engineering proceedings

MDPI

Proceeding Paper

Development of Digital NDT Methodology: Data Augmentation for Automated Fluorescent Penetrant Inspection of Aircraft Engine Blades †

Milan T. Bril, Daniel Friesen * and Konstantinos Stamoulis

Faculty of Technology, Amsterdam University of Applied Sciences, Rhijnspoorplein 2, 1091 GC Amsterdam, The Netherlands

* Correspondence: d.friesen@hva.nl

† Presented at the 14th EASN International Conference on "Innovation in Aviation & Space towards sustainability today & tomorrow", Thessaloniki, Greece, 8–11 October 2024.

Abstract: Fluorescent Penetrant Inspection (FPI) is a widely used inspection technique in the aerospace industry. Because of the aging aerospace sector, and because of the safety-criticality of the inspection, aerospace companies aim to automate (parts of this) inspection process to support inspectors. This paper focuses on a model that can assist inspectors by detecting (possible) defects. YOLOv8 is selected as the object detection model. For training such models, a dataset of sufficient size and variety is necessary to ensure good performance and to prevent overfitting. Because data acquisition is still in its beginning stages, an insufficient amount of data has been acquired. In this paper, we propose a data augmentation technique named Mosaic to artificially create more training data. This technique is tested by applying it to the acquired dataset numerous times and using the resulting dataset to train models with a static architecture (YOLOv8), after which the trained models are evaluated. The best trained model had a 0.834 mAP(50-95) performance, which is an increase of 0.666 mAP(50-95) over its baseline (the model trained on the dataset without data augmentation applied). The results show that, by using this Mosaic technique, promising object detection performance via Deep Convolutional Neural Networks (DCNNs) can be achieved even when the data are limited.

Keywords: data augmentation; automation; NDT digitization

check for updates

Academic Editors: Spiros Pantelakis, Andreas Strohmayer and Nikolaos Michailidis

Published: 18 March 2025

Citation: Bril, M.T.; Friesen, D.; Stamoulis, K. Development of Digital NDT Methodology: Data Augmentation for Automated Fluorescent Penetrant Inspection of Aircraft Engine Blades. *Eng. Proc.* **2025**, *90*, 63. https://doi.org/10.3390/engproc2025090063

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

'Fluorescent Penetrant Inspection (FPI) is the most widely used Non-Destructive Testing (NDT) method in the aerospace industry' [1]. Aerospace companies utilize this inspection technique to locate defects, such as cracks and corrosion, in their aircraft components. FPI is a safety-critical inspection signified by the documentation of multiple engine failures, which resulted from defects being missed during FPI [2–4].

The inspection is currently conducted by human inspectors, which presents certain challenges. First, the aviation industry faces an aging workforce and the impending retirement of a large number of highly experienced and qualified inspectors. Second, the task of component inspection through FPI can be repetitive and presents a high workload; for example, an FPI inspection of a specific engine blade can necessitate the inspection of 72 identical components one after another. Third, this method's consistency and reliability is highly dependent on human factors such as the skills and experience of the operator [5,6].

To investigate potential options and to support FPI inspectors in their task, Deep Convolutional Neural Networks (DCNNs) can be employed for object detection purposes. As the automation process is still in its beginning stages, the current aim of these object detection models is to support human inspectors rather than serve as fully automated systems. Large quantities of data with a large variety are required to train such models to prevent them from overfitting. In the current project, FPI data are limited as image acquisition is still in development. This paper proposes and evaluates the data augmentation technique Mosaic as a way to resolve the problem of sparse data.

2. Background

2.1. Fluorescent Penetrant Inspection (FPI)

FPI is a visual inspection technique first introduced in the rail industry in 1941 [1] and makes use of a penetrant dye that lights up under a specific type of lighting (UV for the dye used in this FPI setup). This dye is applied onto the component and the surface of the component is cleaned afterward (also known as pre-cleaning). The idea of this is that the dye should only reside in openings and cavities of the material after this pre-cleaning process. Then, the inspection can be divided into two phases:

1. Locating possible defects (also known as *indications*): During the first phase, an inspection takes place in a room excluded from any lighting but UV light (that the inspector carries), under which the dye lights up. This causes possible defects to be more easily detectable. In this phase, the inspector looks only for parts of the component that look like defects (i.e., indications) and marks them by drawing a circle around them.
2. Assessing the indications: In the second phase, the indications are further expected by a different department to determine whether the indications are defects, which may lead to the component needing replacement.

