INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
B. Arslan, Vol.15, No.1, pp.1-24.
https://doi.org/10.55859/ijiss.1826477

Causal Image Processing: A Mechanism-Based Framework for Secure Visual Intelligence

Bilgehan Arslan[1]

[1]Department of Computer Engineering, Gazi University, Ankara, Turkey
Corresponding Author: bilgehanarslan@gazi.edu.tr

Research Paper Received: 19.11.2025 Revised: 29.12.2025 Accepted: 09.01.2026

Abstract—This study introduces Causal Image Processing (CIP), a mechanism based conceptual framework designed to address fundamental weaknesses of contemporary deep learning systems in security critical visual intelligence. Unlike conventional convolutional networks, vision transformers, or diffusion models that rely on observational correlations, CIP models the image formation process through four independent causal mechanisms: content, domain conditions, sensor characteristics, and identity factors. These mechanisms are formalized within a directed acyclic graph that provides a structured representation separating the physical and semantic processes underlying image generation. CIP integrates three forms of inference within a unified architecture: predictive inference, interventional inference, and counterfactual inference. Mechanism faithful representations enforce invariance to sensor and domain variations, while counterfactual reasoning enables principled evaluation of identity consistency under hypothetical acquisition conditions. The framework also introduces learning principles based on mechanism fidelity, sparse intervention sensitivity, and causal invariance, and defines evaluation criteria centered on counterfactual consistency and tamper resistance. Overall, CIP offers a theoretical foundation for developing robust, explainable, and tamper resistant vision systems capable of operating reliably across heterogeneous sensors, environments, and acquisition conditions. The framework establishes a mechanism centered alternative to correlation driven deep learning pipelines and provides a pathway for next generation secure visual intelligence.

Keywords—causal image processing, causal representation learning, counterfactual inference, mechanism-faithful modeling, secure visual intelligence, sensor and domain robustness

1. Introduction

Artificial intelligence and deep learning methods play a decisive role in many security-critical applications today. These applications include biometric authentication, fraud detection, intrusion detection, and forensic image analysis. Although CNN- and Vision Transformer–based systems achieve high accuracy rates, an increasing number of studies show that these models rely not on the mechanisms that generate images, but on superficial and distribution-specific correlations. This creates a direct risk in the context of information security. Correlation-based models are structurally vulnerable to distribution shifts, spoofing attempts, attacks, and adversarial manipulations.

Training deep learning systems under the i.i.d.

assumption is especially problematic in security contexts. When models encounter sensor variation, differences in acquisition protocols, environmental conditions, or attacker-generated manipulated samples, their performance drops sharply. These failures show that the i.i.d. assumption does not hold in real-world security scenarios. The "shortcut learning" phenomenon demonstrated by Geirhos et al. reveals that deep networks often gravitate toward statistical cues that are irrelevant to the task but easily exploitable by an attacker [1]. This leads to serious vulnerabilities in biometric systems against fake fingerprint generation, adversarial noise, domain-shift-based attacks, and presentation attacks.

CIP interprets the ImageNet degradation results of Recht et al. [2] as direct evidence that correlation-based representations lack mechanism fidelity, even under minimal distribution changes. Such fragility poses a direct risk in security-related domains such as biometric systems, forensic image analysis, deepfake detection, and surveillance cameras.

Vision Transformer models are also not immune to correlation dependency. From a CIP standpoint, the findings of Raghu et al. [3] show that even globally attentive architectures fail to separate mechanism-level factors, reinforcing the need for explicit causal decomposition. As a result, small manipulations introduced by an attacker can produce unexpected shifts in the representation space. This means that ViT-based models are structurally fragile from a security perspective.

The failure of deep models to internalize the mechanisms that generate images offers a fundamental explanation for why deep learning-based security systems often break down under real-world conditions. The causal representation learning perspective introduced by Schölkopf and colleagues highlights the necessity of a framework that models not only observational correlations but also independent causal mechanisms, intervention structure, and counterfactual relations [4].

This work introduces a new conceptual framework aimed at addressing this structural deficiency in security-focused image processing systems: Causal Image Processing (CIP). CIP treats an image not merely as a data object but as an observation of physical and biological processes that can be manipulated by an attacker or influenced by external conditions. This approach seeks to establish a mechanism-centric foundation for biometric security, digital forensics, sensor-spoofing detection, adversarial robustness, and trustworthy artificial intelligence.

Overall, this study presents a novel theoretical paradigm for rethinking image processing in a security context. By integrating causal representation learning, intervention-based reasoning, counterfactual analysis, and invariance principles, CIP aims to develop a mechanism-based understanding of visual intelligence that can be employed in safety-critical applications.

Modern deep learning models predominantly learn statistical correlations rather than the physical or biological mechanisms that generate visual observations. This correlation-driven paradigm makes such models fragile under sensor and domain shifts, shortcut learning, and adversarial manipulations in security-critical applications. To address this root limitation, this work introduces Causal Image Processing (CIP), a mechanism-based framework that models image formation through a structured causal DAG, learns mechanism-faithful representations, and supports predictive, interventional, and counterfactual inference. CIP provides a mechanism-centric foundation for robust and explainable security-focused vision systems across heterogeneous sensors and acquisition conditions.

Importantly, the encoders and mechanism vari-

ables in CIP are presented as conceptual abstractions that operationalize the framework, rather than guarantees of perfect causal factor recovery from pixels. The feasibility and limitations of reliably extracting or disentangling these mechanisms from observational data are discussed explicitly in the later "Challenges and Open Problems of CIP" section.

2. Causal Motivation: Why Classical and Modern Approaches Fail Mechanistically

When examining the historical development of the image processing field, it becomes clear that, despite significant changes in methods, a common shortcoming has persisted throughout its evolution. At no point have the physical, biological, or sensor-based mechanisms that generate images been directly represented. This gap reflects the shortcut learning behavior described by Geirhos et al. [1] and the distribution shift vulnerability demonstrated by Recht et al. [2]. In security-focused domains such as biometric authentication, forensic analysis, and fraud detection, the lack of mechanism fidelity in correlation-based representations expands the attack surface and reduces model stability. The foundations of this historical approach can also be seen in Marr's theory of visual processing. Marr's hierarchical computational model does not internalize the mechanisms that produce images; therefore, it inherently contains the mechanism-fidelity deficiency that CIP later identifies [5].

Classical image processing methods attempt to approximate physical processes using analytical filters and deterministic transformations. However, their fundamental weakness from a security perspective lies in their sensitivity to small and controlled manipulations of the input. Simple changes such as modifying illumination conditions, altering sensor optics, adding surface reflections, or varying the contact pattern in biometric images can directly affect the output. Because the true mechanism of image formation is not represented within the computational model, these variations create a predictable attack surface for an adversary.

With the introduction of hand-crafted features in the machine learning era, representational capacity expanded, but the causal gap remained. Features such as LBP, SIFT, SURF, and minutiae-based descriptors capture statistical patterns related to the appearance of the image. However, they contain no inherent information about the mechanisms by which the image was produced. As a result, these features remain vulnerable to both physical and digital forgery attacks. The model's decision often changes unpredictably when the appearance of the image is altered. Examples include changes in the pressure distribution on a fingerprint, the lighting direction in a facial photograph, or the atmospheric conditions in remote sensing data. As a result, the critical principle of preserving the "essence" of the image in security contexts is not fulfilled.

Although performance increased dramatically in the deep learning era, correlation dependency became structurally stronger. CNN and Vision Transformer architectures are capable of learning rich representations from high-dimensional data. However, most of these representations reflect frequent superficial patterns in the dataset rather than the causal factors that generate the image. This leads to two major consequences for security. First, models become highly sensitive to small perturbations introduced by attackers and can be easily deceived by adversarial examples. Second, when sensor-dependent or environment-dependent noise cues dominate the learned representation, an attacker can bypass authentication systems by simulating these

cues. Therefore, high accuracy does not imply a mechanism-based representation, and the robustness of such models in security-critical settings remains limited.

When this historical trajectory is considered as a whole, it becomes evident that the fundamental shortcoming in the field of image processing has critical implications not only for performance-oriented machine learning techniques but also for security. The inability to represent the mechanisms by which images are formed leaves models vulnerable to attacks and unstable behavior in real-world security scenarios such as distribution shift, spoofing, sensor migration, and environmental changes. For this reason, the motivation for Causal Image Processing (CIP) is not merely an academic concern about representation. Instead, it reflects a structural necessity for information security.

