatial domain, x_i^{aug}, is then obtained by applying the inverse Fourier transform: $x_i^{\text{aug}} = \mathcal{F}^{-1}(X_i^{\text{aug}})$. The model training involves a joint optimization of an AFA-augmented image and a non-AFA-augmented image.

2.4. Quantitative Evaluation

We segment all structures separately, namely LV, MYO, RV in cardiac cine MRI, and TZ and PZ in prostate MRI. Results are reported as average Dice Similarity Coefficients (DSC) and 95$^{\text{th}}$ percentile Hausdorff Distances (HD95), over all structures, and frames (ED, ES) in cine MRI. For all settings, we perform a 5-fold cross-validation and test for statistical significance using the paired t-test at $p < 0.05$.

3. Experiments and Results

3.1. Synthetically Corrupted Images

We train a total of sixteen nnU-Net models with different combinations of base augmentations, MixUp, and AFA, for both the ACDC cardiac cine MR dataset and the P158 prostate

cancer dataset. For each dataset, we evaluated results on the original test set and the test set with the transformations described in Sec. 2.2. Note that we do not apply any of these corruptions to the training sets. Tab. 1 lists DSC and HD95 values for this experiment. Fig. 2 shows the relation between the severity of individual transformations and DSC values obtained for ten nnU-Net models.

Cardiac cine MRI For cardiac cine MRI, when not using any data augmentation, there is a large performance gap between the original ACDC test set and the transformed test set, i.e., DSC 0.891 vs. 0.755, indicating poor generalization to out-of-distribution data. We find that adding either MixUp or AFA to this model improves performance on the transformed test set, to DSC 0.760 and 0.801, respectively. Moreover, the combination of both augmentation strategies improves performance further to DSC 0.804. The gain on the transformed test set exceeds the performance drop on the original test set.

A similar pattern unfolds when we combine MixUp and AFA with base augmentations. Here, we see a performance increase in all settings for both the original and transformed data. Notably, base augmentations lead to a DSC of 0.801 on the transformed test set, compared to DSC 0.755 when no augmentation is used. This indicates that these augmentations are able to improve performance on some of the out-of-distribution samples. However, we find that MixUp and AFA can lead to significant ($p < 0.05$) performance gains on top of these augmentations, up to DSC 0.862 and HD95 7.02 when both are used. Moreover, when MixUp and AFA are used in combination with base augmentations, the performance drop on the original test set is smaller and not significant.

The results in Fig. 2 indicate that adding MixUp and AFA improves robustness to *all* imaging variations. Moreover, while these techniques do not explicitly target variations like k-space subsampling and Rician noise, they are very effective in overcoming challenges posed by these corruptions. We also find that the performance increase is largest at higher severity levels. Furthermore, while the bias field can be reasonably simulated using intensity shifts, which is a base augmentation, only adding MixUp or AFA further enhances the performance.

Prostate MRI Our findings in the prostate MRI data set match those found in the cardiac cine MRI data set to a large extent. We find that in the model without any augmentations, there is a performance gap between the original (DSC 0.789) and transformed (DSC 0.705) data sets. Using AFA, either stand-alone or in combination with MixUp, narrows this gap. However, we also find that using *only* MixUp has a detrimental effect on model performance. Similar to the cardiac cine MRI set, we find that MixUp and AFA reduce performance on the original data set when used without base augmentations; however, they only significantly reduce when using both MixUp and AFA. In contrast to cardiac cine MRI, adding MixUp and AFA to base augmentations not only leads to a significant performance increase on the transformed data set but also on the original dataset. Results for prostate MRI in Fig. 2 show similar trends as for cardiac cine MRI, confirming the general nature of our findings and added value of MixUp and AFA over base augmentations.

3.2. Real-World Distribution Shifts

In our previous experiments, we synthesized corruptions in data to explore the robustness of models with and without augmentation. We now consider generalization between datasets

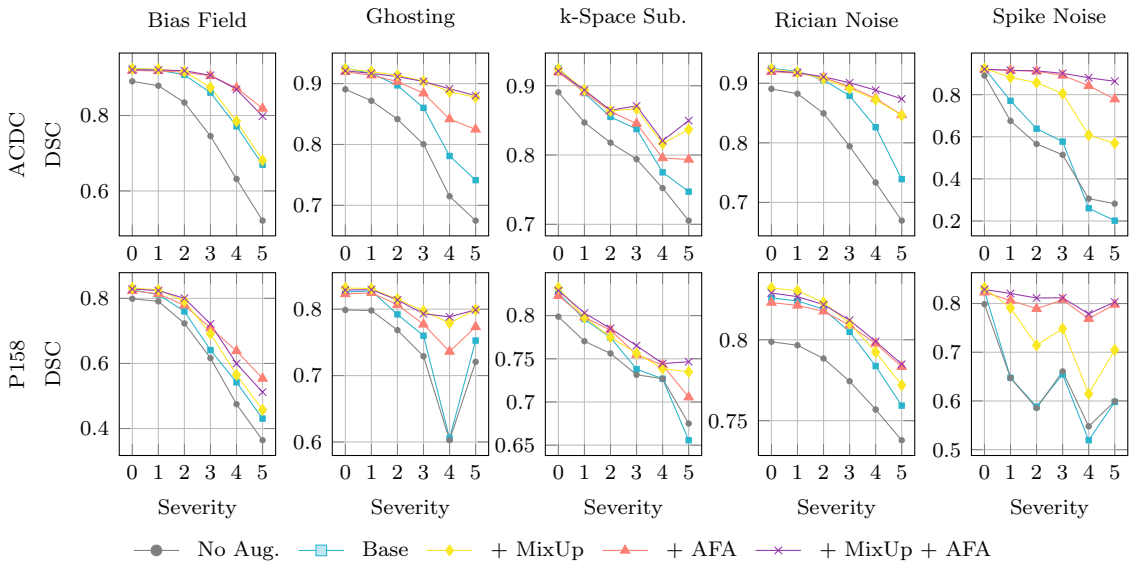

Figure 2: Trend of DSC per severity for test sets corrupted with bias field, ghosting, k-space subsampling, Rician noise and spike noise. We notice that MixUp and AFA are effective in mitigating the challenges posed by complex image corruptions.

originating from similar yet different distributions. Differences may be a result of different demographics, protocols, or scanner vendors. As in our previous experiment, we train with combinations of base, MixUp and AFA augmentation. For clarity, we omit results trained without base augmentations. Tab. 2 lists results for cardiac cine MRI and prostate MRI.

