XAI-Enhanced Semantic Segmentation Models for Visual Quality Inspection

Tobias Clement*†, Truong Thanh Hung Nguyen*†§, Mohamed Abdelaal‡, Hung Cao§
†Friedrich-Alexander-University Erlangen-Nürnberg, Germany
‡Software AG, Germany
§Analytics Everywhere Lab, University of New Brunswick, Canada
Email: {tobias.clement, hung.tt.nguyen}@fau.de, mohamed.abdelaal@softwareag.com, hcao3@unb.ca

***Abstract*—Visual quality inspection (VQI) systems are essential in various industries, including manufacturing, logistics, and semiconductor. By leveraging computer vision technology and machine learning algorithms, manufacturers can identify defects and anomalies in their products with high precision and speed. The lack of explainability in VQI systems can lead to an inability to identify biases or errors, a lack of trust, and difficulty in improving the system. In this work, we introduce a novel framework for enhancing VQI systems via leveraging CAM-based model explanations to improve the performance of semantic segmentation models. Our framework entails four components, including 1) Model Training, 1) Model Explanation with XAI, 3) XAI Evaluation, and 4) Model Enhancement via annotation augmentation guided by explanations and domain experts. Our extensive evaluation reveals that the XAI-enhanced models outperform the original DeepLabv3-ResNet101 models, particularly in complex object segmentation tasks.**

***Index Terms*—Explainable AI, Visual Quality Inspection, Annotation Augmentation**

I. Introduction

Visual Quality Inspection (VQI) systems are automated mechanisms, typically engineered to scrutinize and continuously monitor the status of hardware assets. They harness the capabilities of Artificial Intelligence (AI) to automate the quality inspection process, thereby mitigating human error and augmenting efficiency. For instance, car manufacturers commonly employ VQI for real-time monitoring of paint robots to detect paint flaws [1]. With the advent of advanced AI models, such as Deep Neural Networks (DNNs), the precision of numerous VQI systems has been significantly enhanced. However, this advancement has engendered a trade-off between accuracy and interpretability [2]. These models, often perceived as "black boxes," lack transparency, thereby posing challenges for users in comprehending their decision-making processes. This opacity becomes a significant concern in sensitive domains [3], where decisions have profound implications.

Explainable Artificial Intelligence (XAI) endeavors to bridge this gap, offering human-understandable explanations for AI models' decisions [4]. XAI not only bolsters user trust but also facilitates model debugging, fairness assessment, and regulatory compliance [5]. Despite the growth of various XAI techniques, a comprehensive framework that enhances the transparency, plausibility, and fairness of models in VQI systems, especially with semantic segmentation models, is currently absent. To address this challenge, we propose a novel XAI-enhanced VQI framework that exploits the CAM-based explanations to improve the performance of semantic segmentation models, e.g., DeepLabv3-ResNet101. With our framework, we aim to harmonize the trade-off between highly accurate AI models and their interpretability, offering users meaningful explanations to refine the model's performance. To sum up, the paper offers the following notable contributions:

1) Enhanced VQI Framework (Section III): We introduce a comprehensive framework that integrates XAI into conventional VQI systems, comprising four building blocks: model training, model explanation, XAI evaluation, and model enhancement.
2) Evaluation of CAM-based Explanations (Section IV-A): We assess the faithfulness and plausibility of CAM-based explanations, providing valuable guidance for selecting suitable XAI methods for model enhancement.
3) XAI-guided Performance Improvement (Section IV-B): We optimize the performance of the DeepLabv3-ResNet101 model through annotation augmentation, directed by CAM-based explanations and domain experts.

The remainder of the paper is structured as follows: Section II discusses the related work in the realm of VQI, semantic segmentation, and XAI methods. Section III introduces our use case of VQI and the XAI-enhanced VQI framework. In Section IV, we discuss the results of our experimental evaluations, before Section V draws a conclusion of the obtained results.

II. Background & Prior Research

In this section, we provide an overview of four key areas related to our study: visual quality inspection, semantic segmentation, XAI, and model enhancement with XAI.

Visual Quality Inspection (VQI): In general, quality control is a crucial, yet often costly and time-consuming process in manufacturing and similar environments [6]. The VQI system, an AI-driven solution, can transform the inspection process into a more consistent and reliable procedure [7]. VQI systems are applicable across various industries, including automotive, electronics, and hardware assets via industrial, surveillance,

*Equal Contribution

or aerial cameras, which can bring productivity enhancement, quality assurance, and cost reduction [7]–[9]. Deep Learning (DL) models, e.g., YOLO [10], ResNet [11], are increasingly employed in VQI systems [12] and have demonstrated exceptional results across several inspection applications.