By using the outlet guide vane (OGV) (see Figure 1) as a component, our focus in this paper lies solely on aiding inspectors during the first phase of FPI inspections of this component. Thus, the model tries to detect possible defects (also known as indications) on images of components treated with fluorescent dye (e.g., see Figure 2). Indications are divided into four categories: corrosion, cracks, dents, and greensea, as shown in Figure 3.



Figure 1. The outlet guide vane (OGV).

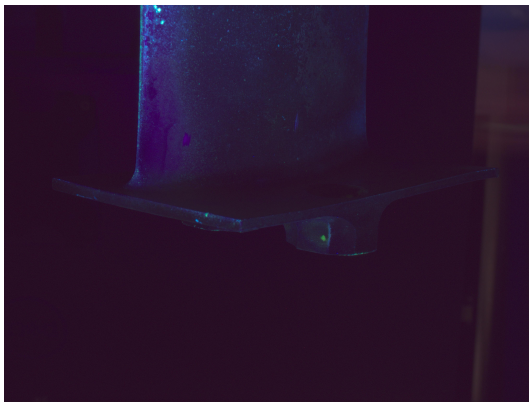
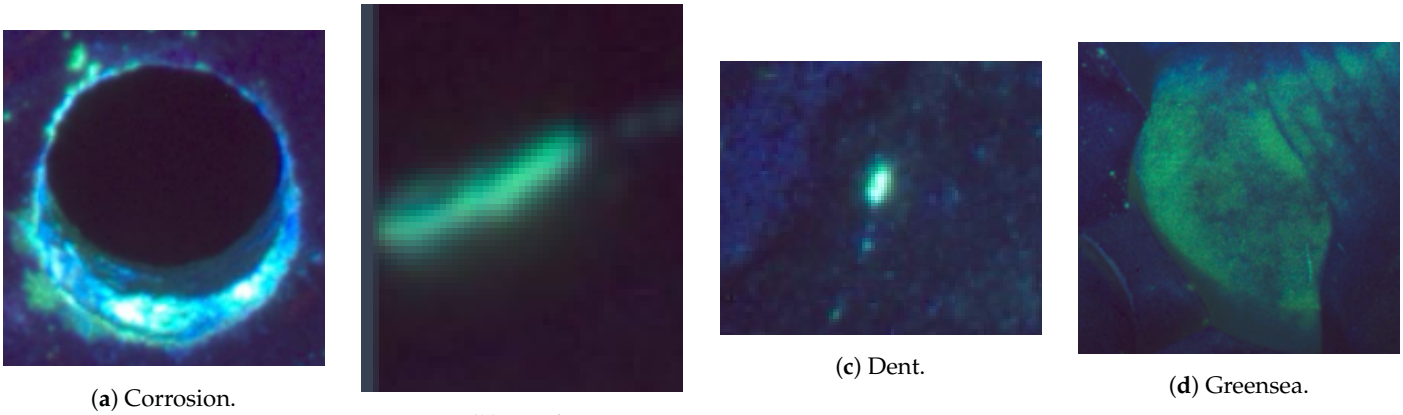


Figure 2. Example of an OGV during FPI. The bright green fluorescent dye is visible on the dark blue component.



(**a**) Corrosion.

(**b**) Crack.

(**c**) Dent.

(**d**) Greensea.

Figure 3. Types of indications: (**a**) corrosion; (**b**) crack; (**c**) dent; (**d**) greensea.

The first three categories cover indications that the inspector is trying to detect. The last category, greensea, covers parts of the component where excess penetrant resides on the surface. This means that the model cannot detect indications in their respective positions. In these locations, the inspector is required to apply extra cleaning to enable the detection of corrosion, cracks, or dents.

2.2. Data Augmentation

The data consist of images of OGVs with aforementioned indications, taken during FPI. As discussed earlier, these data are limited as the image acquisition is still in development. Data augmentation refers to a set of techniques used to enlarge and diversify datasets by creating and adding more data to add to the dataset the model is trained on. Large and diverse datasets yield good performances, even when using less sophisticated detection algorithms [7]. Training a system on a limited amount of data makes it prone to suffer from a phenomenon called overfitting. Overfitting means that a system gets too focused on the data it was trained on, disabling it to adapt well to new situations (represented as validation data during training). This can be seen during training when the validation loss fails to decrease along with the training loss. An example of this can be seen in Figure 4. An example of a system that did not overfit during training can be seen in Figure 5.