Cardiac cine MR We train segmentation models on the ACDC data set and use M&Ms as a test set. We find that a model trained on ACDC with only base augmentations experiences a performance drop compared to a model trained on M&Ms with only base augmentations, with Dice coefficients of 0.870 and 0.882. However, when combinations of MixUp with AFA are added, we find consistent performance improvements across all augmentation combinations. Furthermore, we find that for models that use a combination of base augmentations, MixUp, or both MixUp and AFA, we approach the DSC and HD95 score of the model that was trained on M&Ms itself, bridging the generalization gap.

Prostate MRI The segmentation of the prostate bpMRI presented a more significant domain shift challenge. We train the segmentation model on P158 and use the PX dataset as the test set. As noted earlier, the large variability in prostate glands poses a difficult challenge and substantially impacts model performance. Despite this challenge, a model trained with base augmentations, along with MixUp (DSC 0.737, HD95 6.60), is a significant improvement over using only the base augmentations (DSC 0.705, HD95 7.87), indicating improved generalization capabilities in this setting where there is large variance in anatomy. Models in combination with AFA also showcase significantly improved performance.

Table 2: DSC and HD95 performance under distribution shift for Cardiac Cine MR, testing on M&Ms, and Prostate bpMRI, testing on PX, with various data augmentation strategies. Highlighted numbers denote significantly improved results when models using MixUp or AFA compared to model only using base augmentation ($p < 0.05$).

Augmentation			Cardiac Cine MR			Prostate bpMRI		
Base	MixUp	AFA	Trained On	DSC	HD95 (mm)	Trained On	DSC	HD95 (mm)
✓			ACDC	0.870	7.97	P158	0.705	7.87
✓	✓			0.880	5.74		**0.737**	**6.60**
✓		✓		0.874	6.92		0.718	7.67
✓	✓	✓		**0.880**	**5.70**		0.732	7.52
			M&Ms†	0.882	5.02	PX+	0.826	4.77

Best model trained on the test dataset from † (Campello et al., 2021) and + (Xu et al., 2023).

3.3. Model Interpretation

To investigate whether our empirical findings reflect characteristics of the trained models under data augmentation, we quantify the separability and compactness of their learned features using k-variance gradient-normalized margins (kVGM) (Chuang et al., 2021). A positive higher value for this metric indicates that the model has learned more separable and compact clusters of representations, while negative values mean that the model misclassifies. Fig. 3 visualizes the position of voxels from the transformed test set in the feature space of ten of our trained models (dimensionality reduced via PCA), along with the kVGM of each model. These plots show that the absence of augmentation leads to poor feature separation while using only base augmentations leads to better clustering features that are not easily separable. Adding AFA alone improves separability, and MixUp alone enhances compactness, and when combined, they appear to promote both compactness and separability. The kVGM metric, in increasing order for generalisability, ranks the models starting from no augmentations, followed by base augmentations, base with AFA, base with MixUp, and finally base with both AFA and MixUp. This supports our finding that these augmentations enrich feature representations, leading to enhanced out-of-distribution generalization.

4. Discussion and Conclusion

In this study, we have demonstrated how non-standard augmentation techniques that do not target specific variations, specifically MixUp and Auxiliary Fourier Augmentation (AFA), can enhance the robustness of state-of-the-art segmentation frameworks like nnU-Net against many variations in MRI.

While MixUp has been known to be an effective augmentation for various tasks (Eaton-Rosen et al., 2018; Thulasidasan et al., 2019; Gazda et al., 2022), we find it is very well suited for overcoming challenging medical image conditions as well. However, we also observe that without base augmentations on P158, MixUp alone leads to a (non-significant) decline in performance. Our results highlight the advantage of combining augmentation strategies that intrinsically exploit different mechanisms. For example, AFA follows a fun-

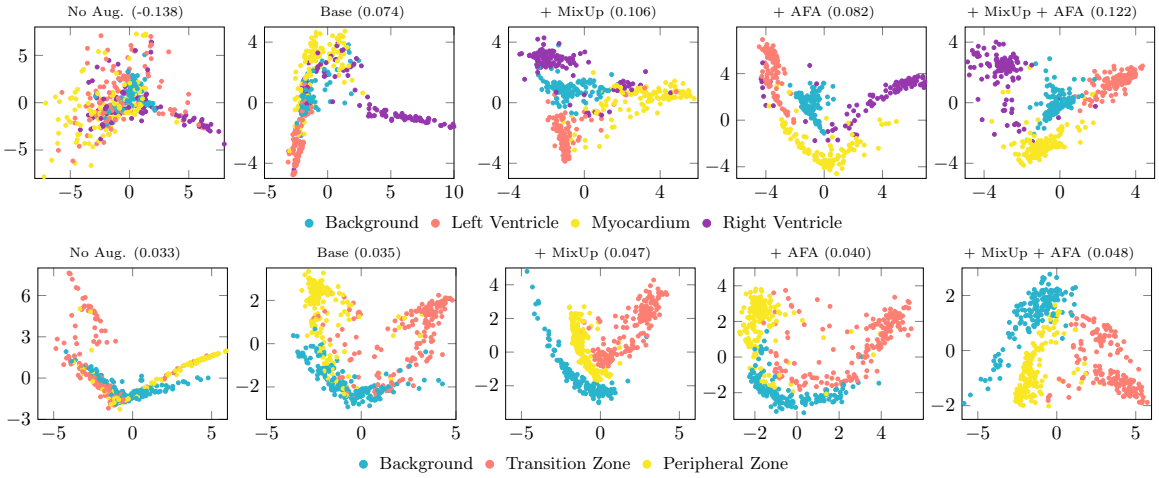

Figure 3: PCA projection of learned features, for final features from nnU-Net trained with different augmentation techniques for samples from the transformed test sets (top: ACDC, bottom: P158) with the corresponding kVGM metric.

damentally different strategy than MixUp by directly perturbing k-space data. The effect of this is shown in our results, in which the combination of both always improves over their individual use. This is corroborated by an evaluation of the feature space using k-variance gradient-normalized margins. We consider this metric a promising tool for studying model generalizability.

Our results align with studies incorporating general augmentation strategies into nnU-Net (Atya et al., 2021), and MixUp and AFA are straightforward additions with a lot of benefits. However, augmentations cannot address all generalization gaps. We find that prostate zonal segmentation remains challenging due to significant inter-subject variability. One example is the effect of age, where younger cohorts have sharper tissue boundaries (Allen et al., 1989; Situmorang et al., 2012) and such differences are difficult to address without prior knowledge, regardless of augmentations.

