Interpretability To Enhance Transparency in Computer Vision System

IBNAZHIFI. NEVINDRA[1], NILMA. MUFID[2]
[1, 2] *Departement of Business Information System, Gunadarma University, Jakarta, Indonesia*

Abstract—The rise of AI adoption in diverse fields has raised concerns about the transparency and accountability of decision-making processes. Computer vision companies face challenges in data transparency, bias mitigation, traceability, and AI decision explication. This research addresses these issues by investigating the implementation of explainability in computer vision systems. This study aims to provide valuable insights for improving the understanding and trustworthiness of AI models in computer vision applications by emphasizing explainability. The findings potentially foster wider acceptance of AI solutions across various industries.

Indexed Terms—Computer Vision, Explainability, Transparency

I. INTRODUCTION

The increasing adoption of artificial intelligence (AI) in many industries, including computer vision, has led to significant advancements that can rival human intelligence in complex problem-solving. Setting AI apart from other technologies, as many cognitive tasks traditionally performed by humans can be replaced and outperformed by machines (Taeihagh, 2021).

The growing adoption of AI in multiple industries has raised concerns about the transparency and accountability of these systems (Burr & Leslie, 2022). The public and governments demand greater transparency from AI systems (European Commission, 2020).

While the technology can yield positive impacts for humanity, its AI applications can also generate unexpected and unintended consequences and pose new forms of risks that need to be effectively managed by governments (European Commission, 2020; Chen et al., 2023). In recent years, numerous studies have documented that biases inherent in AI systems and the data they are trained on have led to discriminatory and unethical consequences for individuals across various domains (Wehrli et al., 2021; Helbing, 2018). This lack of transparency has led to increased calls for greater accountability and traceability in AI systems and has raised questions about the ethics and reliability of these systems (Kroll, 2021; European Commission, 2020).

European Commission presents a European AI strategy focusing on achieving excellence and trust. The paper highlights the potential benefits of AI for European society and the economy, such as improved healthcare, transport, and energy efficiency. However, it also recognizes the potential risks and challenges associated with AI, such as the impact on employment, the potential for bias and discrimination, and threats to fundamental rights and freedoms. The EC aims to address these challenges by promoting ethical and trustworthy AI and ensuring that the technology is developed and used in a human-centric way.

The paper proposes a three-pronged approach to achieve the EU's vision for AI:

- Excellence: The EU aims to promote world-class research and innovation in AI and foster a vibrant ecosystem for startups and SMEs. The EC plans to invest in AI research and development, establish a European AI research network, and encourage the development of AI skills through education and training programs.
- Trust: The EU aims to develop a framework for trustworthy AI that ensures the technology is developed and used ethically, responsibly, and transparently. The EC proposes establishing a European AI Alliance, developing AI ethics guidelines, and establishing a certification framework for AI products and services.
- Deployment: The EU aims to ensure that AI is deployed in a way that benefits society and

respects fundamental rights and freedoms. The EC proposes to develop a regulatory framework for AI, establish a European AI support center, and promote international cooperation on AI.

Linking it to transparency in AI, the paper stresses the importance of transparency and explainability in AI systems to ensure that the decision-making processes are accountable and can be audited. It also highlights the need to address the potential biases and discrimination in AI systems through transparency and explainability.

The primary objective of this study is to develop a comprehensive understanding of how MLOps can be leveraged to improve transparency in computer vision systems. The research will investigate various aspects, including data collection, model training, deployment, and maintenance, to identify key factors and methodologies contributing to increased transparency. By examining the implementation of MLOps techniques, the study aims to provide practical insights and guidelines for implementing transparent and ethically sound computer vision systems.

II. LITERATUR REVIEW

A. Transparency in AI System

Transparency is a complex and dynamic concept that has been defined in various ways. Transparency refers to the extent to which information about the activities of an organization or government is available to the public (Meijer, 2013). This information can take many forms, including documents, reports, meetings, and decisions.

Albert Meijer argues that transparency is not a static concept but rather a dynamic process that is constantly evolving. Various factors, including strategic interactions, cognitive factors, and institutional rules, influence the construction of transparency.

Generally, transparency refers to how individuals and organizations provide access to information about their activities, decisions, and outcomes. Proponents of transparency argue that it can clean up corruption, build trust, and increase accountability (Douglas & Meijer, 2016). However, some detractors argue that transparency can have adverse effects, such as limiting privacy and creating a culture of suspicion (Douglas & Meijer, 2016).

In the context of artificial intelligence (AI), transparency is essential because AI systems can make automated decisions affecting individuals and organizations. Thus, there is a need for transparency in the design and operation of AI systems to ensure that they are fair, unbiased, and accountable. Meijer and his colleagues have developed the concept of "Transparency by Design," which provides practical guidance for promoting the beneficial functions of transparency while mitigating its challenges in automated-decision making environments (Felzmann et al., 2020).

