A Benchmark & Dataset for Detecting AI-Manipulated Visual Evidence in the Court System

Kelly McConvey
Sajad Ebrahimi
Jalehsadat Mahdavimoghaddam
University of Toronto

Nima Jamali
Matina Mahdizadeh Sani
Wentao Zhang, Yuntian Deng
Maura Grossman
University of Waterloo

Maksym Taranukhin
Vered Shwartz
University of British Columbia

Jaquelyn Burkell
University of Western Ontario

Karen Eltis
University of Ottawa

Ebrahim Bagheri
University of Toronto

Abstract

Photographic evidence is becoming increasingly vulnerable to forms of alteration and fabrication that existing legal and technical workflows are not well equipped to evaluate. Surveillance frames, dashcam stills, and phone photographs may be used to establish presence, sequence, causation, damage, or identity, yet contemporary generative systems allow non-experts to alter or fabricate such images through ordinary prompt-based interfaces. Existing image-forensics benchmarks provide important resources for face manipulation, classical tampering, and general synthetic-image detection, but they are not organized around the forms of visual evidence submitted in courts, the localized edits that can change what an exhibit appears to prove, or the consumer-tool threat model now facing the justice system. We introduce the CIFAR Synthetic Evidence Corpus for Detecting AI-Manipulated Images, a benchmark for evidentiary image authentication in court and justice-system contexts. The corpus contains 1,505 photographic items, including 720 authentic controls and 785 manipulated or fabricated images, spanning surveillance, dashcam, and consumer-photo imagery. Manipulations are organized into scene-condition edits, localized element edits, and full fabrications produced with contemporary generative systems. Each item is released with structured metadata covering source provenance, manipulation tier, subtype, generator, prompt template, and scene attributes, enabling controlled evaluation beyond aggregate binary detection. We also establish state-of-the-art baselines with publicly available image-manipulation detectors, showing that current systems exhibit error profiles that remain problematic for evidentiary use. The dataset, prompts, metadata manifest, code, and baseline evaluation scripts are released to support research on visual evidence authentication, information integrity, and trustworthy AI for the justice system.

1 Introduction

Courts increasingly receive photographs whose authenticity can no longer be taken for granted, including surveillance frames offered to show that a person was present at a location, dashcam stills used to reconstruct traffic incidents, and phone photographs submitted to document damage, injury, possession, or the condition of a scene. In each case, the image is not merely visual content, but evidence whose value depends on whether the depicted scene can be trusted, whether relevant details have been altered, and whether the image can be situated within a reliable chain of provenance. This paper addresses that problem as one of evidentiary image authentication, where the task is to evaluate whether photographic evidence is authentic, manipulated, or synthetic under conditions that resemble legal and institutional use.

Photographs have never been self-authenticating records of reality, since courts have long required them to be connected to testimony, context, and procedure before they can function as evidence [12]. Historical accounts of photographic admissibility show that courts have repeatedly had to reconcile the apparent objectivity of photographs with the possibility of staging, selection, distortion, and later digital alteration [17]. The current challenge is therefore not that visual evidence has suddenly become uncertain, but that the practical conditions under which courts have managed that uncertainty have changed. Evidence that once required specialized skill to fabricate can now be altered or produced through widely available generative systems.

Fabricated and misleading evidence is not new, yet generative AI changes its cost, accessibility, and plausibility in ways that are directly relevant to evidentiary workflows. Earlier forms of photographic manipulation often required specialized tools, technical expertise, or substantial effort, whereas contemporary image-generation systems allow non-experts to produce plausible visual alterations through ordinary prompt-based interfaces. A litigant, witness, or third party can change scene conditions, insert or remove salient objects, or generate a plausible image from scratch without the expertise that previously constrained visual forgery. The same technological shift also creates a second evidentiary risk, since genuine evidence may be dismissed as fabricated precisely because fabrication itself has become credible. This is the dynamic Chesney and Citron [3] describe as the liar's dividend, and it places pressure on courts and related institutions not only to detect false evidence, but also to preserve confidence in authentic evidence.