While the augmentations we treat in this work are valuable tools for improving robustness, there remains substantial potential for further advancements in this domain. For example, we have here used a base version of MixUp which has been previously shown to be sufficient (Atya et al., 2021; Liu et al., 2024), but there are many variants (Cao et al., 2024). There are other data-agnostic augmentations as well which are out of scope of discussion, and would need to be adapted for medical domain, like PRIME (Modas et al., 2022) which seeks to maximise image entropy. Therefore, our work should be viewed as a stepping stone toward broader research on using simple general augmentations for out-of-distribution generalization in medical imaging, as opposed to more complicated methods like model-based methods like GANs and diffusion models (Garcea et al., 2023).

In conclusion, we find that adding non-standard data-agnostic augmentation to a state-of-the-art nnU-Net model can consistently and significantly increase segmentation performance under various generalisation challenges, for cardiac cine MRI and prostate MRI.

Acknowledgments

This publication is part of the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), Siemens Healthineers, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023.

References

Lisa C. Adams, Marcus R. Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M. Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K. Bressem. Prostate158 - an expert-annotated 3T MRI dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148: 105817, September 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.105817. URL https://www.sciencedirect.com/science/article/pii/S0010482522005789.

Ks Allen, Hy Kressel, Ph Arger, and Hm Pollack. Age-related changes of the prostate: evaluation by MR imaging. *American Journal of Roentgenology*, 152(1):77–81, January 1989. ISSN 0361-803X. doi: 10.2214/ajr.152.1.77. URL https://www.ajronline.org/doi/10.2214/ajr.152.1.77. Publisher: American Roentgen Ray Society.

Samuel G. Armato, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S. Kirby, Nicholas Petrick, George Redmond, Maryellen L. Giger, Kenny Cha, Artem Mamonov, Jayashree Kalpathy-Cramer, and Keyvan Farahani. PROSTATEx challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *Journal of Medical Imaging (Bellingham, Wash.)*, 5(4):44501, October 2018. ISSN 2329-4302. doi: 10.1117/1.JMI.5.4.044501.

Hadas Ben Atya, Ori Rajchert, Liran Goshen, and Moti Freiman. Non parametric data augmentations improve deep-learning based brain tumor segmentation. In *2021 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)*, pages 357–360, November 2021. doi: 10.1109/COMCAS52219.2021.9629083. URL https://ieeexplore.ieee.org/document/9629083.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, Gerard Sanroma, Sandy Napel, Steffen Petersen, Georgios Tziritas, Elias Grinias, Mahendra Khened, Varghese Alex Kollerathu, Ganapathy Krishnamurthi, Marc-Michel Rohé, Xavier Pennec, Maxime Sermesant, Fabian Isensee, Paul Jäger, Klaus H. Maier-Hein, Peter M. Full, Ivo Wolf, Sandy Engelhardt, Christian F. Baumgartner, Lisa M. Koch, Jelmer M. Wolterink, Ivana Išgum, Yeonggul Jang, Yoonmi Hong, Jay Patravali, Shubham Jain, Olivier Humbert, and Pierre-Marc Jodoin. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, November 2018. ISSN 1558-254X. doi: 10.1109/TMI.2018.2837502. URL https://ieeexplore.ieee.org/document/8360453. Conference Name: IEEE Transactions on Medical Imaging.

Lyndon Boone, Mahdi Biparva, Parisa Mojiri Forooshani, Joel Ramirez, Mario Masellis, Robert Bartha, Sean Symons, Stephen Strother, Sandra E. Black, Chris Heyn, Anne L. Martel, Richard H. Swartz, and Maged Goubran. ROOD-MRI: benchmarking the robustness of deep learning segmentation models to out-of-distribution and corrupted data in MRI. *Neuroimage*, 278:120289, September 2023. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2023.120289. URL https://www.sciencedirect.com/science/article/pii/S1053811923004408.

Víctor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martín-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, Mario Parreño, Alberto Albiol, Fanwei Kong, Shawn C. Shadden, Jorge Corral Acero, Vaanathi Sundaresan, Mina Saber, Mustafa Elattar, Hongwei Li, Bjoern Menze, Firas Khader, Christoph Haarburger, Cian M. Scannell, Mitko Veta, Adam Carscadden, Kumaradevan Punithakumar, Xiao Liu, Sotirios A. Tsaftaris, Xiaoqiong Huang, Xin Yang, Lei Li, Xiahai Zhuang, David Viladés, Martín L. Descalzo, Andrea Guala, Lucia La Mura, Matthias G. Friedrich, Ria Garg, Julie Lebel, Filipe Henriques, Mahir Karakas, Ersin Çavuş, Steffen E. Petersen, Sergio Escalera, Santi Seguí, José F. Rodríguez-Palomares, and Karim Lekadir. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, December 2021. ISSN 1558-254X. doi: 10.1109/TMI.2021.3090082. URL https://ieeexplore.ieee.org/document/9458279. Conference Name: IEEE Transactions on Medical Imaging.

Chengtai Cao, Fan Zhou, Yurou Dai, Jianping Wang, and Kunpeng Zhang. A survey of mix-based data augmentation: taxonomy, methods, applications, and explainability. *ACM Computing Surveys*, 57(2):1–38, October 2024. doi: 10.1145/3696206. URL https://dl.acm.org/doi/10.1145/3696206. Publisher: Association for Computing Machinery.

Ching-Yao Chuang, Youssef Mroueh, Kristjan Greenewald, Antonio Torralba, and Stefanie Jegelka. Measuring generalization with optimal transport. In *Advances in Neural Information Processing Systems*, volume 34, pages 8294–8306. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4607f7fff0dce694258e1c637512aa9d-Abstract.html.

Zach Eaton-Rosen, Felix Bragman, Sebastien Ourselin, and M. Jorge Cardoso. Improving data augmentation for medical image segmentation. In *MIDL Abstract*, April 2018. URL https://openreview.net/forum?id=rkBBChjiG.

Fabio Garcea, Alessio Serra, Fabrizio Lamberti, and Lia Morra. Data augmentation for medical imaging: a systematic literature review. *Computers in Biology and Medicine*, 152:106391, January 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2022.106391. URL https://www.sciencedirect.com/science/article/pii/S001048252201099X.

