National University of Singapore

Department of Mathematics

DSA 5201 Report - AI Image Detection

Junshi Wang

Supervisor: Prof. Qianxiao Li

A report submitted in partial fulfilment of the requirements of
the National University of Singapore for the degree of
Master of Science in *Data Science and Machine Learning*

November 17, 2025

Abstract

The rapid advancement of generative AI has created new opportunities for malicious actors to fabricate convincing visual evidence, posing serious risks for industries that rely on photographic documentation. This report investigates the detection of AI-generated car damage images, a problem of growing importance for insurance fraud prevention. Using a curated real-image dataset and synthetic images produced by several modern generative systems (Stable Diffusion 2, Kontext/FLUX, and Qwen-VL), we experiment and evaluate a series of detection models, compare their performance across architectures, and explore strategies to improve robustness and transferability. Our results show that synthetic car-damage images leave identifiable forensic traces, enabling reliable detection within individual generative domains. However, these detectors exhibit significant overfitting and struggle to generalize across different generative architectures. Domain adaptation methods offer partial improvement, particularly between closely related transformer-based models, but fail between fundamentally different generative families. Finally, our interpretability analysis highlights that current tools provide only limited insight into the features driving detection decisions, underscoring the need for more transparent and generalizable forensic frameworks.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid development of generative artificial intelligence has transformed how visual content is created, edited, and disseminated. Diffusion models, flow-matching transformers, and multimodal vision–language systems now produce images of striking realism, enabling new forms of creativity but also introducing serious risks when misused. Recent incidents demonstrate the severity of these risks: fabricated car-damage images submitted for insurance fraud, AI-generated documents used in financial scams, and realistic voice or identity impersonations leading to substantial personal and institutional losses. More broadly, the increasing prevalence of synthetic visual content threatens public trust and complicates our ability to distinguish genuine digital evidence from algorithmically produced fabrications.

Despite growing awareness, existing countermeasures remain limited in practice. Generative models evolve rapidly, producing artifacts that differ sharply from those targeted by earlier forensic detectors. Open-source model releases make centralized regulation or watermark enforcement unreliable, as malicious users can remove safety mechanisms or deploy modified versions privately. Under realistic adversarial conditions, detection systems must therefore rely solely on the observable properties of images, without assuming cooperation from the model creators or the image producers.

Within this landscape, a critical but understudied form of AI misuse has begun to emerge: object-centric synthetic images used for financial gain. While human-centric deepfake detection has received substantial academic, industrial, and policy attention, the manipulation of everyday physical objects—cars, receipts, household items, merchandise—has attracted comparatively little research. Yet the consequences can be severe. Fraudulent car-damage images, for example, can directly inflate insurance claims, generate monetary losses, and undermine trust in digital workflows that rely heavily on photographic documentation.

This study focuses specifically on synthetic car-damage images as a representative and societally relevant case of object-level AI misuse. By curating a high-quality real-image dataset and generating synthetic counterparts using modern diffusion and transformer-based models, we systematically examine how these images differ from authentic ones and whether those differences can be detected reliably. One key finding is that detection is indeed feasible within individual generative domains: all models we evaluate leave subtle but learnable forensic traces, and even simple convolutional networks can achieve high precision and recall—provided that high-resolution inputs are available. These results suggest that the relevant cues are embedded in fine-grained frequency patterns or error-level artifacts produced during rendering or reconstruction.

However, our analysis also reveals that these cues are fragile. Detection performance collapses when images are downsampled or compressed, indicating that the statistical signatures distinguishing real from generated images lie primarily in high-frequency components that are

easily lost. Moreover, generalization across generative models is uneven. Whereas some modern systems, such as flow-matching transformers and multimodal editors, produce sufficiently similar artifacts to allow cross-domain transfer, others—particularly architectures based on latent diffusion—exhibit distinct and non-overlapping forensic fingerprints. As a result, detectors trained on one model may fail entirely when confronted with another, highlighting the structural dependence of forensic signals on generative design choices.

A further challenge lies in interpretability. Although attribution methods such as Grad-CAM and neuron-level analysis indicate that detectors do attend to meaningful forensic cues—often focusing on altered textures, structural discontinuities, or localized artifact clusters—existing explanation tools provide only a coarse and sometimes unstable view of the underlying decision process. For high-stakes applications such as insurance assessment or digital evidence verification, such limited transparency is insufficient. Understanding not only *whether* an image is synthetic but also *why* a detector reaches its conclusion is essential for building trust and ensuring responsible deployment.

Taken together, these observations illustrate both the promise and the current limitations of AI-image forensics in object-centric settings. They show that robust detection is achievable under controlled conditions, yet highly sensitive to image fidelity, generative lineage, and interpretability constraints. These insights motivate the methodological developments and analyses presented in the remainder of this thesis, and they point toward future directions aimed at improving generalizability, robustness, and forensic transparency in next-generation detection systems.

Chapter 2

Literature Review

This chapter reviews the existing research landscape from three perspectives: the taxonomies used to characterize AI misuse, the technical approaches developed for detecting AI-generated or manipulated images, and the role of interpretability in determining whether these systems can be deployed in high-stakes settings. Together, these perspectives reveal a notable gap in current research. While prior work overwhelmingly focuses on human-centric generative misuse and early-generation models, very few studies examine object-level synthetic images that carry real financial implications. Our work is situated precisely in this underexplored space.

2.1 Taxonomy of AI Misuse

Advances in generative AI have substantially lowered the barrier to producing convincing synthetic content, prompting a range of studies that attempt to categorize different forms of misuse. These taxonomies provide an essential foundation for understanding societal and security risks, yet they remain heavily oriented toward human-centric harms and leave object-level falsification largely unaddressed.

2.1.1 Human-Centric Focus in Existing Taxonomies

Generative AI Misuse: A Taxonomy of Tactics and Insights from Real-World Data [1] offers one of the most comprehensive examinations of malicious use cases, organizing them into realistic representations of humans, realistic representations of non-human content, and harmful downstream applications. Despite this broad framing, empirical examples and research attention are overwhelmingly concentrated on impersonation, identity fraud, voice cloning, facial deepfakes, and political manipulation. These threats are serious, but the taxonomy allocates comparatively little focus to object-level forgery that can yield immediate financial harm.

A similar tendency appears in *A Technological Perspective on Misuse of Available AI* [2] and in the early but influential report *The Malicious Use of Artificial Intelligence* [3]. Both works survey digital, physical, and political security domains, yet the most extensively studied risks again relate to human identity, social influence, or critical infrastructure. Object-level AI manipulation, particularly for profit-driven fraud, receives far less systematic treatment.

