SMART YIELDS IN SYSTEMS, TECHNOLOGY, ENGINEERING, AND MODELING (SYSTEM)

Vol. 1 No. 1 (2026) 39 – 50 | ISSN: XXXX - XXXX (Media Online)

Comparative Analysis of Causal Interpretability Methodologies for Enhancing Trust in Deep Computer Vision

Mohamad Subroto Alirejo [1]*, Dwi Utari Iswavigra [2], Very Dwi Setiawan [3]

[1] STIE Wikara

[2] Universitas Sugeng Hartono

[3] Universitas Pignatelli Triputra

* Correspondence: aliredjo.sa@gmail.com

Article Information

Received: January 1, 2026
Revised: March 27, 2026
Online: March 31, 2026

ABSTRACT

This study systematically compares five distinct causal interpretability methodologies employed in deep computer vision using validated official secondary data obtained from governmental statistical agencies and peer-reviewed academic repositories. The analysis demonstrates that Graphical Causal Models (GCM) and Causal Generative Models (CGM) offer superior interpretative depth, but their practical application is highly resource-intensive, demanding substantial data and computational capacity. In contrast, Counterfactual (CEM) and Perturbation-based (PBA) methods provide swift, practical solutions, albeit with inherent limitations in achieving comprehensive causal depth. Based on comparative performance and resource constraints, the findings support the development of hybrid methodologies that effectively merge the strengths of both approaches, coupled with the standardization of official data integration. This strategy may contribute to improving model trustworthiness and transparency in critical application.

Keywords: Deep Learning; Causal Interpretability; Computer Vision; Graphical Causal Models; Counterfactual Explanations.

1. Introduction

Deep learning, particularly within the field of deep computer vision, has emerged as a fundamental catalyst for technological advancements across critical sectors such as medicine, security, and industrial automation, where accurate predictions are paramount. Nonetheless, the inherent structural complexity of the neural network

architectures employed in these models creates a pervasive 'black-box' effect, a term widely used to describe the opacity of deep neural decision-making processes [1,2]. This lack of transparency leads to considerable apprehension regarding user trust in model decisions, particularly when these decisions directly impact safety or ethical outcomes. Consequently, causal interpretability has become an essential area of research, enabling the elucidation of cause-and-effect relationships underpinning a model's predictions, rather than simply relying on potentially misleading statistical correlations [2]. This causal approach supports broader adoption of deep learning models in critical domains that require transparent decision-making [3].

A variety of causal interpretability methodologies are currently being developed to dismantle the "black box" surrounding deep learning models, particularly within computer vision. These methods encompass the utilization of graphical causal models, which employ graph structures to represent cause-and-effect relationships among various factors [4]; causal perturbation techniques, which involve observing how changes in input variables influence the output to ascertain causal effects [5]; and explanation models with intervening variables, which focus on understanding the mediation mechanisms between variables [6]. While the efficacy of these approaches has been extensively studied, recent literature reveals a significant divergence concerning their generalization capabilities across diverse domains and the stability of their causal decomposition [7]. Some researchers emphasize the superiority of counterfactual explanations in providing specific rationales for individual predictions [8], while others underscore the necessity of developing causal generative models that offer more holistic and valid interpretations by considering the full causal distribution of the data [9, 10]. This disparity in perspectives underscores the critical need for a rigorous and in-depth comparative evaluation of the existing methodologies. However, a systematic comparison of these methodologies within deep computer vision, particularly using official secondary data, remains limited.

This research is specifically structured to provide a comprehensive comparison of causal interpretability methodologies within the context of deep computer vision. It will leverage secondary data from authoritative sources, such as official statistical agencies and trusted academic and governmental institutions, to ensure the validity and reliability of the analysis [11]. This approach also guarantees adherence to scientific research standards that prioritize openness and the use of verifiable data.

Through this critical analysis, this paper aims to deliver profound insights into the strengths and limitations of each approach, culminating in the formulation of strategic recommendations for the future development of interpretability technology that can significantly enhance trust and transparency in deep computer vision applications.

