ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Improving Robustness of Defect Detection models using Adversarial-based Data Augmentation

Daniel García[1,2], Aleix García[2], Diego García[1] and Ignacio Díaz[1] *

1- University of Oviedo - Department of Electrical Engineering
Gijon 33204 Asturias - Spain

2- CIN Advanced Systems - Machine Learning Department
Gijon 33211 Asturias - Spain

Abstract.

We propose an adversarial-based data augmentation method to improve the robustness of object detection models, specifically for industrial defect detection. Unlike prior approaches focused on classification or synthetic datasets, our method generates adversarial examples that target both classification and localization outputs. We further introduce controlled white noise to these examples, enhancing robustness against environmental variations. Empirical evaluation on a real-world dataset of defective laser welding images shows that our approach outperforms standard data augmentation and existing adversarial training methods, improving both model accuracy and resilience to diverse perturbations encountered in real-world settings.

1 Introduction

The implementation of *Deep Learning* (DL) models in real-world settings, especially in critical fields like industrial defect detection, faces significant challenges related to generalization arising from the discrepancy between their performance on controlled test datasets and their operation in diverse, unpredictable environments. This issue is exacerbated in industrial applications where the robustness of detection is critical; models must produce consistent results under varying conditions to ensure reliability and safety.

Data augmentation techniques, such as rotations, flips, and color adjustments, have been employed to improve model generalization by artificially increasing the diversity of training data [1]. However, these methods often fall short in capturing the complex variations encountered in real-world scenarios [2]. To address the problem of *adversarial attacks*, where models are vulnerable to inputs intentionally perturbed to cause errors, *adversarial training* has been adopted. This technique trains models on artificially generated adversarial examples to improve robustness against these perturbations [3].

Previous works have explored adversarial attacks and defenses in the context of image classification [4, 5], semantic segmentation [6], and object detection [7, 8]. In object detection, methods like the *Dense Adversary Generation* (DAG)

*This work is part of Grant PID2020-115401GB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

[6] and *Targeted Adversarial Objects* (TAO) [7] have been used to generate adversarial examples that manipulate both classification and localization outputs. While these approaches have improved robustness to some extent, they often rely on synthetic datasets or focus solely on the attack mechanisms rather than defense strategies.

In this paper, we introduce a novel adversarial-based data augmentation method specifically designed to achieve robust object detection in industrial applications. Our contributions are threefold:

1. **Adversarial training framework for object detection**: We propose an adversarial training framework that generates adversarial examples targeting both classification and localization outputs in object detectors. This framework integrates realistic adversarial examples, specifically designed to enhance robustness in industrial applications. Unlike previous methods, which often focus mainly on classification tasks [4, 5], our approach addresses challenges specific to object detection [7, 8].

2. **Controlled white noise augmentation**: We enhance the adversarial examples by adding controlled white noise, which simulates environmental variations such as sensor noise or lighting changes [12]. This augmentation further improves the model's ability to generalize to unseen conditions.

3. **Empirical evaluation on real-world industrial data**: Unlike previous studies that often rely on synthetic or publicly available datasets, we conduct extensive experiments on a real-world dataset of defective laser welding images. We compare our method against standard data augmentation techniques and existing adversarial training methods, demonstrating significant improvements in both accuracy and robustness.

By addressing the specific challenges of industrial object detection and providing a comprehensive evaluation, our work offers practical solutions for deploying robust DL models in critical applications.

2 Materials and Methods

2.1 Adversarial Sample Generation

Adversarial examples are crafted by introducing perturbations to input images that lead the model to make incorrect predictions. For object detection, these perturbations can affect both the classification and localization outputs. We build upon the *Targeted Adversarial Objects* (TAO) method [7], extending it to generate adversarial examples that are more representative of the perturbations encountered in industrial environments.

Given an input image $\boldsymbol{x}$ and a pre-trained object detection model with parameters $\boldsymbol{\mathcal{W}}$, we aim to generate an adversarial example $\boldsymbol{x}'$ that causes the model to misclassify or mislocalize objects. The adversarial example is generated using Eq. 2, where $\mathcal{L}$ combines classification and localization losses. Here,