2.1. Classical Image Processing: Physical Assumptions and Security Weaknesses

Classical image processing methods relied on deterministic rules such as illumination models, edge operators, and analytic filtering. These methods made intuitive assumptions about how the image was formed, but they did not incorporate these mechanisms into the model's internal representation. As a result, small manipulations of the image could directly affect the output and were easily exploited. Examples include changes in light angle, reflections, variations in sensor optics, and differences in fingerprint contact pressure. From a security perspective, this period was the most vulnerable to image manipulation because the system processed only the "effect" and did not represent the "cause." This corresponds to an early form of the behavior that modern literature refers to as spurious correlation and shortcut learning [1], [6].

2.2. The Machine Learning Era: Learning Correlations and Expanding the Attack Surface

With the advent of machine learning, it became possible to learn highly discriminative statistical features from images. However, these features did not represent the physical or biological processes that produced the image. They merely encoded frequently recurring patterns in the data. Consequently, models tended to learn random correlations specific to the dataset rather than the underlying structure of the class. Within the CIP perspective, these observations indicate that appearance-level correlations lack mechanism fidelity, which clarifies why domain generalization studies [7], [8] consistently report failures when such correlations are transferred across environments.

Predictive features such as minutiae density in biometric systems, specific shadow patterns in facial recognition, and regional spectral noise in remote sensing could be easily spoofed by attackers. This era made the attacker's job easier because the system relied on superficial statistics rather than deep mechanisms within the image. Although Distributionally Robust Optimization (DRO) studies [9] proposed regularization-based solutions to distribution shifts, these methods cannot achieve the level of structural generalization targeted by CIP unless mechanism fidelity is ensured.

2.3. The Deep Learning Era: Performance Enhancement, Causal Blindness, and Adversarial Vulnerability

Deep learning radically increased image processing performance, but this success relied heavily on stronger detection of correlations. CNNs became dependent on local patterns, while Vision Transformers relied on global patch statistics. This structure caused sensor-dependent artifacts to seep

into the representation, inconsistent decisions when the environment changed, and subtle adversarial perturbations to mislead the model.

The high parametric capacity of deep models allowed them to memorize data instead of learning the true generative factors. This is directly related to the "deep memorization" phenomenon demonstrated by Zhang and colleagues [10]. Raghu and colleagues also showed that Vision Transformer models cannot decompose underlying mechanisms [3].

This is dangerous from a security perspective because representations that depend on superficial factors make the model's behavior predictable and susceptible to exploitation. These superficial factors are easy for an attacker to manipulate. In addition, the CAUSAL3D benchmark study shows that modern models produce inconsistent and inexplicable outputs when exposed to physical manipulations [11]. Recent studies have explored the integration of causal reasoning into deep visual models, particularly within CNN- and ViT-based architectures. For instance, Li et al. [12] proposed Causal-ViT, which enhances robustness by introducing causal interventions within the attention mechanism to mitigate spurious correlations. Similarly, Xie et al. [13] presented ViT-CX, focusing on causal explanation by attributing model predictions to causal visual concepts. While these approaches leverage causal principles to improve robustness or interpretability, they primarily operate at the feature or attention level and do not explicitly model independent data-generating mechanisms such as content, domain, and sensor processes. In contrast, the proposed CIP framework formulates causal reasoning at the mechanism level, enabling unified predictive, interventional, and counterfactual inference over disentangled causal factors.

2.4. The Common Missing Piece: The Lack of a Causal Model of Image Formation

Across all historical periods, a common gap becomes evident. None of these eras represented the physical, biological, or sensor-based cause-and-effect mechanisms that generate an image. As a result, the model learns only the "effect," ignores the "cause," and behaves very inconsistently in security contexts.

Schölkopf and colleagues identified this gap as a mismatch between learned representations and the true generative factors. They demonstrated that systematic generalization is impossible unless causal mechanisms are modeled [4]. Zhang's sheep–grass example [6] provides a practical illustration of this missing mechanism.

This gap also explains why modern systems are vulnerable to distribution shifts, spoofing attacks, sensor variation, adversarial examples, and manipulated images.

2.5. Towards CIP: The Necessity of Mechanism-Integrity for Security

A security-robust image processing system should achieve high accuracy while also producing representations that align with the underlying "cause." Causal Image Processing (CIP) is a conceptual framework developed with this motivation. The goal of CIP is to model the mechanisms underlying changes in the image rather than the correlations present in the data.

This approach draws on Schölkopf's principle of independent causal mechanisms [4], the intervention-based assessments of the CAUSAL3D benchmark [11], and TraCE's counterfactual consistency analyses [14]. Together, these insights outline key security objectives:

- generating sensor-invariant representations,
- improving resistance to spoofing and adversarial attacks,
- increasing stability under domain changes,
- enhancing intervention-based and counterfactual explainability,
- ensuring mechanism-based reliability in critical security applications.

Therefore, CIP is not merely a new technical approach. It represents a necessary direction for the future of information security.

3. Learning in Vision: Security-Focused Theoretical Foundations

One of the fundamental problems facing image processing systems in the security context is the excessive reliance of models on observational statistics and their inability to internally represent the mechanisms that generate the image. Findings such as shortcut learning [1] and distribution shift vulnerability [2] demonstrate that deep models substitute appearance for mechanism. In security-critical tasks such as biometric verification, fraud detection, forensic analysis, sensor spoofing, and adversarial defense, systems that rely on superficial cues are inherently vulnerable and unstable. To address this shortcoming, causal learning places the image-generating process and the mechanisms involved in this process at the core of the model. The CIP theoretical framework provides an integrated approach that spans causal representations, intervention logic, counterfactual analysis, and invariance principles. Schölkopf's causality approach to machine learning [15] explains conceptually that this vulnerability originates from a lack of mechanism representation.

Recent causal representation learning and causal generative modeling studies provide complementary tools for disentangling latent factors and constructing counterfactual samples. Representative examples include CausalGAN [16], CausalVAE [17], and flow-based causal models such as Causal Autoregressive Flows [18]. CIP does not aim to replace these model families; instead, it provides a higher-level mechanism-based framework that can incorporate such generative components to realize intervention-consistent and counterfactual image formation in security-critical settings.

3.1. Causal Representations: Learning the Generating Mechanism of an Image

Causal representation learning aims to represent not only the appearance of an image but also the independent mechanisms that generate it. Schölkopf's principle of independent causal mechanisms [4] argues that an image should be understood as a combination of physical, biological, and environmental processes. This distinction is critical for security. The appearance of an image may change, but the model's decision is expected to remain consistent as long as the generating mechanism remains unchanged. Traditional models, however, suffer from the "spurious correlation" problem introduced by Zhang et al. [6], where frequently occurring—but mechanism-irrelevant—appearance cues dominate the representation space. CIP's causal-representation-based approach improves stability, particularly in challenging environments such as cross-sensor biometrics [19] and multi-source image analysis.

3.2. Intervention Logic: Causal Modeling of Attacks and Changes

Pearl's intervention logic [20] is highly important from a security perspective because many attacks involve forcibly modifying a variable from the outside. Artificial fingerprint materials used in presentation attacks, adversarial perturbations applied

in facial recognition, and manipulations of sensor optics are all examples of interventions. In the CIP perspective, these are formalized through "do-operators." Traditional models interpret such interventions as observational variations, which leads to instability under attack. CIP interprets the failures reported in CAUSAL3D [11] as evidence that models lacking mechanism-level structure cannot maintain consistency under controlled interventions. CIP explicitly represents interventions within an internal causal structure, making the impact of attacks more traceable and manageable.

3.3. Counterfactual Reasoning: A Fundamental Dimension for Explainability and Reliability

Counterfactual analysis examines how the behavior of a model would change if a specific component of an image were modified. This method is not only a tool for explainability but also a critical testing mechanism for security. From a CIP perspective, the counterfactual inconsistencies reported by TraCE [14] and SharpShooter [21] confirm that deep models lack mechanism-faithful structure. This raises concerns about the reliability of decisions in image forensics and biometric systems. CIP enables the generation and evaluation of counterfactual scenarios in a mechanism-consistent manner, allowing security analysts to assess not only how a model behaves but also why it behaves the way it does.

3.4. Invariance Principles: Mechanism-Consistency for Domain Security

The ability of image processing models to make consistent decisions across different sensors, environments, and quality conditions is a fundamental requirement in security-critical applications. However, invariant learning (IRM) [7] and domain generalization studies [8] demonstrate that appearance-oriented models are inherently vulnerable to such changes. The invariance principle proposed by CIP is grounded in mechanism consistency rather than appearance consistency. Even if appearance-based factors in the image change, the representation should remain stable as long as the generating mechanism remains the same. Large performance drops during sensor transitions [19] and domain-shift issues in remote sensing [22] clearly demonstrate the importance of this principle. CIP aims to increase stability in security systems through mechanism-based invariance.