2.1.2 Financial and Object-Level Misuse: An Underexplored Category

One partial exception is the discussion of financially motivated misuse in *Protecting Society from AI Misuse* [4], which highlights risks such as automated phishing and document forgery.

Even here, however, most examples involve text-based attacks rather than image-based manipulation. As a result, falsification of physical objects—such as forged receipts, edited medical records, or fabricated car damage images—remains significantly underrepresented in existing frameworks despite the clear and measurable financial losses these falsifications can produce. Many industries rely heavily on photographic documentation but lack reliable forensic tools for authenticating visual evidence, leaving a vulnerable attack surface that is not captured in current academic taxonomies.

This gap in the literature directly motivates our study. Object-centric synthetic images constitute an actionable and increasingly accessible vector for fraud, yet they remain insufficiently examined both conceptually and empirically.

2.2 Technical Approaches to AI Image Forgery Detection

While misuse taxonomies offer conceptual structure, a large body of research investigates concrete technical methods for detecting manipulated or fully synthetic images. Existing approaches fall broadly into three categories: image manipulation localization, synthetic image classification, and watermark-based detection. Despite substantial progress, most methods depend on outdated generative models or low-resolution datasets and have not been rigorously evaluated against modern diffusion or vision–language model (VLM) based generators.

2.2.1 Limitations of Current Datasets and Generative Models

A central observation from the literature is that many foundational forensic methods were developed using early GAN architectures such as ProGAN, older diffusion models, or low-resolution datasets like CIFAR and CelebA-HQ. Consequently, a significant portion of prior work benchmarks detection performance only on ProGAN-derived images or early DDPM variants. This raises concerns about whether conclusions drawn from these settings generalize to modern latent diffusion systems such as Stable Diffusion 2.x, to flow-matching generators such as FLUX or Kontext, or to multimodal image-editing systems such as Qwen-VL. These limitations motivate our decision to construct a more realistic and representative dataset combining high-resolution real car-damage images (CarDD) with synthetic images produced by a diverse set of contemporary open-source models.

2.2.2 Existing Methods: Strengths and Weaknesses

Technical approaches can be grouped into three major methodological directions. Image manipulation localization methods such as PSCC-Net [5], SPAN [6], TruFor [7], and MMFusion [8] attempt to identify specific tampered regions within an image by modeling multi-scale feature relationships, forensic cues, noise fingerprints, or correlation structures. These techniques are powerful when only part of an image has been altered, but they are less suited to fully synthetic images, where no pristine region exists to serve as a reference.

A second direction centers on synthetic image classification, where detectors attempt to identify global forgeries. Representative methods include DIMD [9], which captures diffusion-model fingerprints; UnivFD [10], which improves generalization by operating in CLIP's feature space; and RINE [11], which leverages intermediate encoder representations. Although these classifiers achieve strong results on specific datasets or generator families, they tend to be sensitive to distribution shifts and may overfit to low-level artifacts produced by particular pipelines.

Watermark-based detection offers a third strategy. Approaches such as TreeRing [12], StegaStamp [13], and ImageDetectBench [14] embed identifiable signals into generated images, enabling robust detection under controlled conditions. However, these methods generally fail when facing open-source generation workflows, adversarial manipulation, watermark removal, or diffusion-based resynthesis [15, 16]. As such, watermarking is ill-suited for adversarial scenarios, which are precisely the cases of interest in this work.

2.2.3 Methodological Gaps

Across these categories, several limitations appear consistently. Many existing methods focus on early GANs or early diffusion systems and therefore do not reflect the capabilities of modern open-source generators. The datasets used for evaluation are often narrow in scope, typically low-resolution or human-centric, which restricts their relevance to object-level financial fraud. Finally, many detectors rely on artifacts that are specific to a particular model family, resulting in weak generalization to unseen generators. These methodological constraints motivate our exploration of cross-model generalization using more contemporary generative pipelines and domain-specific real-world datasets.

2.3 Interpretability as a Prerequisite for Practical Use

Although prior work in synthetic image forensics evaluates detectors largely in terms of accuracy and robustness, real-world deployment requires more than predictive performance. A forensic system is only usable if its decisions can be interpreted in a credible and actionable manner. Without meaningful interpretability, even high-performing detectors may fail to earn trust from insurance examiners, investigators, auditors, or other analysts responsible for assessing visual evidence.

2.3.1 Insufficient Interpretability in Current Forensic Methods

Interpretability is one of the least developed components of modern forensic pipelines. Most methods treat explanation as an optional add-on and do not meaningfully analyze the internal reasoning processes of their models. For instance, CNN-based detectors seldom incorporate CAM or Grad-CAM analysis, leaving their decision-making opaque. Systems relying on CLIP or other vision–language models can produce natural-language rationales, as demonstrated by ForenX [17], but these explanations stem primarily from the linguistic priors of large language models rather than from the internal representations used by the forensic classifier itself. This mismatch creates what we refer to as an interpretability gap: the entity providing the explanation is not the entity making the classification decision.

Moreover, many classifiers rely on high-frequency generative artifacts that are non-semantic in nature and difficult to articulate. Such cues may be invisible to humans and resistant to explanation even with advanced interpretability tools. This creates a broader problem: forensic models often depend on features that cannot be translated into human-understandable terms, thereby limiting their usefulness in operational contexts.

2.3.2 Interpretability as a Requirement for Any Real-World Use

We emphasize a fundamental point: a forensic system without meaningful interpretability is effectively unusable in real operational environments. Human oversight requires an understanding of why a detection decision was made. Insurance examiners, investigators, and risk analysts must be able to justify their judgments, and unexplained false positives or false

negatives would undermine trust in any automated tool. Furthermore, high-stakes decisions involving financial claims or legal evidence demand transparency. Regulators increasingly require explainability for automated systems, particularly when outcomes affect liability or compensation. Thus, even detectors that perform well under controlled academic benchmarks may fail to meet real-world standards if their reasoning processes remain opaque.

2.4 Summary of Literature Gaps

Across misuse taxonomies, technical methodology, and interpretability, three major gaps emerge. First, existing taxonomies underrepresent object-level and financially motivated forgery, focusing predominantly on human impersonation and political manipulation while overlooking commercial forgeries such as synthetic car damage. Second, technical evaluations depend heavily on outdated generative models and limited datasets, rarely testing against recent diffusion or vision–language models that produce more realistic output. Third, interpretability remains significantly underdeveloped. Most detectors operate as black boxes, and existing explanation tools often rely on auxiliary VLMs that do not reflect true model reasoning. Together, these gaps motivate our work, which focuses on object-level synthetic image detection, evaluates a diverse set of modern generative models, and positions interpretability as an essential component of forensic system design.