Matej Gazda, Peter Bugata, Jakub Gazda, David Hubacek, David Jozef Hresko, and Peter Drotar. Mixup augmentation for kidney and kidney tumor segmentation. In Nicholas Heller, Fabian Isensee, Darya Trofimova, Resha Tejpaul, Nikolaos Papanikolopoulos,

and Christopher Weight, editors, *Kidney and Kidney Tumor Segmentation*, pages 90–97, Cham, 2022. Springer International Publishing. ISBN 978-3-030-98385-7. doi: 10.1007/978-3-030-98385-7_12.

Evgin Goceri. Medical image data augmentation: techniques, comparisons and interpretations. *Artificial Intelligence Review*, 56(11):12561–12605, November 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10453-z. URL https://doi.org/10.1007/s10462-023-10453-z.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*. arXiv, March 2019. doi: 10.48550/arXiv.1903.12261. URL http://arxiv.org/abs/1903.12261. arXiv:1903.12261 [cs].

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, February 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. URL https://www.nature.com/articles/s41592-020-01008-z. Publisher: Nature Publishing Group.

Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *International Journal of Computer Vision*, 129(2):462–483, February 2021. ISSN 1573-1405. doi: 10.1007/s11263-020-01383-2. URL https://doi.org/10.1007/s11263-020-01383-2.

Chang Liu, Fuxin Fan, Annette Schwarz, and Andreas Maier. Cut to the mix: simple data augmentation outperforms elaborate ones in limited organ segmentation datasets. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 145–154, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72111-3. doi: 10.1007/978-3-031-72111-3_14.

Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. PRIME: a few primitives can boost robustness to common corruptions. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 623–640, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19806-9. doi: 10.1007/978-3-031-19806-9_36.

Efrat Shimron, Jonathan I. Tamir, Ke Wang, and Michael Lustig. Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences*, 119(13):e2117203119, March 2022. doi: 10.1073/pnas.2117203119. URL https://www.pnas.org/doi/full/10.1073/pnas.2117203119. Publisher: Proceedings of the National Academy of Sciences.

Gerhard Reinaldi Situmorang, Rainy Umbas, Chaidir A. Mochtar, and Rachmat Budi Santoso. Prostate cancer in younger and older patients: do we treat them differently? *Asian Pacific journal of cancer prevention: APJCP*, 13(9):4577–4580, 2012. ISSN 2476-762X. doi: 10.7314/apjcp.2012.13.9.4577.

Sunil Thulasidasan, View Profile, Gopinath Chennupati, View Profile, Jeff Bilmes, View Profile, Tanmoy Bhattacharya, View Profile, Sarah Michalak, and View Profile. On mixup training. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13911–13922, December 2019. doi: 10.5555/3454287.3455533. URL https://dl.acm.org/doi/10.5555/3454287.3455533.

Puru Vaish, Shunxin Wang, and Nicola Strisciuglio. Fourier-basis functions to bridge augmentation gap: rethinking frequency augmentation in image classification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17763–17772, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.01682. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.01682.

Lili Xu, Gumuyang Zhang, Daming Zhang, Jiahui Zhang, Xiaoxiao Zhang, Xin Bai, Li Chen, Qianyu Peng, Ru Jin, Li Mao, Xiuli Li, Zhengyu Jin, and Hao Sun. Development and clinical utility analysis of a prostate zonal segmentation model on T2-weighted imaging: a multicenter study. *Insights into Imaging*, 14(1):44, March 2023. ISSN 1869-4101. doi: 10.1186/s13244-023-01394-w. URL https://doi.org/10.1186/s13244-023-01394-w.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Appendix A. More Examples of Variations

We show an example of all the image variations that we make part of the transformations. In Fig. 4 we show an example image from each transformed test set for ACDC and in Fig. 5 we show an example image from each transformed test set of P158.

The images show while the object of interest remains discernable to the human eye, the difference to the original sample is large. Some variations are not diagnostically relevant, like smoothing, severe random motion, but they serve as interesting examples of where human expertise might still outperform sophisticated deep learning methods, and through this study we explore how to bridge this gap to unknown variations without explicitly using them as augmentations.

Appendix B. Performance per Severity for All Corruptions

We plot the trend of DSC metric per severity for each transformation considered under image variations. Trends for DSC for ACDC are shown in Fig. 6 and HD95 in Fig. 7. Trends for DSC for P158 are shown in Fig. 8 and HD95 in Fig. 9.

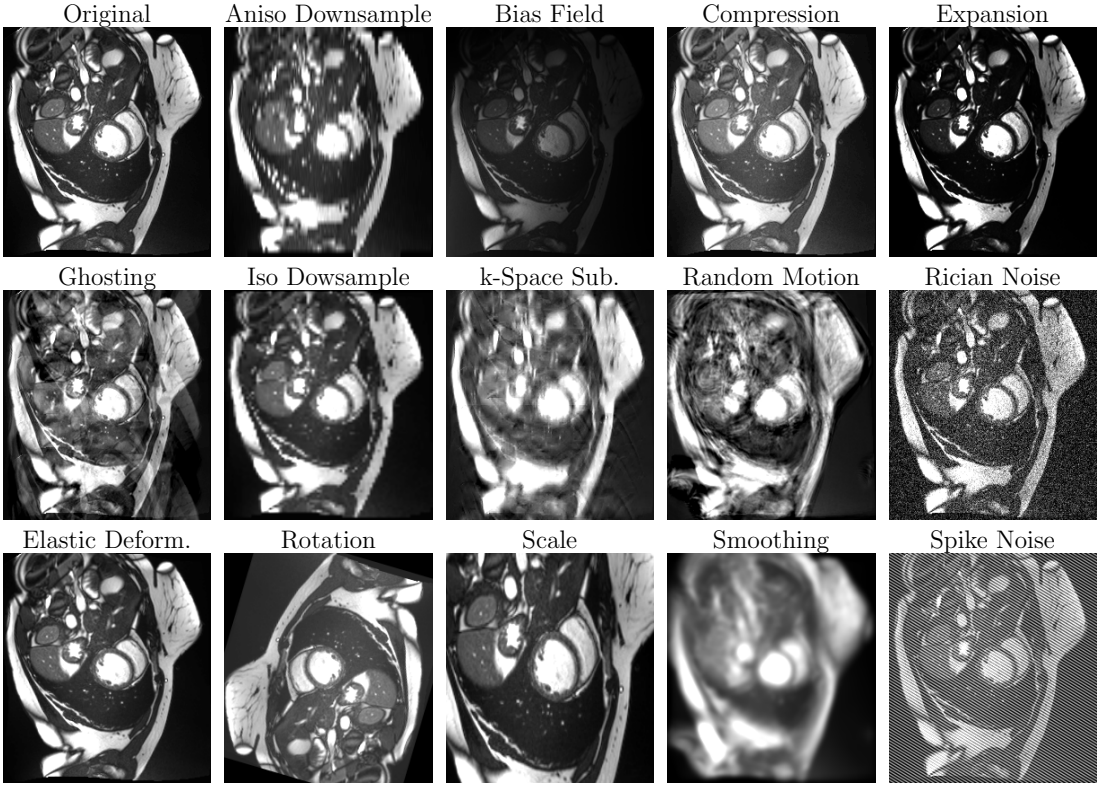

Figure 4: Visualization of the 14 data variations, alongside the original image (top-left) for a test sample in the ACDC dataset. All transforms visualised at severity 3.

