Image Augmentation and Model Selection for Defect Detection in Automotive Heat Shields

Saber Yu*,†, Mengchen Liu*, Timothy Reinhart*,†, Seshasai Srinivasan*, Zhen Gao*, Majan Alavi*

*W Booth School of Engineering Practice and Technology, McMaster University, Hamilton, ON, Canada
†Ascension Automation Solutions Ltd, Cambridge, ON, Canada

Emails: {yux8, liu2583, reinhat, ssriniv, gaozhen, alavis2}@mcmaster.ca

***Abstract*—The issue of limited samples and low data variation poses a fundamental challenge in developing deep learning-based models for industrial defect detection applications. Through evaluating various augmentation techniques on YOLO11-Nano using a custom dataset containing 249 images of automotive heat shields, we showed that a carefully designed augmentation pipeline improved the model's average $mAP_{0.5}$ from 0.863 to 0.952, corresponding to a relative increase of approximately 10.3% over the no-augmentation baseline. Comparative benchmarking of YOLO11 against YOLO12, RT-DETR, RF-DETR, and Faster R-CNN indicated that RF-DETR was the architecture most recommended for detection precision, with the highest $mAP_{0.5}$ (1.000) and $mAP_{0.5:0.95}$ (0.505) achieved at an average inference speed of 41 FPS, while YOLO11 demonstrated superior consistency (0.946 $mAP_{0.5}$ and 0.476 $mAP_{0.5:95}$ in average) across multiple training runs at 108 FPS. These findings offer practitioners with specific augmentation guidelines and model selection criteria for similar tasks.**

***Index Terms*—Computer Vision, Object Detection, Defect Detection, Data Augmentation, You Only Look Once (YOLO), DEtection TRansformer (DETR).**

I. Introduction

Specialized industrial datasets often contain only a few defective classes and limited samples. This data scarcity poses significant challenges for model training and generalization, highlighting the value and necessity of image augmentation [27]. Traditional augmentation techniques include geometric transformations like translation, scaling, and rotation, and photometric transformation such as hue, brightness, and saturation adjustments. More advanced methods include structural rearrangement such as mix-up [25], cut-mix [23], and mosaic [1]. Beyond manual augmentations, researchers have invented automatic algorithms like AutoAugment [5], RandAugment [6], and TrivialAugment [12], all of which have proven effective in image classification. The latest research in this field explores generating completely synthetic images using Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) [19], [22] and performing augmentation combined with domain knowledge [7], [11]. As a general rule, effective augmentation focuses on creating realistic variations of training data while avoiding unrealistic transformations that deviates the model from practical use-cases [22], [27].

Another piece of the puzzle for developing vision applications is the selection of a proper pre-trained model before attempting to fine-tune it on the custom dataset. As illustrated in Figure 1, the rapid evolution of computer vision has created a large pool of options for model selection. While the latest architectures have showcased promising results on large datasets like ImageNet and COCO, their performance on industrial datasets remains unclear. Our work aims to address this gap while providing valuable insights for data augmentation. The models we evaluated include YOLO12-Nano [21], YOLO11 (Nano-XL) [9], RT-DETR-Large [26], RF-DETR-Base [16], and Faster R-CNN (ResNet50-FPN) [15].

II. Methods

A. Data Acquisition

Heat shields are used to isolate heat, noise, and vibration from high-temperature automotive components such as engines and turbos. We acquired a total of 20 defective samples from Dana Inc, including 12 with splits, 6 with crushed edges, and 2 with both types of defect. These heat shields were transported to our laboratory for controlled imaging with a **Cognex IS3816M** camera and **MORITEX ML-U1217SR-18C** lens, producing monochrome images of 5320 x 3032 pixels. Multiple pictures were taken of each sample at varying angles and distances to create a dataset of 249 images, consisting of 2 annotated classes (splits and crushed edges), with 110 splits, 66 crushed edges, and 79 background instances. We performed a 70-20-10% split and balanced our training data with 40% splits, 30% crushed edges, and 30% background, while ensuring all classes were evaluated across validation and test. An example of our data is illustrated by Figure 2.