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

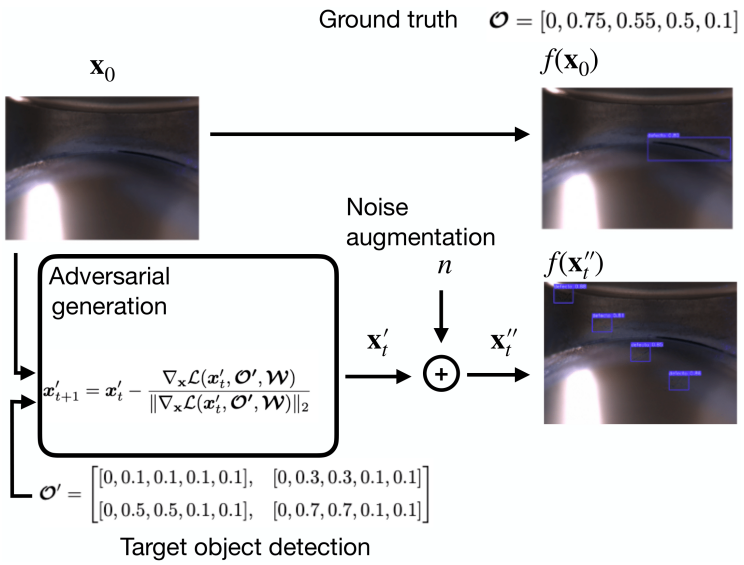

Fig. 1: Example of adversarial sample generation. O and O' correspond to the target of an object detection model, represented by [class, x, y, w, and h].

$\boldsymbol{\mathcal{O}}$ represents the ground truth annotations, while $\boldsymbol{\mathcal{O}'}$ is simplified to represent no detections—reflecting a primary concern in industrial settings where missed defect detections pose significant risks. Although this single-class setup limits the adversarial space, future work should extend it to more complex scenarios. The step size α controls perturbation magnitude, and the L2 norm normalizes gradients to keep perturbations realistic.

$$\boldsymbol{\mathcal{O}'} = \emptyset \tag{1}$$

$$\boldsymbol{x}'_{t+1} = \boldsymbol{x}'_t - \alpha \cdot \frac{\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}'_t, \boldsymbol{\mathcal{O}'}, \boldsymbol{\mathcal{W}})}{\|\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}'_t, \boldsymbol{\mathcal{O}'}, \boldsymbol{\mathcal{W}})\|_2} \tag{2}$$

The adversarial generation process continues iteratively until $\mathcal{L}(\boldsymbol{x'_t}, \boldsymbol{\mathcal{O}'}) < \tau$, ensuring that $\boldsymbol{x_t}'$ effectively misleads the model.

2.2 White Noise Augmentation

To further enhance robustness, we introduce a controlled amount of white Gaussian noise to the adversarial examples as shown in Eq 3.

$$\boldsymbol{x''} = \boldsymbol{x'} + \boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

The standard deviation σ is chosen to simulate sensor noise and environmental variations, introducing a layer of realism to the perturbation.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

The standard deviation σ is chosen to simulate sensor noise and environmental variations common in industrial settings.

2.3 Comparison with Standard Data Augmentation and Adversarial Training

To evaluate the effectiveness of our method, we compare it against:

- **Baseline Model**: Trained both with and without standard data augmentation techniques (e.g., flipping, rotation).
- **Adversarial Training (AT)**: Incorporating adversarial examples generated without white noise, similar to methods in [3, 8].
- **Enhanced Adversarial Training (AT + WN)**: Our comprehensive method that integrates various enhancements, including the addition of white noise and selective adversarial example generation.

In the adversarial training methods (AT and AT+WN), adversarial examples are generated from the original dataset by selecting samples where the model's accuracy is below a specific threshold. These samples, close to the decision boundary, are identified by filtering out those where $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\mathcal{O}}) > 3\sigma$, being σ the standard deviation of the loss, as these are considered outliers. The adversarial examples are then generated from these selected samples and added to the original dataset, retaining the same labels as the corresponding original samples.

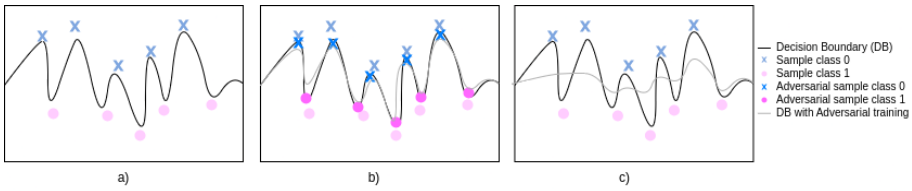

Fig. 2: Schematic representation of the effect of adversarial samples on the decision boundary. a): Original decision boundary. b): Effect of an adversarial iteration on the decision boundary. c): Final decision boundary after fine-tuning.

For the retraining, we perform two-stage training: first, the baseline model is trained, and then the adversarial training is applied in a second stage, using the augmented dataset consisting of both original and adversarial samples. This method ensures the model is exposed to challenging examples without overfitting to outliers.