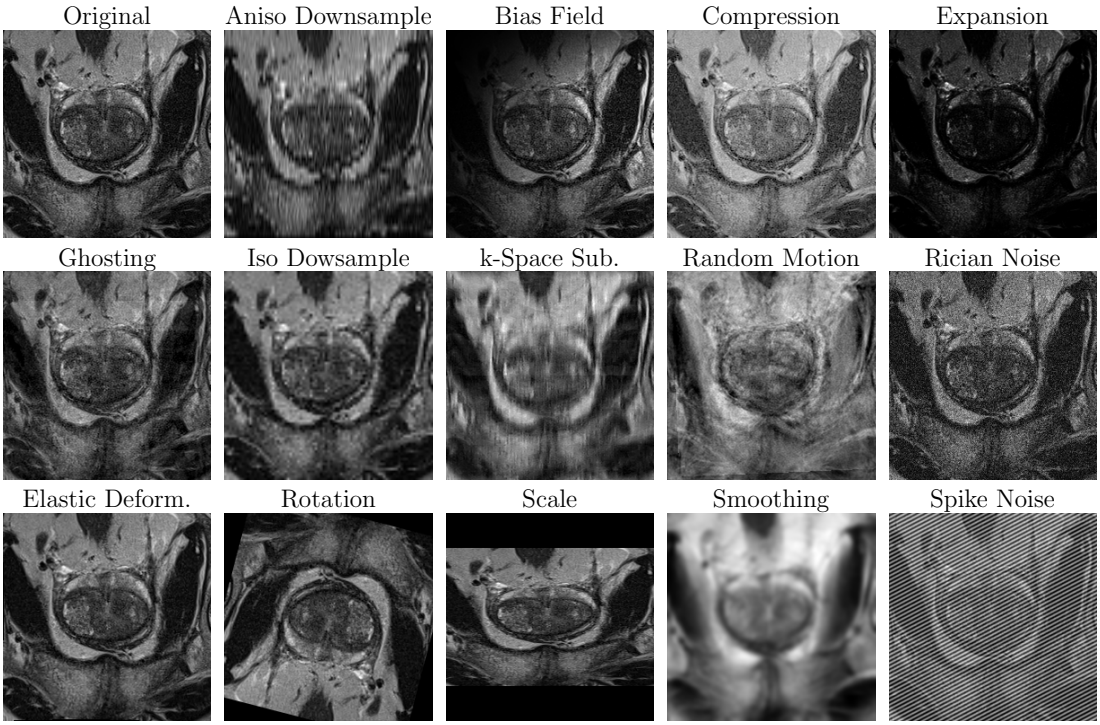

Figure 5: Visualization of the 14 data variations, alongside the original image (top-left) for a test sample in the P158 dataset. All transforms visualised at severity 3.

We see that while base augmentations are really effective in improving generalisation to some transformations, rotation, scale contrast compression and expansion, and iso-downsample. These are transformations that base augmentations also overlap with. However, more complicated transformations are not completely overcome, for instance, bias field, ghosting, rician noise, k-space subsampling, spike noise, smoothing, aniso-downsampling and random motion.

The augmentations MixUp and AFA also seem to complement each other. Using AFA without MixUp can reduce performance on HD95 metric on some transformations, which can be understood by the fact that regularising frequency components leads to difficulty delineating boundaries (typical high frequency changes) while MixUp does not deteriorate boundary delineation, but it is unable to regularise frequency components leading to performance decline on various variations. However, using both MixUp and AFA in general leads to the best performance for each metric, and this pattern is consistent between both test datasets.

Appendix C. Qualitative Segmentation Performance

Fig. 10 visualizes the effect of using base, MixUp, and AFA augmentations on the transformed test set. Notably, while intensity shift augmentation is among the base augmenta-