Chapter 3

Methodology and Research Questions

Building upon the literature review and the gaps identified in AI misuse taxonomy, forensic image detection, generalization, and interpretability, this chapter formalizes the key problems motivating our study and outlines the methodological framework adopted to investigate them. Whereas prior research typically focuses on broad image domains or on human-centric deepfakes, our work instead targets a narrower yet practically significant domain: maliciously generated car damage images. This subdomain presents unique challenges that existing forensic methods have not fully addressed.

3.1 Problems to Be Solved

The literature reveals three fundamental challenges that remain unresolved. In the context of our object-level domain, these challenges manifest as three concrete problems.

Problem 1 — Detectability in a Narrow Object-Level Domain

Most existing studies evaluate forensic detectors on broad, diverse datasets or on human-centric manipulations. By contrast, our focus lies entirely on car damage images, a highly homogeneous object-level subdomain. At first glance, such homogeneity might simplify detection, as the geometric structure of cars is relatively consistent. However, it is unclear whether this expectation holds in practice. Domain consistency may indeed stabilize certain forensic cues, and generative models may exhibit systematic biases when producing car-damage patterns. Yet car surfaces also contain complex textures, reflections, and lighting effects that modern diffusion models often reproduce convincingly, potentially weakening the detectability of synthetic artifacts. Further complicating the problem, subtle signals may be highly sensitive to image quality, and operations such as compression or resizing can easily destroy detectable cues. The required model capacity is also uncertain: strong cues may be captured by shallow CNNs, whereas subtle cues may necessitate deeper architectures such as ResNet-18. Thus, whether restricting the problem to a single object-level domain makes detection easier or harder remains an open empirical question and forms the basis of our first research inquiry.

Problem 2 — Model Overfitting and Limited Cross-Model Generalization

A recurring theme across the literature is that forensic detectors tend to overfit to artifacts from specific generative models, leading to poor performance on unseen generators. Prior

work usually studies this issue across heterogeneous domains, leaving open the question of how it manifests within a single, narrow subdomain such as car damages. Malicious actors today have access to a diverse set of tools, including latent diffusion models such as Stable Diffusion 2, flow-matching transformers such as FLUX Kontext, and instruction-following vision–language models such as Qwen-VL for image editing. Whether a detector trained on one of these models generalizes to the others in a constrained domain remains unknown. Our second problem therefore concerns the magnitude of cross-model overfitting in a domain-specific setting and whether detectors can reliably identify synthetic images generated by unseen systems.

Problem 3 — Interpretability and Trustworthiness

As highlighted in the literature review, interpretability remains one of the least developed aspects of synthetic image detection. Existing explanation tools, such as Grad-CAM, filter activation visualizations, or VLM-based rationales, often fail to reflect the true reasoning processes of the underlying classifiers. This creates a deep interpretability gap in safety-critical settings. Real-world applications—particularly in insurance or financial systems—require not only high accuracy but also transparent and credible explanations that support human verification, institutional trust, and robustness against adversarial misuse. Our third problem is therefore to determine whether a meaningful interpretability framework can be developed for object-level forgery detection and whether such a framework can provide stable, domain-relevant insights into model behavior.

3.2 Research Questions

These three problems give rise to three focused research questions. The first, RQ1, asks how difficult it is to detect fully synthetic images within a narrow object-level domain such as car damages. We investigate whether domain homogeneity amplifies or weakens forensic cues and whether detection reliability requires lightweight CNNs or deeper architectures, especially under realistic image-quality degradations. The second, RQ2, examines the extent to which detectors overfit to particular generative models even within a constrained domain. We study cross-generator robustness by training on one system and evaluating on others, and we explore whether strategies such as multi-source training or adversarial domain adaptation can reduce overfitting. The third question, RQ3, concerns interpretability: we ask how the decisions of synthetic-image detectors can be meaningfully explained and whether interpretability can help assess a model's credibility, reveal reliance on spurious cues, or illuminate transferability across generators.

3.3 Methodology Overview

To investigate these research questions, we adopt a structured methodological pipeline. We begin by assembling a real-image dataset consisting of car damage and non-damage scenes collected from publicly available sources. To simulate realistic adversarial conditions, we generate synthetic car damage images using three representative systems accessible to malicious users: Stable Diffusion 2, FLUX Kontext, and Qwen-VL Image Editing. These generators span different architectural families, enabling controlled comparisons of cross-model generalization.

We then train multiple detection models (different feature extractor with same label classifier), including a shallow CNN (3 layer) and a deeper ResNet-18 architecture, on combinations

of real and synthetic data. This enables us to measure baseline detection difficulty and understand how model capacity interacts with domain-specific cues. Generalization is evaluated by testing detectors on synthetic images produced by generators unseen during training, as well as through experiments involving multi-source joint training (such as training on SD2 and Kontext and evaluating on Qwen-VL). To further explore techniques for improving robustness, we incorporate adversarial domain adaptation approaches such as DANN, analyzing whether feature alignment helps mitigate overfitting across generator families.

This methodological design provides a coherent framework for assessing detectability, quantifying cross-model overfitting, and developing an interpretability approach suited for applications where trustworthiness and transparency are essential, such as insurance fraud detection.

Chapter 4

Dataset Collection

A reliable dataset is essential for evaluating the detectability of synthetic car damage images. This chapter describes the selection of generative models, the construction of the real-image baseline dataset, and the full pipeline used to collect and generate the data used in our experiments. Together, these components form a controlled yet realistic foundation for studying both in-domain detection and cross-model generalization.

4.1 Model Selection

To simulate realistic adversarial conditions, we select three modern open-source generative systems that represent distinct families of image-generation architectures. Our goal is to cover latent diffusion models, transformer-based flow-matching models, and multimodal editing diffusion systems. These families reflect the most relevant threat vectors in contemporary AI-misuse scenarios [18].

Stable Diffusion 2 (Latent Diffusion Model)

Stable Diffusion 2 (SD2) is built upon the Latent Diffusion Model (LDM) framework [19], where the diffusion process occurs in a compressed latent space. Images are encoded and decoded through a VAE, and denoising is performed by a U-Net trained using DDPM objectives [20]. This architecture achieves a strong balance between efficiency and fidelity. However, the VAE introduces characteristic reconstruction artifacts, which prior forensic studies have shown to be highly informative when differentiating synthetic images from real ones [21]. Because SD2 remains one of the most widely deployed open-source generators, it serves as a realistic and impactful adversarial model for our investigation.

