A Study on Data Augmentation Techniques for Visual Defect Detection in Manufacturing

Lars Leyendecker[1][0000−0001−8692−3188], Shobhit Agarwal[2], Thorben Werner[2], Maximilian Motz[1], and Robert H. Schmitt[3]

[1] Fraunhofer Institute for Production Technology IPT, Aachen, Germany
[2] Information Systems and Machine Learning Lab (ISMLL), Hildesheim, Germany
[3] Laboratory for Machine Tools and Production Engineering (WZL) of RWTH Aachen, Germany

Abstract. Deep learning-based defect detection is rapidly gaining importance for automating visual quality control tasks in industrial applications. However, due to usually low rejection rates in manufacturing processes, industrial defect detection datasets are inherent to three severe data challenges: data sparsity, data imbalance, and data shift. Because the acquisition of defect data is highly cost-intensive, and Deep Learning (DL) algorithms require a sufficiently large amount of data, we are investigating how to solve these challenges using data oversampling and data augmentation (DA) techniques. Given the problem of binary defect detection, we present a novel experimental procedure for analyzing the impact of different DA-techniques. Accordingly, pre-selected DA-techniques are used to generate experiments across multiple datasets and DL models. For each defect detection use-case, we configure a set of random DA-pipelines to generate datasets of different characteristics. To investigate the impact of DA-techniques on defect detection performance, we then train convolutional neural networks with two different but fixed architectures and hyperparameter sets. To quantify and evaluate the generalizability, we compute the distances between dataset derivatives to determine the degree of domain shift. The results show that we can precisely analyze the influences of individual DA-methods, thus laying the foundation for establishing a mapping between dataset properties and DA-induced performance enhancement aiming for enhancing DL development. We show that there is no one-fits all solution, but that within the categories of geometrical and color augmentations, certain DA-methods outperform others.

Keywords: Deep Learning · Defect Detection · Data Augmentation · Manufacturing · Visual Quality Control

1 Introduction

Manufacturing processes have been optimized in recent decades to achieve minimum reject rates and high product qualities. However, as product and process complexities increase, the importance of reliable quality continues to grow.

Defects such as internal holes, pits, abrasions, and scratches on workpieces or knots, broken picks, and broken yarn in fabrics [28] negatively impact both visual and functional product properties [36]. Defects also contribute to the additional wastage of resources, safety hazards, and can have severe economic consequences for a company. Therefore, reliably assuring the quality of manufactured products is of paramount importance in manufacturing. One of the famous and contemporary solutions towards achieving the goal of a fully automated quality control system is through deep learning (DL)-based computer vision. DL algorithms improve over existing rule-based systems in terms of generalization and performance, while requiring less domain expertise [9, 19, 22]. However, a major disadvantage of data-driven approaches compared to rule-based techniques lies in the strong dependency of model precision on data quantity, data quality, and the evolution of the data over time (data drift) [31]. While the focus in recent years has been on the development of advanced network architectures (e.g., ResNet-50 [8] or Inception-v3 [32]), the progress that is being made in model-space is increasingly diminishing. As a result, the development is shifting more towards data-centric approaches, especially in real-world domains like for example manufacturing or medical diagnostics. Table 1 provides an overview of the main data challenges that are characteristic for image data acquired from production processes. These properties form a strong contrast to the ones of (research) datasets (e.g., ImageNet [3], COCO [17], MNIST [16]) used for developing and benchmarking of deep neural network architectures and DL-algorithms, which is why the approaches from research are difficult to transfer one-to-one to such complex defect detection use-cases.

Data Quality Issue	**Description**
Amount of data	Difficulty in collecting sufficiently large amounts of data.
Label inconsistencies	Labor-intensive task that is oftentimes ambiguous and usually requires multiple domain-experts
Data imbalance	Defective parts tend to be significantly underrepresented compared to non-defective ones
Changing lightning conditions	Contrasts and brightness changes across different work shifts
Exposure issues	Reflections and shadows cast by complex components
Sensor failure	Image failures or high noise-levels due to sensor degradation amplified by harsh environments
Changing object poses	Especially in mass production often different orientation of components
Changing appearances	Changes in the appearance of a product from time to time can make the data previously collected unusable.

Table 1: Causes of data quality issues in DL-based visual defect detection in terms of data sparsity, data imbalance and data shift

Data augmentation (DA) represents a data-space solution addressing the above mentioned data quality challenges. There are various DA techniques that aim for changing both the geometrical and visual appearance of images to improve both performance and robustness properties of deep neural networks. The most common DA techniques are geometric transformations, color augmentations, kernel filters, mixing images and random erasing [39]. Even though DA is already an integral part of DL pipelines, different DA-methods are often blindly

applied based on empirical knowledge and require elaborate tuning for specific datasets. To analyze the impact of different DA-methods on both precision and generalization for the task of visual defect detection, this paper introduces our experimental procedure in Section 3.3, presents the results in Section 4.2 and finally derives insights about the studied DA-methods in Section 5. Section 3.2 introduces the three real-world datasets which we work with. Our DA-methods are chosen according to a preliminary study of related papers that is summarized in Sections 2 and 3.3.

2 Related Work

This section provides a brief overview of work that addresses the generalization problem, DA approaches, and its impact on real-world DL tasks. One central drawback of real-world datasets is that the models trained on them do not generalize well as these datasets are prone to domain shift [40]. In recent years model-centric techniques such as dropout [29], transfer learning [34], and pre-training [4] have tried to address the issues of generalization, particularly in deep neural networks. DA tries to avoid poor generalization by solving the root problem of training data [27] rather than changing the model or training process. Applications of DA can be found in various works across multiple domains such as natural language processing [6], computer vision [27], and time series classification [11]. Particularly in computer vision tasks DA has been applied to address the domain generalization problem [24, 33, 35]. Many papers exist that apply and analyze basic DA-techniques (e.g., oversampling and data warping on histopathological images [5]) and advanced methods (e.g., stacked DA on medical images [38], style-transfer augmentations [12], cGan, and geometric transformations [21]) for specific use cases and datasets.

Fewer papers exist that provide an overview of DA-methods and try to examine their influences on model accuracy. The survey of Shorten et al. [27] presents a comprehensive overview of DA and present the impact examination of individual methods on well-known datasets (e.g., CIFAR-10, MNIST, Caltech101) in an isolated manner of pairwise comparisons. Shijie et al. [26] explore the impact of various DA-methods on image classification tasks with CNNs. On subsets of CIFAR10 and ImageNet, they conduct pair and triple comparisons to identify best-performing DA-techniques and to draw general conclusions. Yang et al. [37] systematically review different DA-methods and propose a taxonomy of reviewed methods. For semantic segmentation, image classification, and object detection, they compare the performances of different model architectures on datasets (e.g., CIFAR-100, SVHN) with and without pre-defined set of DA-techniques. The survey paper of Khosla et al. [15] presents an overview of selected DA-methods without conducting further effect analyses. In addition to generic studies on scientific datasets, a few domain-specific approaches exist. The only related work on DA in defect detection is provided by Jain et al. [13]. They propose a DA-

framework utilizing GANs which they use to investigate data synthetization for classification of manufacturing datasets.