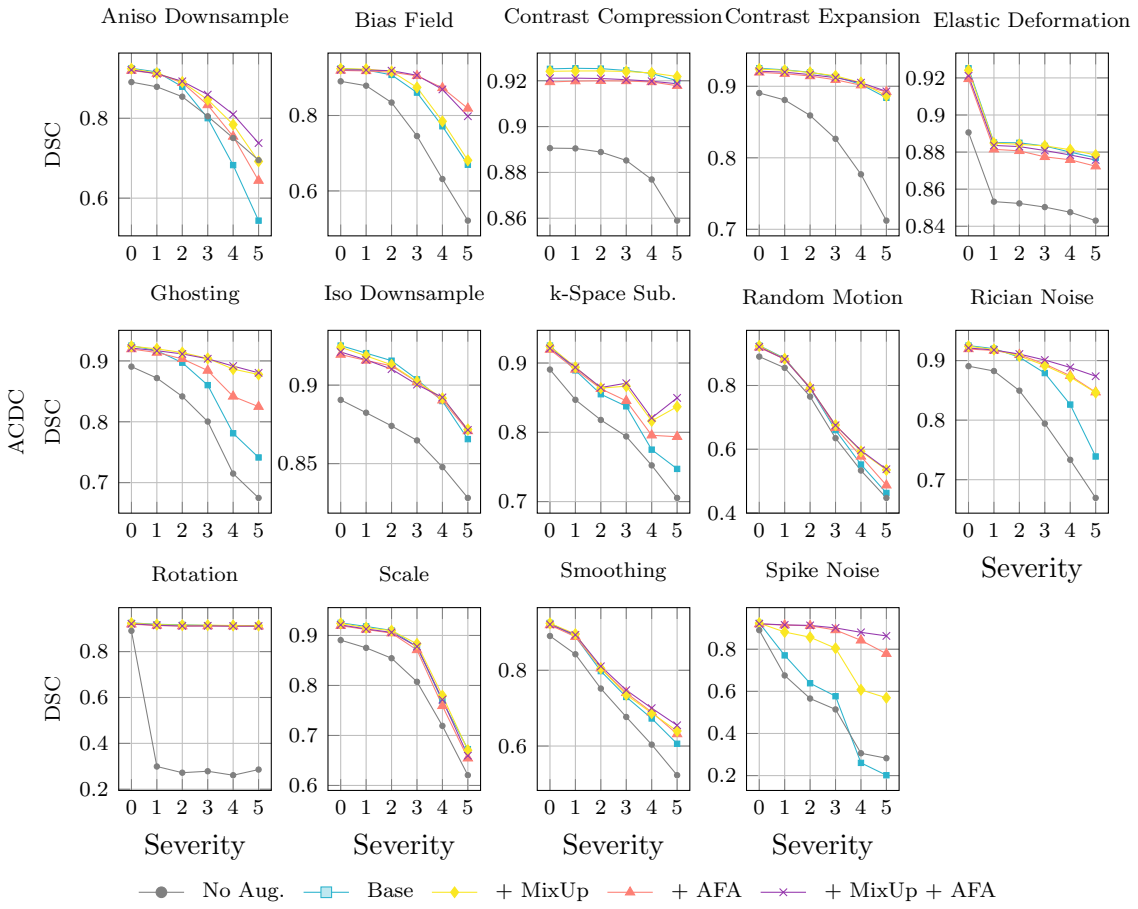

Figure 6: Trend of DSC per severity for test sets transformed with 14 different transformations for the ACDC dataset, including from the 5 repeated from the main paper.

tions, the top row shows that AFA and MixUp are required to correct for the bias field in the data.

Appendix D. Evidence of Regularisation

L2 regularisation is often used in training deep neural networks to reduce overfitting to noise by promoting learnt weights to have a lower l2-norm. We see that while we do not explicitly regularise the models using l2-norm, base augmentations and MixUp lead to substantially lower norms of the learnt convolutional kernels on models trained for both cardiac cine MR (Fig. 11) and prostate bpMRI (Fig. 12).

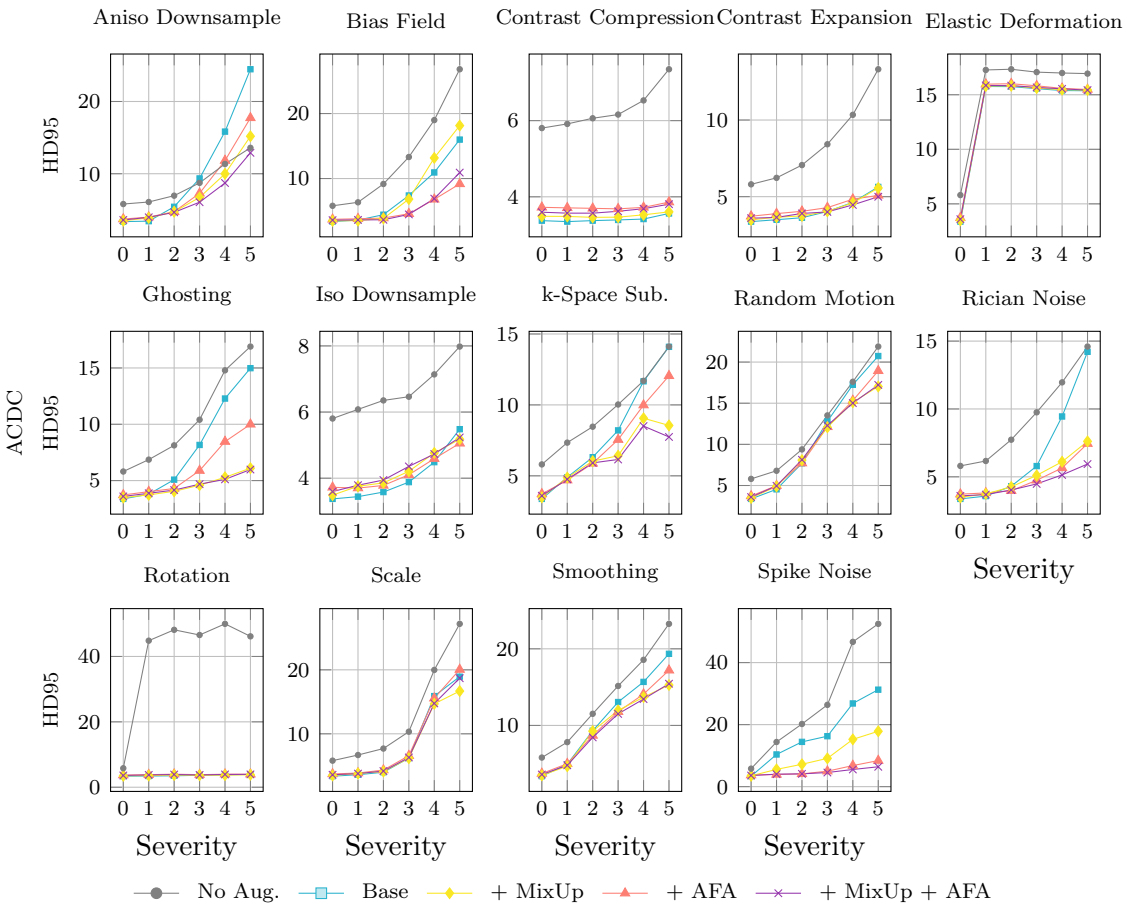

Figure 7: Trend of HD95 per severity for test sets transformed with 14 different transformations for the ACDC dataset.