FLUX.1 Kontext (Flow-Matching Diffusion Transformer)

Kontext belongs to the emerging class of rectified-flow and flow-matching models [22], which learn a continuous transport map between noise and images rather than relying on stepwise denoising. The model uses Diffusion Transformers (DiT) [23] and dual-stream conditioning to support high-resolution image synthesis and in-context editing. Within the open-source ecosystem, the FLUX.1 series represents one of the most capable and flexible transformer-based generators [24]. Flow-matching models tend to reduce traditional diffusion artifacts, making Kontext a challenging test case for passive forensic detection.

Qwen-VL / Qwen Image Edit (Multimodal Editing Diffusion)

Qwen-VL [25] integrates a ViT-based vision encoder [26] with a large language model to perform multimodal understanding and generation. The Qwen Image Edit extension [27] introduces a multimodal diffusion module and VAE-based reconstruction, enabling fine-grained, instruction-driven edits such as adding scratches, dents, or cracks. These targeted manipulations closely mirror realistic fraudulent modifications found in insurance submissions, making Qwen an essential model for studying localized tampering.

Rationale for Model Selection

These three systems collectively span the major paradigms of modern generative modeling, yielding diverse types of artifacts: SD2 introduces VAE-driven latent-space biases, Kontext reduces classical diffusion noise but introduces transformer-flow inconsistencies, and Qwen produces localized edit-induced anomalies. All three models are freely available and executable on consumer-grade hardware, aligning with realistic adversarial conditions. This selection allows us to analyze both detectability and cross-generator robustness in a comprehensive and technically relevant manner.

4.2 Baseline Datasets

High-quality real images are required to establish a reliable ground truth for training and evaluation. We explored several candidate datasets and curated the most suitable option according to resolution, realism, and annotation completeness.

Evaluation of Initial Dataset Candidates

Initial attempts using datasets such as CIFAR-10 and Fake-CIFAR proved unsuitable. Their extremely low resolution eliminates meaningful texture and damage detail, making diffusion artifacts nearly impossible to detect. Furthermore, their scene diversity and damage realism are limited, preventing meaningful forensic evaluation. We also considered scraping large collections of car images from platforms such as Copart.com. Although such images are realistic, they vary dramatically in quality and resolution, lack reliable annotations, and pose potential copyright concerns.

CarDD: The Final Baseline Dataset

After evaluating multiple options, we adopt the Car Damage Detection dataset (CarDD), the first public large-scale dataset specifically designed for car damage recognition and segmentation. CarDD contains more than four thousand high-resolution images and over nine thousand annotated damage instances spanning six major damage categories, including scratches, dents, cracks, glass damage, tire damage, and others. The images are collected through a standardized and carefully curated pipeline, resulting in high-quality, diverse, and reproducible data. This level of quality is crucial for distinguishing subtle generative artifacts and for evaluating detection performance under realistic conditions.

4.3 Dataset Pipeline

To study synthetic car-damage forgery under realistic adversarial conditions, we construct a hybrid dataset combining real images from CarDD with synthetic variants generated using

SD2, FLUX.1 Kontext, and Qwen Image Edit. For each real image, one synthetic counterpart is produced by each generative model, yielding a balanced dataset suitable for controlled comparisons across generators.

Unified Prompt Design

To ensure consistency across models, we design a unified prompt framework based on a taxonomy of damage descriptors spanning minor, moderate, and severe damage, as well as collision or vandalism-related alterations. These descriptors are embedded into short contextual templates such as "on this car, ..." or "modify the damaged parts of the car to have ..." to maintain alignment across different generation systems. This unified design ensures that differences in output reflect architectural properties of the models rather than inconsistencies in prompt formulation.

Synthetic Image Generation

All three generators are applied using the same prompt set, with model-specific adaptations only when absolutely necessary. SD2 tends to edit entire images when prompted directly, so we apply a car-body mask to constrain edits to relevant regions and produce realistic inpainting-style modifications. Kontext does not reliably support localized inpainting in batch settings, and therefore we apply global prompt-based editing. Qwen, despite its "Image Edit" designation, also applies changes globally in practice under similar settings and is thus treated in the same way. This approach ensures that observed differences between generators arise from their intrinsic mechanisms rather than extraneous configuration details.

Visual Examples

To mirror realistic adversarial misuse, we retain all synthetic samples produced by the generators without applying quality filtering or post-processing. Representative examples of SD2 inpainting, Kontext global editing, and Qwen-based modifications are included in Appendix A. These examples illustrate typical texture patterns, artifact structures, and damage characteristics produced by each system.

Final Assembly

All images are standardized to a resolution of 512×512 and partitioned into independent training, validation, and test sets with no overlap across splits. This dataset construction pipeline produces a controlled yet realistic benchmark for evaluating detectability, cross-model robustness, and interpretability in synthetic car-damage forensics.

Chapter 5

Experiment Results

5.1 Single-Domain Detection

We first evaluate baseline detection performance under a controlled setting in which training and testing are conducted within the same generative domain. All detectors are trained for 100 epochs under identical optimization settings, and the results are summarized in Tables 5.1, 5.2, and 5.3. Two dominant findings emerge from these experiments.

Resolution Has a Dominant Impact

Across both architectural choices—a shallow three-layer CNN and a ResNet-18—image resolution consistently proves to be the primary determinant of detection accuracy. When inputs are downsampled to 224×224, all models degrade to near-random performance, with accuracies typically ranging between 52% and 61% for SD2-only experiments, 54% to 63% for Kontext-only settings, and 52% to 58% in cross-model evaluations. This collapse occurs regardless of the generative domain or detector depth, suggesting that downsampling removes almost all meaningful forensic cues.

In stark contrast, when operating at 512×512, both architectures achieve substantial improvements. Even the shallow CNN reaches 93.85% accuracy on SD2 and 97.19% on Kontext, with ResNet-18 achieving similarly strong results. These findings indicate that the critical indicators of synthetic content—such as high-frequency noise patterns, VAE-driven distortions, and flow-matching inconsistencies—are almost entirely eliminated at low resolutions. Once images are reduced to 224×224, SD2 and Kontext outputs become nearly indistinguishable from real photographs, severely weakening detectability.

Model Depth Has Limited Influence

Although ResNet-18 generally outperforms the shallow CNN, the performance gap is small compared to the dramatic influence of image resolution. At 512×512, both models converge rapidly within 5–20 epochs and reach high detection accuracy, with the shallow CNN even matching or surpassing ResNet-18 in several cases. This pattern suggests that, within a narrow object-level domain such as car damage, architectural complexity is not the primary driver of performance. The decisive factor remains image fidelity rather than model depth.