Scientific Impact Existing studies are almost exclusively conducted on scientific datasets and no reference is made to specific application domains (with the exception of [13]). To the best of our knowledge, there is currently no preliminary work, that examines the impact of DA-methods specific to DL-based visual quality control in manufacturing datasets in an unconstrained setting (i.e. only pairwise evaluations).

3 Approach

In this section, we present our approaches and procedures. Section 3.1 defines the mathematical problem of binary defect detection. Section 3.2 introduces the datasets considered in this study and their properties. The experimental procedure, the domain shift measure, and the evaluation metrics are presented in Section 3.3.

3.1 Binary Defect Detection Problem Definition

For binary visual defect detection, the input feature space is denoted by $\mathcal{X}$ and $\mathcal{Y}$ denotes the target space. We define the domain as a joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$ and the dataset as $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}_i^N$, where N is the number of training examples. In this work, $\mathcal{X}^1$, $\mathcal{X}^2$, $\mathcal{X}^3$ comprises images from three datasets, namely: (1) AITEX fabric defects [28], (2) Magnetic tile defects [10], and (3) TIG Aluminium 5083 welding defects [1]. We define the binary classification problem where $\mathcal{Y} \in$ {Defected, Non-defected}. Furthermore, the DL model is defined as $f : \mathcal{X} \rightarrow \mathcal{Y}$, where the primary objective is to learn a mapping from the input space $\mathcal{X}$ to target space $\mathcal{Y}$. In this work $f \in$ {ResNet-50 [8], Inception-V3 [32]}. The predictions generated using model f are denoted as $\hat{\mathcal{Y}}$. The categorical cross entropy loss function is defined as $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \rightarrow [0, \infty)$. Each dataset $\mathcal{D} = \{(\mathcal{X}_i, \mathcal{Y}_i)\}_i^N$ is augmented using various DAs, where θ denotes the list of all DAs, and a new augmented dataset is generated as $\mathcal{D}^1 = \theta(\mathcal{D})$. For each dataset, ten DA-pipelines with varying DAs are constructed to create ten different data sets $\mathcal{D}^1$ .. $\mathcal{D}^{10}$.

3.2 Presentation of the Datasets

Three real-world industrial-grade datasets are used in this work. An overview of examplary images is provided in Figure 1. The Magnetic tile defects dataset (MagTile) contains a total of 1,344 images of magnetic tiles with five defect types: blowhole, crack, fray, break, uneven (grinding), and free (no defects). AITEX is a fabric production dataset containing 246 images of 4,096 x 256 pixels that capture seven different fabric structures. In total, there are 140 defect-free images, 20 for each type of fabric, and there are a total of 105 images with

defects. The TIG Aluminium 5083 welding seam dataset (TIG5083) contains 33,254 images of aluminium weld seams and the surrounding area of the weld seam, with six classes: good weld, burn through, contamination, lack of fusion, misalignment, and lack of penetration. We convert the multi-class classification task of all datasets into a binary classification problem by merging all individual defect types into a single defect class.

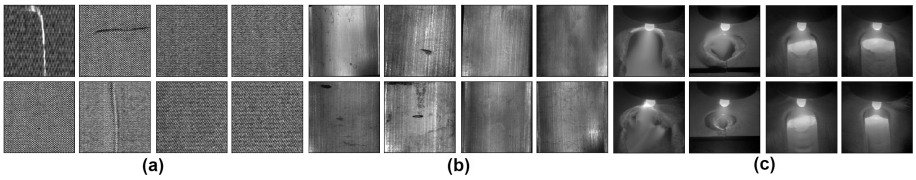

Fig. 1: Exemplary raw images of the datasets studied: AITEX (a), MagTile (b), and TIG5083 (c)

3.3 Experiment Procedure

To evaluate the impact of DA-techniques we propose a three-stage process: First, for each dataset, apply a DA-pipeline and evaluate model performance on different test sets. Second, measure the domain shift between the train set and the test sets. Third, correlate the achieved performance with the domain shift. This framework provides insight into the effects of different DAs on model performance, domain shift, and, through the correlation of both, the generalization capabilities of the trained model. An overview of our algorithm can be found in Figure 2. We assume a standard train-test split of 80/20 and further a validation split of 60/20 (based on the 80% train split). Additionally, we create a hold-out test set by splitting off one of the defect classes per dataset before they are merged (see Section 3.2). This hold-out set serves as an additional out-of-distribution test set to measure the generalization capabilities of the model. We apply DA in two different settings. For AITEX and MagTile, augmented datapoints were added as new instances, retaining the original ones. This was done to increase the overall number of instances in the dataset and stabilize training. For TIG5083, augmented datapoints replace the originals since the dataset already contains enough images for training. The hold-out class for the AITEX data set was 'Broken end', the hold-out class for Magnetic tile defects was 'Crack', and the hold-out class for TIG Aluminium 5083 was 'contamination' class.

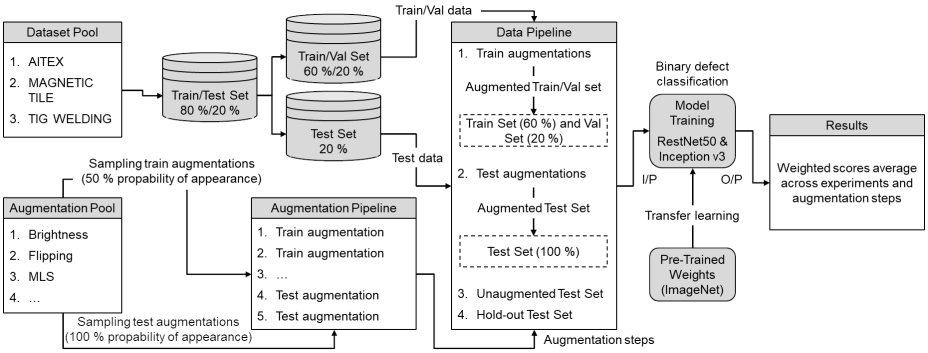

Fig. 2: Experiment protocol for constructing the DA-pipelines, training, and evaluating the defect detection model

Data Augmentation Pipelines In order to pre-select the DA-steps for this paper, a survey was conducted across 24 papers dealing with 6 major industrial image data sets. Table 2 describes all available augmentations for each dataset. From these augmentation pools, different pipelines for each dataset were constructed. For each pipeline, two of the augmentations are reserved for the test set and are later referred to as test augmentations. The remaining DAs have a 0.5 chance of being applied to the training set. This process is repeated ten times (see Table 3). Appendix A.1 provides an overview of selected un-augmented and augmented images for all three datasets.