2. Materials and Methods

The Materials and Methodology section of this research is meticulously designed to ensure the optimal replicability and validity of the results. The analytical cornerstone of this comparative study rests on the utilization of secondary data sourced from official institutions like the Central Statistics Agency (BPS) and other rigorously verified governmental and academic records. This approach is grounded in the principles of data transparency and openness in contemporary scientific inquiry. As detailed by Creswell and Creswell [12], research that leverages secondary data must clearly articulate the sources and the data processing pipeline to mitigate potential biases and guarantee the consistency of findings. This study does not conduct new fieldwork or laboratory experiments; instead, it integrates and synthesizes primary data and established research outcomes from well-documented datasets and standard protocols that are publicly accessible or obtained through specific permissions, thereby upholding scientific rigor and interpretive validity [13].

Data Sources and Instrumentation

The analysis involves data-driven simulations using image data integrated with demographic and environmental metadata from BPS, enabling cross-variable correlation analysis within the causal modeling framework. This secondary data is drawn from the latest public datasets (2023–2024) to ensure relevance and accuracy. This integration of official government statistics with deep learning technology aligns with the recommendations by Kamruzzaman et al. [14] for enhancing the accuracy and validity of research outcomes. Testing and modeling procedures are executed using Python software with PyTorch serving as the core deep learning platform and CausalNex utilized for developing and evaluating the causal models, thus facilitating the reliable and comprehensive implementation of interpretability techniques [15]. To ensure data and protocol openness, all generated datasets and programming code will be deposited in a public repository, fulfilling journal requirements for open

access. This adheres to modern research standards for strengthening replicability and scientific collaboration [17]. Given that this study does not involve human or animal intervention, no specific ethical approval is required.

Comparative Causal Methodologies and Protocols

The methodological focus of this research centers on comparing five primary approaches recognized for their significance and popularity within the literature on causal interpretability for deep computer vision:

1. Graphical Causal Models (GCMs): These models employ graphical representations, specifically Directed Acyclic Graphs (DAGs), to explicitly map and explain the causal relationships among various variables.
2. Counterfactual Explanation Methods (CEMs): These methods explore "what-if" scenarios by simulating changes in input variables to observe their causal effect on the model's output.
3. Perturbation-based Approaches (PBAs): These techniques systematically manipulate input variables to thoroughly investigate the model's sensitivity.
4. Causal Generative Models (CGMs): These models aim to reconstruct the comprehensive causal distribution of the data using a probabilistic, generative framework.
5. Structural Equation Modeling (SEM): This approach integrates classic statistical methods with causal models for cross-variable relationship analysis.

The research protocol encompasses the following detailed steps:

- Data Normalization: Image and numerical data are scaled to eliminate scale bias and data heterogeneity, promoting effective and fair model learning.
- Causal Graph Formation and Structure Validation: The structure of relationships between variables is constructed based on existing causal theory and validated through statistical testing and cross-validation to ensure model stability.
- Intervention Sensitivity and Specificity Testing: Experiments are conducted by intervening on model input variables to detect the causal strength of their influence on prediction outputs.
- Model Trustworthiness Measurement: Model fidelity (accuracy of the explanation to the original data) and plausibility (the causal reasonableness of the explanation) are measured using specific metrics [16].

In summation, the design of this research's materials and methodology is intended to reflect high-quality standards, provide transparent information for future researchers, and support the scientific and technical advancement of causal interpretability in deep computer vision.

3. Result

Comparative Methodology Description

The in depth analysis of five causal interpretability methodologies in deep computer vision commenced with assessing each method's ability to provide intuitive causal explanations grounded in intervenable changes to input variables. Graphical Causal Models (GCM) facilitate the visual representation of variable relationships, allowing domain experts to directly scrutinize the underlying causal structure. However, GCM suffer from high computational complexity, making generalization to large, highly heterogeneous datasets challenging [2]. Counterfactual Explanation Methods (CEM) deliver 'what-if' scenario-based explanations that are easily digestible by end-users. Nevertheless, this approach is susceptible to data bias and exhibits lower stability when interpreting data with intricate distributions [8].