Recent legal developments show that this concern has moved beyond speculation, as courts and judicial organizations are now confronting acknowledged and suspected AI-generated material, including fabricated legal authorities, synthetic media, and allegedly manipulated exhibits. In Mata v. Avianca [22], generative AI produced plausible but non-existent legal authorities that were submitted in litigation, illustrating how synthetic material can enter formal legal processes when verification fails. More directly for visual evidence, judicial and legal scholarship has begun to examine how courts should handle acknowledged and unacknowledged AI-generated evidence, while court-focused organizations have

warned that synthetic evidence may affect public trust in judicial proceedings [4]. These developments make evidentiary image authentication a timely problem for legal institutions and for computational researchers working on information integrity, provenance, verification, and trustworthy decision support.

The evidentiary risk is not limited to fully synthetic images, because in many legal settings the more consequential manipulation is a localized or contextual edit to an otherwise ordinary photograph. A person may be removed from a surveillance frame, a road sign or traffic signal may be changed in a dashcam still, or visible damage may be added to or removed from a phone photograph. Such edits can alter what an image appears to prove while preserving most of its visual content, which makes them difficult for lay observers [1, 19] and also difficult for detectors trained mainly on whole-image synthesis or face-centered deepfakes. For evidentiary authentication, the relevant question is therefore not simply whether an image was generated by AI, but whether verification systems can identify the forms of manipulation that change the informational and evidentiary meaning of a visual item.

Existing datasets do not adequately support this domain. Face-deepfake corpora have enabled important progress in synthetic-media detection, but they primarily address identity, expression, and facial manipulation rather than scene-level photographic evidence. General image-tampering datasets capture classical operations such as splicing, copy-move, and removal, but many predate prompt-driven consumer generation and do not reflect the artifacts, workflows, or editing patterns now available to non-experts. Other public resources provide authentic surveillance, driving, or consumer imagery, yet they do not provide controlled AI-generated manipulations, prompt metadata, generator metadata, or benchmark-ready labels. What is missing is a public resource for evaluating authentication systems on evidentiary image families, realistic manipulation types, consumer-interface generation, structured provenance, and detector behavior under conditions relevant to institutional use.

We introduce the CIFAR Synthetic Evidence Corpus for Detecting AI-Manipulated Images, a dataset and benchmark for evidentiary image authentication. The corpus contains 1,505 photographic items, including 720 authentic controls and 785 manipulated or fabricated images. It spans three families of visual evidence, namely surveillance imagery, dashcam stills, and consumer photographs. Manipulations are organized into three tiers covering scene-condition changes, localized element edits, and complete fabrication. The manipulated images are produced using contemporary generative systems from multiple organizations, and each item is paired with metadata describing its source provenance, evidentiary family, manipulation tier, generator, prompt template, and scene attributes.

This paper makes four contributions. First, it releases a benchmark corpus of authentic and AI-manipulated photographic evidence across surveillance, dashcam, and consumer-photo imagery. Second, it introduces a manipulation taxonomy grounded in evidentiary practice, with particular attention to localized edits that can alter the meaning of a visual exhibit while leaving most of the scene intact. Third, it provides a generation pipeline based on a realistic non-expert consumer-tool threat model. Fourth, it reports a zero-shot benchmark of off-the-shelf detectors, showing that current systems exhibit error profiles that remain problematic for evidentiary use. The corpus, manifest, and code are available at https://github.com/UofT-CIFAR/Synthetic-Evidence-Image-Corpus.

2 Distinction from Existing Resources

Our corpus is positioned not as a replacement for existing image-forensics benchmarks but as a complementary resource for a different evaluation setting. Prior resources have been essential for studying face manipulation, generic image tampering, and the robustness of synthetic-image detectors, yet they do not directly support evaluation under the evidentiary conditions that motivate this work. The distinction lies in the object of authentication, the manipulation model, the source imagery, and the kinds of metadata needed for controlled analysis.