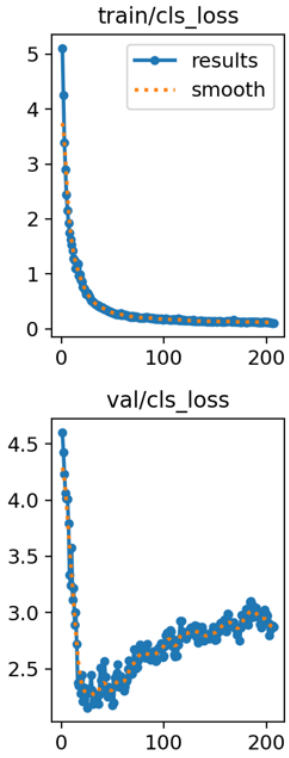


Figure 4. An example of a model that is overfitting during training.

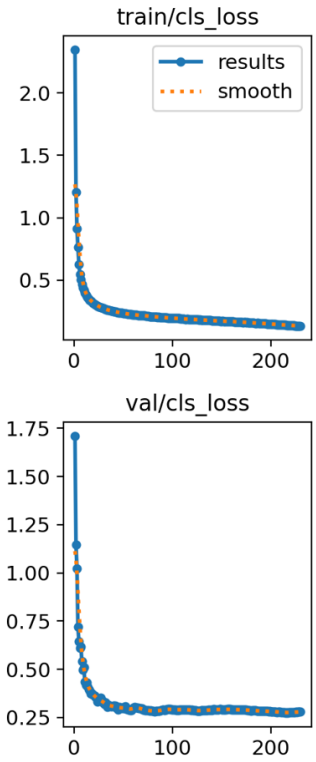


Figure 5. An example of a model that is not overfitting during training.

Preliminary research and experiments showed that the Mosaic data augmentation technique is the most effective data augmentation technique for this use-case [8]. Mosaic is a data augmentation technique first introduced by the authors in [9], where the technique was referred to as *Random Image Cropping and Patching (RICAP)*. This technique takes four

random images from the original dataset, crops them, and patches them together to create a new image. An example of an FPI Mosaic-generated image can be seen in Figure 6.

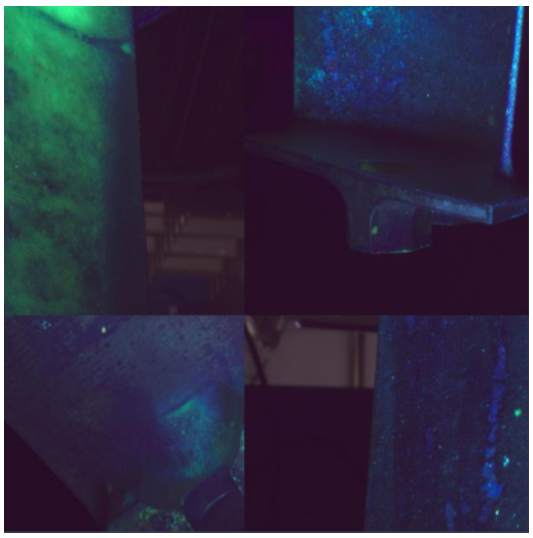


Figure 6. An example of a Mosaic-generated image.

Model training benefits from FPI Mosaic-generated images. There are two probable reasons for this: 1. Object distribution gets increased. Mosaic makes objects appear anywhere in the generated images rather than sticking to the original positions that were captured in the original images. This is a relevant simulation for the real world as defects can also occur anywhere in the frame. 2. Bounding boxes can get fractured. When this happens, the Mosaic technique automatically draws the bounding box around the new object, as seen in the Mosaic image. This forces the system to also learn fractured objects, forcing it to learn from less prominent object features.

3. Method Section

The setup in which the images of the dataset were acquired was realized at an aviation MRO partner, and can be seen in Figure 7. In this setup, a camera is attached to a robotic arm, where the arm can rotate around the inspected OGV. The arm is programmed so that it automatically takes pictures of the OGV while rotating around it. This setup was transferred to the FPI environment (a dark room), where pictures of the OGVs with defects were taken during inspections. In total, 403 pictures were taken and distributed over 14 different OGVs, with a total of 1413 indications.