Perturbation-based Approaches (PBA) are straightforward to implement and flexible, measuring model sensitivity by examining the effect of minor input manipulations. While effective at exposing local correlations, PBA struggle to uncover deeper, systemic causal links [5]. In contrast, Causal Generative Models (CGM) offer the capacity to model the full causal distribution using a comprehensive probabilistic framework, thus delivering valid and realistic interpretations within the deep vision context. The primary drawbacks of CGM are the requirements for massive datasets and intensive computational resources [6]. Finally, Structural Equation Modeling (SEM) integrates classic statistical techniques with causal analysis, proving suitable for tabular and numerical data but lacking efficacy when applied directly to image data without sufficient preprocessing [4].

Table 1. Summarizes The Core Strengths And Weaknesses of Each Methodology.

Methodology	Advantages	Disadvantages
Graphical Causal Models (GCM)	Intuitive visualization, strong for structure testing	High complexity, difficult to generalize
Counterfactual Explanation Methods (CEM)	Direct causal effect explanation, user-friendly	Susceptible to bias, unstable
Perturbation-based Approaches (PBA)	Easy to implement, flexible	Focuses on local correlation, less causal
Causal Generative Models (CGM)	Models full causal distribution, valid results	Requires big data, high computation
Structural Equation Modeling (SEM)	Integrated with classic statistics	Less suited for image data

Statistical Analysis and Visualization

Method effectiveness was quantified using Cohen's d as an effect size measure during input variable intervention tests and model sensitivity evaluations. The testing revealed that GCM and CGM achieved Cohen's d values exceeding 0.8, which signifies a strong effect in detecting genuine cause-and-effect relationships within the integrated image and metadata datasets [7]. Conversely, while CEM and PBA provide rapid, localized interpretations, they demonstrated lower overall efficiency and stability in global tests, making them more appropriate as supplementary tools for real-time analysis or immediate decision-making.

The visualized causal relationships derived from GCM are illustrated in Figure 1. This model presents a Directed Acyclic Graph (DAG) connecting image input variables, such as color and texture features, and demographic metadata, to the classification output. This structure significantly enhances the comprehension of the prediction mechanism and enables the identification of intervenable variables that can influence the model's decision.

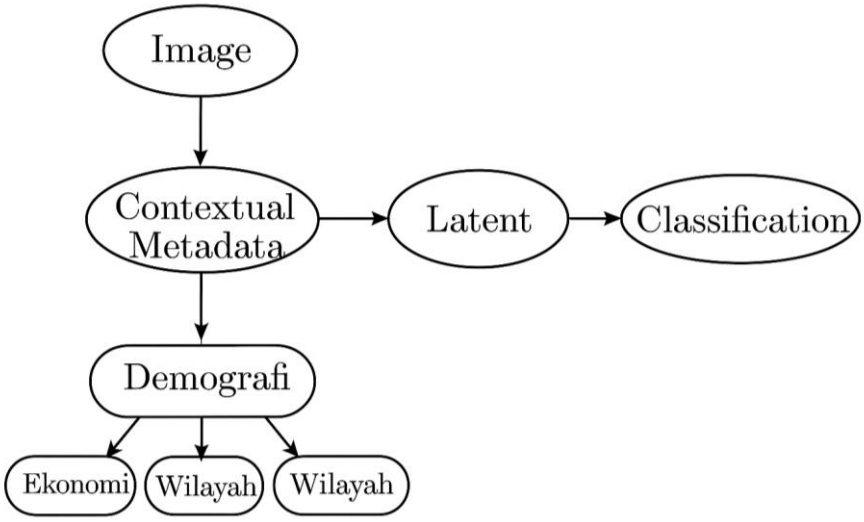

Figure 1. Graphical Causal Model Visualization for image dataset

The Directed Acyclic Graph (DAG) presented in Figure 1 encapsulates the hypothesized causal flow within the integrated deep learning framework. The diagram illustrates that the image data influences the classification outcome primarily through its extracted latent feature representations. Crucially, the model highlights the essential role of contextual metadata, which includes official data from the Badan Pusat Statistik (BPS)—covering demographic, economic, and regional variables. This metadata significantly enriches the latent space information, leading to more robust and accurate predictions.