4. Framework: Proposed Conceptual Paradigm

A fundamental weakness of security-focused image processing systems is that they treat images solely as observational data objects. However, in fields such as biometric verification, fraud analysis, sensor-spoofing detection, and forensic investigation, an image is the output of physical, biological, and device-based mechanisms. Causal Image Processing (CIP) is a new paradigm that aims to model images within the framework of these generative mechanisms. By incorporating representation, inference, and learning processes within a causal structure, CIP seeks to provide stability, explainability, and resistance to attacks in security-critical applications. This section introduces the proposed CIP architecture as an integrated framework that spans causal graphs of image formation, counterfactual inference, and mechanism fidelity.

4.1. Causal DAG for Image Formation: Content–Domain–Label–Sensor Mechanisms

In classical image processing, a two-dimensional observed image can be represented as a spatial function

Algorithm 1 CIP: Mechanism-Based Visual Reasoning Framework

Require: Observed image X, Image formation function f (content–domain–sensor), Task mapping function g (identity/class)

Ensure: Y_{pred} (Predictive decision), Y_{do} (Interventional response), Y_{cf} (Counterfactual prediction)

Step 1: Causal Factor Extraction
$X_c \leftarrow \text{Encoder}_{content}(X)$
$X_d \leftarrow \text{Encoder}_{domain}(X)$
$X_s \leftarrow \text{Encoder}_{sensor}(X)$
$X_l \leftarrow \text{Encoder}_{identity}(X)$
Step 2: Mechanism-Based Disentanglement
$Z_c \leftarrow \text{Disentangle}(X_c)$
$Z_d \leftarrow \text{Disentangle}(X_d)$
$Z_s \leftarrow \text{Disentangle}(X_s)$
$Z_l \leftarrow \text{Disentangle}(X_l)$
$R \leftarrow \text{Combine}(Z_c, Z_d, Z_s, Z_l)$
Step 3: Predictive Inference
$Y_{pred} \leftarrow g(Z_c, Z_l)$ ▷ Compute prediction using observational representation
Step 4: Interventional Inference (do-operator)
$Z_s^* \leftarrow \text{Intervention}(Z_s)$
$Y_{do} \leftarrow g(Z_c, Z_l \mid \text{do}(Z_s = Z_s^*))$
Step 5: Counterfactual Inference
$Z_s' \leftarrow \text{Counterfactual_Modify}(Z_s)$
$X_{cf} \leftarrow f(Z_c, Z_d, Z_s')$
$Y_{cf} \leftarrow g(Z_c, Z_l)$
Return Y_{pred}, Y_{do}, Y_{cf}

$$X(u,v) = \mathcal{F}\big(X_c(u,v), X_d(u,v), X_s(u,v)\big) + \epsilon(u,v) \tag{1}$$

where (u, v) denote spatial coordinates. Here, X_c represents the intrinsic content-related structure of the scene, X_d models domain-dependent effects such as illumination variations, atmospheric conditions, and surface interactions, X_s captures sensor-specific characteristics including optics, device artifacts, and acquisition noise, and $\epsilon(u, v)$ denotes additive or structured noise. In Equation (1), the term $\epsilon(u, v)$ represents stochastic perturbations affecting the observed image. Importantly, CIP does not treat all additive noise as a homogeneous phenomenon. Instead, additive noise is interpreted according to its causal origin. When noise arises naturally from the acquisition process—such as thermal noise, electronic readout noise, or sensor-specific Gaussian noise—it is considered an inherent part of the sensor mechanism X_s. In contrast, when additive noise is deliberately injected into the image by an external agent (e.g., adversarial noise injection or malicious signal corruption), it is modeled as an intervention on the sensor mechanism, formalized as $do(X_s)$.

This distinction is critical from a security perspective. While both phenomena may be additive at the pixel level, their causal interpretation within CIP differs fundamentally. Natural sensor noise reflects a stable stochastic component of the acquisition process, whereas injected noise corresponds to an external manipulation that violates the assumed sensor mechanism. Unlike classical formulations that treat these effects as inseparable nuisance factors, CIP explicitly models each component as an independent causal mechanism.

The fundamental assumption of CIP is that an image is produced not by a single source but by four main mechanisms. These mechanisms are content, domain conditions, label or identity factors, and sensor characteristics. Although these mechanisms are independent, they interact with one another. Their interrelationships are represented by a Directed Acyclic Graph (DAG) that aligns with Schölkopf's principle of independent causal mechanisms [4], Kocaoglu's causal generative modeling approach [16], and causal intervention studies in visual reasoning [23].

Let us define:

- X: observed image
- X_c, X_d, X_s, X_l: independent causal mechanisms (content, domain, sensor, label)
- f: physical–sensor image formation function
- g: task mapping (identity/class)

CIP is formally defined as the tuple:

$$\mathcal{M}_{CIP} = (\mathcal{G}, f, g) \tag{2}$$

where $\mathcal{G}$ is a DAG over $\{X_c, X_d, X_s, X_l, X, Y\}$ satisfying

$$X = f(X_c, X_d, X_s), \quad Y = g(X_c, X_l) \tag{3}$$

This formal structure distinguishes CIP from conventional deep learning, where no explicit factorization of mechanisms exists.

Let the generative process of an image be defined by:

- X_c: content mechanism
- X_d: domain mechanism
- X_s: sensor mechanism
- X_l: identity mechanism

The image is generated via:

$$X = f(X_c, X_d, X_s), \quad Y = g(X_c, X_l) \tag{4}$$

Causal dependencies:

$$X_c \to X, \quad X_d \to X, \quad X_s \to X, \quad X_c \to X_l \to Y \tag{5}$$

CIP assumes independent causal mechanisms, giving:

$$\begin{aligned} p(X, X_c, X_d, X_s, X_l, Y) = {} & p(X_c) \times p(X_d) \times p(X_s) \\ & \times p(X_l \mid X_c) \\ & \times p(X \mid X_c, X_d, X_s) \\ & \times p(Y \mid X_c, X_l) \end{aligned} \tag{6}$$

Equation (6) provides the unique and complete causal factorization implied by the CIP DAG and serves as the master probabilistic formulation of the framework. This DAG abstracts the causal structure of an image in the following way:

- Content determines the structural properties of the scene or biometric sample.
- Domain conditions determine external factors such as ambient lighting, surface interaction, or atmospheric effects.
- The sensor mechanism generates device-based transformations and artifacts.
- The label factor carries the identity or class properties of the content.

Proposition 1 (Mechanism-Faithful Stability) *Let the representation R satisfy*

$$R = h(X_c, X_d, X_s, X_l) \tag{7}$$

If h is mechanism-faithful such that

$$\frac{\partial R}{\partial X_s} = 0, \quad \frac{\partial R}{\partial X_d} = 0 \tag{8}$$

for identity-preserving transformations, then the model remains stable under sensor changes and domain shifts. If a model depends only on content and identity mechanisms, then environmental or sensor variations cannot alter its decision—a requirement CIP explicitly enforces.

Both instances of R denote the same unified mechanism-based representation.

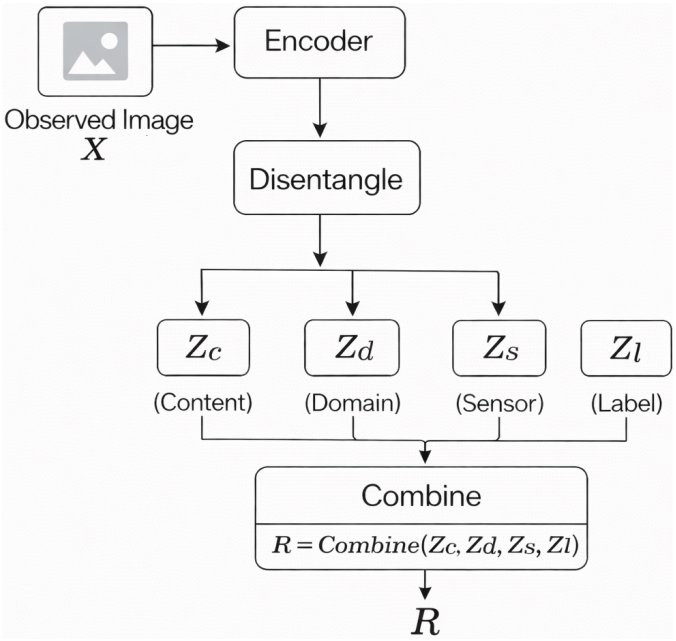


Figure 1. CIP System Diagram showing the decomposition of observed image into four independent causal mechanisms and the formation of unified representation.