TIG5083	**AITEX**	**MagTile**
1. Gaussian Noise	1. Gaussian Noise	1. Salt & Pepper Noise
2. Transpose Image	2. Transpose Image	2. Transpose Image
3. Flip Image	3. Flip Image	3. Flip Image
4. Perspective Transformation	4. Random Perspective	4. Random Perspective
5. Add Brightness	5. Color Jitter	5. Color Jitter
6. Affine Transformation	6. Moving Least Squares (MLS)	6. MLS [25]
	7. Random Erase [39]	7. Retinex [14, 23]
	8. Random Rotate	

Table 2: Preselected set of DA-methods for TIG5083, AITEX, and MagTile

Nr.	**Train & Validation augmentations**	**Test augmentations**
1	Random Perspective, Flip Image, Color Jitter	Random Rotate, Transpose Image
2	Flip Image, Random Perspective	Transpose Image, Random Erase
3	Gaussian Noise, Color Jitter, Random Perspective, Random Rotate, Flip Image	Random Erase, MLS
4	Random Erase, MLS, Gaussian Noise	Color Jitter, Random Perspective
5	Random Rotate, MLS, Gaussian Noise	Transpose Image, Flip Image
6	Random Perspective, Random Rotate	Gaussian Noise, MLS
7	Random Erase, AddNoise, Color Jitter	Random Rotate, Random Perspective
8	Random Rotate, Random Erase, Flip Image, Color Jitter	Random Perspective, Gaussian Noise
9	Color Jitter, Random Rotate, Gaussian Noise	MLS, Random Perspective
10	Random Perspective	Random Rotate, MLS

Table 3: Train, validation, and test set DA-Pipelines (AITEX)

Domain Shift Measures We use an algorithm proposed by [30] for measuring the domain shift between datasets. In computer vision tasks, calculating domain shift can be seen as calculating the difference in representation by a model given the source and target domain. Given that a source domain is distant from the target domain, the representation of the domains in the learned space for a specific model tends to diverge. The authors used the activation values from the model's last layers to quantify the domain shift. Specifically, by creating a statistical distribution using each kernel's activation value in those layers, we can measure the distance between the datasets using the Wasserstein distance.

Evaluation Metrics To evaluate the results of the binary classification problem, various metrics such as F1-Score, precision, recall, Jaccard similarity [7], Cohen's kappa score [20], and Matthews correlation coefficient (MCC) [2] are used. Since the datasets are imbalanced even after applying DA, all metrics (Jaccard, precision, recall, and F1-Score) are weighted by the class distribution. We use multiple different evaluation metrics, as they all slightly deviate from each other. In this way, we circumvent the difficulties due to the sensitivity of individual metrics and obtain a more conclusive evaluation. Since all these scores are bound between $[0, 1]$ we average all of them for our reporting of final performance values.

4 Results

In this section, we present the results. Section 4.1 defines the training and implementation procedure. Section 4.2 provides an overview of the protocol followed to evaluate the results at the example of the AITEX dataset. Section 4.3 presents the results of our ablation study.

4.1 Training and Implementation

For controlling the model training, a validation set is split of from the augmented training set. The model is evaluated on the original test set, augmented test sets (using the two reserved test augmentations) and the hold-out set as described in Section 3.3. The hold-out class for the AITEX data set was 'Broken end', the hold-out class for MagTile defects was 'Crack', and the hold-out class for TIG5083 was 'contamination'. As models for our experiment, ResNet-50 and Inception-v3 were chosen, as both are widely used in the literature about industrial applications. The learning rate for both models is set to 10^{-3}, the Adam optimizer [18] is used and the first-layer input shape of the networks is set to 224 and 299 respectively. We initialize the networks using pre-trained weights (ImageNet) for both architectures. DL is enhanced via transfer learning with 50 epochs of frozen weights in the encoder (shallow training) and additional 30 epochs of fine-tuning the entire model (deep training). Similarly to the evaluation metrics, the class-balanced version of the loss function was employed to stabilize the learning process. The data for each experiment was normalized according to the statistics of the train set after applying DA.

4.2 Results for the AITEX Dataset

Figure 3 depicts the average F1-Score across both the models and across the DA steps for each test set. The values are obtained by averaging the performance of each pipeline that contains the respective augmentation. We observe that the performance on the original test and, to a lesser extent, the augmented test set remains stable, but on the hold-out set (highest amount of domain shift) the model performance has significantly deteriorated. The top three DA-steps for AITEX dataset are MLS, Gaussian noise and random rotating. As stated in Section 3.3, we also averaged the performance across multiple other metrics, since they all slightly differ from each other. Similar trends can be observed in Figure 4.

Next, the distance between the train set (source domain) and the test sets (target domain) was calculated for all the models and datasets. Table 4 contains the mean and standard deviation across all the pipelines for the AITEX dataset and ResNet-50 model. The domain shift increases from the original test set to the augmented test set to the hold-out set. Finally, the domain shift is correlated to the respective F1-Scores, as Wasserstein distance alone lacks interpretability.

Train/Test	**Train/Aug_test**	**Train/Hold_out**
0.0764±0.0831	0.0808±0.0804	0.1841±0.1981

Table 4: Domain shift measure averaged across DA-pipelines for the last layer of ResNet-50 (AITEX)

A negative correlation means that with increasing domain shift the performance of the model on the test data decreases. Therefore, a greater correlation is desirable. Table 5 contains the Pearson correlations between the distance measure and F1-Scores across all test sets. Since the domain shift is measured based on a single layer of the model we evaluated the last three layers of each model and reported the values separately in the columns. The correlation values don't change depending on the layer used, but we observe two outliers in the pipelines that display a weaker correlation between domain shift and model performance. Further information can be found in Appendix A.2. The same evaluation protocol was followed for evaluating the results across the other two datases as well and similar trends were observed. The results TIG5083 and MagTile can be found in Appendix A.3.