Semantic Segmentation: To realize VQI systems, semantic segmentation is a crucial tool, where it involves assigning semantic labels to each pixel in an image, enabling VQI systems to focus on critical parts of an image while ignoring irrelevant regions [13]. Examples of semantic segmentation models, achieving notable results, include FCN [14], LRASPP [15], and DeepLabv3 [16]. In this paper, we utilize the DeepLabv3 [16], equipped with the ResNet101 backbone [11]. Such a model represents a significant advancement in the field of semantic segmentation due to its applicability and performance on mobile devices. It incorporates atrous convolutions and spatial pyramid pooling modules to effectively capture multi-scale contextual information without the need for multiple input scales [11].

Explainable AI: Adopting XAI tools in Computer Vision (CV) usually provide insights into deep Convolutional Neural Network (CNN) models. These methods can be classified based on the mechanism of generating explanation maps, which highlight influential regions in the model's prediction, including Backpropagation-based, Class Activation Mapping (CAM)-based, and Perturbation-based methods [17]. Backpropagation-based methods use the backpropagation algorithm to identify each neuron's contribution to the final prediction [18], [19]. CAM-based methods generate heatmaps to highlight influential regions in the input image [20]–[24]. Perturbation-based methods modify the input data and observe changes in the model's output, providing insights into the model's decision-making process [25]–[27].

In fact, the abundance of XAI methods can create confusion among end-users seeking appropriate techniques for their systems [4]. Therefore, a systematic evaluation of explanation methods is crucial to validate such XAI methods. Specifically, XAI can be evaluated based on various metrics, which can be grouped based on their logical similarity [28]. These metrics include plausibility, faithfulness, robustness, localization, complexity, randomization, and axiomatic metrics [3], [28]. In this paper, we assess XAI methods based on plausibility and faithfulness. Plausibility measures whether the AI system's explanations align with human intuition, while faithfulness quantifies how well the explanation reflects the model's decision-making process [3], [28].

Model Enhancement with Explainable AI: XAI explanations can significantly contribute to the improvement of model's performance, robustness, efficiency, reasoning capabilities, and fairness [29]. Several ways of enhancing the performance of CV models using XAI explanations have been proposed, including:

- *Data augmentation*: label rectification and systematic retraining can enhance model robustness and performance [30]. Techniques like Guided Zoom [31] can improve model performance by eliminating irrelevant information and refining model predictions. SHapley Additive exPlanations (SHAP) values [32] and synthetic samples [33] generated from XAI explanations can also be used to enhance model performance.
- *Feature augmentation*: Relevance-based feature masking [34] can enhance model performance by focusing on the most relevant features. Other techniques, like feature transformations, can improve model performance by identifying and eliminating biases and artifacts [35].
- *Loss augmentation*: augmenting the loss function with regularization terms or scaling derived from XAI explanations can enhance a model's reasoning, robustness, performance, and convergence. The Attention Branch Network (ABN) [36] and attribution priors [37] are examples of loss augmentation techniques that incorporate XAI insights into the learning process.
- *Gradient augmentation*: the optimization method that employs Layer-wise Relevance Propagation (LRP) [38] can enhance model performance by selecting the most pertinent gradients during backpropagation.
- *Model augmentation*: pruning and quantization [39], [40] can reduce the model's complexity and storage requirements without compromising performance. Knowledge transfer techniques [41] can create a different model with similar behavior and beneficial properties.

III. Methodology

This section outlines our methodological approach toward the creation and evaluation of an enhanced VQI framework leveraging XAI for increased performance and interpretability.

Use Case: Our work assumes a cloud-based AI solution for VQI that monitors field-installed hardware assets. Such a VQI system aids field engineers in capturing asset images via edge devices, e.g., a mobile application, processed by an AI-driven VQI module in the cloud. This AI-powered module identifies the asset type and assesses its health. The asset health estimates are updated in an asset management system, facilitating maintenance planning and providing field-level insights. However, VQI systems usually face challenges such as model calibration [42], out-of-distribution generalization [43], and adversarial examples [44]. Moreover, end-users may struggle to evaluate the models' adequacy [45].