Most deep models mix these four mechanisms together. For example, they may correlate sensor artifacts with identity information, which results in spurious relationships. By explicitly separating these mechanisms on the DAG, CIP enables the detection of which mechanism is being manipulated during attacks in security-critical tasks. In the CIP framework, spoofing materials affect the sensor mechanism, adversarial perturbations alter the interaction between content and sensor, and domain shifts modify the environmental component.

Figure 1 illustrates the causal mechanism decomposition and representation formation in the proposed Causal Image Processing (CIP) framework. The observed image is decomposed into four independent causal mechanisms—content, domain, sensor, and label—representing the physical and semantic factors that give rise to the visual signal. These mechanisms are disentangled in the representation layer, producing mechanism-faithful latent variables (Z_c, Z_d, Z_s, Z_l) that isolate identity from sensor- or environment-specific variations. The combined representation R serves as the unified mechanism-based embedding from which all subsequent inferences are derived. The three inference routes supported by CIP (predictive, interventional, and counterfactual) are illustrated in Figure 2.

4.1.1 Generality of CIP's Interventional Framework

Sensor-level interventions are used throughout the paper as illustrative examples due to their intuitive interpretation in image acquisition scenarios. However, the proposed CIP framework is not restricted to sensor-level interventions. In general, interventions can be applied to any causal mechanism in the image formation process. Domain-level interventions can be formulated as $do(X_d)$, capturing changes in environmental or acquisition conditions such as illumination or atmospheric effects, while noise-level interventions correspond to $do(X_s)$, modeling variations in stochastic perturbations. Moreover, simultaneous or multi-factor interventions, such as $do(X_d, X_s)$, are naturally supported within the same framework. Counterfactual inference follows an analogous formulation, enabling queries such as "what would the image or prediction have been if the noise level were different while the scene content and domain conditions remained unchanged," which is implemented by fixing Z_c, Z_d, Z_l and varying Z_s accordingly.

Figure 2 illustrates the inference layer of the proposed Causal Image Processing (CIP) framework. Building upon the unified representation R derived through the mechanism decomposition shown in

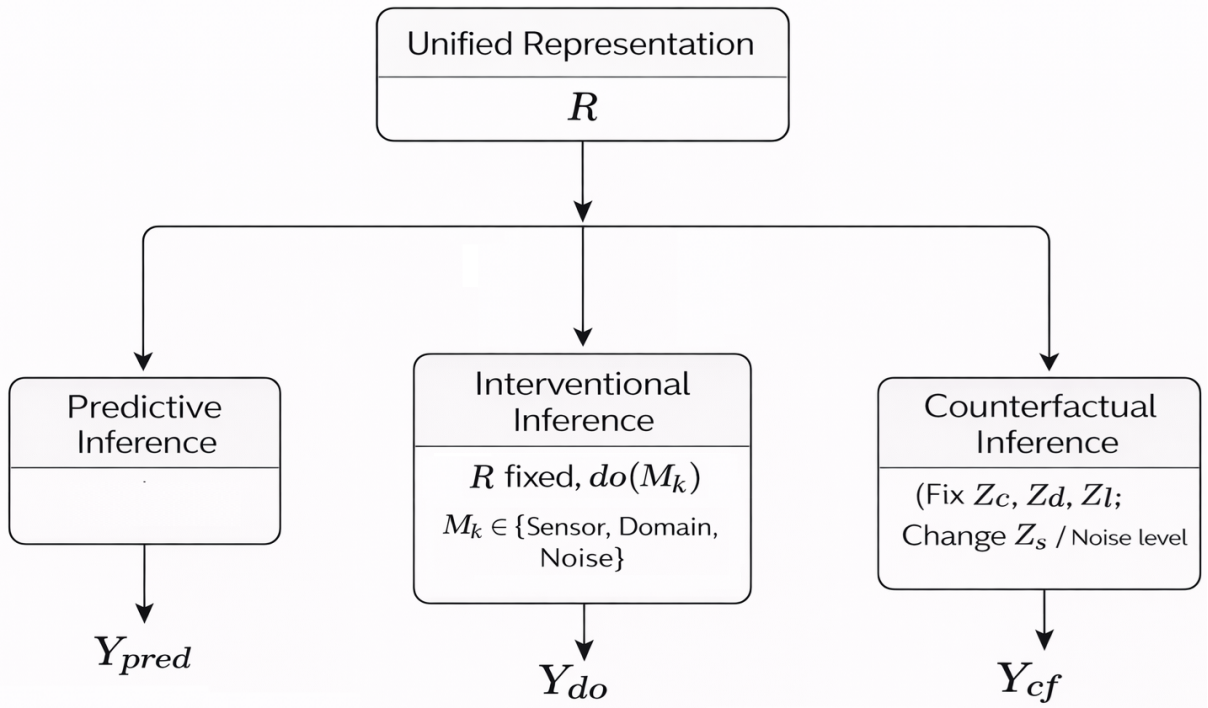


Figure 2. CIP Inference Pipeline showing three unified reasoning routes enabled by the mechanism-based representation.

Figure 1, this diagram shows how CIP enables three distinct inference routes: predictive, interventional, and counterfactual reasoning. Each route produces specific outputs (Y_{pred}, Y_{do}, Y_{cf}) that enable mechanism-aware decision analysis, attack behavior evaluation, and cross-sensor stability assessment. The interventional inference can target any causal mechanism (sensor, domain, or noise-level perturbations), while counterfactual inference evaluates identity consistency under hypothetical acquisition conditions.

4.2. Representation Layer: Separation of Causal Factors

The representation layer of CIP aims to separate causal factors instead of relying on observational features in the image. This approach addresses the problem known as observational entanglement, where modern deep models mix content, sensor, environment, and identity into a single representation. Although studies on causal generative models such as CausalVAE [17], Causal Autoregressive Flows [18], and CausalGAN [16] have demonstrated that causal factor separation is feasible in generative models, this capability has not yet been fully adopted in safety-critical applications.

Causal Autoregressive Flows and CausalGAN represent early evidence that causal factors can be explicitly separated during the generative process. Kocaoglu et al.'s CausalGAN model further shows that this type of separation ensures mechanism fidelity and is conceptually compatible with the representation layer of CIP.

CIP promotes three principles in the representation layer:

1 **Mechanism fidelity**: The representation should reflect the true generating mechanism rather than the superficial appearance of the image.
2 **Sensor independence**: Identity, content, and task-related factors should be separated from sensor-dependent features.
3 **Domain disentanglement**: Environmental and sensor-induced variations should not alter the identity factor.

This type of representation increases stability in cross-sensor matching and provides a clear explanation of which factor an attack has modified during adversarial perturbations.

4.3. Inference Layer: Prediction, Intervention, and Counterfactual Inference

CIP transforms not only the representation but also the inference process by redefining it within a causal framework. Traditional models perform only predictive inference, whereas CIP supports three distinct types of inference:

- Mechanism-Level Interpretation of Additive Noise and Attacks: CIP does not aim to explicitly classify additive noise as benign or malicious at the pixel level. Instead, it evaluates mechanism consistency within the latent causal representations. Additive noise is interpreted according to its causal origin: stochastic noise arising from the acquisition process is treated as part of the sensor mechanism X_s, whereas externally injected noise corresponds to an intervention on the sensor mechanism, formalized as $do(X_s)$. Consequently, CIP does not "detect additive-noise attacks" through ad-hoc classifiers; rather, it localizes intervention-induced inconsistencies in the sensor-related causal variables. This distinction allows CIP to differentiate benign sensor noise from adversarial noise injection, even when both appear additive at the pixel level.
- Predictive Inference: The model predicts the class or identity based on observational data. This prediction is performed only on representations that comply with the causal structure defined in the DAG.
- Interventional Inference: Based on Pearl's do-operator [20], this form of inference evaluates how the model behaves when specific variables are externally modified. Attack types naturally fit into the intervention framework. For example, one can ask: Does the identity prediction change when sensor noise is increased? For example, mild Gaussian noise that statistically matches the sensor's natural noise profile is treated as part of X_s, and does not constitute an attack. However, Gaussian noise with abnormal variance, frequency characteristics, or spatial structure—when inconsistent with the sensor's generative mechanism—is interpreted as an intervention $do(X_s)$. This causal distinction allows CIP to differentiate between benign acquisition noise and adversarial noise injection, even when both appear additive at the pixel level.
- Counterfactual Inference: CIP revisits the counterfactual inconsistency problems highlighted by TraCE [14] and SharpShooter [21] and applies them directly to security-critical tasks. For example, CIP can address questions such as: Would the model make the same decision if this image belonged to the same person but had been captured with a different sensor? Counterfactual inference is particularly powerful for forensic verification and deepfake analysis.