These results have practical implications: lightweight CNNs may be sufficient for real-world deployment as long as high-resolution images can be collected or requested from users. Ensuring access to 512px images is therefore a more impactful design consideration than employing deeper or more complex neural architectures.

Table 5.1: Single-Domain Detection Performance (Shallow CNN)

Train Domain	Test Domain	Acc.	Prec.	Recall	F1
Resolution: 224×224					
SD2	SD2	60.96	0.6005	0.6551	0.6266
	Kontext	54.41	0.5460	0.5241	0.5348
	Qwen	54.58	0.5478	0.5256	0.5365
Kontext	SD2	54.28	0.5597	0.4011	0.4673
	Kontext	63.10	0.6467	0.5775	0.6102
	Qwen	58.22	0.6027	0.4825	0.5359
Resolution: 512×512					
SD2	SD2	93.85	0.9713	0.9037	0.9363
	Kontext	51.74	0.6970	0.0615	0.1130
	Qwen	52.56	0.7436	0.0782	0.1415
Kontext	SD2	52.01	0.7586	0.0588	0.1092
	Kontext	97.19	0.9809	0.9626	0.9717
	Qwen	92.05	0.9785	0.8598	0.9154

5.2 Domain-Invariant Features

We next examine whether detectors can learn domain-invariant forensic features that transfer across different generative models. To investigate this, we conduct six domain adaptation experiments by varying the source and target domains as well as the detector architecture, with results summarized in Tables 5.4 and 5.5. Two main findings emerge.

Detector Architecture Has Minimal Effect

In line with the single-domain results, architectural depth exerts only a modest influence on domain adaptation outcomes. Both the shallow CNN and ResNet-18 exhibit nearly identical transferability patterns: whenever a given source–target pairing is fundamentally incompatible in terms of forensic cues, neither architecture is able to learn meaningful domain-invariant representations, even under adversarial alignment via DANN. This observation reinforces the conclusion that the nature of the generative models themselves, rather than the complexity of the detector, determines which forensic cues can transfer and which cannot.

Dataset Choice Strongly Determines Transferability

The dominant factor in cross-domain performance is the combination of source and target generative models. Our experiments reveal pronounced asymmetries in transferability. The SD2→Kontext and Kontext→SD2 settings fail entirely: both architectures collapse to near-random predictions, and adversarial domain alignment offers no improvement. These failures suggest that SD2 and Kontext produce largely incompatible or non-overlapping forensic fingerprints.

In contrast, the Kontext→Qwen setting demonstrates strong positive transfer. A classifier trained solely on Kontext images already achieves substantial accuracy on Qwen-generated samples, and adversarial adaptation further improves this performance. This indicates that

Table 5.2: Single-Domain Detection Performance (ResNet-18)

Train Domain	Test Domain	Acc.	Prec.	Recall	F1
Resolution: 224×224					
SD2	SD2	58.96	0.5938	0.5668	0.5800
	Kontext	49.33	0.4912	0.3743	0.4249
	Qwen	49.06	0.4875	0.3693	0.4202
Kontext	SD2	51.87	0.5380	0.2647	0.3548
	Kontext	64.97	0.6986	0.5267	0.6006
	Qwen	55.93	0.6028	0.3477	0.4410
Resolution: 512×512					
SD2	SD2	97.46	0.9759	0.9733	0.9746
	Kontext	61.76	0.9151	0.2594	0.4042
	Qwen	66.58	0.9362	0.3558	0.5156
Kontext	SD2	52.14	0.8333	0.0535	0.1005
	Kontext	98.80	0.9893	0.9866	0.9880
	Qwen	93.67	0.9880	0.8841	0.9331

Table 5.3: Mixed-Domain Detection Performance

Train Domain	Test Domain	Acc.	Prec.	Recall	F1
Shallow CNN (512×512)					
SD2 + Kontext	SD2	90.78	0.9345	0.8770	0.9048
	Kontext	92.91	0.9373	0.9198	0.9285
	Qwen	88.27	0.9303	0.8275	0.8759
ResNet-18 (512×512)					
SD2 + Kontext	SD2	98.80	0.9765	1.0000	0.9881
	Kontext	98.66	0.9764	0.9973	0.9868
	Qwen	97.98	0.9759	0.9838	0.9799

Kontext and Qwen share partially aligned artifact distributions that enable meaningful domain-invariant learning.

Taken together, these results suggest that SD2 and Kontext reflect fundamentally different generative families with distinct forensic traces, whereas Kontext and Qwen exhibit substantial overlap in their artifacts.

5.3 Interpretability Gap

Although synthetic-image detectors can achieve strong performance under controlled settings, a major challenge remains: existing models offer limited transparency regarding *why* an image is classified as real or AI-generated. This interpretability gap is problematic for safety-critical applications, where human reviewers must be able to trust and justify algorithmic decisions. When synthetic images are visually indistinguishable from real ones, the absence of clear, stable, and human-understandable explanations becomes a fundamental limitation. This sec-

Table 5.4: Domain Adaptation Performance (DANN, Shallow CNN)

Source → Target	Test Domain	Acc.	Prec.	Recall	F1
SD2 → Kontext					
Shallow, 512	SD2	65.51	0.9203	0.3396	0.4961
	Kontext	50.94	0.6207	0.0481	0.0893
Kontext → SD2					
Shallow, 512	Kontext	58.16	0.5721	0.6471	0.6073
	SD2	50.27	0.5027	0.4893	0.4959
Kontext → Qwen					
Shallow, 512	Kontext	97.33	0.9538	0.9947	0.9738
	Qwen	91.78	0.9506	0.8814	0.9147

Table 5.5: Domain Adaptation Performance (DANN, ResNet-18)