B. Applying Augmentation

All geometric and photometric transformations were applied using Albumentations [2]. Mosaic was implemented in our code with 2x2 grids, mimicking the original approach [1]. We performed post-augmentation visibility checks to ensure at least 70% of the original object remained visible because we have observed circumstances where crushed edges appeared somewhat normal under a 30% or more visibility reduction. In cases where an API has been made available for model training, we simply adjusted augmentation by modifying arguments

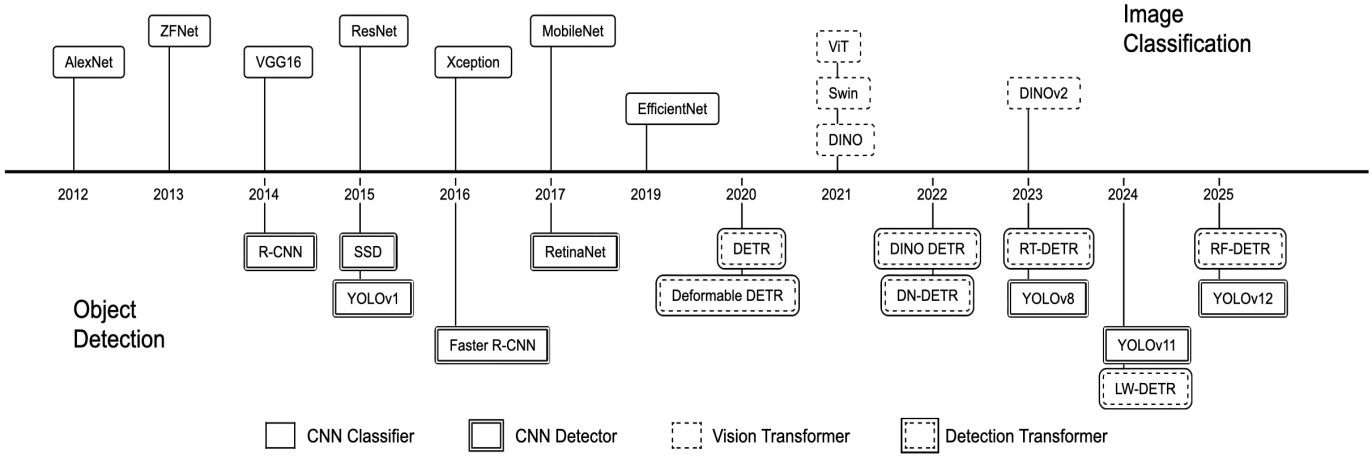

Fig. 1: The recent advancement of vision models from 2012 to 2025. Key breakkthroughs include AlexNet [10], ResNet [8], EfficientNet [20], Faster R-CNN [15], YOLO series [14], DETR [3], DINO DETR [24], RT-DETR [26], and RF-DETR [16].

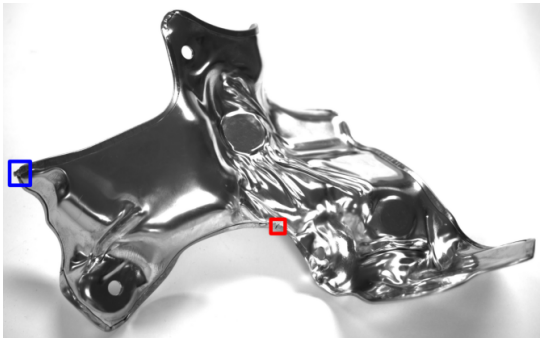

Fig. 2: Heat shield with a split (red) and crushed edge (blue).

TABLE I: Data Augmentation Techniques and Parameters.