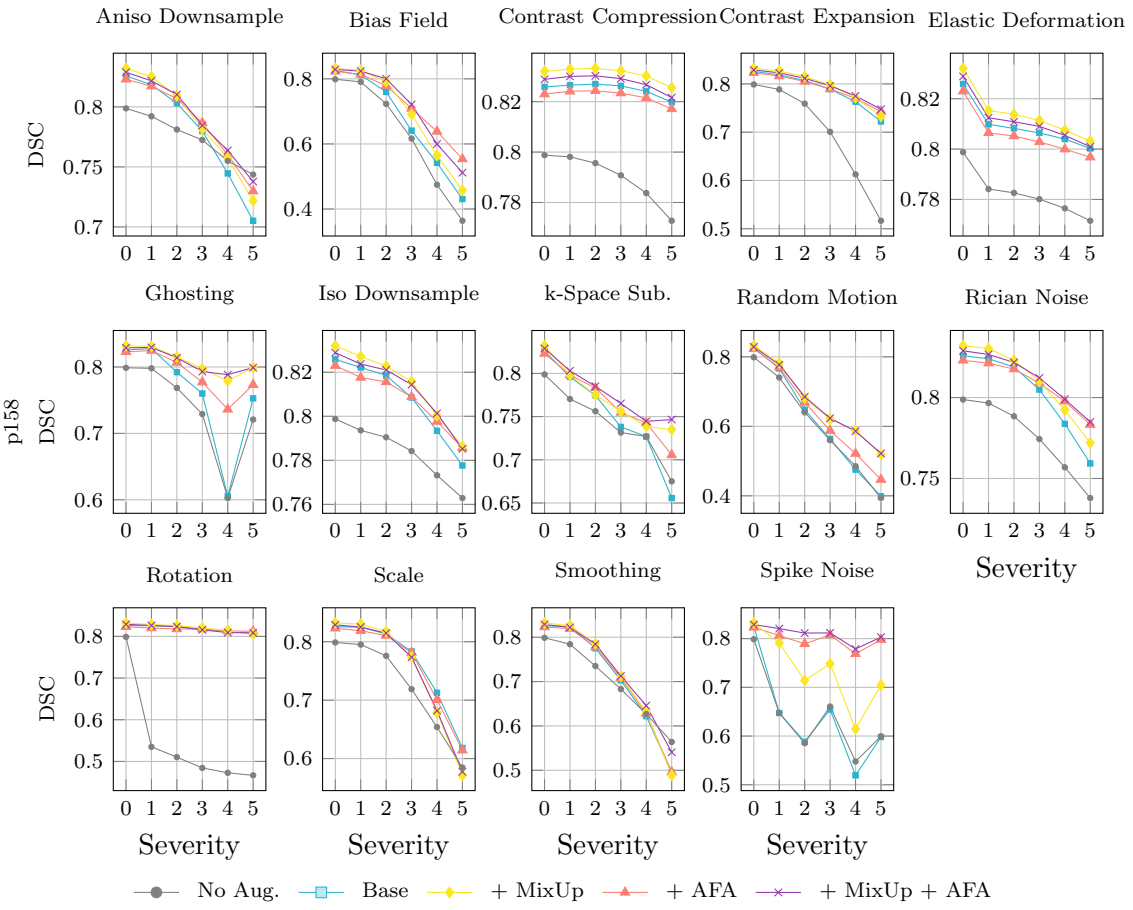

Figure 8: Trend of DSC per severity for test sets transformed with 14 different transformations for the P158 dataset, including the 5 presented from the main paper.

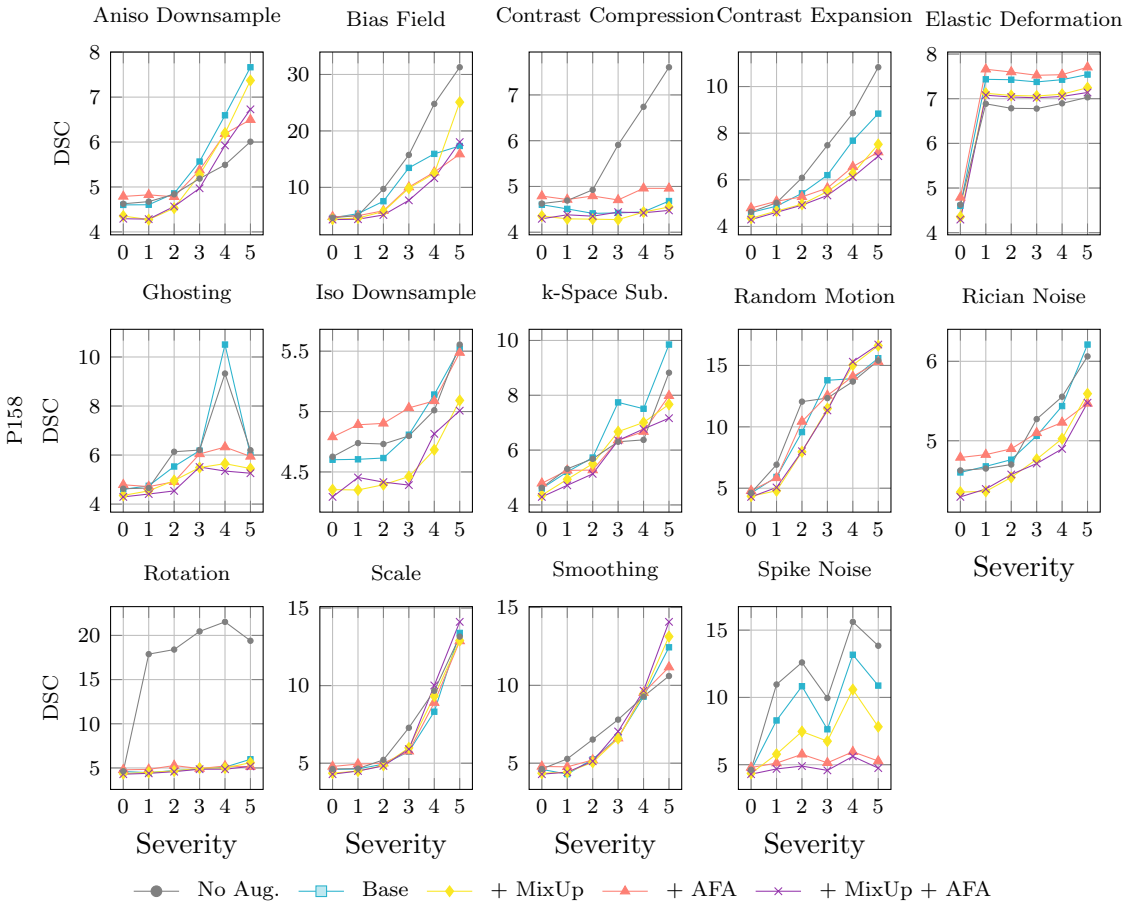

Figure 9: Trend of DSC per severity for test sets transformed with 14 different transformations for the P158 dataset.