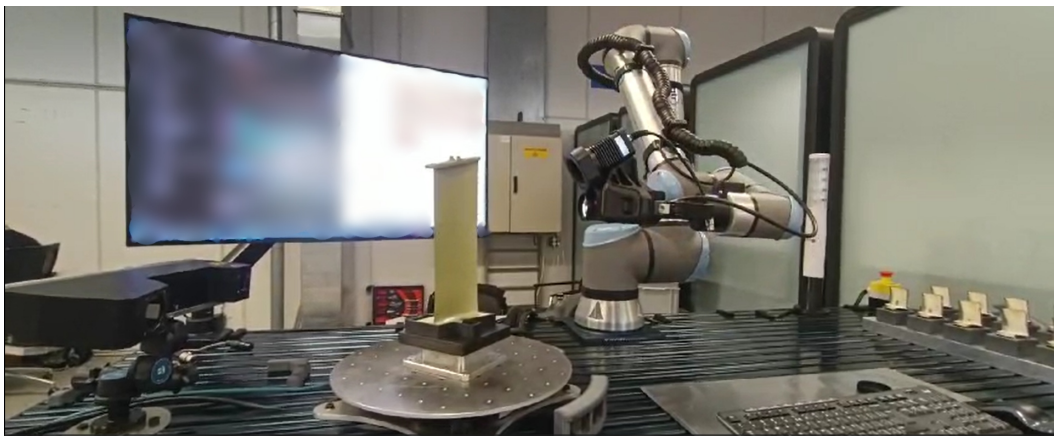


Figure 7. The image acquisition setup (outside the FPI environment).

To teach the model what defects look like, all the images then underwent an annotation process. This means that for every image, indications were manually located, and bounding boxes were drawn around them with the names of the indications as their respective labels (e.g., see Figure 8).

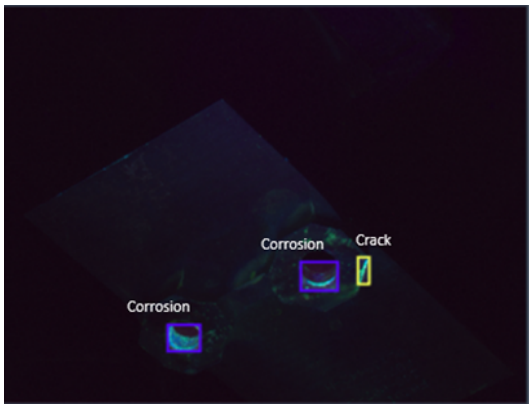


Figure 8. Example of an annotated image.

Afterward, the images are divided into training, validation, and testing images (70%, 20%, and 10%, respectively). As data augmentation is only used to improve the training of the model, it is important to note that data augmentation is only applied to the training dataset and not the datasets used to validate or test the model.

The original training images were then merged with the generated images to form new training datasets. Each new dataset had the exact same images as its predecessor plus uniquely generated Mosaic images, where each new dataset had twice the number of generated images as its predecessor (except for the third dataset, which had three times the number of generated images as the second dataset).

All resulting datasets (including the original dataset with no generated images, called *Baseline*) were then used individually to train models that all had the same model architecture (YOLOv8), resulting in different object detection models. This allowed for a solid evaluation, as the only difference in the resulting models was the datasets that were used to train the models. After training, the models were evaluated by using the Mean Average Precision (50-95) (mAP(50-95)) metric, which is a very common evaluation method for object detection models (e.g., used [10,11]).

To determine reproducibility, a second test was conducted in which the original labeled dataset was reshuffled (in training, validation, and test data). After the redistribution, the same tests were conducted on the reshuffled datasets, where models were again trained and evaluated. In addition to the determination of reproducibility, for this test, larger datasets were forged in an attempt to extend the peak performance.

4. Results

4.1. First Test Results

As seen in Figure 9, the performance of the system seems to be positively correlated with the number of generated Mosaic images in the training dataset, as is the training time. With a performance of 0.824 mAP(50-95), there is a substantial performance increase (0.666 mAP(50-95) over Baseline) by merely generating Mosaic images and adding them to the training dataset.