To address these challenges, we propose an XAI-enhanced VQI framework integrating XAI methods, providing transparency and interpretability to the AI decision-making processes. The XAI integration into the VQI system allows for the application of XAI methods to the Machine Learning (ML) models, offering explanations for their predictions and decisions. These explanations can be used to refine the models, leading to more accurate and reliable inspections. This improved VQI system enhances inspection accuracy and reliability and boosts end-users trust in the system.

Dataset: We employ the public TTPLA dataset, a key resource for detecting and segmenting power-grid hardware

assets [46]. The dataset comprises 1242 high-resolution images with 8987 instances of transmission towers and power lines, classified into four categories: cable, tower_wooden, tower_lattice, tower_tucohy (Figure 1). The images, manually annotated in the COCO format, present unique challenges due to the nature of the objects and diverse backgrounds, lighting conditions, and object sizes. The dataset supports both detection and semantic segmentation, as well as instance segmentation, enabling the identification and differentiation of individual towers and lines.

(a) (b) (c) (d)

Fig. 1. Samples from the TTPLA dataset represent the main objects of categories (a) cable, (b) tower_wooden, (c) tower_lattice, (d) tower_tucohy.

Enhanced VQI Framework: The XAI-enhanced VQI framework, depicted in Figure 2, consists of four main components. First, it involves training semantic segmentation models. Second, XAI methods are integrated into these models to generate interpretable explanations for their predictions. The third component assesses the XAI methods using both qualitative and quantitative metrics to ensure accuracy and comprehensibility of explanations. Lastly, the framework enhances the model's performance by augmenting annotations with XAI explanations, thereby facilitating improved learning data. Additionally, a user-friendly web application has been developed to facilitate seamless interaction with the enhanced VQI framework. Below, we elaborate on the four components.

1) Model Training: This component focuses on the training of core models for the VQI module using the training set, where the original dataset is split into an 80%-20% training-test set, with all images resized to 500×500 pixels. Corresponding COCO annotations are transformed into masks, serving as the ground truth. The DeepLabv3-ResNet101 [16] is employed as the core segmentation model due to its applicability and performance on mobile devices. The Dice loss function is used for training the ResNet101 model, which is particularly useful for imbalanced classes in the image segmentation task, as it considers the overlap between the predicted and ground truth masks [47].

2) Model Explanation with XAI: In this component, the explanation maps of all methods are extracted from the predictions of the segmentation model on the test set, which will be used for the evaluation step. We utilize five notable CAM-based XAI methods: GradCAM [22], GradCAM++ [21], XGradCAM [48], HiResCAM [49], ScoreCAM [50] due to their applicabilities and plausibility in the semantic segmentation task. The explanation maps can be delivered to end-users via a web application where they can upload input images to verify the model's behavior.

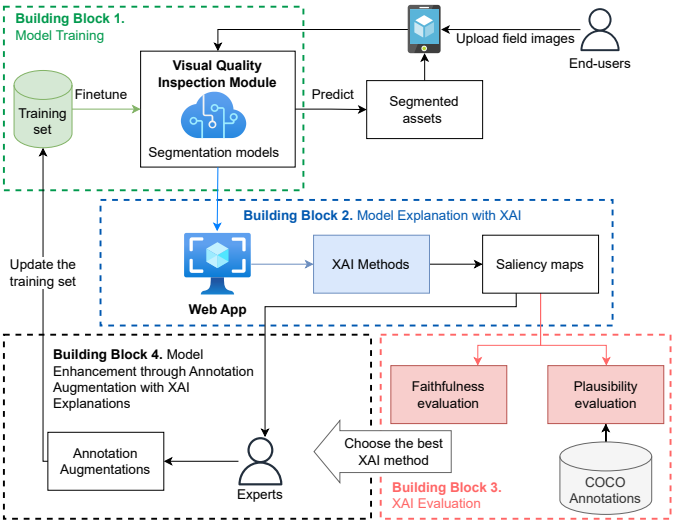


Fig. 2. The enhanced Visual Quality Inspection (VQI) framework integrated with XAI methods with 4 building blocks: (1) Training models, (2) Model Explanation with XAI, (3) XAI Evaluation, and (4) Model Improvement by XAI with Human-in-the-loop. The end-users interact with the framework via a web application.