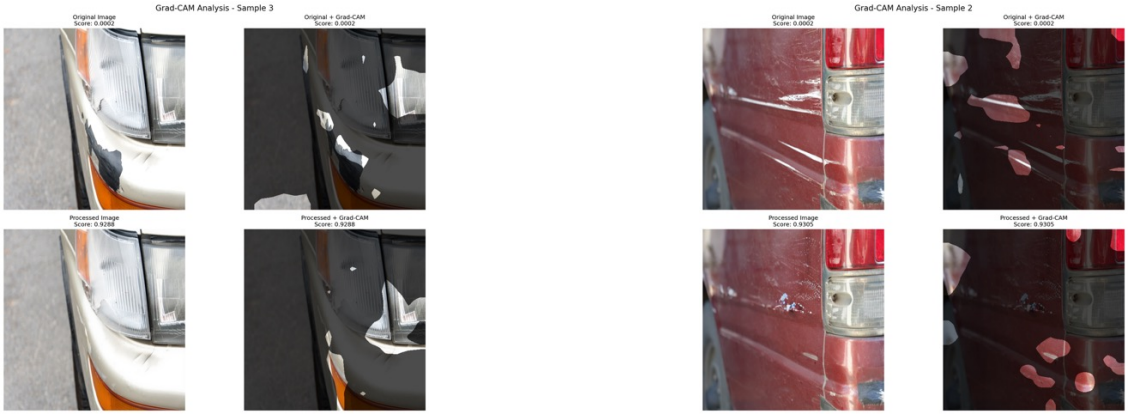
Source → Target	Test Domain	Acc.	Prec.	Recall	F1
SD2 → Kontext					
Deep, 512	SD2	88.50	0.8830	0.8877	0.8853
	Kontext	50.67	0.5269	0.1310	0.2099
Kontext → SD2					
Deep, 512	Kontext	54.28	0.5611	0.3930	0.4623
	SD2	50.67	0.5106	0.3209	0.3941
Kontext → Qwen					
Deep, 512	Kontext	96.66	0.9509	0.9840	0.9671
	Qwen	94.88	0.9488	0.9488	0.9488

tion reviews existing interpretability approaches, discusses our empirical observations, and motivates the need for more principled forensic-explanation frameworks.

5.3.1 Interpretability in Existing Literature

Most existing forensic methods prioritize classification or localization accuracy while paying relatively little attention to explaining model behavior. Classical interpretability tools such as Class Activation Maps (CAM) [28], the Network-in-Network global averaging mechanism [29], Grad-CAM [30], and early filter-visualization techniques [31] are sometimes used as diagnostics but rarely integrated into forensic pipelines. More advanced approaches, including guided backpropagation [32] and neuron-level concept analysis frameworks such as NetDissect [33] and its extensions [34], attempt to illustrate or label the concepts encoded in intermediate layers. Recent developments such as ForenX incorporate multimodal LLMs to verbalize forensic cues by combining localization heatmaps with linguistic reasoning. Automated neuron-discovery methods similarly aim to map feature channels to human-interpretable concepts.

Despite these advances, several limitations remain. Many interpretability tools were originally developed for semantic tasks like object recognition and thus do not align well with the low-level statistical cues used in forensic detection. Synthetic-image detectors often rely



(a) Kontext (b) SD2

Figure 5.1: Grad-CAM visualizations on Kontext and SD2 detectors. Both highlight modified regions without supervision, but the heatmaps remain fragmented and only partially interpretable.

on distributed, high-frequency, non-semantic features, which are inherently more difficult to interpret. Moreover, attribution maps can be unstable and may highlight spurious regions when models rely on subtle statistical anomalies rather than semantic structure. For these reasons, interpretability in forensic models remains an open and challenging research area.

5.3.2 Case Study: What the Detector Actually Looks At

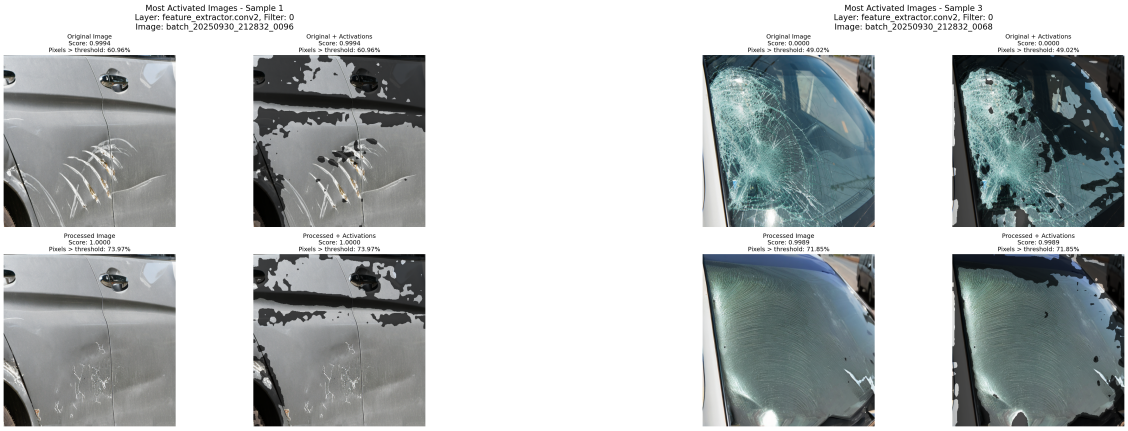
To better understand how our detector makes decisions, we analyze model behavior using two complementary approaches: image-level attribution with Grad-CAM and neuron-level analysis inspired by NetDissect. Together, these methods help reveal where the model focuses and what types of patterns individual filters respond to.

Grad-CAM: Localizing Model Attention

Figure 5.1 shows representative Grad-CAM visualizations for samples generated by SD2 and Kontext. Even without localization supervision, the heatmaps frequently highlight modified regions, such as added scratches, dents, or altered textures. However, the explanations remain limited. The highlighted regions often appear fragmented or diffuse, spreading across multiple unrelated areas. The intensity and structure of the heatmaps vary substantially across similar images, which undermines interpretability stability. Even when the correct region is localized, the underlying forensic cue—often a subtle statistical discrepancy—cannot easily be articulated by human reviewers. Thus, Grad-CAM provides useful spatial information but does not reveal the nature of the features being exploited.

NetDissect-Style Filter Analysis

To complement Grad-CAM, we analyze individual convolutional filters by identifying examples that maximally activate each unit and examining their activation maps. This allows us to observe the types of patterns each filter responds to across many images, as shown in Figure 5.2. Three consistent behaviors emerge. First, many filters show a strong preference for high-frequency micro-textures such as fine scratches, shattered glass patterns, or chipped paint, which differ subtly between real and generated images. Second, numerous units are sensitive



(a) Example 1 (b) Example 2

Figure 5.2: Neuron-level activation patterns for a selected filter, showing strong sensitivity to high-frequency textures and local structural discontinuities. Only pixels with top 1% are activated.

to material boundaries such as door seams, panel edges, and window contours; edited images often introduce slight inconsistencies in lighting or compression near these boundaries. Third, most discovered features are non-semantic: filters do not correspond to meaningful concepts like "door" or "wheel" but instead capture distributed statistical cues that are stable across samples yet difficult for humans to interpret.

Overall, the detector relies heavily on subtle low-level patterns rather than semantic understanding. While this enhances detection sensitivity, it complicates interpretability and underscores the need for more principled explanation frameworks.