Adversarial attacks, spoofing attempts, and sensor manipulations are naturally expressed as interventions on X_s. Within the CIP framework, undesired

conditions such as sensor tampering, artificial noise injection, or domain manipulation are not detected through ad-hoc classifiers, but through mechanism inconsistency. Specifically, attacks manifest as abnormal variations in mechanism-specific latent representations.

For instance, artificial noise injected into an image primarily affects the sensor mechanism X_s. Statistical deviations in Z_s, such as abnormal variance, frequency-domain inconsistencies, or entropy shifts, can indicate sensor-level interventions. Similarly, images acquired from heterogeneous sensors, such as optical versus infrared devices, introduce systematic changes in X_s while preserving X_c and X_l, allowing CIP to localize the source of variation without explicit sensor labels.

Therefore, CIP enables principled localization of intervention-induced mechanism inconsistencies by identifying which causal mechanism exhibits anomalous behavior, rather than relying on handcrafted attack detectors.

Formally, the image formation process in CIP is expressed as

$$X = f(X_c, X_d, X_s) \tag{9}$$

A counterfactual image corresponding to a modified sensor mechanism X_s' is then defined as

$$X_{cf} = f(X_c, X_d, X_s') \tag{10}$$

The corresponding counterfactual prediction is:

$$Y_{cf} = g(X_c, X_l) \tag{11}$$

It is important to note that Y_{cf} does not explicitly depend on X_{cf}. This is intentional and reflects the core assumption of CIP: counterfactual reasoning is performed at the mechanism level rather than at the pixel level.

Although X_{cf} represents a hypothetical image generated under a modified sensor mechanism, the decision function $g(\cdot)$ operates only on the identity-preserving causal variables (X_c, X_l). As long as these mechanisms remain unchanged, the prediction must remain invariant. This formulation provides a formal test for identity-preserving counterfactual consistency, which is a defining property of CIP.

A model is CIP-consistent if:

$$Y_{cf} = Y \quad \text{whenever } X_c \text{ and } X_l \text{ remain unchanged} \tag{12}$$

This provides the mathematical test for identity-preserving stability under hypothetical sensor changes. The integration of all three inference types within a single architecture is one of the most defining characteristics that distinguishes CIP from classical deep learning models.

4.3.1 Illustrative Architectural Instantiations for Counterfactual Generation

While CIP defines counterfactual reasoning at the mechanism level, the counterfactual image generator can be instantiated using different generative backbones. For example, a VAE-based design may encode (X_c, X_d, X_s) into structured latent variables and decode counterfactual samples by substituting the sensor-related latent component $(X_s \rightarrow X_s')$ while preserving identity-related mechanisms. Alternatively, a GAN-based design (e.g., causal GAN variants) can be used to generate X_{cf} by conditioning on mechanism-specific codes, enabling realistic counterfactual rendering under modified acquisition conditions. Flow-based architectures provide another option by learning invertible map-

pings that allow controlled latent manipulations and likelihood-consistent counterfactual sampling. Importantly, CIP is intentionally agnostic to the specific generative backbone; the central requirement is that counterfactual generation respects the causal DAG and preserves identity mechanisms when only nuisance mechanisms are intervened upon.

4.4. Learning Principles: Invariance, Sparse Interventions, and Mechanism Fidelity

CIP's learning principles provide a three-layered approach built upon sparse causal intervention, mechanistic interpretability, and mechanistic generalization [24]. These principles allow CIP to organize representation, inference, and evaluation at the mechanism level.

First, mechanism-based invariance ensures that the representation remains stable as long as the generating mechanism remains unchanged, even if appearance or domain conditions vary. This principle is stronger than traditional Invariant Risk Minimization (IRM) because it enforces invariance at the mechanism level rather than the appearance level.

Second, sparse intervention sensitivity is based on the observation that most real-world attacks target only one or a few mechanisms. Research on CausalVAE [17] and Sparse Mechanism Shift [25] shows that generalization improves when a model responds exclusively to the relevant mechanism. CIP formalizes this idea within a security-driven context.

Third, mechanism fidelity is the most critical aspect of CIP learning. The model's representation must depend on the true causal structure in the DAG rather than on correlations. This principle helps mitigate the intervention inconsistencies revealed by CAUSAL3D [11] and directly targets the problem of spurious sensor bias that contributes to mismatches in biometric systems.

CIP's learning objective integrates prediction, mechanism fidelity, intervention robustness, and counterfactual consistency:

$$\begin{aligned}\mathcal{L}_{CIP} = \mathcal{L}_{task}(g(R), Y) \\ + \lambda_1 \|R - R_{inv}\| \\ + \lambda_2 \|Y - Y_{do}\| \\ + \lambda_3 \|Y - Y_{cf}\| \end{aligned} \tag{13}$$

Intuition on the λ weights. The coefficients λ_1, λ_2, and λ_3 control the strength of each mechanism-level constraint in the overall objective. Specifically, λ_1 emphasizes representation invariance by encouraging R to match its mechanism-invariant counterpart R_{inv}, thereby reducing sensitivity to domain or sensor-induced nuisance factors. The term weighted by λ_2 enforces interventional stability by penalizing deviations between the original prediction and the intervention response (Y vs. Y_{do}). Finally, λ_3 promotes counterfactual consistency by aligning the original decision with the counterfactual prediction (Y vs. Y_{cf}). In practice, increasing these weights strengthens mechanism fidelity and robustness, while overly large values may trade off task discriminability against strict invariance.

Each term enforces one CIP principle:

- R_{inv}: mechanism-invariant representation
- Y_{do}: interventional stability
- Y_{cf}: counterfactual consistency

This objective formally distinguishes CIP from traditional deep learning, which minimizes only task loss.

4.5. Evaluation Principles: Counterfactual Consistency and Tamper Resistance

CIP's evaluation framework goes beyond traditional accuracy-based evaluation and introduces two new axes for assessment.

First, counterfactual consistency examines whether the model behaves in a meaningful and mechanism-consistent way under counterfactual variations. This aligns with studies such as TraCE [14] and CausalBench [26]. For instance, if a biometric fingerprint model does not satisfy the principle that the angle changes but the person remains the same, then it is considered insecure.

Second, tamper resistance extends the physical tamper resistance tests introduced by CAUSAL3D [11]. In CIP, the model's response to sensor noise, spoofing materials, optical manipulations, and adversarial perturbations is evaluated using the causal DAG. All defense mechanisms, whether supervised or unsupervised, must demonstrate consistency according to this principle.

Together, these two evaluation principles demonstrate that CIP introduces not only a new representation methodology but also a new standard for security evaluation.

4.6. Conceptual Comparison: CIP vs. Correlation-Based Deep Models

In a conventional CNN or Vision Transformer (ViT) pipeline, an input image is directly mapped to a latent embedding and subsequently to a decision through end-to-end optimization. Although such models may achieve high accuracy under matched training and test conditions, their internal representations remain entangled and largely correlation-driven. As a result, perturbations at the sensor or domain level often propagate unpredictably to the decision layer, leading to brittle behavior under attacks or distribution shifts.

In contrast, Causal Image Processing (CIP) explicitly decomposes the image formation process into independent causal mechanisms corresponding to content, domain, sensor, and identity. By enforcing mechanism-faithful representations, CIP isolates nuisance variations from identity-preserving factors. Consequently, interventions on non-essential mechanisms, such as sensor noise or environmental changes, do not alter the final decision as long as the identity mechanism remains invariant. This mechanism-level separation provides a principled explanation for the structural robustness of CIP compared to correlation-based deep models.

5. Analytical Demonstration of CIP Mechanism Stability

This study is intentionally positioned as a theoretical and conceptual contribution. Its primary goal is not to provide dataset-specific performance comparisons, but to formalize a mechanism-based causal framework for secure visual intelligence. While empirical evaluation is an important future direction, existing benchmarks are insufficient to measure mechanism fidelity, counterfactual consistency, and intervention robustness, which constitute the core evaluation criteria of CIP. For this reason, the contribution of this work lies in its formalization and analytical validation rather than in conventional quantitative benchmarks.