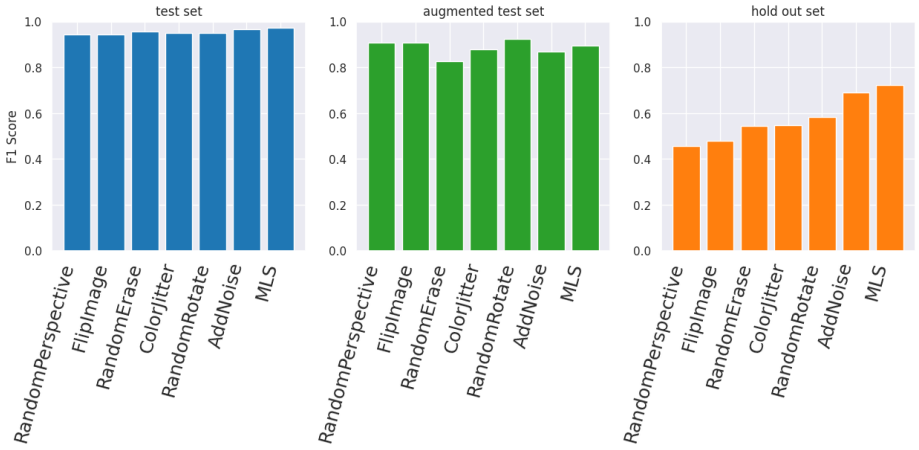

Fig. 3: F1-Score averaged across models for each DA-method (AITEX). Sorted by hold-out performance.

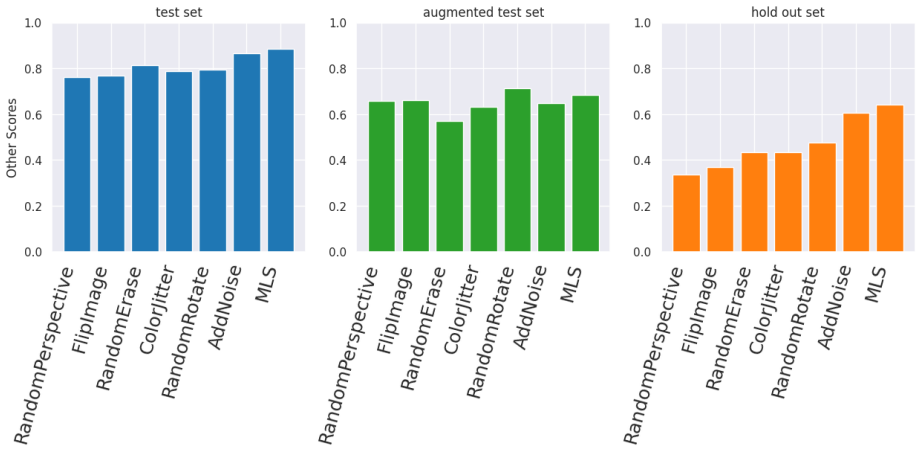

Fig. 4: Averaged Jaccard, precision, recall, kappa and MCC scores across models for each DA-method (AITEX). Sorted by hold-out performance.

Pipeline	Inception v3			ResNet-50			Mean
	Last layer	2nd Last layer	3rd Last layer	Last layer	2nd Last layer	3rd Last layer	
1	-0.996	-0.996	-0.998	-0.999	-0.999	-0.999	-0.998± 0.002
2	-0.727	-0.734	-0.505	-0.941	-0.940	-0.979	-0.804± 0.167
3	-0.989	-0.990	-0.971	-0.960	-0.967	-0.986	-0.977± 0.012
4	-0.970	-0.972	-0.995	-0.352	-0.317	-0.530	**-0.689±0.297**
5	-0.916	-0.916	-0.986	-0.900	-0.905	-0.979	-0.934± 0.035
6	-1.000	-1.000	-1.000	-0.917	-0.931	-0.996	-0.974±0.036
7	-0.999	-0.996	-0.988	-0.935	-0.946	-0.826	-0.948±0.060
8	-0.999	-0.998	-1.000	-0.955	-0.916	-0.974	-0.974±0.0307
9	-0.952	-0.955	-1.000	-0.341	0.121	-0.785	**-0.652±0.411**
10	-0.994	-0.994	-0.991	-0.989	-0.987	-0.994	-0.991±0.003
	-0.954 ± 0.084	-0.955 ± 0.082	-0.943±0.154	-0.829±0.256	-0.779±0.375	-0.905±0.152	

Table 5: Pearson correlations between the domain shift and model F1-Scores (AITEX). The bold values represent the largest negative mean correlations value.

4.3 Results of the Ablation Study

In addition to the average score presented in Section 4.2, we draw additional insights from comparing performances across all models and datasets. Figure 5 depicts the stacked bar plot of weighted F1-Scores averaged across all datasets and models for each augmentation that was available for the dataset. Across all the experiments, affine transformations, moving least squares (MLS) and random rotation DA techniques performed the best. Similarly, Figure 6 depicts the average of the scores across all other evaluation metrics. We can observe similar trends where on average across experiments affine transformations, perspective transformation and MLS perform the best.

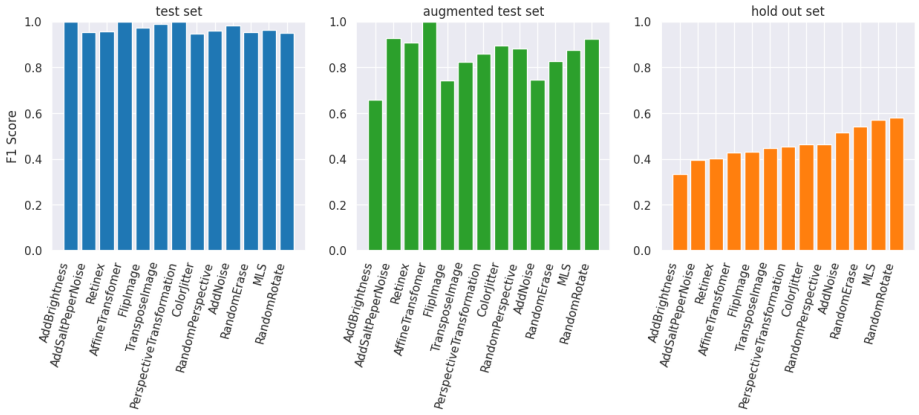

Fig. 5: F1-Scores averaged across all augmentations steps in the train sets

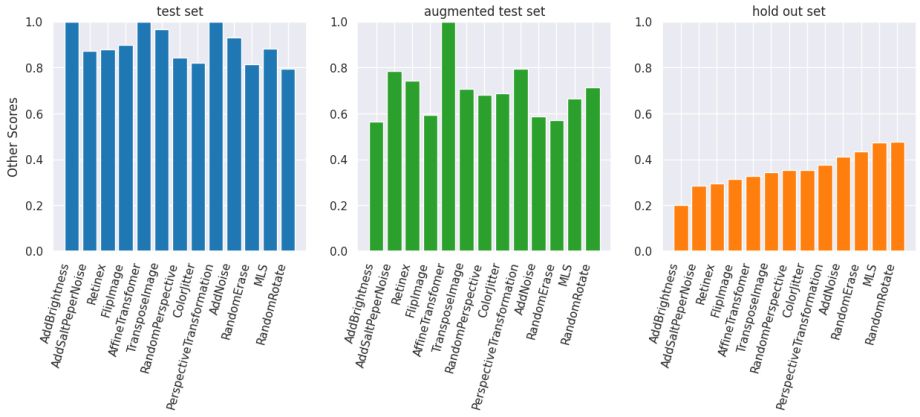

Fig. 6: Jaccard, precision, recall, kappa and MCC scores averaged across all augmentations steps in the train sets