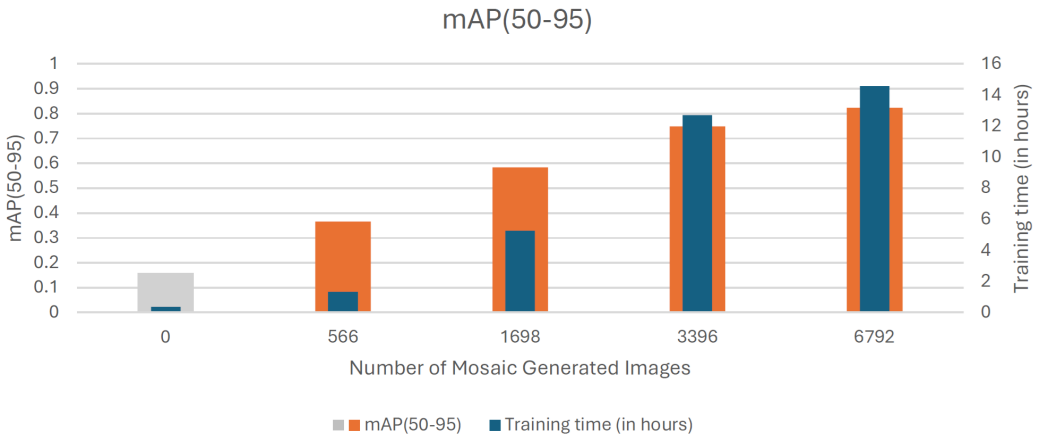


Figure 9. mAP(50-95) scores of the trained models for the first test.

4.2. Second Test Results

As can be seen in Figures 9 and 10, there is a performance difference between the different dataset splits, but this difference is quite small (0.054 mAP(50-95) at 6792 Mosaic-generated images). This small performance difference can have multiple causes. First, the training images could have been more diverse, which would allow the Mosaic-generated images to have more training value. A second reason might have been the quality of the testing dataset. Despite the small performance difference, the results seem to be reproducible as the trends are very similar.

After reproducibility was determined, another iteration was added, which was a training dataset of 13,584 generated Mosaic images, leading to the best performing model with a performance of 0.834 mAP(50-95). The maximal performance of this strategy seems to stagnate around this performance level. However, this asymptote could not be fully explored due to hardware constraints.

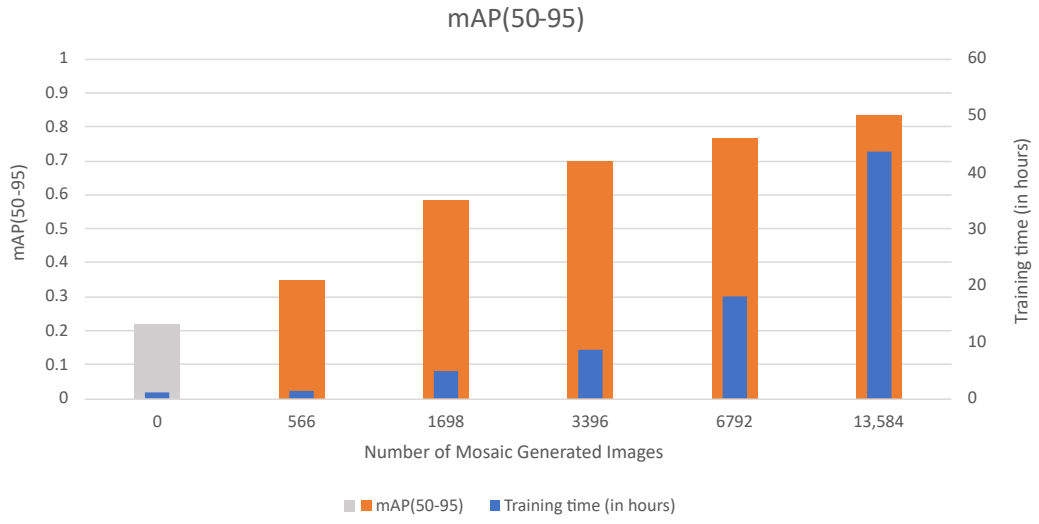


Figure 10. mAP(50-95) scores of the trained models for the second test.

4.3. Best Model Performance per Indication

Figure 11 shows the best model's performance on the individual objects. As can be seen, the best system has the worst performance on detecting dents and greenseas.