2.4 Dataset and Experimental Setup

We use a proprietary dataset comprising 5,000 images of defective laser welds on common-rails, capturing various types of defects under different conditions. The dataset is split into training (70%), validation (15%), and testing (15%) sets.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

We employ the YOLOv8 object detection model due to its efficiency and accuracy in real-time applications [9]. All models are trained using the same hyperparameters to ensure a fair comparison, with additional training steps for adversarial training as required.

Model performance is evaluated using Mean Average Precision (mAP) at 0.5 and [0.5:0.95] IoUthresholds. To assess robustness, we apply the Robust Detection Benchmark (RDB) framework [11] on test images with simulated real-world perturbations.

3 Results

3.1 Performance Comparison

Table 1 presents the performance of different training methods on the test set without perturbations.

Training Method	**mAP50**	**mAP50:95**
Baseline Without Augmentation (WO Aug)	0.81	0.57
Baseline With Augmentation (W Aug)	0.84	0.61
Adversarial Training (AT)	0.82	0.60
Adversarial Training + White Noise (AT+WN)	**0.86**	**0.63**

Table 1: Performance comparison between different training methods

Our method (AT+WN) outperforms both the baseline and the standard adversarial training (AT) methods, achieving the highest mAP scores.

3.2 Robustness to Corruptions

Table 2 demonstrates that the adding adversarial samples with targeted white noise consistently achieves higher performance across all severity levels, with better mAP50 compared to other models. This highlights the effectiveness of combining Adversarial Training and White Noise in improving model robustness under varying perturbation levels.

Experiment	**1**	**2**	**3**	**4**	**5**
Baseline W Aug	0.85	0.71	0.53	0.48	0.35
Baseline W Aug + AT	0.85	0.72	0.54	0.49	0.36
Baseline W Aug + AT + WN	**0.87**	**0.73**	**0.56**	**0.52**	**0.37**
Baseline WO Aug	0.82	0.68	0.51	0.45	0.30
Baseline WO Aug + AT	0.83	0.69	0.52	0.46	0.32
Baseline WO Aug + AT + WN	0.86	0.71	0.55	0.50	0.36

Table 2: Comparison of mAP50 values across different experiments and severity levels of corruptions according to RBA benchmark

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

4 Conclusions and Future Work

Our experiments demonstrate that the proposed adversarial-based data augmentation method, combined with controlled white noise addition, significantly enhances both the accuracy and robustness of object detection models in industrial settings. The method outperforms standard data augmentation and existing adversarial training techniques, as evidenced by higher mAP scores and better performance under various image corruptions.

By targeting both classification and localization outputs, our adversarial examples expose the model to challenging scenarios that are representative of real-world industrial conditions. The addition of white noise simulates environmental factors such as sensor noise, lighting variations, and other perturbations, further enhancing model generalization.

While our method shows promising results on a specific industrial dataset, further research is needed to generalize the approach to other domains and datasets. Testing on diverse datasets, including publicly available benchmarks, would strengthen the validity of the method. Additionally, exploring other forms of perturbations and adversarial attacks could provide deeper insights into model robustness.

References

[1] C. Shorten and T. M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning, *Journal of Big Data*, 2019.

[2] R. Geirhos et al., Shortcut Learning in Deep Neural Networks, *Nature Machine Intelligence*, 2020.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards Deep Learning Models Resistant to Adversarial Attacks, *International Conference on Learning Representations (ICLR)*, 2018.

[4] N. Carlini and D. Wagner, Towards Evaluating the Robustness of Neural Networks, In *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and Harnessing Adversarial Examples, *International Conference on Learning Representations (ICLR)*, 2015.

[6] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, Adversarial Examples for Semantic Segmentation and Object Detection, *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[7] K.-H. W. Chow, L. Liu, M. E. Gursoy, S. Truex, W. Wei, and Y. Wu, Targeted Adversarial Objectness Gradient Attacks on Real-time Object Detection Systems, *arXiv preprint arXiv:2004.04320*, 2020.

[8] H. Zhang and J. Wang, Towards Adversarially Robust Object Detection, *arXiv preprint arXiv:1907.10310*, 2019.

[9] G. Jocher et al., YOLOv8, `https://github.com/ultralytics/ultralytics`, 2023.

[10] T.-Y. Lin et al., Microsoft COCO: Common Objects in Context, *European Conference on Computer Vision (ECCV)*, 2014.

[11] C. Michaelis et al., Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming, *arXiv preprint arXiv:1907.07484*, 2019.

[12] C. M. Bishop, "Regularization and complexity in neural networks," *Network: Computation in Neural Systems*, vol. 5, no. 4, pp. 451–474, 1995.