5 Conclusion

DL offers enormous potential to automate complex visual quality control tasks that cannot be solved using rule-based methods. However, manufacturing applications entail three severe data challenges: data sparsity, data imbalance and data shift. DA-methods have become an integral part of DL-pipelines to improve both performance and generalization. To provide precise assistance for the selection of DA-methods for developing DL-based quality control in the future, in this paper, we present an experiment protocol. Thereby, we aim to evaluate the impact of individual DA-methods on defect detection performance depending on dataset characteristics. We apply this protocol to three defect detection use-cases, present and interpret the results.

Using our approach, we can evaluate the influences of each DA method on the model metrics in detail. We show how to determine the domain shift between genuine and augmented dataset derivatives and therefore providing a measure and interpretability for choosing the degree of DA. By correlating this domain shift with F1-Scores, the strength of the positive influence of a DA-pipeline on bridging the domain shift can be determined. Applying our protocol to the datasets, we obtain the three best DA-methods MLS, Gaussian noise, random rotating (AITEX), image transpose, random perspective, salt & pepper noise (MagTile), and affine transformation, perspective transformation, image transpose (TIG5083). Thereby we confirm that the performance improvement of DA-methods depends on dataset characteristics, the DL-task to be solved and the degree of DA. This shows that there is no one-fits-all solution, but at the same time makes it all the more clear that establishing a mapping between dataset properties (e.g., degree of imbalance, defect sizes, positional variance of defects) and DA-induced performance enhancement will enable tailor-made and precise

DL-pipeline development, especially in real-world applications.

Correlating the found performances with the respective domain shift revealed additional insights. The two pipelines for the AITEX dataset that induced the weakest negative correlation between domain shift and performance were mainly composed of our three best-performing augmentations for that dataset (see table 5 pipeline 4,9). Additionally, we found that the worst performing pipelines either had very few augmentations or contained badly performing augmentations in them (mainly "random rotate" for AITEX), further highlighting the need for tailor-made DA-pipelines for each dataset. Our ablation study showed that (in contrast), by averaging the results over all datasets and models, at least some augmentations do perform better than others **on average**. The better-performing augmentations are the more complex ones, showcasing their versatility and robustness, while simple of-the-shelf augmentations display the least amount of lift in model performance. Figure 6 can serve as a benchmark of augmentation techniques for new industrial-grade datasets, or those with unknown properties.

With the proposed protocol, we lay the foundation for determining the appropriateness of DA-methods for specific data properties in an analytical approach. We will include also more advanced DA-methods and extend the study to additional domain-specific datasets to provide more validity to the results. By establishing a catalog of dataset properties to which we can map the results of the study, we aim to develop a domain-specific decision support system for choosing optimal DA-pipelines for DL-enhanced visual quality control applications in manufacturing.

Acknowledgements

The results presented in this paper originate from the master's thesis of Shobhit Agarwal titled *Analysis of data augmentation techniques for deep learning-based visual defect detection in manufacturing*. This work was part of a collaboration between Fraunhofer Institute for Production Technology IPT and the Information Systems and Machine Learning Lab (ISMLL) of University of Hildesheim.

The research in this paper was partially supported by the European Commission through the H2020 project EPIC (https://www.centre-epic.eu/) under grant No. 739592.

Funded by the Lower Saxony Ministry of Science and Culture under grant number ZN3492 within the Lower Saxony "Vorab" of the Volkswagen Foundation and supported by the Center for Digital Innovations (ZDIN).

References

1. Bacioiu, D., Melton, G., Papaelias, M., Shaw, R.: Automated defect classification of aluminium 5083 tig welding using hdr camera and neural networks. Journal of Manufacturing Processes **45**, 603–613 (2019), https://www.sciencedirect.com/science/article/pii/S1526612519302245
2. Chicco, D., Jurman, G.: The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics **21** (2020). https://doi.org/10.1186/s12864-019-6413-7
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
4. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research **11**(19), 625–660 (2010), http://jmlr.org/papers/v11/erhan10a.html
5. Faryna, K., van der Laak, J., Litjens, G.: Tailoring automated data augmentation to h&e-stained histopathology. In: Medical Imaging with Deep Learning (2021), https://openreview.net/forum?id=JrBfXaoxbA2
6. Feng, S.Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., Hovy, E.: A survey of data augmentation approaches for NLP. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 968–988. Association for Computational Linguistics, Online (2021), https://aclanthology.org/2021.findings-acl.84
7. Hancock, J.: Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient) (2004). https://doi.org/10.1002/9780471650126.dob0956
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
9. Hssayeni, M., Saxena, S., Ptucha, R., Savakis, A.: Distracted driver detection: Deep learning vs handcrafted features. Electronic Imaging **2017**, 20–26 (2017). https://doi.org/10.2352/ISSN.2470-1173.2017.10.IMAWM-162
10. Huang, Y., Qiu, C., Guo, Y., Wang, X., Yuan, K.: Surface defect saliency of magnetic tile. In: 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE). pp. 612–617 (2018). https://doi.org/10.1109/COASE.2018.8560423
11. Iwana, B.K., Uchida, S.: An empirical survey of data augmentation for time series classification with neural networks. PLoS ONE **16** (2021)
12. Jackson, P.T.G., Abarghouei, A.A., Bonner, S., Breckon, T.P., Obara, B.: Style augmentation: Data augmentation via style randomization. CoRR **abs/1809.05375** (2018), http://arxiv.org/abs/1809.05375
13. Jain, S., Seth, G., Paruthi, A., Soni, U., Kumar, G.: Synthetic data augmentation for surface defect detection and classification using deep learning. Journal of Intelligent Manufacturing **33**(4), 1007–1020 (2022). https://doi.org/10.1007/s10845-020-01710-x
14. Jobson, D., Rahman, Z., Woodell, G.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Transactions on Image Processing **6**(7), 965–976 (1997). https://doi.org/10.1109/83.597272
15. Khosla, C., Saini, B.S.: Enhancing performance of deep learning models with different data augmentation techniques: A survey. In: 2020 International Conference

