Comparative Analysis of Generative Data Augmentation Techniques for Aircraft Damage Detection Algorithms: A Case Study

Salvatore MEROLA[1,a *], Aditi MHATRE[2,b], Ann-Kathrin KOSCHLIK[2,c], Michele GUIDA[1,d], and Francesco MARULO[1,e]

[1]University of Naples Federico II – Dept. of Industrial Engineering, Naples, Italy

[2]German Aerospace Center (DLR) - Institute of Maintenance, Repair and Overhaul, Hamburg, Germany

[a]salvatore.merola@unina.it, [b]aditi.mhatre@dlr.de, [c]Ann-Kathrin.Koschlik@dlr.de, [d]michele.guida@unina.it, [e]francesco.marulo@unina.it

Keywords: Aircraft Maintenance, Generative AI, Data Augmentation, Computer Vision, Digital Optics

Abstract. This study addresses the challenge of limited annotated data in aircraft surface damage detection by evaluating generative models for data augmentation. Conducted within the CINNABAR 2 project with DLR MRO institute in Hamburg (DE), it compares Generative Adversarial Networks (GANs) and Diffusion Models for producing realistic synthetic images. Real data are collected from smartphones, DSLRs, and robotic camera systems. Image quality is assessed using the Learned Perceptual Image Patch Similarity (LPIPS) metric and visual inspection. Results indicate that Diffusion Models outperform GANs, achieving a lower LPIPS score and better detection metrics, demonstrating superior realism, diversity, and suitability for enhancing deep learning model training.

Introduction

A major limitation affecting the accuracy and generalization of deep learning algorithms in object detection tasks, such as aircraft surface damage detection, is the scarcity of annotated training data [1]. In industrial applications, particularly in safety-critical fields like aerospace, creating large, high-quality datasets is both costly and time-consuming. The diversity of defect types, materials, lighting conditions, and imaging perspectives further complicates data collection, as these factors significantly impact model performance [2]. When datasets are insufficient or unbalanced, AI algorithms often fail to generalize beyond their training domain. Therefore, developing comprehensive datasets that capture a broad spectrum of damage appearances and acquisition conditions—such as variations in illumination, resolution, surface texture, and viewing angle—is essential for effective model training. However, achieving such diversity manually is often impractical. To address this challenge, recent research has explored the use of generative models to address the lack of annotated data in structural damage detection [3] to augment limited datasets for multiclass structural damage classification, obtaining limited results in terms of accuracy in the subsequent testing phase for damage detection algorithms.

Methodology

This study investigates two different generative strategies for augmenting training datasets to make a comparison between them. First, Generative Adversarial Networks (GANs) [4] employ a dual-network architecture where a generator produces synthetic images, and a discriminator learns to distinguish between real and fake samples. The process begins with configuration and initialization, where the generator (G), discriminator (D), and perturbation module are initialized along with key hyperparameters (learning rate, number of steps, perturbation strength). The training loop iterates until convergence; within each iteration, the discriminator update phase

Materials Research Proceedings 69 (2026) 21-24 https://doi.org/10.21741/9781644904251-4

(performed for multiple steps per loop) involves generating fake images, sampling real images, performing adversarial training on real data, and computing the discriminator loss on both real and fake samples. The adversarial training component introduces perturbations to real images to enhance discriminator robustness and reuses the perturbed data efficiently across discriminator updates. In the generator update phase, the generator produces fake images to fool the discriminator, and its weights are updated based on the discriminator's feedback. The loss and stability emphasize training balance, combining discriminator and generator losses without requiring additional normalization or large batch sizes. The convergence/termination stage monitors performance metrics such as FID and IS, stopping training once convergence or saturation is reached.