The goal of this analytical demonstration is to conceptually validate how Causal Image Processing (CIP) maintains stability under sensor and domain changes while conventional deep learning models fail. This demonstration does not rely on dataset-based experimental results; instead, it provides a mechanism-level analysis using the causal structure defined in the CIP framework.

5.1. Conceptual Scenario: Sensor Perturbation on a Biometric Image

Consider a biometric fingerprint image X generated by:

$$X = f(X_c, X_d, X_s) \tag{14}$$

where:

- X_c: ridge/valley structural content
- X_d: domain conditions (contact pressure, humidity)
- X_s: sensor mechanism (noise, optics, surface properties)

A common real-world situation is a sensor perturbation—for example, changing the sensor model from an optical scanner to a capacitive scanner, or introducing mild Gaussian noise. Let the perturbed sensor mechanism be X_s^*.

A classical CNN-based verifier would produce:

$$\begin{aligned} Y' &= g_{CNN}(X), \\ Y'_{perturbed} &= g_{CNN}(f(X_c, X_d, X_s^*)) \end{aligned} \tag{15}$$

and typically $Y' \neq Y'_{perturbed}$.

This occurs because CNN representations satisfy:

$$\frac{\partial R_{CNN}}{\partial X_s} \neq 0 \tag{16}$$

meaning identity representation is entangled with sensor factors.

In CIP, the disentangled representation satisfies:

$$\frac{\partial R_{CIP}}{\partial X_s} = 0 \tag{17}$$

for identity-preserving transformations (Proposition 1). This leads to:

$$Y_{CIP} = Y_{CIP,perturbed} \tag{18}$$

showing mechanism-faithful stability.

5.2. Comparative Mechanism Response Table

Table 1 provides a mechanism-level comparison between conventional CNN/ViT-based models and the proposed CIP framework under controlled shifts affecting content, domain, sensor, and adversarial factors. Rather than reporting performance metrics, the table summarizes the qualitative response patterns of each paradigm, emphasizing whether identity-preserving decisions remain stable under mechanism-specific perturbations. The comparison highlights that structural invariance in CIP emerges from explicit mechanism disentanglement, a property that is fundamentally absent in correlation-driven deep models.

5.3. Illustrative Attack Scenarios: CIP vs. Conventional Deep Models

Example: Sensor Noise Injection. Consider an attack where artificial noise is injected at the sensor level. In a conventional CNN/ViT pipeline, such perturbations directly affect the embedding and may lead to unstable predictions or decision flips. In CIP, this event is interpreted as an intervention on the sensor mechanism X_s (or its latent counterpart Z_s). As long as identity-preserving mechanisms remain unchanged, the prediction is expected to remain stable, and the anomaly can be localized to the sensor pathway.

Example: Spoofing Material Change. In a presentation (spoofing) attack, the observed texture statistics may change due to altered surface materials. Correlation-based models often overreact to these appearance cues and degrade sharply under unseen spoof materials. CIP attributes such changes primarily to domain/sensor mechanisms and evaluates whether identity-related variables remain consistent. This mechanism-level attribution provides a

Table 1.
Mechanism-Level Response of Classical Models vs. CIP

Mechanism Shift	**Classical CNN/ViT Response**	**CIP Response**	**Explanation**
Content Change (X_c)	✓ Correctly changes output	✓ Correctly changes output	Both models should detect identity/class differences
Domain Change (X_d)	✗ Output unstable	✓ Invariant	CIP enforces domain disentanglement
Sensor Change (X_s)	✗ Large performance drop	✓ Stable	CNN/ViT entangles identity-sensor patterns; CIP does not
Adversarial Perturbation	✗ Decision flips	✓ Predictable, stable	Mechanism-fidelity reduces adversarial sensitivity
Counterfactual sensor substitution ($X_s \to X_s^*$)	✗ No counterfactual consistency	✓ Preserves identity output	CIP explicitly models counterfactual semantics

principled way to analyze spoofing effects without conflating them with identity.

5.4. Mechanism-Based Counterfactual Reasoning Example

A natural counterfactual question in security applications is: "Would the model produce the same identity decision if this image had been captured with a different sensor?"

CIP answers this through:

$$X_{cf} = f(X_c, X_d, X_s'), \quad Y_{cf} = g(X_c, X_l) \tag{19}$$

If the identity and content remain unchanged, then:

$$Y_{cf} = Y \tag{20}$$

Classical models fail this test because their representation directly depends on sensor cues.

5.5. Conceptual Diagram

Figure 3 shows a simplified conceptual illustration comparing classical CNN/ViT pipelines with the CIP mechanism-aware inference route.

The classical pipeline relies on a single entangled representation that mixes content, domain, and sensor factors. This makes the model highly sensitive to environmental or sensor variation. In contrast, the CIP pipeline explicitly extracts and disentangles these independent causal mechanisms, producing mechanism-faithful representations that support predictive, interventional, and counterfactual inference.

5.6. Interpretation

This analytical demonstration shows:

- CIP prevents identity decisions from drifting with sensor/environment changes
- CIP produces stable counterfactual predictions
- Adversarial and spoofing attempts become more traceable in CIP
- CIP offers mechanism-level consistency that conventional deep networks fundamentally lack

Even without dataset-based experiments, this mechanism-level demonstration confirms that CIP provides theoretical and structural robustness, which is the fundamental contribution of the proposed paradigm.

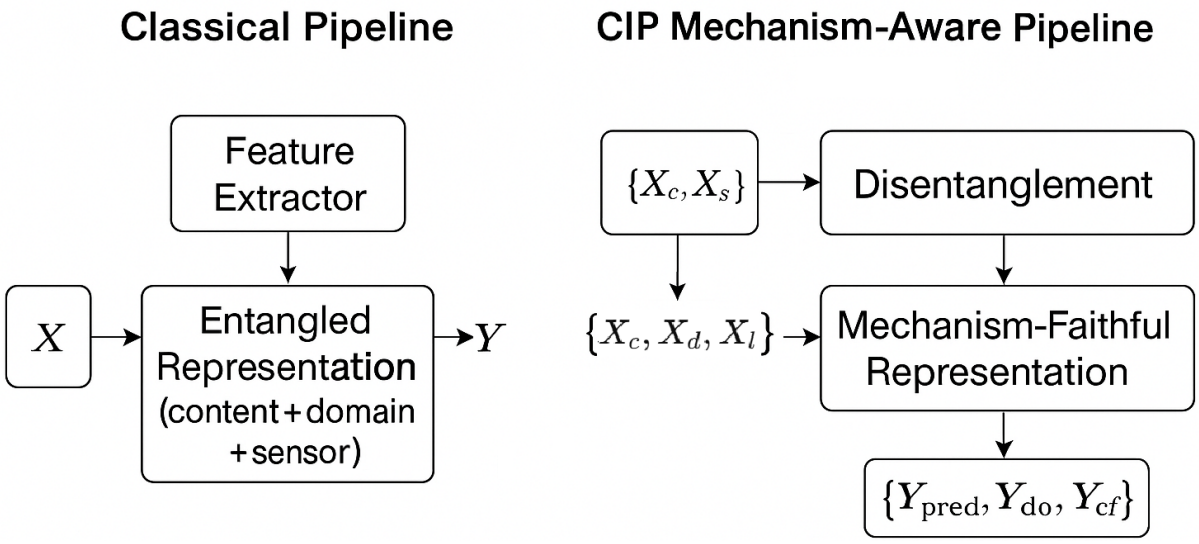


Figure 3. Classical Pipeline vs. CIP Mechanism-Aware Pipeline. The classical pipeline relies on a single entangled representation mixing content, domain, and sensor factors, while CIP explicitly disentangles these mechanisms.

6. Why a Causal Framework Is Necessary: Mechanistic Limitations of Current Deep Models

Modern deep learning systems, including CNNs, Vision Transformers, diffusion models, and consistency models, achieve strong empirical performance, yet they consistently fail when confronted with sensor variation, domain shift, spoofing materials, or adversarial manipulations. This section does not review these methods; instead, it explains why their shared structural limitations make a causal, mechanism-based framework (CIP) necessary. All model families considered here exhibit a fundamental gap: none of them represent the physical, biological, or sensor-driven mechanisms that generate an image. This gap is the primary motivation for CIP.