In computer vision, transparency refers to the ability of stakeholders to understand the processes, data, and algorithms used to make decisions and produce results. This includes understanding how data is collected, processed, and used to train computer vision models and how these models are deployed and make predictions in real-world scenarios.

B. Interpretability in Computer Vision System

Explainability in computer vision systems refers to the ability to understand and interpret the decisions and processes of the system. It is important for several reasons: transparency, accountability, and trustworthiness. In order to ensure that computer vision systems are reliable and fair, it is crucial to be able to explain how they arrive at their conclusions and to identify any biases or errors that may be present.

One approach to achieving interpretability in computer vision systems is using explainers for deep neural networks. Buhrmester et al. (2021) conducted a survey that analyzed explainers of black box deep neural networks for computer vision tasks (Buhrmester et al., 2021). The survey provides a comprehensive overview of the mechanisms and properties of explaining systems, and it compares several survey papers that deal with explainability in general. The authors identify the drawbacks and gaps in the field and propose further research ideas (Buhrmester et al., 2021).

In addition to deep neural networks, other methods can be employed for interpretability in computer vision.

Tizhoosh & Pantanowitz (2018) highlight the role of incorporating handcrafted features in computer vision schemes (Tizhoosh & Pantanowitz, 2018). These well-established feature extraction methods, such as local binary patterns and encoded local projections, can provide interpretable results and do not require excessive computational resources for learning. These methods allow the system's behavior to be fully understood, and humans can interpret the results (Tizhoosh & Pantanowitz, 2018).

C. Eigen-CAM

Eigen-CAM (Class Activation Map using Principal Components) is an extension of the Class Activation Map (CAM) technique, which aims to provide interpretability for convolutional neural networks (CNNs) in computer vision tasks. Eigen-CAM was introduced in the research paper "Eigen-CAM: Class Activation Map via Eigen Decomposition" by A. Chattopadhyay et al. (2019).

CAM is a method used to visualize the important regions in an image that contribute to the predictions made by a CNN. It is commonly used in tasks like image classification and object detection to gain insights into what regions of an image the model focuses on when making a prediction. The final convolutional layer's output generates the class-specific activation map in CAM. The final convolutional layer typically has a spatial resolution that matches the input image's size. CAM takes the weights of the final fully connected layer (or the global average pooling layer) and combines them with the activation map to obtain a weighted sum. This process highlights the regions in the input image most important for the specific class prediction.

Eigen-CAM builds upon the idea of CAM and enhances its interpretability by using principal components obtained through eigende composition. The primary motivation behind Eigen-CAM is to identify more interpretable patterns and features that CNN focuses on during decision-making.

III. PROPOSED METHODOLOGY

The primary goal of this research is to enhance the interpretability of Computer Vision models for image classification tasks, specifically for differentiating between various classes of vehicles. Interpretable models are crucial for building trust in AI systems and enabling better decision-making in real-world applications. To achieve this objective, we will explore and implement the EigenCAM (Eigen Class Activation Map) technique, a popular approach for interpreting deep learning models.

A. Training Model

In this subsection, we outline the process of training a robust and accurate Computer Vision model capable of distinguishing between different classes of vehicles.

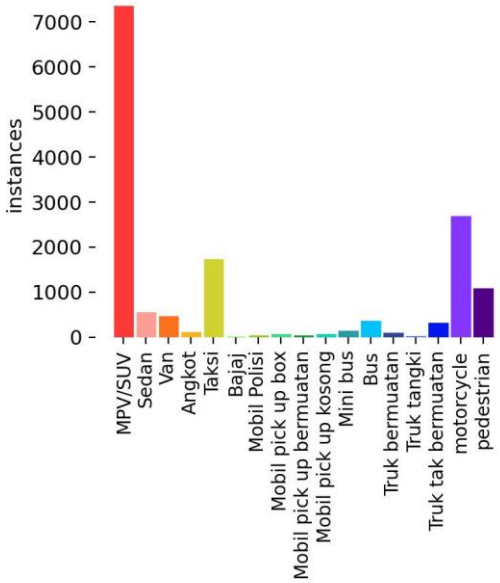

Figure 1. Class distribution visualization.

This study will use internal data from a total of 17 classes. The number of images whose distribution can be seen in Figure 1. The YOLOv8 model, which is pre-trained on a large-scale dataset, is used as the starting point for retraining. The pre-trained model provides a strong foundation with learned weights and features that can be fine-tuned on the specific dataset.

B. Implementation of EigenCAM

This subsection details the implementation of the EigenCAM technique, which enables the visualization and understanding of the regions of an image that contribute most to the model's classification decision.

By following this proposed methodology, we aim to develop an interpretable Computer Vision model for

vehicle classification, contributing to the advancement of explainable AI and fostering trust in AI systems.