Category	Code	Description	Strength	Probability
Geometric	G1	Horizontal Flip	NA	50%
	G2	Scale	± 10%	100%
	G3	Translation	up to 5%	100%
	G4	Shear	± 5 deg	100%
	G5	Rotate	± 20 deg	100%
	G6	Perspective	up to 5%	100%
Photometric	P1	Brightness	± 20%	100%
	P2	Hue	± 20%	100%
	P3	Saturation	± 20%	100%
Structural	S1	Mosaic	NA	100%

under model.train(). YOLO's default Albumentations settings, which included a minor blur, togrey, and CLAHE, was kept on at all times. We performed 10 runs for each experiment and fixed the training seeds used across trials.

C. Model Evaluation

When vision applications are developed in the industry, practitioners are often required to screen for potential model candidates with out-of-the-box configurations. To provide practical guidance that reflects this screening approach, we evaluated each model using their default augmentation settings, with the exception of Faster R-CNN, on which we adopted YOLO's settings due to the absence of default augmentation in TorchVision. All images were processed at their original resolution with model-level resizing. Our experiment was conducted on an **NVDIA RTX 4070 Laptop** GPU and **Intel Core i7-14700HX** CPU.

III. Results

A. Image Augmentation

We began our experiments without any augmentation and gradually added geometric, photometric, and image-structural transformations (in this order). Details for our augmentation parameters are captured in Table 1. The **No Augment** baseline, as expected, showed a relatively low performance, with an average $mAP_{0.5}$ of 0.863 and $mAP_{0.5:0.95}$ of 0.378. Most augmentation techniques provided substantial improvements over this baseline, except for **G1235**, which involved rotation. Variances in our experiments tended to shrink as the number of augmentations applied increased, with the addition of shear being the only outlier. The best result in $mAP_{0.5}$ (0.995) was achieved with the **G1234** and **G123P123** combinations, while **YOLO Default**, consisting of horizontal flip (50% chance), scale (0.5), translate (0.1), brightness (0.4), hue (0.015), saturation (0.7), and mosaic (always applied), yielded the best $mAP_{0.5:0.95}$ (0.538). Lastly, we found perspective augmentation (G6) highly detrimental to model convergence and therefore excluded it from the results shown in Figure 3.

B. Model Performance

As illustrated in Figure 4, we benchmarked the performance of multiple state-of-the-art architectures, including YOLO12-Nano, YOLO11 (Nano-XL), RT-DETR-Large, RF-DETR-Base, and Faster R-CNN (ResNet50-FPN). Among these models, RF-DETR-Base achieved the highest $mAP_{0.5}$ (1.000) and $mAP_{0.5:0.95}$ (0.505) in its best-performing training run, with a moderate average inference speed of 24.33 ms per image (41

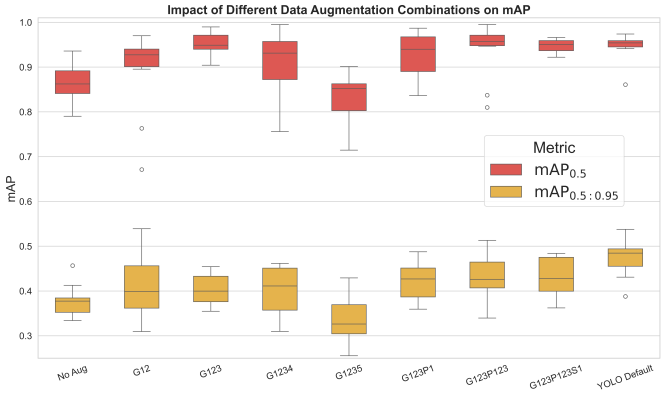

Fig. 3: Effects of data augmentation strategies on YOLO11-Nano. Notation: **G** denotes geometric, **P** denotes photometric, and **S** denotes structural augmentation. Example: an **G123P1** trial combines horizontal flip, scale, and translation with brightness changes.