3) XAI Evaluation: This component evaluates the XAI methods with plausibility and faithfulness (Section IV-A) metrics on their explanations. Plausibility measures how well the explanations align with human intuition and understanding, while faithfulness measures how accurately the explanations reflect the underlying model's decision-making process. By evaluating both plausibility and faithfulness, we can ensure that the chosen XAI method provides explanations that are both understandable to humans and accurately represent the model's behavior. Eventually, the method achieving the highest scores in most metrics will be chosen as the core XAI method of the model enhancement step.

4) Model Enhancement via Annotation Augmentation with XAI Explanations: This component enhances the DeepLabv3-ResNet101 model's performance on the TTPLA dataset. Data augmentation strategies, such as altering data distribution or adjusting data and labels, have been used to enhance model performance [30]. The XAI method having the highest faithfulness and plausibility from the XAI evaluation step will be used to guide the annotation augmentation process. The COCO annotations from the TTPLA dataset are relabeled based on expert recommendations. The model is then retrained on the enhanced training dataset with augmented annotations. The original test set is used to compare the performance of the conventional and enhanced models, demonstrating the potential of annotation augmentation, supported by XAI explanations, in enhancing semantic segmentation models. After acquiring the final improved model, we deploy it on mobile devices via PyTorch mobile framework [51].

IV. Performance Evaluation

As stated in our contributions, this section details the results derived from our evaluation of CAM-based XAI techniques.

Additionally, we discuss their use in improving model performance, specifically for applications on mobile devices.

A. XAI Evaluation

Evaluation Metrics: In the following, we introduce two relevant metrics, including plausibility and faithfulness of XAI explanations. Plausibility, the alignment of explanations with human intuition, is assessed using measures like *Energy-Based Pointing Game (EBPG)* [50], *Intersection over Union (IoU)* [52], [53], and *Bounding Box (Bbox)* [54]. These measures, based on human annotations, validate the model by assessing the statistical superiority of explanations. Specifically, EBPG evaluates the precision and denoising ability of XAI methods to identify the most influential region in an image for a given prediction [50]. Whereas, IoU assesses the localization capability and the significance of the attributions captured in an explanation map [52], [53]. Finally, Bbox is a variant of the IoU metric that adapts to the size of the object of interest [54].

Faithfulness, the alignment of explanations with the model's predictive behavior, is evaluated using the *Drop* and *Increase* measures [48]. These measures quantify the degree to which the explanations align with the predictive behavior of the model. Drop [48] measures the average model prediction decrease when the explanation is used as input. Alternatively, the Increase measure [48] quantifies the frequency at which the model's confidence increases when the explanation is used as input.

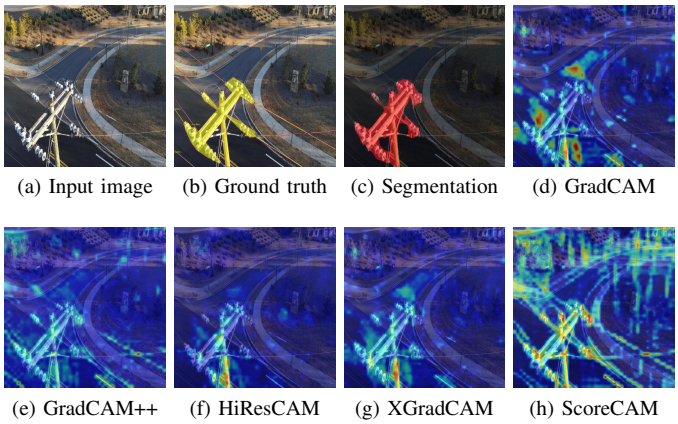
Evaluation Results: The explanation maps of implemented XAI methods are demonstrated in Figure 3. The plausibility and faithfulness of XAI methods are quantitatively evaluated to find the most suitable XAI method, which can act as the core method of the model enhancement step. As shown in Table I, HiResCAM achieves not only the best performance in the faithfulness evaluations, such as Drop and Increase but also the shortest computational time. While GradCAM++ has the highest scores with BBox and IoU for plausibility, HiResCAM still performs plausibly with the highest score in EPBG. Hence, we choose HiResCAM as the core XAI method for the model enhancement step.

TABLE I
THE QUANTITATIVE EVALUATIONS OF XAI METHODS. FOR EACH METRIC, THE ARROW ↑ / ↓ INDICATES HIGHER/LOWER SCORES ARE BETTER. THE BEST IS IN BOLD.