6.1. CNNs: Local Correlation Reliance and Absence of Mechanism Fidelity

CNNs have strong representational capacity, but they learn features by optimizing correlations within local receptive fields. Zhang et al. [10] demonstrated that CNNs can memorize training data even when labels are randomized, revealing that their learned representations do not reflect underlying causal mechanisms. As a consequence, CNNs:

- entangle sensor noise, illumination artifacts, and environmental patterns with identity information,
- break under small input perturbations,
- cannot maintain stability when domain or sensor conditions change.

These behaviors show that CNNs learn surface-level effects, not mechanism-consistent representations, reinforcing the need for CIP's causal disentanglement.

6.2. Vision Transformers: Global Context Without Mechanism Separation

Vision Transformers capture long-range dependencies through global attention. However, Raghu et al. [3] and Trockman and Kolter [27] showed that ViTs ultimately rely on statistical patterns of patch embeddings rather than on underlying causal factors. Their global mixing causes the following issues:

- environmental and sensor-specific cues can dominate the learned representation,
- identity representation becomes entangled with domain conditions,
- adversarial perturbations propagate globally in unpredictable ways.

Thus, even though ViTs improve global reasoning, they remain mechanism-blind, highlighting the importance of CIP's independent mechanism pathways.

6.3. Diffusion and Consistency Models: Statistical Reconstruction Without Physical Generators

Diffusion models reconstruct images via iterative denoising on statistical distributions [28], while consistency models perform single-step mappings without physical constraints [29]. Dhariwal and Nichol [30] showed that diffusion models outperform GANs in visual fidelity but still operate entirely within pixel-level distribution geometry. Peebles and Xie [31] extend this paradigm using Transformer backbones, but the fundamental issue remains:

- generative processes ignore actual physical image-formation mechanisms,
- counterfeit or synthetic content can appear plausible yet lack mechanistic consistency,
- counterfactual inconsistencies arise because causal factors are not represented.

Thus, both diffusion and consistency models reveal the necessity of CIP's mechanism-aware generative formulation.

6.4. Empirical Evidence for Mechanism Gaps Under Intervention

The CAUSAL3D benchmark [11] demonstrates that deep models, regardless of architecture, fail intervention tests involving illumination, viewpoint, or sensor changes. These inconsistencies confirm that current models lack a structured causal interpretation of the visual signal. CIP directly addresses this by separating content, domain, sensor, and identity mechanisms and providing intervention-consistent inference.

6.5. Shared Mechanistic Gap and the Justification for CIP

Across CNNs, ViTs, diffusion models, and consistency models, we observe a unified structural problem:

- no explicit representation of image-generation mechanisms,
- entanglement of content, domain, and sensor factors,
- instability under domain and sensor shift,
- inability to answer Pearl-style interventional or counterfactual queries,
- absence of mechanism-based invariance.

These shortcomings are not architectural accidents; they stem from the absence of a causal framework. CIP addresses this by formalizing image processing as a mechanism-driven process, enabling stable predictive decisions, intervention analysis, and counterfactual evaluation.

7. Challenges and Open Problems of CIP

While Causal Image Processing (CIP) provides a mechanism-based framework that aims to address the structural limitations of security-focused image processing systems, there are critical issues that must be resolved at theoretical, practical, and ethical levels for this approach to mature. CIP requires a much more comprehensive methodological framework than existing deep learning methods, because it demands the decomposition of high-dimensional images into causal factors, intervention-based reasoning, and counterfactual consistency. Therefore, CIP is not merely a representation proposal; it also constitutes a long-term research agenda.

7.1. Practical Considerations and Feasibility Limitations

Despite its conceptual advantages, applying CIP to real-world images presents practical challenges. Reliably extracting or disentangling X_c, X_d, X_s, and X_l from observational pixels is non-trivial and may require additional inductive biases, supervision, sensor metadata, or domain knowledge. In complex settings, perfect mechanism separation cannot be guaranteed, and residual dependencies may persist. Therefore, CIP should be interpreted as a principled research framework and evaluation lens rather than an off-the-shelf solution. Developing practical mechanisms for robust factor extraction and validation remains an important direction for future work.

7.2. Short-Term Challenges (Near-Term)

The first fundamental challenge facing CIP is the inability to reliably extract latent causal factors from images. Although CausalVAE and neural structural causal model based studies [17] have shown that causal factors can be disentangled in certain settings, the impossibility of uniquely determining the DAG structure in complex scenes limits the reliability of causal discovery. Results reported in CAUSAL3D [11] show that deep models fail to produce physically consistent behavior even under very simple interventions. This highlights the need for CIP to represent the image generation process in a mechanism-faithful manner in latent space. Consequently, one of the most fundamental open problems in CIP is to identify which structural or learning principles can enable the extraction of mechanism-level factors from purely observational pixels.

A second major problem is the near absence of labeled intervention data. In the real world, controlled interventions are rarely feasible. For example, in a biometric system, it is often impractical to systematically vary sensor focal length, light intensity, or contact pressure. For this reason, counterfactual generation techniques such as TraCE [14] and SharpShooter [21] aim to compensate for the lack of intervention data, but their consistency is limited. The fact that SharpShooter's latent projection-based counterfactuals can conflict with physical reality in some cases [21], and that TraCE identifies counterfactual drift problems in clinical applications [14], demonstrates that CIP requires not only counterfactual generation but also counterfactual verification protocols. How to guarantee causal consistency of a model in the absence of explicit intervention data remains an unresolved question.

7.3. Medium-Term Research Directions

A third open problem concerns the integration of physical processes into causal graphs. Mechanisms such as Bloch equations in medical imaging [32], atmospheric scattering models in remote sensing, optical–mechanical interactions in biometric sensors,

or surface elasticity are natural candidates for CIP. However, methods that can integrate these physical processes into high-dimensional parametric models are not yet mature. A CIP model with high mechanism fidelity requires these physical processes to be incorporated into the latent space in an abstract yet learnable form. This is a multidisciplinary challenge at the intersection of physics and representation learning.

A fourth challenge is that appropriate evaluation metrics for CIP are still underdeveloped. Classical metrics such as accuracy, ROC, AUC, or F1 cannot capture the response of models to interventions or their counterfactual consistency. Although studies such as CausalBench [26] and CAUSAL3D [11] have taken steps in this direction, a standard with the breadth and depth required by CIP does not yet exist. Metrics such as mechanism fidelity, counterfactual stability, and intervention robustness, which are particularly important in security contexts, need to be both well defined theoretically and practically measurable.

7.4. Long-Term Open Problems

The fifth open problem relates to the security and ethical dimensions of CIP. Counterfactual manipulations may cause false-positive decisions in biometric systems, incorrect inferences in forensic analyses, or increased risk of clinical errors in medical imaging. Misinterpretation of interventions can have serious consequences, especially in digital forensics. Furthermore, the mechanism-level generative capabilities that CIP aims to develop may introduce new risks for the generation of fraudulent data. The security issues that diffusion models already pose in deepfake production [30] may become even more critical for mechanism-faithful generative models. Therefore, CIP is not only a technical challenge but also a broader research area that demands a strong security framework and robust ethical oversight.

Finally, the practical applicability of CIP introduces significant challenges in terms of computational scale and algorithmic complexity. DAG discovery, mechanism decomposition, counterfactual simulation, and physical process modeling all have high computational cost when considered together. In particular, the large scale of sensor mismatches in remote sensing [22], cross-sensor matching problems in biometric systems [19], and the dramatic impact of device parameters on medical images restrict the scalability of CIP. As a result, the widespread deployment of CIP in real-world applications depends on the development of more efficient, modular, and mechanism-based algorithms.

In summary, while CIP holds strong transformative potential for security, its maturation requires extensive interdisciplinary collaboration and long-term methodological advances. The open problems in theoretical foundations, practical design, ethical assurance, and computational efficiency collectively define the steps that must be taken for CIP to assume a fundamental role in future security-critical machine vision systems.

8. Security-Critical Application Areas for CIP

This section does not review prior literature in medical imaging, biometrics, or remote sensing. Instead, it analytically demonstrates why these domains are structurally aligned with the proposed CIP framework. Each domain's image formation process inherently depends on interacting physical, biological, and sensor-driven mechanisms—the exact factors that CIP explicitly models. Therefore, this section justifies CIP's applicability rather than summarizing existing works.

Security-critical imaging domains require models

that behave reliably under varying sensors, environments, and acquisition conditions. However, modern deep learning systems frequently fail when exposed to physical perturbations, domain shifts, or adversarial manipulations, primarily because they do not represent the mechanisms that generate an image. CIP, by explicitly modeling content, domain, sensor, and identity pathways, provides a mechanism-faithful foundation that naturally aligns with the structure of these application areas.