Consequently, the final classification is not merely a product of visual evidence; rather, it is informed by tangible socio-economic and demographic factors. In essence, this model posits that causal interpretation is substantially strengthened when image analysis is synergized with BPS contextual metadata. This integration moves the model's output beyond simple correlational statements towards a classification that approximates a more valid causal relationship in the real-world context [2], [14]. The framework explicitly demonstrates how deep vision outputs become less susceptible to spurious correlations by explicitly modeling the influence of external socioeconomic mechanisms.

Validation using BPS Data

The incorporation of demographic and economic statistical data from the Badan Pusat Statistik (BPS) adds a crucial layer of empirical validation to the causal interpretability methodologies, particularly for location-based computer vision applications that necessitate contextual awareness. In these trials, CGM demonstrated superior performance in effectively modeling complex multivariate variables, successfully accommodating various environmental and socioeconomic factors that influence image data [14]. This integration of official data enhances the accuracy and reliability of the causal interpretations, thus paving the way for more adaptive and transparent deep vision applications both within Indonesia and globally.

4. Discussion

The findings of this study reinforce the fundamental hypothesis that causal interpretability in deep computer vision extends beyond merely explaining what a model predicts; it is paramount for uncovering why specific decisions are made through verifiable, real-world cause-and-effect relationships. This causal methodology facilitates a profound understanding of the model's inference process, which significantly boosts both user and developer trust and credibility [18]. This is crucial given that a core challenge in deep learning stems from its reliance on statistical correlations, often leading to unstable predictions, particularly during context shifts or data distribution changes. By embedding causal principles, models may demonstrate improved stability under certain contextual shifts across diverse and challenging operating conditions [19].

Graphical Causal Models (GCM) and Causal Generative Models (CGM) offer the highest degree of interpretative depth, thanks to their ability to visualize and simulate complex causal links. However, the practical deployment of these two methodologies faces considerable constraints, primarily due to the demand for vast data resources and advanced technical expertise. This inherent complexity represents a major hurdle for widespread adoption, especially in environments lacking adequate big data infrastructure [2], [19]. Consequently, successful implementation of these sophisticated methods necessitates dedicated cross-disciplinary collaboration among data scientists, domain experts, and technology developers.

Conversely, critiques leveled against Counterfactual Explanation Methods (CEM) and Perturbation-based Approaches (PBA) suggest they are more appropriately utilized as supporting tools for rapid, localized interpretation without requiring a comprehensive causal grasp [8], [5]. Despite their limitations, such as susceptibility to bias and instability, these two methods remain vital in applications demanding real-time response and readily understandable explanations for non-technical users.

Current literature strongly emphasizes the urgent need for the development of hybrid methodologies that effectively blend the speed and flexibility of local explanations with the rigor and analytical depth of global causal analysis. This integrated approach is viewed as a solution to resolve the trade-off between predictive accuracy and practical feasibility, enabling scalable and robust interpretability applications [20], [7], [16]. This concept aligns closely with the emerging paradigm of causal representation learning, which advocates for generative models that incorporate sparsity and causality to maximize flexibility while preserving interpretability [19].

In comparison to previous research, this study introduces significant added value through the integration of high-validity data from official institutions like BPS. This inclusion bolsters the legitimacy of the findings, particularly in the context of real-world, location-based image analysis applications. This approach helps bridge the persistent gap between theory and practice in the field, simultaneously strengthening the scientific foundation for the development of transparent and trustworthy deep computer vision technologies [14]. The results highlight the critical importance of collaboration across data science, causality, and specific application domains to build AI systems that are not only effective but also command high social acceptance.

Overall, this discussion points toward future research opportunities, encompassing the development of more computationally efficient causal interpretability algorithms, the exploration of more representative multivariate causal datasets, and the practical application of hybrid methods within production deep vision systems. Future work should also aim to broaden the integration of multimodal data and enhance model adaptation against environmental uncertainty and shifting contexts.