The second approach, Diffusion Models [5], synthesizes images by iteratively denoising random noise, effectively reversing a diffusion process that gradually corrupts data. As the previous algorithm setup, the process begins with configuration, including specification of training parameters, dataset paths, and device selection. The pretrained model is then loaded, comprising components such as the Stable Diffusion pipeline, CLIP tokenizer, U-Net network with Low Ranking Adaption (LoRA) layers, Variational Auto Encoder (VAE), frozen text encoder, and DDPM scheduler. Next, LoRA configuration defines which attention layers to adapt and sets hyperparameters such as rank and dropout. The dataset and dataloader stage involve loading and augmenting image–prompt pairs for training. The training setup specifies optimization settings, including the AdamW optimizer, learning rate scheduler, and early stopping. During the training loop, images are encoded using the VAE, prompts are tokenized, noise is added and predicted by the U-Net, and model weights are updated based on MSE loss between predicted and target noise. Finally, the model saving stage stores the best-performing model, including the trained U-Net and LoRA configuration files.

As a case study, this work is conducted within the scope of the CINNABAR 2 [6] project, in collaboration with the DLR Institute of Maintenance, Repair and Overhaul – Department of Process Optimization and Digitalization in Hamburg, Germany. The test study focuses on a representative structural component of a commercial aircraft — specifically, a section of a medium-range commercial aircraft wing — used to simulate real-world inspection scenarios. To assess the impact of different image acquisition conditions, three imaging systems are used to collect real data:

- *Smartphone camera*, representing low-cost, flexible field inspections.
- *Digital Single-Lens Reflex (DSLR) camera*, providing high-resolution, professional-grade imagery.
- *Robotic arm* with a mounted camera

The images collected are used to build a dataset of 50 images with different subjects; the main goal is to have an *Image-to-Image (*Img2img*)* model able to generate new damage starting from existing ones in different poses, colors, illumination, and scenes. To do so, both algorithms previously described are trained using similar training options (Table 1).

Table 1 - Training option for Diffusion algorithm (left) and GAN algorithm (right)

	Diffusion model	**GAN model**
N. of epochs	100	1000
Batch size	4	4
Image size [pixel]	512x512	512x512
Learning rate	0.0001	0.0002

Results

The LPIPS metric is used to compare the performance of the two algorithms. The Learned Perceptual Image Patch Similarity (LPIPS) metric is a perceptual measure designed to quantify the similarity between two images in a way that aligns more closely with human visual judgment than traditional pixel-based metrics. LPIPS evaluates differences in feature representations extracted from intermediate layers of convolutional neural networks pretrained on large-scale image datasets. These deep features capture structural and semantic content that humans are sensitive to, such as textures, edges, and object parts. A lower LPIPS score indicates higher perceptual similarity between two images, while higher values suggest greater visual differences. This makes LPIPS widely applicable in the evaluation of generative models, image restoration, and data augmentation tasks. The results obtained are reported in Fig. 1 and Fig.2. The synthetic images generated by Diffusion Models are more suitable for data augmentation purposes, as indicated by their LPIPS average score of 0.2707 (versus GAN model LPIPS average score: 0.5797). For this reason, it is decided to use a diffusion model to generate 100 synthetic images to compare the performance of a damage detection algorithm trained on a mixed dataset (real images + synthetic images) with the performance of the same algorithm trained only on real images. YOLOv11 is chosen as the damage detection algorithm. YOLOv11 is the latest iteration in the You Only Look Once (YOLO) series, developed by Ultralytics [7]. It is designed to offer high performance across a range of computer vision tasks, including object detection and image classification. Two datasets are therefore prepared: the first containing 50 real images (real dataset) and the second containing 50 real images and 100 synthetic images (real + fake dataset). Both are split into 80%/20% for the training/validation phase, using only real images for the validation phase in both cases. The results are evaluated in terms of Precision, Recall, mean Average Precision at 50% Intersection over Union (mAP50), and mAP50-95 (Table 2). More details about its use for Damage Detection can be found in [8].