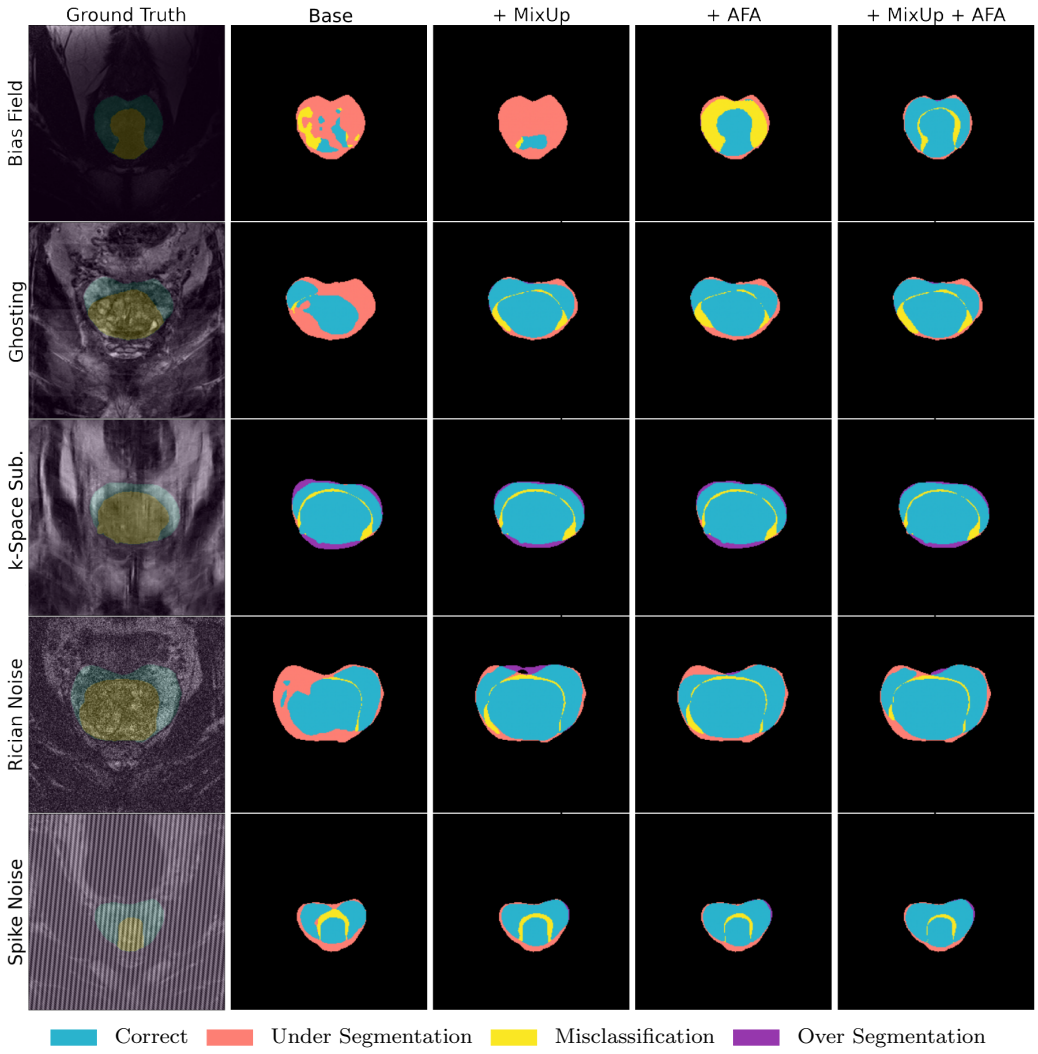

Figure 10: In the first column, we show the ground truth with the expert annotated mask and in the following columns, we show the model predictions made by nnU-Net with augmentations, alone or in combination with MixUp and/or AFA. In presence of AFA and/or MixUp we see a substantial improvement in segmentation coverage under difficult imaging conditions.

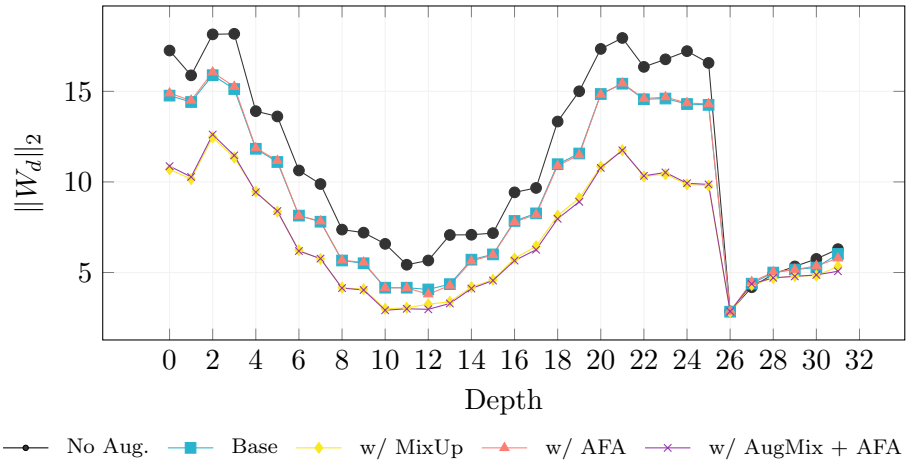

Figure 11: The norm of the weights of convolutional kernels W at different depths, d, for different nnU-Nets trained on ACDC with different augmentation techniques. The plot highlights the regularisation effect the methods have on the model weights.

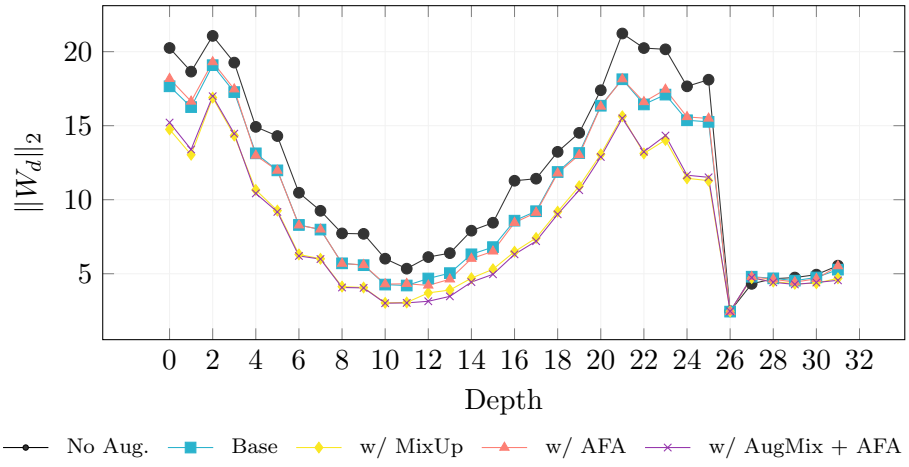

Figure 12: The norm of the weights of convolutional kernels W at different depths, d, for different nnU-Nets trained on P158 with different augmentation techniques. The plot highlights the regularisation effect the methods have on the model weights.