5.3.3 Limitations of Current Approach

Our interpretability analysis sheds light on internal model behavior but also reveals fundamental limitations. Attribution maps such as Grad-CAM tend to be coarse and sometimes unstable, highlighting broad regions rather than precise forensic cues. The forensic signals themselves are usually non-semantic and cannot be easily verbalized, which undermines the usefulness of explanations for human reviewers. Finally, filter-level analysis methods like NetDissect rely on the assumption that a filter's role can be inferred from a finite set of activating examples. This assumption is restrictive, inherently sample-dependent, and non-scalable, offering only a partial view of how filters behave across the full input space. Thus, while such analyses provide valuable insights, they are insufficient for capturing the full reasoning dynamics of forensic detectors.

Chapter 6

Conclusions and Future Work

6.1 Discussion

This study examines whether synthetic car-damage images generated by diffusion-, flow-matching-, and multimodal-based models can be reliably detected, whether such detection generalizes across generators, and how interpretable the resulting detectors are. Through a curated dataset and a series of controlled experiments, we derive clear answers to all three research questions. The main findings are discussed below.

RQ1: Detectability in a Narrow Object-Level Domain

Our results demonstrate that detecting synthetic car-damage images is highly feasible when working within a single generative domain. All models—including SD2, Kontext, and Qwen—leave identifiable forensic traces that even a shallow CNN can reliably learn. High precision and recall are consistently achievable, suggesting that subtle high-frequency or error-level signals introduced during VAE reconstruction, denoising, or flow-matching processes serve as stable discriminative features.

However, this detectability is extremely sensitive to input resolution. Downsampling to 224×224 almost completely eliminates forensic cues, causing all detectors to collapse to near-random performance. This indicates that the relevant signals operate at fine frequency bands that vanish under low-resolution conditions. Practical forensic systems must therefore rely on high-quality inputs to preserve these delicate statistical differences.

RQ2: Cross-Model Generalization and Domain Adaptation

Our domain adaptation experiments reveal that cross-model generalization is achievable but fundamentally generator-dependent. When using Kontext as the source domain and Qwen as the target, substantial positive transfer emerges: a plain Kontext-trained classifier already performs strongly, and adversarial alignment improves performance further from 83% to 95%. This implies that generative models with similar architectural lineages—here, transformer-based modern systems—may share partially aligned artifact distributions that enable domain-invariant learning.

In contrast, attempts to transfer between SD2 and Kontext fail in both directions. Neither model architecture nor adversarial alignment helps. We hypothesize that this is due to architectural divergence: SD2's latent diffusion and VAE reconstruction pipelines differ sharply from Kontext's flow-matching transformer architecture, yielding non-overlapping forensic fingerprints. These findings highlight the structural nature of generative artifacts and suggest

that generalization will depend on understanding when two generators produce similar statistical traces.

RQ3: Interpretability and Forensic Transparency

Our interpretability analysis shows that detectors often attend to meaningful forensic cues, such as localized inconsistencies, high-frequency noise irregularities, and structural discontinuities. Grad-CAM visualizations and neuron-level analyses both support this conclusion. However, current interpretability tools remain limited. Attribution maps are frequently unstable, while filter-level analyses expose largely non-semantic cues that humans cannot easily articulate. These limitations underscore the need for more principled, domain-specific interpretability frameworks—especially for high-stakes environments such as insurance, auditing, or legal review.

6.2 Future Work

The results confirm that detecting AI-generated images is feasible and highlight methods that could strengthen generalizability and interpretability in real-world settings. Yet the limitations observed in our experiments also expose unresolved challenges, motivating the following directions for further inquiry.

Generalizability Across Generative Models

The divergent generalization outcomes observed between SD2, Kontext, and Qwen highlight the need to understand *when* and *why* forensic features transfer between models. Future work may explore theoretical and empirical characterizations of generator similarity, develop architectures that explicitly encourage generator-agnostic feature learning, or incorporate pretrained encoders such as CLIP. Because CLIP is trained on large-scale real-image distributions rather than synthetic fingerprints, its feature space may provide a more stable foundation for cross-model generalization.

Robustness Under Image Compression

Given the extreme sensitivity of detection performance to image resolution, a central open question concerns robustness under real-world degradations. Future research should systematically examine how compression, resizing, re-encoding, and noise reduction affect forensic cues. Techniques such as adversarial degradation training, feature-invariance regularization, or generative augmentation may help detectors remain reliable when images have undergone unknown transformations.

Advancing Interpretability in High-Stakes Settings

The interpretability gap identified in our analysis suggests a need for more reliable and transparent explanation tools. Future work may explore LLM-assisted interpretability pipelines that translate non-semantic statistical cues into human-understandable narratives, or adopt methodologies from mechanistic interpretability to analyze detectors at the neuron or circuit level. Building principled, faithful interpretability frameworks will be essential for deploying forensic detectors responsibly in high-stakes domains.

References

[1] J. Zhuo, P. Karamolegkos, I. Shumailov, N. Papernot, and B. Li, "Generative ai misuse: A taxonomy of tactics and insights from real-world data," *arXiv preprint arXiv:2406.13843*, 2024.

[2] P. Lorenz-Spreen *et al.*, "A technological perspective on misuse of available ai," *arXiv preprint arXiv:2403.15325*, 2024.

[3] M. Brundage, S. Avin, J. Clark *et al.*, "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[4] L. Carrieri, T. Florence, R. Heersmink *et al.*, "Protecting society from ai misuse: when are restrictions on capabilities warranted?" *AI & Society*, 2024.

[5] Z. Liu, Q. Liu, Y. Chen *et al.*, "Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[6] J. Hu *et al.*, "Span: Spatial pyramid attention network for image manipulation localization," *arXiv preprint arXiv:2009.00726*, 2020.

[7] T. Bianchi *et al.*, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Advances in Neural Information Processing Systems*, 2023.

[8] D. Afchar *et al.*, "Mmfusion: Combining image forensic filters for visual manipulation detection and localization," *arXiv preprint arXiv:2312.01790*, 2023.

[9] L. Corvi *et al.*, "On the detection of synthetic images generated by diffusion models," *arXiv preprint arXiv:2211.00680*, 2022.

[10] U. Ojha *et al.*, "Towards universal fake image detectors that generalize across generative models," *arXiv preprint arXiv:2302.10174*, 2023.

[11] A. Kamath *et al.*, "Leveraging representations from intermediate encoder blocks for synthetic image detection," *arXiv preprint arXiv:2402.19091*, 2024.

[12] Z. Luo *et al.*, "Treering: Robust image watermarking via neural networks," *arXiv preprint arXiv:2012.06454*, 2020.

[13] M. Tancik *et al.*, "Stegastamp: Invisible watermarking in images," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[14] Y. Kuo *et al.*, "Ai-generated image detection: Passive or watermark?" *arXiv preprint arXiv:2411.13553*, 2024.