Method	EPBG ↑	BBox ↑	IoU ↑	Drop ↓	Inc ↑	Time(s) ↓
GradCAM	50.49	48.39	47.94	5.21	52.57	3.21
GradCAM++	58.13	**52.24**	**53.22**	5.17	54.66	4.20
HiResCAM	**60.81**	41.69	52.19	**5.01**	**55.93**	**3.13**
XGradCAM	57.94	47.81	53.09	5.94	55.01	4.43
ScoreCAM	54.01	43.95	51.94	7.34	47.19	52.50

B. Model Enhancement

This section presents the experimental results of enhancing the DeepLabv3-ResNet101 model's performance using annotation augmentation guided by XAI methods and a domain expert. The process begins with the XAI method generating explanations for each training data sample. The domain expert, knowledgeable in semantic segmentation models and XAI algorithms, analyzes the saliency maps to guide annotation augmentation. We select samples of increasing complexity from the training set and use HiResCAM to generate explanations.



(a) Input image (b) Ground truth (c) Segmentation (d) GradCAM

(e) GradCAM++ (f) HiResCAM (g) XGradCAM (h) ScoreCAM

Fig. 3. The qualitative evaluation of implemented XAI methods on the segmentation result of the DeepLabv3-ResNet101 model on a sample from the test set. The category for the segmentation is the tower_wooden denoted under the yellow box shown in the ground truth. The IoU value between the segmentation and the ground truth is 0.9085.

As shown in Figure 4, the model effectively segments the cable from a clean or mixed-objects background. However, when the background contains objects resembling the target object, the model's performance decreases. The explanations reveal that the model's attention is directed at the object and the surrounding background. However, the model lacks contextual attention to surrounding objects and background in complex cases. This behavior is due to the ability of models to leverage local and global contextual information from the original annotations [26].

To enhance the model's performance, a domain expert suggests annotation augmentation for each sample. Two approaches are proposed, namely *Annotation Enlargement* and *Adding Annotations for Perplexed Objects* (cf. Figure 5). The enhanced DeepLabv3-ResNet101 model demonstrates improved segmentation of thin objects from the background and perplexing objects (Figure 6). The IoU of the enhanced model is also higher than that of the conventional version, particularly noticeable in the `cable` IoU, which increased from 55.06 to 58.11, leading to a higher overall IoU (from 83.94 to 84.715), as shown in Table II.

TABLE II
QUANTITATIVE RESULTS OF DEEPLABV3-RESNET101 BEFORE AND AFTER APPLYING THE ENHANCING MODEL BY ANNOTATION AUGMENTATION WITH XAI METHODS IN IOU (%) ON EACH CATEGORY AND IN AVERAGE. THE BETTER IS INDICATED IN BOLD.

Model	cable	tower_wooden	tower_lattice	tower_tucohy	Overall
Original	55.06	94.75	95.31	90.63	83.94
Enhanced	**58.11**	**94.78**	**95.32**	**90.65**	**84.715**

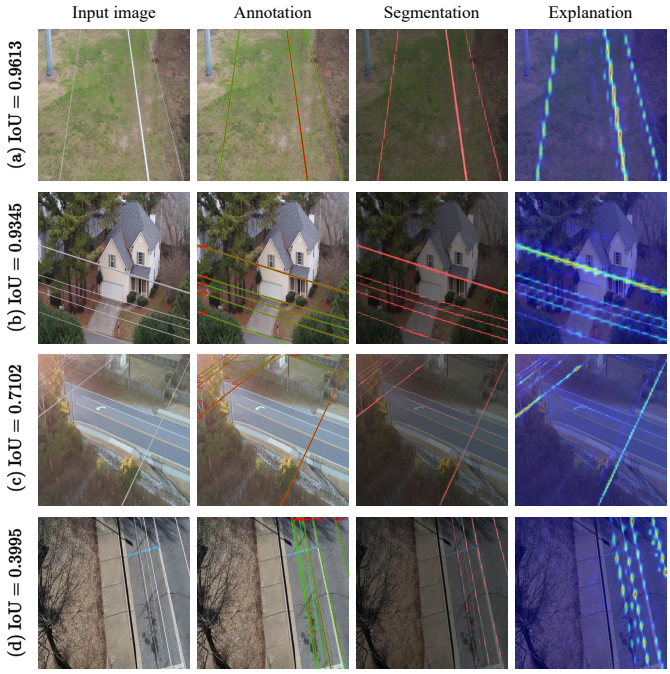


Fig. 4. List of input images, COCO annotations (ground truth), segmentation results of the DeepLabv3-ResNet101 model, and the HiResCAM explanations in increasing order of complexity.