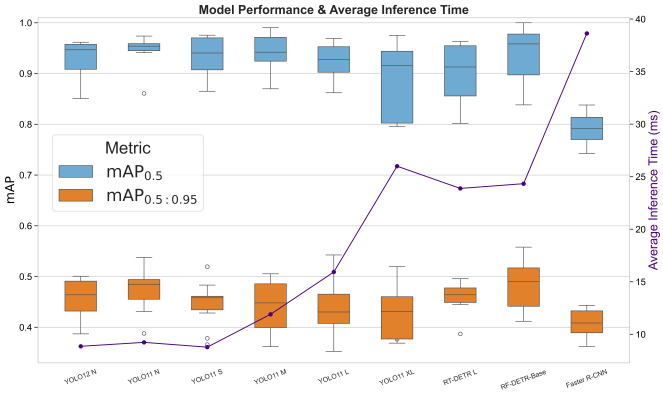

Fig. 4: Model Performance (top: $mAP_{0.5}$, bottom: $mAP_{0.5:0.95}$) and averaged inference speeds (indicated by the purple line). Faster R-CNN represents the ResNet50-FPN variant.

FPS). YOLO11-Nano demonstrated strong overall stability, with an average $mAP_{0.5}$ of 0.946 and $mAP_{0.5:0.95}$ of 0.476 at an excellent inference speed of 9.24 ms (108 FPS). Faster R-CNN (ResNet50-FPN) lagged in inference speed (38.64 ms or 25 FPS) and produced the lowest mAP scores in all models, while RT-DETR achieved moderate performance with relatively balanced metrics.

IV. Discussion

Most augmentations at tested intensities improved model performance, but there were several exceptions. Perspective, for instance, is found to be highly detrimental. We believe this is because the transformation alters the semantic representation of the features extracted by the convolutional blocks and causes training instability. Rotation is another example. Although it has been shown beneficial for general object detection tasks [28], it leads to reduced performance in our experiments, most likely because the augmentation biases the model toward unrealistic rotated orientations that are absent from the test data. This highlights the importance of a domain-specific augmentation design, as discussed in [18], and suggests that rotation might be unnecessary for industrial detection involving fixed sample orientations. Finally, shear appears to have widened the performance variance, particularly toward lower mAPs, and is therefore not recommended.

Increasing the size of the model within the YOLO11 family should intuitively suggest performance gains, but this is not the case in our experiments. In fact, the average $mAP_{0.5:0.95}$ tends to decrease with increasing model size, indicating diminishing generalization abilities despite higher model capacity. Although performance peaks are occasionally observed in individual runs for YOLO11-Medium and YOLO11-Large, they are generally not sustained. In addition, larger variants within the same model family usually accompany trade-offs in latency and inference speeds. For example, YOLO11-Large requires nearly twice the inference time of YOLO11-Small. These findings suggest that scaling up the size of the model without a proportional increase in data may result in negligible improvement or even degraded performance, particularly under stricter evaluation thresholds. Our results further indicate that RF-DETR, which utilizes the LW-DETR architecture [4] and DINOv2 [13] backbone, is the model most recommended for detection precision. This architecture outperforms RT-DETR (with a convolutional backbone) and recent CNN benchmarks such as YOLO11 and YOLO12 despite having higher performance variances across different runs. These findings align with [16] and [17] and confirm the potential of RF-DETR for industrial defect detection applications.

V. Conclusion

This study provided comprehensive guidance for fine-tuning vision models for industrial defect detection applications with limited data. Through a systematic review of augmentation techniques applied on our heat shield dataset, we showed that a carefully designed augmentation pipeline increased the average $mAP_{0.5}$ of YOLO11-Nano from 0.863 to 0.952, representing a relative improvement of approximately 10.3% over the baseline without augmentation. Our results further suggested that perspective augmentation should be avoided for similar applications, whereas rotation and shear required additional consideration. With the evaluation of multiple state-of-the-art models, including YOLO12-Nano, YOLO11 (Nano-XL), RT-DETR-Large, RF-DETR-Base, and Faster R-CNN (ResNet50-FPN), we showed that the best single-run results were achieved by RF-DETR, while YOLO11-Nano demonstrated superior consistency across multiple training runs at quicker inference speeds.