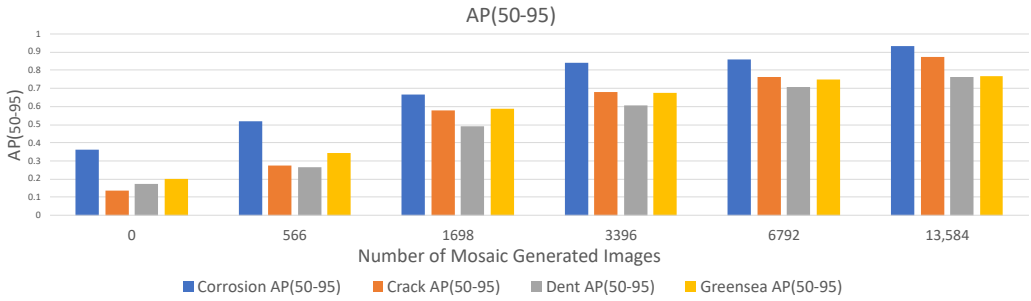


Figure 11. Best model's performance scores per object.

The poorer performance of greenseas is explainable as labeling these objects is very prone to subjectivity; on a lot of OGVs, there was quite a lot of penetrant remaining due to poor pre-cleaning, and deciding when to label greenseas proved to be a very subjective task (e.g., see Figure 12).

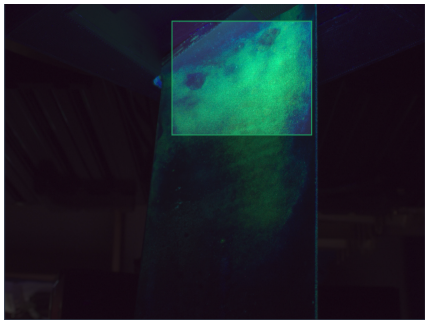


Figure 12. Example of an image where the labeling subjectivity of greenseas is visible.

The same goes for dents. As can be seen in Figures 13 and 14, dents are easily confusable with spats of penetrant, making it difficult for the annotator to decide when to label something as a dent.

For mentioned reasons, the poorer performance on these objects is more likely due to poor cleaning and labeling rather than the system's inability to locate dents and greenseas when cleaning and labeling would be performed properly.

Lastly, Figure 11 shows that the best model performs best on corrosion. This is explainable as corrosion is the only defect with a distinct blue color, which is the most probable reason why the model has an easier job detecting these objects.

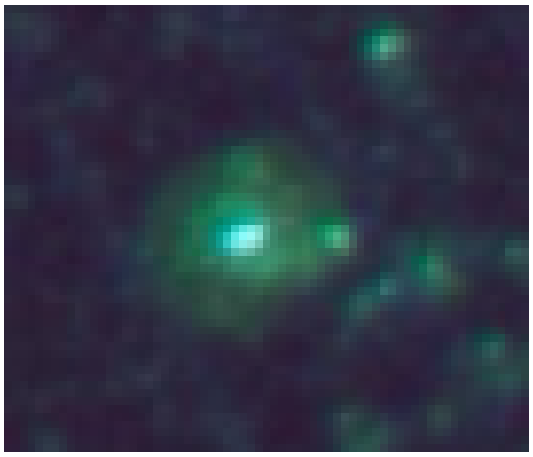


Figure 13. Instance of a dent.

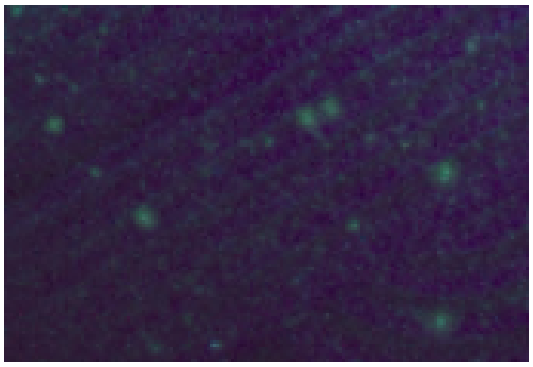


Figure 14. Instances of spats of penetrant.

5. Discussion

5.1. Limitations

When interpreting the results, a number of limitations should be considered.

1. Limited testing data: Although the results were reproducible over two experiments, the test dataset only consisted of 40 images with 137 total indications for both experiments.
2. Greensea as extra indication: The model performance was calculated by including the model's performance on greenseas. As discussed earlier, this annotation is not shared with the inspectors; they do not consider this an indication. Because the model was trying to detect an extra indication, model performance meant something slightly different than inspector performance.
3. Mosaic's peak performance: Although this study aimed to find Mosaic's peak performance, this could not be found due to hardware constraints.