(a) Annotation enlargement (b) Adding annotations for perplexed objects

Fig. 5. Annotation augmentation approaches: (a) Annotation enlargement where the size of the annotation for thin objects like cables is increased, (b) Adding annotations for perplexed objects like the road surface marks to guide the model in differentiating between white cables and perplexed objects.

Fig. 6. Qualitative results of DeepLabv3-ResNet101 before and after applying the enhancing model performance by annotation augmentation with XAI methods procedure.

V. Conclusion

This paper presents an improved VQI system, employing XAI techniques to enhance interpretability and performance in semantic segmentation tasks on mobile devices. We utilized a public dataset to illustrate the potential of XAI in model improvement and explanation. A variety of XAI methods were evaluated qualitatively and quantitatively, providing insights for users to select suitable XAI methods. The model enhancement procedure, guided by XAI's explanation maps, effectively improved model performance in complex object segmentation and detection, especially in challenging contexts where objects and backgrounds are indistinguishable. In our future research, we plan to expand the scope of our framework to encompass additional image tasks, such as object detection and instance segmentation. Furthermore, we aim to enhance the usability of our model by developing a more intuitive interface for end-users, thereby reducing the reliance on human-in-the-loop intervention. This will facilitate the adoption of our framework by a wider audience and enable its application in a broader range of contexts.

Acknowledgment

This work was supported by the German Federal Ministry of Education and Research through grants 01IS17045 (Software Campus project), 02L19C155, 01IS21021A (ITEA project number 20219).

References

[1] Software AG, "Duerr customer story, cumulocity iot, software ag," (Accessed on August 3, 2023). [Online]. Available: https://www.softwareag.com/en_corporate/customers/customer-stories/duerr.html

[2] G. Baryannis, S. Dani, and G. Antoniou, "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability," *Future Generation Computer Systems*, vol. 101, pp. 993–1004, 2019.

[3] T. T. H. Nguyen, V. B. Truong, V. T. K. Nguyen, Q. H. Cao, and Q. K. Nguyen, "Towards trust of explainable ai in thyroid nodule diagnosis," *arXiv preprint arXiv:2303.04731*, 2023.

[4] T. Clement, N. Kemmerzell, M. Abdelaal, and M. Amberg, "Xair: A systematic metareview of explainable ai (xai) aligned to the software development process," *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 78–108, 2023.

[5] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[6] H. Tang, *Manufacturing system and process development for vehicle assembly*. SAE International, 2017.

[7] X. Sun, J. Gu, S. Tang, and J. Li, "Research progress of visual inspection technology of steel products—a review," *Applied Sciences*, vol. 8, no. 11, p. 2195, 2018.

[8] A. Q. Md, K. Jha, S. Haneef, A. K. Sivaraman, and K. F. Tee, "A review on data-driven quality prediction in the production process with machine learning for industry 4.0," *Processes*, vol. 10, no. 10, p. 1966, 2022.

[9] Y. D. Yasuda, F. A. Cappabianco, L. E. G. Martins, and J. A. Gripp, "Aircraft visual inspection: A systematic literature review," *Computers in Industry*, vol. 141, p. 103695, 2022.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] S. Sundaram and A. Zeid, "Artificial intelligence-based smart quality inspection for manufacturing," *Micromachines*, vol. 14, no. 3, p. 570, 2023.

[13] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[15] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017.

[17] S.-A. Rebuffi, R. Fong, X. Ji, and A. Vedaldi, "There and back again: Revisiting backpropagation saliency methods," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8839–8848.

[18] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[19] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[21] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[23] P. X. Nguyen, H. Q. Cao, K. V. Nguyen, H. Nguyen, and T. Yairi, "Secam: Tightly accelerate the image explanation via region-based segmentation," *IEICE TRANSACTIONS on Information and Systems*, vol. 105, no. 8, pp. 1401–1417, 2022.

[24] Q. K. Nguyen, T. T. H. Nguyen, V. T. K. Nguyen, V. B. Truong, and Q. H. Cao, "G-came: Gaussian-class activation mapping explainer for object detectors," *arXiv preprint arXiv:2306.03400*, 2023.

[25] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421*, 2018.