[15] A. Rohit *et al.*, "Evading watermark-based detection of ai-generated content," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2023.

[16] Z. Li *et al.*, "Waves: Benchmarking the robustness of image watermarks," *arXiv preprint arXiv:2401.08573*, 2024.

[17] H. Li *et al.*, "Forenx: Towards explainable ai-generated image detection with multimodal large language models," *arXiv preprint arXiv:2405.XXXXX*, 2024.

[18] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, and et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CVPR*, 2022.

[20] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, 2020.

[21] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems*, 2021.

[22] R. Liu, C. Gong, B. Zhou, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022.

[23] W. Peebles and J.-Y. Xie, "Scalable diffusion models with transformers," *arXiv preprint arXiv:2212.09748*, 2023.

[24] B. F. Labs, "Flux.1 kontext: Flow matching for in-context image generation and editing in latent space," *arXiv preprint arXiv:2403.XXXX*, 2024, technical report; official model card on HuggingFace.

[25] Q. Team, "Qwen2.5-vl: A next-generation vision-language model," *arXiv preprint arXiv:2409.12168*, 2024.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[27] Q. Team, "Qwen image edit," HuggingFace Repository, 2024. [Online]. Available: https://huggingface.co/Qwen/Qwen-Image-Edit

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," *arXiv preprint arXiv:1512.04150*, 2016.

[29] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2014.

[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *arXiv preprint arXiv:1610.02391*, 2017.

[31] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *International Journal of Computer Vision*, vol. 120, pp. 233–255, 2016.

[32] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2015.

[33] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," 2017.

[34] C. Olah, S. Carter, and L. e. a. Schubert, "Understanding the role of individual units in a deep neural network," *arXiv preprint arXiv:2009.05041*, 2020.

Appendix A

Synthetic Image Samples

This appendix provides several representative examples of synthetic car-damage images generated by the three models used in our study: Stable Diffusion 2 (inpainting), FLUX.1 Kontext (prompt-based editing), and Qwen Image Edit (multimodal editing). These examples supplement the dataset pipeline, and illustrate the typical visual characteristics produced by each generator.

Note that no explicit quality filtering or manual correction was applied during dataset construction. The images shown here reflect the raw outputs under realistic adversarial conditions.

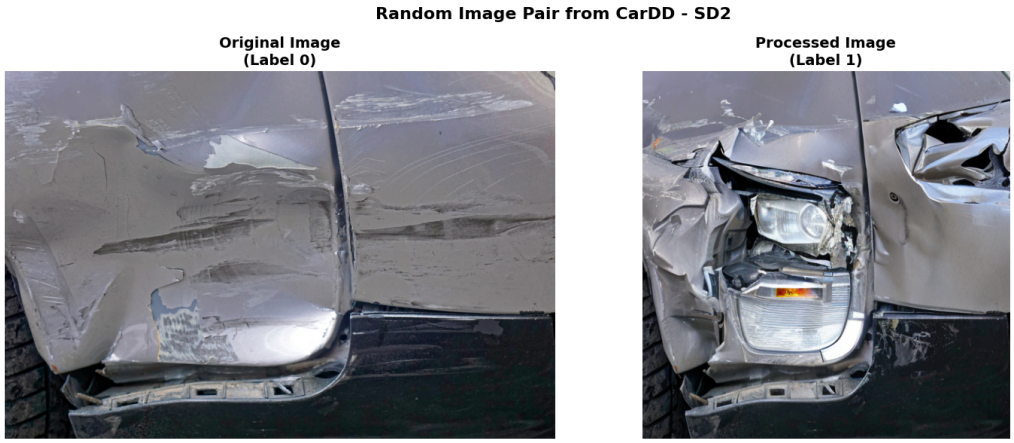


Figure A.1: **Stable Diffusion 2 (inpainting)**
Prompt: on the vehicle, collision damage, impact marks, crash damage
Negative Prompt: None.

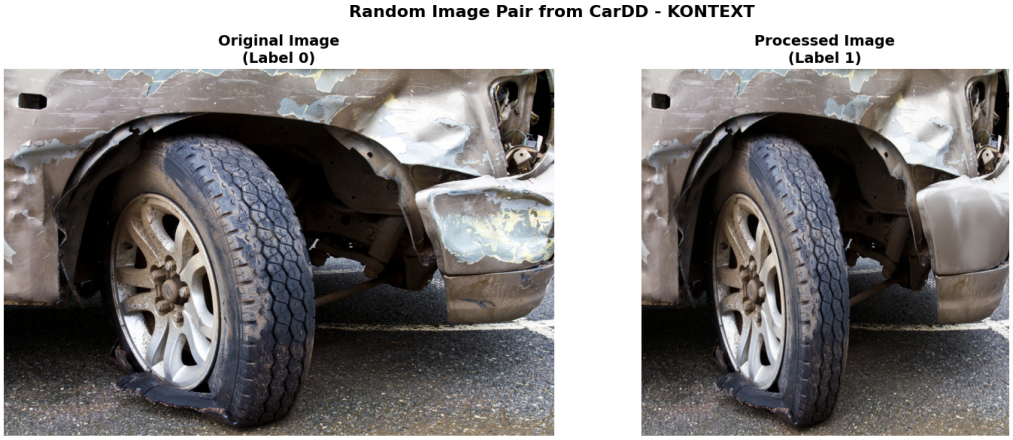


Figure A.2: **FLUX.1 Kontext**.
Prompt: on this damaged car, find the damaged parts of the car and modify them to have, medium-sized dent, noticeable indentation
Negative Prompt: no change to the undamaged parts of the car in the image, and should not keep the original damage

Figure A.3: **Qwen Image Edit**.
Prompt: on this damaged automobile, find the damaged parts of the car and modify them to have, minor scratch marks, light surface damage, subtle wear
Negative Prompt: no change to the undamaged parts of the car, and should not keep the original damage

Appendix B

Supplementary Code Repository

All code, experimental logs, dataset generation scripts, and trained model checkpoints used in this project are publicly available at the following repository:

ResponsibleAI — Synthetic Car Damage Detection
https://github.com/wjshku/ResponsibleAI

The repository includes:

- complete data pipeline for SD2, FLUX Kontext, and Qwen Image Edit generation,
- training scripts for shallow CNN and ResNet-18 architectures,
- cross-model evaluation and domain adaptation experiments,
- interpretability tools (Grad-CAM, activation visualization, neuron attribution),
- configuration files and model checkpoints for reproducibility.

This repository serves as the primary reference for reproducing all results reported in the main text.