5.2. Future Work

Future work includes the following:

1. Discovering Mosaic's peak performance: By simply adding more computing power, Mosaic's peak performance on this testing setup should be explored.
2. Enlarging the testing dataset: By either acquiring more FPI data or choosing a different distribution ratio for training, validation, and testing data, a bigger and more diverse testing dataset should be constructed, which would lead to a more valid model evaluation.
3. Optimizing the pre-cleaning process: If it would be possible for aerospace companies to update their pre-cleaning processes, the subjectivity of the image indications would decrease. Additionally, if updated pre-cleaning would allow for the disappearance of greenseas, the models' indications and the inspectors' indications would match, allowing for a more valid model evaluation.
4. Integrating into existing maintenance processes: Although this paper showed promising results regarding models that could potentially support inspectors, no model has been implemented in an MRO environment. Research into AI legislation plays an integral role in achieving this, and this is an already ongoing process.

6. Conclusions

This paper aimed to improve the performance of object detection DCNNs, which can support inspectors by detecting (possible) defects when limited data are available for training such models. By using the Mosaic data augmentation technique to generate and add more images to the model's training dataset, we trained a model to a performance of 0.834 mAP(50-95). These results show that the Mosaic data augmentation technique can

be a very powerful tool for improving object detection DCNNs for FPI when limited data are available.

Author Contributions: Conceptualization, methodology, validation, formal analysis, investigation, M.T.B. and D.F.; software, data curation, visualization, writing—original draft preparation, M.T.B.; resources, writing—review and editing, M.T.B. and D.F. and K.S.; supervision, project administration, D.F. and K.S.; funding acquisition, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: The authors wish to acknowledge funding from the Netherlands Enterprise Agency (RVO; Dutch: Rijksdienst voor Ondernemend Nederland) Subsidy Scheme R&D Mobility Sectors (RDM) within the Bright Sky project, grant number MOB21010.

Data Availability Statement: The data has been collected under non-disclosure agreements close to industrial operation; sharing the full raw dataset is not possible.

Acknowledgments: The data have been collected in cooperation with KLM—Royal Dutch Airlines and NLR—Netherlands Aerospace Centre in the context of project Brightsky.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Karigiannis, J.; Liu, S.; Harel, S.; Bian, X.; Zhu, P.; Xue, F.; Bouchard, S.; Cantin, D.; Beaudoin-pouliot, M.; Bewlay, B.P.; et al. Multi-Robot System for Automated Fluorescent Penetrant Indication Inspection with Deep Neural Nets. *Procedia Manuf.* **2021**, *53*, 735–740. [CrossRef]
2. Anonymous. *Aircraft Accident Report Uncontained Engine Failure/Fire Valujet Airlines Flight 597*; Technical Report; National Transportation Safety Board: Atlanta, GA, USA, 1995.
3. Anonymous. *Aircraft Accident Report United Airlines Flight 232 McDonell Douglas DC-10-10*; Technical Report; National Transportation Safety Board: Sioux City, IA, USA, 1989.
4. Anonymous. *Aircraft Accident Report Uncontained Engine Failure Delta Air Lines Flight 1288*; Technical Report; National Transportation Safety Board: Pensacola, FL, USA, 1996.
5. Wall, M. Human Factors guidance to improve reliability of non-Destructive testing in the Offshore Oil and Gas Industry. In Proceedings of the 7th European-American Workshop on Reliability of NDE, Potsdam, Germany, 4–7 September 2017.
6. Stamoulis, K. *Innovations in the Aviation MRO: Adaptive, Digital, and Sustainable Tools for Smarter Engineering and Maintenance*; Eburon Academic Publishers: Utrecht, The Netherlands, 2022.
7. Hasanpour, S.H.; Rouhani, M.; Fayyaz, M.; Sabokrou, M. Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures. *arXiv* **2023**, arXiv:1608.06037.
8. Bril, M. Data Augmentation for Fluorescent Penetrant Inspection Object Detection Systems. Bachelor's Thesis, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands, 2024.
9. Takahashi, R.; Matsubara, T.; Uehara, K. RICAP: Random Image Cropping and Patching Data Augmentation for Deep CNNs. In Proceedings of the 10th Asian Conference on Machine Learning, Beijing, China, 14–16 November 2018; Volume 95, pp. 786–798.
10. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef] [PubMed]
11. Wang, S.; Li, Q.; Yang, T.; Li, Z.; Bai, D.; Tang, C.; Pu, H. LSD-YOLO: Enhanced YOLOv8n Algorithm for Efficient Detection of Lemon Surface Diseases. *Plants* **2024**, *13*, 2069. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.