on Intelligent Engineering and Management (ICIEM). pp. 79–85. IEEE (2020). https://doi.org/10.1109/ICIEM48762.2020.9160048
16. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010), http://yann.lecun.com/exdb/mnist/
17. Lin, T., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. CoRR **abs/1405.0312** (2014), http://arxiv.org/abs/1405.0312
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
19. Marnissi, M.A., Fradi, H., Dugelay, J.L.: On the discriminative power of learned vs. hand-crafted features for crowd density analysis. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2019). https://doi.org/10.1109/IJCNN.2019.8851764
20. McHugh, M.: Interrater reliability: The kappa statistic. Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB **22**, 276–82 (2012). https://doi.org/10.11613/BM.2012.031
21. Meister, S., Wermes, M.A.M., Stüve, J., Groves, R.M.: Review of image segmentation techniques for layup defect detection in the Automated Fiber Placement process. Journal of Intelligent Manufacturing **32**(8), 2099–2119 (2021), https://ideas.repec.org/a/spr/joinma/v32y2021i8d10.1007_s10845-021-01774-3.html
22. Minhas, M.S., Zelek, J.S.: Defect detection using deep learning from minimal annotations. In: Farinella, G.M., Radeva, P., Braz, J. (eds.) Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, Volume 4: VISAPP, Valletta, Malta. pp. 506–513. SCITEPRESS (2020), https://doi.org/10.5220/0009168005060513
23. Petro, A.B., Sbert, C., Morel, J.M.: Multiscale Retinex. Image Processing On Line pp. 71–88 (2014), https://doi.org/10.5201/ipol.2014.107
24. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. CoRR **abs/2003.13216** (2020), https://arxiv.org/abs/2003.13216
25. Schaefer, S., McPhail, T., Warren, J.: Image deformation using moving least squares. ACM Trans. Graph. **25**(3), 533–540 (2006), https://doi.org/10.1145/1141911.1141920
26. Shijie, J., Ping, W., Peiyi, J., Siping, H.: Research on data augmentation for image classification based on convolution neural networks. In: 2017 Chinese Automation Congress (CAC). pp. 4165–4170 (2017). https://doi.org/10.1109/CAC.2017.8243510
27. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**, 60 (2019), https://doi.org/10.1186/s40537-019-0197-0
28. Silvestre-Blanes, J., Albero-Albero, T., Miralles, I., Pérez-Llorens, R., Moreno, J.: A public fabric database for defect detection methods and results. Autex Research Journal **19**(4), 363–374 (2019), https://doi.org/10.2478/aut-2019-0035
29. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
30. Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: Measuring domain shift for deep learning in histopathology. IEEE Journal of Biomedical and Health Informatics **25**(2), 325–336 (2021). https://doi.org/10.1109/JBHI.2020.3032060
31. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era (2017), http://arxiv.org/pdf/1707.02968v2

32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016). https://doi.org/10.1109/CVPR.2016.308
33. Wan, C., Shen, X., Zhang, Y., Yin, Z., Tian, X., Gao, F., Huang, J., Hua, X.S.: Meta convolutional neural networks for single domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4682–4691 (2022)
34. Weiss, K., Khoshgoftaar, T., Wang, D.: A survey of transfer learning. Journal of Big Data **3** (2016). https://doi.org/10.1186/s40537-016-0043-6
35. 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
36. Yang, J., Li, S., Wang, Z., Dong, H., Wang, J., Tang, S.: Using deep learning to detect defects in manufacturing: A comprehensive survey and current challenges. vol. 13 (2020). https://doi.org/10.3390/ma13245755
37. Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., Shen, F.: Image data augmentation for deep learning: A survey (2022), http://arxiv.org/pdf/2204.08610v1
38. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S.A., Turkbey, B., Roth, H., Myronenko, A., Xu, D., Xu, Z.: When unseen domain generalization is unnecessary? rethinking data augmentation. CoRR **abs/1906.03347** (2019), http://arxiv.org/abs/1906.03347
39. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)
40. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Change Loy, C.: Domain Generalization in Vision: A Survey. arXiv e-prints arXiv:2103.02503 (2021)

A Appendix

A.1 Dataset Illustrations

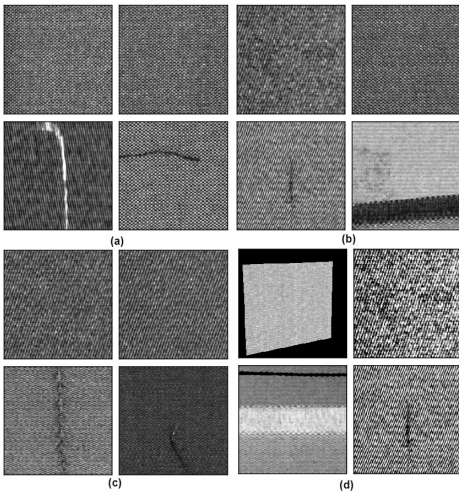

Fig. 7: Selection of AITEX [28] images: train set (a), test set (b), hold-out set (c), and augmented test set (d)

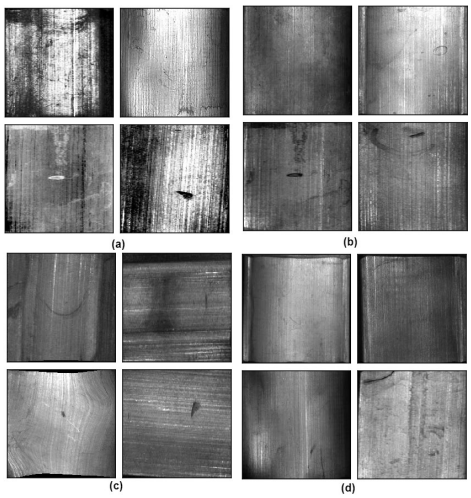

Fig. 8: Selection of MagTile [10] images: train set (a), test set (b), hold-out set (c), and augmented test set (d)

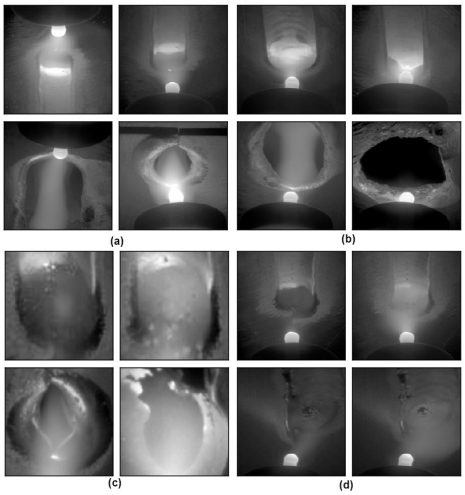

Fig. 9: Selection of TIG5083 [1] images: train set (a), test set (b), hold-out set (c), and augmented test set (d)

A.2 Domain Shift Calculations

The distance measure does not have good interpretability alone. Hence, we correlate the distance measure to the F1-Scores, a negative correlation is expected between them where the distance should be smaller, and the F1-Scores should be higher. Table 6 provides the distance measures for the averagepool layer of the ResNet-50 model across train and test sets, where the first three columns represent the distance and the following three columns represent the F1-score for the same pipelines. We take Pearson correlations along each pipeline, correlating the distance measure with the corresponding performance metric. Similarly, repeating this process for the last layers of both the models gives us Table 5. The same procedure was followed to construct similar tables for MagTile defects and TIG5083 dataset. Furthermore, we take the mean across the last layers of the models.