IV. RESULT AND DISCUSSION

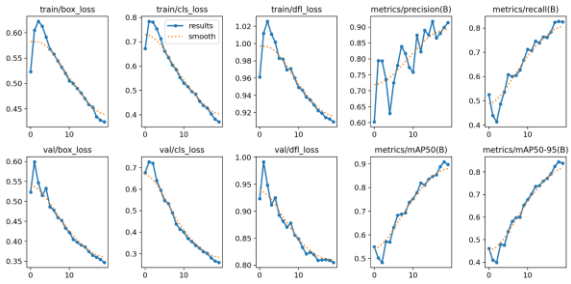

Figure 2. Training Result for YOLOv8 Model

Figure 2 is shown the graph result of the training process. The accuracy improved steadily throughout the training, reaching a final value of around 94%. The precision also improved steadily throughout the training process, reaching a final value of around 85%. The recall also improved steadily throughout the training, reaching a final value of around 90%. The mAP also improved steadily throughout the training, reaching a final value of around 65%.

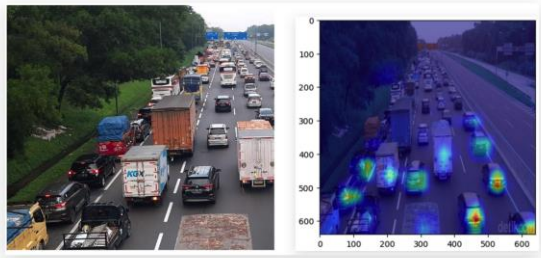

Figure 3. Implementation of Eigen-CAM after model training.

Figure 3 showcases a comparison between a heatmap image and the corresponding original image. The heat map highlights the regions in the image that the model considers most important for the prediction. Researchers can visually identify the areas contributing to the model's decision-making process by overlaying the heat map on the original image.

Eigen-CAM offers valuable insights into the model's inner workings and facilitates transparency by providing visual explanations for the model's predictions. It aids in understanding which image regions the model focuses on when making decisions, supporting subsequent discussions on model biases, potential errors, or highlighting areas for improvement.

CONCLUSION

The research successfully implemented transparency enhancements in computer vision systems, focusing on interpretability. While improvements have been made, there is still potential for further refinement. Specifically, future research about transparency should concentrate on other things, such as explainability and traceability. Eigen CAM offers a visual understanding of the decision-making process, enabling stakeholders to trust and validate model outputs.

Integrating Eigen CAM into the existing system would provide clearer insights into how the computer vision model arrives at its conclusions, aiding bias detection and understanding unintended consequences. Combining interpretability with transparency measures can develop a more robust and trustworthy computer vision system, promoting greater confidence in AI systems. Leveraging open-source tools will continue to be valuable, facilitating the process and ensuring compliance with industry best practices.

REFERENCES

[1] A. Taeihagh, “Governance of Artificial Intelligence,” Policy and Society, vol. 40, no. 2, pp. 137–157, 2021. doi:10.1080/14494035.2021.1928377

[2] C. Burr and D. Leslie, “Ethical assurance: A practical approach to the responsible design, development, and deployment of data-Driven Technologies,” AI and Ethics, vol. 3, no. 1, pp. 73–98, 2022. doi:10.1007/s43681-022-00178-0

[3] Y. Chen, E. W. Clayton, L. L. Novak, S. Anders, and B. Malin, “Human-centered design to address biases in artificial intelligence,” Journal of Medical Internet Research, vol. 25, 2023. doi:10.2196/43251

[4] European Commission, "White Paper on Artificial Intelligence: a European approach to excellence and trust," 2020. [Online]. Available:

https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en. [Accessed: 18 January 2023].

[5] J. A. Kroll, “Outlining traceability,” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. doi:10.1145/3442188.3445937

[6] A. Meijer, "Understanding the Complex Dynamics of Transparency," Public Administration Review, vol. 73, no. 3, pp. 429-439, 2013.

[7] S. Douglas and A. Meijer, "Transparency and public value—analyzing the transparency practices and value creation of public utilities," International Journal of Public Administration, vol. 39, no. 12, pp. 940-951, 2016.

[8] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, “Towards transparency by design for Artificial Intelligence,” Science and Engineering Ethics, vol. 26, no. 6, pp. 3333–3361, 2020. doi:10.1007/s11948-020-00276-4

[9] V. Buhrmester, D. Münch, M. Arens, "Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey", MAKE, vol. 3, no. 4, p. 966-989, 2021. https://doi.org/10.3390/make3040048

[10] H. Tizhoosh, L. Pantanowitz, "Artificial Intelligence and Digital Pathology: Challenges and Opportunities", Journal of Pathology Informatics, vol. 9, no. 1, p. 38, 2018. https://doi.org/10.4103/jpi.jpi_53_18

[11] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class Activation Map using Principal Components," arXiv, 2020. [Online]. Available: https://doi.org/10.1109/IJCNN48605.2020.9206626.