Acknowledgment

The authors would like to thank Patrick Gilbert and Thomas Bachle at Dana Inc. for their assistance with data collection, and Salman Bawa and Richard Allen at McMaster University for coordinating this project with Ascension Automation Solutions Ltd.

REFERENCES

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 163–178, 2020.

[2] Alexander Buslaev, Alex Parinov, Eugene Khvedchenya, Vladimir I Iglovikov, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *arXiv preprint arXiv:1809.06839*, 2018.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[4] Qiang Chen, Xiangbo Su, Xinyu Zhang, Jian Wang, Jiahui Chen, Yunpeng Shen, Chuchu Han, Ziliang Chen, Weixiang Xu, Fanrong Li, Shan Zhang, Kun Yao, Errui Ding, Gang Zhang, and Jingdong Wang. Lw-detr: A transformer replacement to yolo for real-time detection. *arXiv preprint arXiv:2406.03459*, 2024. Available at: https://github.com/Atten4Vis/LW-DETR.

[5] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[6] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[7] Yu Gong, Xiaoqiao Wang, and Chichun Zhou. Human-machine knowledge hybrid augmentation method for surface defect detection based few-data learning. *[Journal name not visible in excerpt]*, pages 1–24, 2023.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[9] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

[11] Adrian Shuai Li, Elisa Bertino, Rih-Teng Wu, and Ting-Yan Wu. Building manufacturing deep learning models with minimal and imbalanced training data using domain adaptation and data augmentation. In *2023 IEEE International Conference on Industrial Technology (ICIT)*, pages 1–8, Orlando, FL, USA, April 2023.

[12] Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. *arXiv preprint arXiv:2103.10158*, 2021.

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 01 2024.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[16] Isaac Robinson, Peter Robicheaux, and Matvei Popov. Rf-detr. https://github.com/roboflow/rf-detr, 2025. SOTA Real-Time Object Detection Model.

[17] Ranjan Sapkota, Rahul Harsha Cheppally, Ajay Sharda, and Manoj Karkee. Rf-detr object detection vs yolov12 : A study of transformer-based and cnn-based architectures for single-class and multi-class greenfruit detection in complex orchard environments under label ambiguity. *arXiv preprint arXiv:2504.13099*, 2025. Submitted to Elsevier.

[18] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(60), 2019.

[19] Chaitanya Singla, Rajat Bhardwaj, Nilesh Shelke, and Gurpreet Singh. Data augmentation: Synthetic image generation for medical images using vector quantized variational autoencoders. In *2025 3rd International Conference on Disruptive Technologies (ICDT)*, pages 1502–1507. IEEE, 2025.

[20] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 6105–6114, 2019.

[21] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. 2025.

[22] Wei Xiong, Janghwan Lee, Shuhui Qu, and Wonhyouk Jang. Data augmentation for applying deep learning to display manufacturing defect detection. In *2020 International Display Workshops (IDW)*, pages 81–1. The Institute of Image Information and Television Engineers, 2020.

[23] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019.

[24] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[25] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

[26] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2024.

[27] Xiaopin Zhong, Junwei Zhu, Weixiang Liu, Chongxin Hu, Yuanlong Deng, and Zongze Wu. An overview of image generation of industrial surface defects. *Sensors*, 23(19):8160, 2023. College of Mechatronics and Control Engineering, Shenzhen University.

[28] Barret Zoph, Ekin D. Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. Learning data augmentation strategies for object detection. *arXiv preprint arXiv:1906.11172*, 2019.