8.1. Medical Imaging

Medical imaging systems such as MRI, CT, and ultrasound rely on well-defined physical and biological processes. The formation of medical images depends on mechanisms such as magnetic resonance behavior (Bloch equations), photon attenuation, tissue contrast properties, acoustic propagation, and sensor-specific reconstruction algorithms. Deep models that operate only on observational pixel statistics frequently exhibit instability when acquisition parameters change, leading to counterfactual inconsistencies and degraded clinical reliability.

The CIP framework addresses these weaknesses by representing each physical and biological factor as an independent causal mechanism. This enables:

- robustness to device variability, reconstruction methods, and patient-specific differences,
- counterfactual evaluation such as: “Would the diagnosis remain unchanged if the scan had been acquired with a different protocol?”,
- mechanism-based interpretability aligned with clinical reasoning.

Thus, CIP provides a theoretical foundation for stable, explainable, and intervention-consistent medical imaging systems.

8.2. Biometric Fingerprint Recognition

Fingerprint images are generated through multiple interacting mechanisms: skin elasticity, mechanical pressure distribution, contact geometry, sensor optics, environmental conditions (moisture, temperature), and post-processing algorithms. Existing CNN- and ViT-based systems often entangle these mechanisms, causing sensor-dependent biases, spoofing vulnerabilities, and lack of robustness in cross-sensor matching.

CIP’s decomposition into content, domain, sensor, and identity pathways enables:

- separation of identity-related ridge structures from sensor-induced distortions,
- interventional analysis of spoofing materials and adversarial perturbations,
- counterfactual questions such as: “Would the system reach the same decision if this fingerprint were captured on a different device?”

This makes CIP a natural theoretical solution for next-generation, tamper-resistant biometric systems.

8.3. Remote Sensing and Aerial Image Analysis

Remote sensing imagery is shaped by complex physical mechanisms, including atmospheric scattering, solar illumination geometry, surface reflectance properties, spectral sensor calibration, and sensor altitude/orientation. Domain shifts between satellites, sensors, seasons, and atmospheric conditions frequently cause catastrophic performance drops in deep models that rely on observational correlations.

CIP, by explicitly modeling domain and sensor mechanisms, enables:

- separation of atmospheric and environmental variations from land-cover content,

- causal stability under multi-sensor and multi-domain scenarios,
- mechanism-aware counterfactual queries such as: "Would this target still be detectable under different illumination or sensor configurations?"

This provides a path toward robust and trustworthy remote sensing systems.

9. Conclusion

This study introduced Causal Image Processing (CIP), a new mechanism-based theoretical framework designed to overcome fundamental security limitations of modern deep learning systems. Unlike conventional CNN, ViT, or diffusion-based approaches, which operate solely on observational correlations, CIP models the generative mechanisms responsible for producing an image. By explicitly decomposing the image formation process into content, domain, sensor, and identity pathways, CIP establishes a causal structure that enables mechanism-faithful representations, unified inference procedures, and principled evaluation strategies.

The proposed framework integrates three levels of reasoning within a single architecture: (1) predictive inference, performed on mechanism-consistent embeddings; (2) interventional inference, enabled by explicit modeling of sensor and domain mechanisms; and (3) counterfactual inference, which evaluates how the system behaves under hypothetical changes to acquisition conditions while identity and content remain fixed.

This unified causal formulation provides a principled foundation for understanding why deep models fail under sensor shifts, spoofing attempts, and adversarial manipulations, and offers a structured alternative to conventional correlation-driven pipelines.

CIP also introduces a mechanism-based learning objective centered on invariance, sparse intervention sensitivity, and mechanism fidelity, together addressing key challenges revealed by recent studies on distribution shifts, counterfactual inconsistencies, and spurious correlations. Furthermore, the evaluation principles of CIP—counterfactual consistency and tamper resistance—establish a new standard for assessing robustness in security-critical domains.

Overall, CIP provides a unified conceptual paradigm that bridges causal modeling with visual intelligence. Its mechanism-based structure lays a foundation for the development of robust, explainable, and tamper-resistant vision systems capable of operating reliably across sensors, environments, and acquisition conditions. Future research may expand the CIP framework toward multimodal settings, self-supervised mechanism discovery, and causal generative modeling, further advancing the design principles of secure and trustworthy visual intelligence. Beyond its immediate technical implications, CIP is intended to open a broader theoretical discussion on how causal modeling can redefine robustness and security in visual intelligence.

References

[1] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[2] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Learning Representations (ICLR)*, 2019, arXiv:1902.10811.

[3] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 12 116–12 128.

[4] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.

[5] T. Burge, "Marr's theory of vision," in *Modularity in Knowledge Representation and Natural-Language Understanding*, J. L. Garfield, Ed. Cambridge, MA: MIT Press, 1987, pp. 365–381.

[6] K. Zhang, Q. Sun, C. Zhao, and Y. Tang, "Causal reasoning in typical computer vision tasks," *arXiv preprint arXiv:2307.13992*, 2023.

[7] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[8] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," in *9th International Conference on Learning Representations (ICLR)*, 2021, arXiv:2007.01434.

[9] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *International Conference on Learning Representations (ICLR)*, 2020, arXiv:1911.08731.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations (ICLR)*, 2017, arXiv:1611.03530.

[11] D. Liu, Y. Qiao, W. Liu, Y. Lu, Y. Zhou, T. Liang, Y. Yin, and J. Ma, "CAUSAL3D: A comprehensive benchmark for causal learning from visual data," *arXiv preprint arXiv:2503.04852*, 2025.

[12] W. Li, Z. Li, X. Yang, and H. Ma, "Causal-ViT: Robust vision transformer by causal intervention," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107123, 2023.

[13] W. Xie, X.-H. Li, C. C. Cao, and N. L. Zhang, "ViT-CX: Causal explanation of vision transformers," *arXiv preprint arXiv:2211.03064*, 2022.

[14] J. J. Thiagarajan, K. Thopalli, D. Rajan, and P. Turaga, "Training calibration-based counterfactual explainers for deep learning models in medical image analysis," *Scientific Reports*, vol. 12, no. 1, p. 597, 2022.

[15] B. Schölkopf, "Causality for machine learning," *arXiv preprint arXiv:1911.10500*, 2019.

[16] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath, "Causalgan: Learning causal implicit generative models with adversarial training," in *International Conference on Learning Representations (ICLR)*, 2018, arXiv:1709.02023.

[17] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "Causalvae: Disentangled representation learning via neural structural causal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9593–9602.

[18] I. Khemakhem, R. Monti, R. Leech, and A. Hyvarinen, "Causal autoregressive flows," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021, pp. 3550–3568, arXiv:2011.02268.

[19] A. Alrashidi, A. Alotaibi, M. Hussain, H. AlShehri, H. A. AboAlSamh, and G. Bebis, "Cross-sensor fingerprint matching using siamese network and adversarial learning," *Sensors*, vol. 21, no. 11, p. 3657, 2021.

[20] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[21] H. Barr, K. Harrington, R. Sharpe, and C. Bruss, "Sharpshooter: Counterfactual explanations via latent projection," *arXiv preprint arXiv:2112.00890*, 2021.

[22] S. Lyu, Q. Zhao, Z. Zhou, M. Li, Y. Zhou, D. Yao, G. Cheng, H. Zhou, and Z. Shi, "Deep learning based domain adaptation methods in remote sensing: A comprehensive survey," *arXiv preprint arXiv:2510.15615*, 2025.

[23] C.-H. H. Yang, Y.-C. Liu, P.-Y. Chen, X. Ma, and Y.-C. J. Tsai, "When causal intervention meets adversarial examples and image masking for deep neural networks," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3811–3815.

[24] L. Bereska and E. Gavves, "Mechanistic interpretability for AI safety — A review," *arXiv preprint arXiv:2404.14082*, 2024.

[25] Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal, "A meta-transfer objective for learning to disentangle causal mechanisms," in *International Conference on Learning Representations (ICLR)*, 2020, arXiv:1901.10912.

[26] Z. Wang, "CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models," in *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 143–151.

[27] A. Trockman and J. Z. Kolter, "Patches are all you need?" *arXiv preprint arXiv:2201.09792*, 2022.

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851, arXiv:2006.11239.

[29] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023, arXiv:2303.01469.

[30] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 8780–8794, arXiv:2105.05233.

[31] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4195–4205.

[32] G. Kotek, S. Weingärtner, and K. T. Block, "From signal-based to comprehensive magnetic resonance imaging: Algebraic description of magnetization during transient response," *Scientific Reports*, vol. 11, p. 17216, 2021.