[26] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, "Black-box explanation of object detectors via saliency maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 443–11 452.

[27] V. B. Truong, T. T. H. Nguyen, V. T. K. Nguyen, Q. K. Nguyen, and Q. H. Cao, "Towards better explanations for object detection," *arXiv preprint arXiv:2306.02744*, 2023.

[28] A. Hedström, L. Weber, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "Quantus: an explainable ai toolkit for responsible evaluation of neural network explanations," *arXiv preprint arXiv:2202.06861*, 2022.

[29] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of xai-based model improvement," *Information Fusion*, 2022.

[30] V. Bento, M. Kohler, P. Diaz, L. Mendoza, and M. A. Pacheco, "Improving deep learning performance by using explainable artificial intelligence (xai) approaches," *Discover Artificial Intelligence*, vol. 1, pp. 1–11, 2021.

[31] S. A. Bargal, A. Zunino, V. Petsiuk, J. Zhang, K. Saenko, V. Murino, and S. Sclaroff, "Guided zoom: Questioning network evidence for fine-grained classification," *arXiv preprint arXiv:1812.02626*, 2018.

[32] H. Sun, L. Servadei, H. Feng, M. Stephan, A. Santra, and R. Wille, "Utilizing explainable ai for improving the performance of neural networks," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 1775–1782.

[33] S. Teso and K. Kersting, "Explanatory interactive machine learning," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 239–245.

[34] D. Schiller, T. Huber, F. Lingenfelser, M. Dietz, A. Seiderer, and E. André, "Relevance-based feature masking: Improving neural network based whale classification through explainable artificial intelligence," 2019.

[35] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Finding and removing clever hans: using explanation methods to debug and improve deep models," *Information Fusion*, vol. 77, pp. 261–295, 2022.

[36] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Attention branch network: Learning of attention mechanism for visual explanation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 705–10 714.

[37] M. Mitsuhara, H. Fukui, Y. Sakashita, T. Ogata, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Embedding human knowledge into deep neural network via attention map," *arXiv preprint arXiv:1905.03540*, 2019.

[38] J. ha Lee, I. hee Shin, S. gu Jeong, S.-I. Lee, M. Z. Zaheer, and B.-S. Seo, "Improvement in deep networks for optimization using explainable artificial intelligence," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 2019, pp. 525–530.

[39] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, and W. Samek, "Pruning by explaining: A novel criterion for deep neural network pruning," *Pattern Recognition*, vol. 115, p. 107899, 2021.

[40] M. Sabih, F. Hannig, and J. Teich, "Utilizing explainable ai for quantization and pruning of deep neural networks," *arXiv preprint arXiv:2008.09072*, 2020.

[41] W. Ha, C. Singh, F. Lanusse, S. Upadhyayula, and B. Yu, "Adaptive wavelet distillation from neural networks through interpretations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20 669–20 682, 2021.

[42] J. M. Rožanec, L. Bizjak, E. Trajkova, P. Zajec, J. Keizer, B. Fortuna, and D. Mladenić, "Active learning and approximate model calibration for automated visual inspection in manufacturing," *arXiv preprint arXiv:2209.05486*, 2022.

[43] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021.

[44] G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial examples that fool both computer vision and time-limited humans," *Advances in neural information processing systems*, vol. 31, 2018.

[45] M. H. Bharati, "Multivariate image analysis and regression for industrial process monitoring and product quality control," Ph.D. dissertation, 2002.

[46] R. Abdelfattah, X. Wang, and S. Wang, "Ttpla: An aerial-image dataset for detection and segmentation of transmission towers and power lines," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[47] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*. Springer, 2017, pp. 240–248.

[48] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based grad-cam: Towards accurate visualization and explanation of cnns," *arXiv preprint arXiv:2008.02312*, 2020.

[49] R. L. Draelos and L. Carin, "Hirescam: Faithful location representation in visual attention for explainable 3d medical image classification," *arXiv preprint arXiv:2011.08891*, 2020.

[50] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[51] "Image segmentation deeplabv3 on ios — pytorch tutorials 2.0.1+cu117 documentation," https://pytorch.org/tutorials/beginner/deeplabv3_on_ios.html, (Accessed on 07/25/2023).

[52] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541–6549.

[53] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," *arXiv preprint arXiv:1807.08024*, 2018.

[54] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," *arXiv preprint arXiv:2001.00396*, 2020.