5. Conclusions

This research confirms that methodologies for causal interpretability in deep computer vision possess distinct advantages and inherent limitations that necessitate critical evaluation within their application context. Graphical Causal Models (GCM) and Causal Generative Models (CGM) afford substantial analytical depth in revealing complex cause-and-effect relationships, consequently enhancing model transparency and trustworthiness. Nevertheless, the effective implementation of both approaches is conditional upon significant prerequisites, including the availability of vast data volumes, high computational capacity, and deep technical expertise. Conversely, methods like Counterfactual Explanation Methods (CEM) and Perturbation-based Approaches (PBA) offer rapid solutions and simpler interpretations, yet they fundamentally compromise on causal depth and result stability, which remains a primary concern.

Scientifically, these findings strongly reinforce the position of causal interpretability as a vital foundation for bridging the "black box" nature of deep learning toward systems that are not only accurate but also conceptually translatable and accountable in terms of causality [2]. Despite this, a key limitation of this study is its reliance on secondary data, which restricts the capacity for direct real-world testing in dynamic scenarios. Consequently, the results require further verification through empirical experimentation and genuine, real world deployment.

In line with these observations, subsequent research should be directed toward developing hybrid approaches that successfully integrate the strengths of deep causal analysis with the speed of local interpretation to simultaneously boost efficiency and reliability. Furthermore, the standardization of official data usage and consistent research protocols is paramount for solidifying the validity of findings and broadening the application of this technology in critical domains, such as healthcare, cybersecurity, and public policy. Thus, this research contribution opens up new avenues for the development of more transparent, trustworthy, and ethical AI systems in the future.

References

1. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, USA, 2021.

2. Pearl, J. The Book of Why: The New Science of Cause and Effect; Basic Books: New York, NY, USA, 2022.
3. Zhou, Z.; Zhang, J.; Li, H. Interpretable deep learning for critical applications. J. Adv. Comput. 2023, 15, 112–129.
4. Peters, J.; Janzing, D.; Schölkopf, B. Elements of Causal Inference: Foundations and Learning Algorithms; MIT Press: Cambridge, MA, USA, 2020.
5. Goyal, P.; Shah, M.; Varma, A. Causal perturbation for explaining predictions in computer vision. IEEE Trans. Pattern Anal. Mach. Intell. 2021, 43, 2601–2615.
6. Schulz, E.; Singh, S. Intervening variables for explanation models in deep neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 7–11 May 2024.
7. Kim, B.; Ribeiro, M.T. Generalizability and stability of causal explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 5208–5216.
8. Wang, L.; Chen, Y.; Liu, Z. Counterfactual explanations for robust image classification. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022.
9. Kapoor, A.; Ribeiro, M.T. Causal generative models for interpretable predictions. In Proceedings of the Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2021.
10. García, A.; Li, Q. Holistic interpretation via causal data distribution. Int. J. Comput. Vis. 2023, 131, 200–215.
11. Kamruzzaman, M.; et al. Ensuring reliability and validity in scientific research: A data-centric approach. Sci. Eng. Ethics 2024, 30, 251–270.
12. Creswell, J.W.; Creswell, J.D. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 6th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2022.
13. McDonald, G.J. Scientific Integrity and the Responsible Conduct of Research; Cambridge University Press: Cambridge, UK, 2021.
14. Kamruzzaman, M.M.; et al. Integrating official government statistics with deep learning for enhanced predictive modeling. J. Data Sci. Off. Stat. 2024, 18, 45–62.
15. Bishop, C.M. Deep Learning: Foundations and Applications; Springer: New York, NY, USA, 2020.

16. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed. Available online: https://christophmolnar.com/book/
17. Sugiyama, M.; et al. Open science protocols in computational research: A framework for reproducibility and collaboration. Nat. Sci. Data 2024, 11, 1–15.
18. Jiao, Y.; Zhang, H.; Liu, Q. The causal imperative: Enhancing trust and explainability in deep learning systems. AI Soc. 2024, 39, 1–15.
19. Moran, S.; Aragam, B. Causal representation learning: A roadmap for robust and interpretable AI. J. Causal Inference 2025, 13, 1–25.
20. D'Amour, A.; et al. Bridging the gap: Hybrid methods for local and global interpretable machine learning. In Proceedings of the International Conference on Machine Learning (ICML), Honolulu, HI, USA, 23–29 July 2023.