Pipeline	**Averagepool layer ResNet-50(Distance)**			**Test results on ResNet-50**		
	Train/Test	Train/Aug_test	Train/Hold_out	Test	Aug_test	Hold_out
1	0.0314	0.04008	0.13468	0.93218	0.91231	0.36298
2	0.30067	0.29725	0.72971	0.94262	0.81365	0.56211
3	0.05871	0.02578	0.14406	0.93388	0.92921	0.56211
4	0.01909	0.09311	0.0724	0.95699	0.90933	0.56211
5	0.07192	0.03547	0.16015	0.95217	0.92163	0.77694
6	0.09642	0.05205	0.16251	0.92431	0.92431	0.36298
7	0.04293	0.09987	0.16797	0.94673	0.87684	0.36298
8	0.02218	0.02751	0.03917	0.92921	0.92262	0.36298
9	0.03592	0.06176	0.0556	0.94262	0.92431	0.6417
10	0.0843	0.07488	0.17482	0.94046	0.89822	0.36298

Table 6: Wasserstein distance between the augmented train set and all test sets for ResNet-50 and corresponding model F1-Scores

A.3 Results

MagTile Dataset

Nr.	**Train & validation Augmentation**	**Test Augmentation**
1	Color Jitter, Salt & Pepper Noise	Flip Image, Transpose Image
2	Random Perspective, Flip Image	Salt & Pepper Noise, Retinex
3	Retinex, MLS	Salt & Pepper Noise, Random Perspective
4	Transpose Image, Random Perspective	MLS, Retinex
5	Color Jitter, Retinex, Salt & Pepper Noise	MLS, Flip Image
6	MLS	Retinex, Salt & Pepper Noise
7	Retinex, Color Jitter, MLS	Flip Image, Transpose Image
8	Random Perspective	MLS, Flip Image
9	Transpose Image	Random Perspective, Flip Image
10	Salt & Pepper Noise, Flip Image, Random Perspective, Retinex	MLS, Transpose Image

Table 7: Train, validation and test set DA-Pipelines (MagTile)

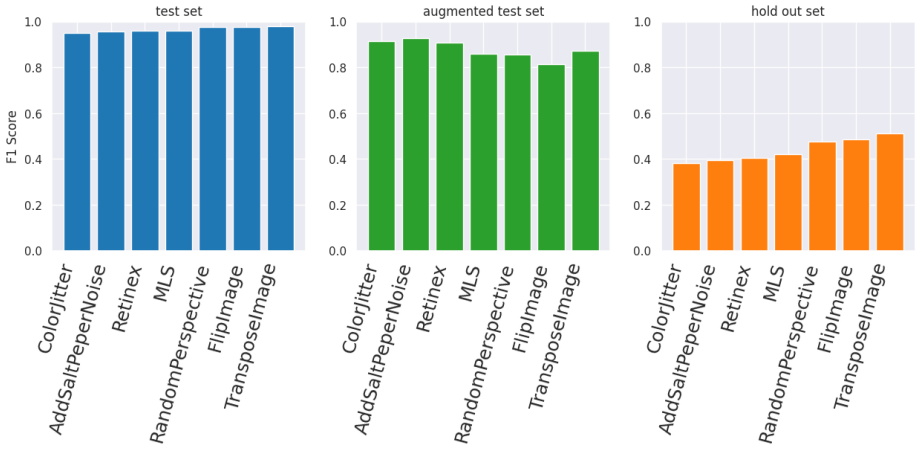

Fig. 10: F1-Scores averaged across models for each DA-method (MagTile). Sorted by hold-out performance.

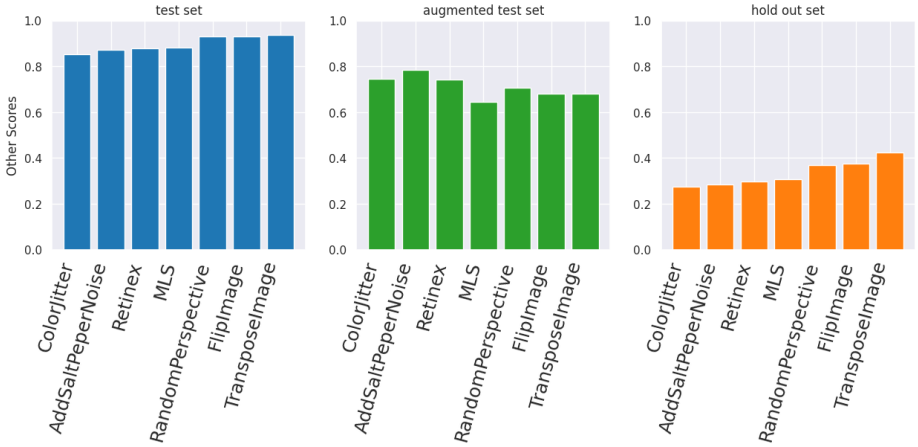

Fig. 11: Jaccard, precision, recall, kappa and MCC scores averaged across models for each DA-method (MagTile). Sorted by hold-out performance.

Pipeline	**Inception v3**			**ResNet-50**			**Mean**
	Last layer	2nd Last layer	3rd Last layer	Last layer	2nd Last layer	3rd Last layer	
1	-0.42134	-0.29361	-0.17805	-0.908	-0.92141	-0.99612	-0.61976±0.3308
2	-0.87905	-0.77159	-0.75817	-0.98297	-0.91635	-0.98849	-0.88277±0.09151
3	-0.99165	-0.86321	-0.89364	-0.99371	-0.95432	-0.97724	-0.94563±0.05
4	-0.07302	-0.64191	-0.95817	0.86934	-0.99643	-0.63085	**-0.40517±0.64512**
5	-0.99206	-0.99545	-0.99109	-0.99631	-0.99184	-0.99959	-0.99439±0.00302
6	-0.27443	-0.28848	-0.41971	-0.85509	-0.91592	-0.90101	**-0.60911±0.28593**
7	-0.99722	-0.99293	-0.9993	-0.40914	-0.5799	-0.98809	-0.82776±0.24077
8	-0.98136	-0.87428	-0.90671	-0.71458	-0.83115	-0.99136	-0.88324±0.09408
9	-0.98295	-0.93676	-0.82706	-0.94623	-0.96887	-0.9424	-0.93404±0.05046
10	-0.98475	-0.97233	0.05444	-0.98154	-0.99635	-0.9711	-0.8086±0.38606
	-0.7578 ± 0.3574	-0.7631 ± 0.2715	-0.6877±0.3741	-0.6918±0.5782	-0.9073±0.1258	-0.9386±0.1123	

Table 8: Pearson correlations between the domain shift and model F1-Scores (MagTile). The bold values represent the largest negative mean correlations value.