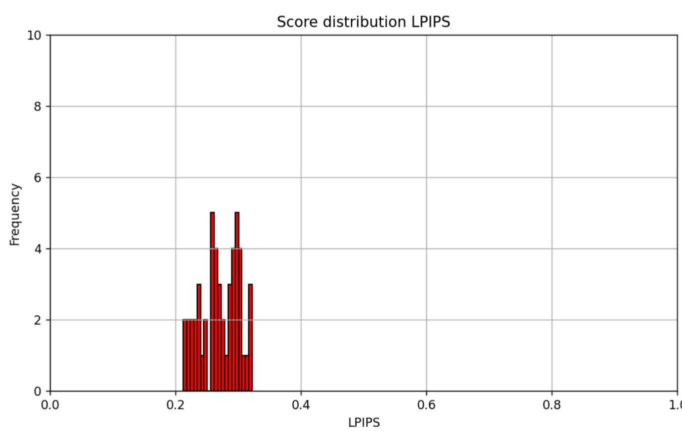

Fig. 1 - Score distribution LPIPS for Diffusion model

Table 2 - Evaluation metrics for real image dataset and real + fake image dataset

	Box P	**Box R**	**mAP50**	**mAP50-95**
Real	0.81	0.85	0.837	0.421
Real + Fake	0.978	0.889	0.951	0.586

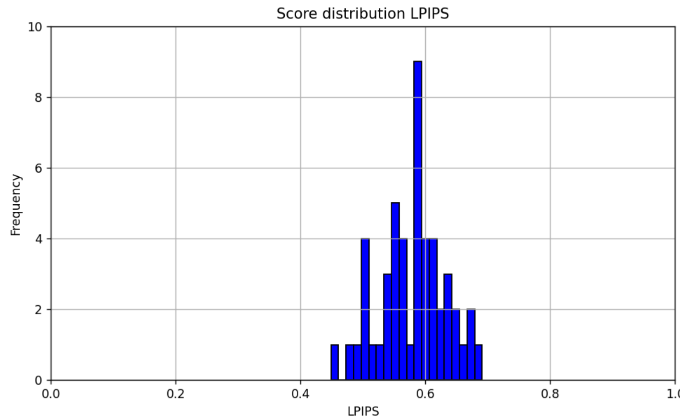

Fig. 2 - Score distribution LPIPS for GAN model

Conclusion

The comparison between the two training strategies demonstrates the clear benefit of augmenting real data with synthetic samples. The model trained exclusively on real images achieved a reasonable balance between precision and recall but struggled with bounding box accuracy. In contrast, the model trained on the mixed dataset showed a notable improvement across all metrics: it is more reliable in its predictions, slightly more sensitive in detecting true cases, and produced more accurate localizations. These results show that synthetic data can effectively complement limited real datasets, minimize false positives while improving overall detection performance, and generalize to real-world circumstances.

References

[1] S.E.Whang, Y. Roh, H. Song, et al. Data collection and quality challenges in deep learning: a data-centric AI perspective. The VLDB Journal 32, 791-813 (2023). https://doi.org/10.1007/s00778-022-00775-9

[2] S.U.G. Hyontai. Performance of machine learning algorithms and diversity in data. In MATEC web of conferences (2018, Vol. 210, p. 04019). EDP Sciences. https://doi.org/10.1051/matecconf/201821004019

[3] K. Dunphy, M. N. Fekri, K. Grolinger, A. Sadhu. Data augmentation for deep-learning-based multiclass structural damage detection using limited information. 2022. Sensors, 22(16), 6193. https://doi.org/10.3390/s22166193

[4] L. Bingchen, et al. Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. ICLR. 2021.

[5] B. Trabucco, K. Doherty, M. Gurinas, R. Salakhutdinov. Effective data augmentation with diffusion models (2023). arXiv preprint arXiv:2302.07944.

[6] A.K. Koschlik, F. Rauscher, M. J. Scott, S. Merola, H. Meyer, W. Verhagen, G. Wende. Digital Twins for Aircraft Structural Inspections: Enhancing Dent Detection (2025). In Proceedings of 21st Australian International Aerospace Congress, 24-26 March 2025, Melbourne & Avalon.

[7] G. Jocher, J. Qiu. Ultralytics YOLO v11 v.11.0.0 (2024).

[8] Mhatre, S. Merola, A.K. Koschlik, R. Rodeck, G. Wende. Deep-Learning-based Dent Detection of Aircraft Surfaces using Synthetic Data (2025). Proceedings of 28th AIDAA International Congress & 10th CEAS Aerospace Europe Conference.