TIG5083 Dataset

Nr.	Train & validation Augmentation	Test Augmentation
1	Add Brightness	Affine Transfomer, Perspective Transformation
2	Add Brightness, Gaussian Noise	Transpose Image, Affine Transfomer
3	Transpose Image, Perspective Transformation, Affine Transfomer	Flip Image, Gaussian Noise
4	Gaussian Noise, Perspective Transformation	Transpose Image, Flip Image
5	Transpose Image, Affine Transfomer, Add Brightness	Gaussian Noise, Flip Image
6		Transpose Image, Add Brightness
7	Gaussian Noise, Transpose Image	Perspective Transformation, Flip Image
8	Gaussian Noise	Add Brightness, Affine Transfomer
9	Transpose Image, Add Brightness, Flip Image	Perspective Transformation, Gaussian Noise
10.	Perspective Transformation, Gaussian Noise	Flip Image, Transpose Image

Table 9: Train, validation and test set DA-Pipelines (TIG5083)

Pipeline	Inception v3			ResNet-50			Mean
	Last layer	2nd Last layer	3rd Last layer	Last layer	2nd Last layer	3rd Last layer	
1	-0.81792	-0.81945	-0.82353	-0.82422	-0.85945	-0.83561	-0.83003±0.01433
2	-0.95836	-0.95362	-0.96833	-0.99884	-0.9493	-0.93576	-0.9607±0.01966
3	-0.92238	-0.90818	-0.24479	-0.91736	-0.9626	-0.90916	-0.81074±0.25376
4	-0.61701	-0.84784	-0.78521	-0.7663	-0.8051	-0.85062	**-0.77868±0.07852**
5	-0.99923	-0.96855	-0.93786	-0.99872	-0.98831	-0.98185	-0.97909±0.02119
6	-0.76129	-0.76871	-0.8904	-0.87209	-0.82126	-0.8541	-0.82798±0.04921
7	-0.98311	-0.98905	-0.97248	-0.47366	-0.33301	-0.36318	**-0.68575±0.29891**
8	-0.9528	-0.96702	-0.93089	-0.95038	-0.99583	-0.96655	-0.96058±0.01985
9	-0.80877	-0.87399	-0.75693	-0.7757	-0.78879	-0.83407	-0.80638±0.03882
10	-0.77332	-0.98553	-0.73991	-0.95172	-0.90796	-0.94447	-0.88382±0.09322
	-0.8594 ± 0.1236	-0.9082 ± 0.0773	-0.805±0.2152	-0.8529±0.1579	-0.8412±0.1942	-0.8475±0.1789	

Table 10: Pearson correlations between the domain shift and model F1-Scores (TIG5083). The bold values represent the largest negative mean correlations value.

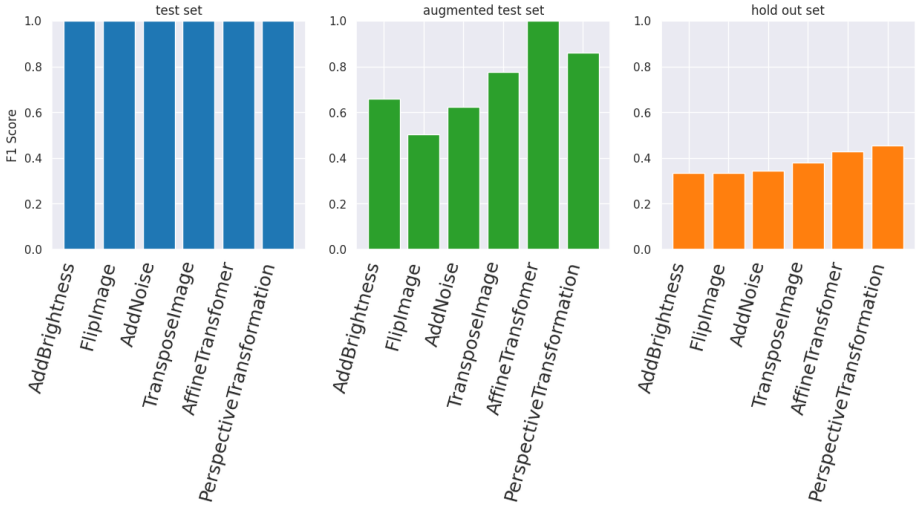

Fig. 12: F1-Scores averaged across models for each DA-method (TIG5083). Sorted by hold-out performance.

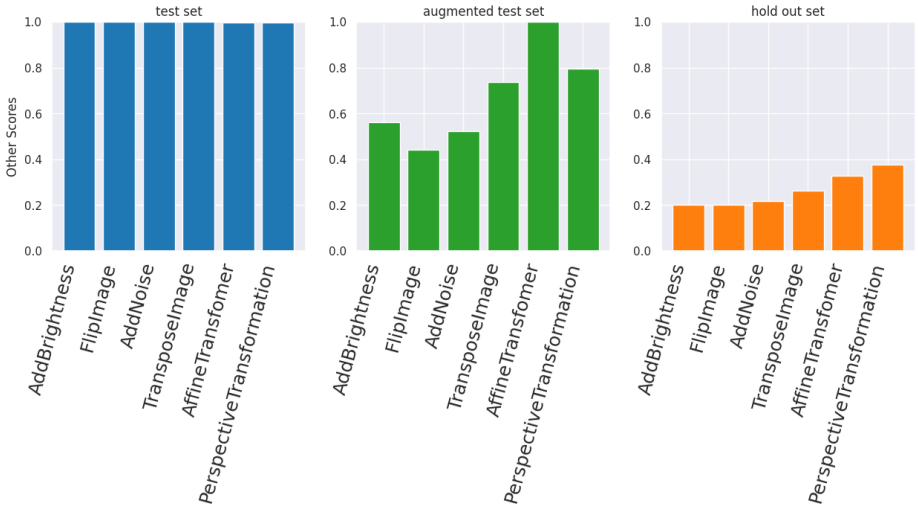

Fig. 13: Jaccard, precision, recall, kappa and MCC scores averaged across models for each DA-method (TIG5083). Sorted by hold-out performance.