Different object of authentication. Most widely used forgery benchmarks define authenticity around either faces or general natural images. Face-manipulation corpora such as FaceForensics++, Celeb-DF, DFDC, ForgeryNet, and OpenForensics [8, 14–16, 20] have enabled substantial progress, especially for identity replacement, expression manipulation, and facial reenactment. These resources remain highly valuable, but their central object is the face rather than the evidentiary scene. Our corpus instead treats the photograph as a visual claim about an event, location, object, or condition. The relevant question is therefore not only whether an image contains synthetic traces, but whether the information carried by a photographic exhibit has been altered in a way that changes what the image can be taken to prove.

Different manipulation model. General image-tampering benchmarks such as CASIA [9] and the NIST Media Forensics Challenge datasets [13] provide important coverage of splicing, copy-move, removal, and other classical manipulation operations. These datasets are closely related to the goals of media forensics, and OpenMFC in particular has legal-adjacent relevance. However, many of these resources reflect an editing paradigm that predates current consumer generative systems. The manipulation process in our corpus is organized around prompt-driven image editing and generation, where a non-expert can request a scene-level change, a localized element edit, or a complete fabrication through an ordinary interface. This matters because detectors may respond differently to artifacts produced by classical editing, GAN or diffusion synthesis, and contemporary instruction-following image models [2, 5]. Our benchmark therefore complements classical tampering resources by covering the consumer-interface threat model that now defines a growing portion of evidentiary risk.

Different evidentiary substrate. Several public datasets provide authentic imagery that resembles material encountered in institutional settings, including VIRAT and UCF-Crime for surveillance video [18, 21], BDD100K for driving scenes [23], and CASIA for consumer photographs [9]. These datasets are useful as source material, but they are not themselves authentication benchmarks for AI-generated evidentiary manipulation because they do not pair authentic controls with controlled synthetic edits, generation metadata, or manipulation labels. Our corpus builds on this ecosystem by converting such authentic sources into a benchmark organized around evidentiary families.

Different unit of analysis. Existing benchmarks often support aggregate detection performance, but evidentiary use requires more

Table 1: Manipulated image corpus by family, source, and manipulation tier. T1 alters scene conditions and T2 inserts or removes a salient element. T3 fabricates a complete image from a single family-specific prompt with no source frame.

Family	Source dataset(s)	T1 (scene conditions)		T2 (element edit)		T3	Total
Surveillance	VIRAT, UCF-Crime	Time of day	30	Swap object	28	45	**264**
		Weather	28	Remove person	26		
		Lighting	29	Remove signage	25		
		Crowd density	30	Add person	23		
Dashcam	BDD100K	Time of day	29	Remove sign	30	45	**255**
		Weather	27	Traffic-light color	27		
		Lighting	29	Remove vehicle	27		
		Crowd density	27	Add hazard	14		
Consumer photos	CASIA	Time of day	30	Add object	29	45	**266**
		Weather	30	Remove damage	28		
		Lighting	30	Swap object	25		
		Crowd density	30	Add damage	19		
Total			**349**		**301**	**135**	**785**

granular analysis. A detector that performs well on fully synthetic images may still fail on a localized edit that removes a person, changes a sign, alters weather, or modifies visible damage. Similarly, a detector with high overall accuracy may be unusable in an evidentiary workflow if it produces too many false positives on authentic images. Our corpus is therefore organized to support analysis by evidentiary family, manipulation tier, manipulation subtype, source dataset, generator, prompt template, and scene attributes. This structure allows the community to ask which kinds of evidentiary changes are detectable, which are missed, and which authentic images are wrongly impugned.

Relationship to legal scholarship. Legal scholarship has increasingly argued that courts need more robust ways to handle acknowledged and unacknowledged AI-generated evidence, including stronger gatekeeping, earlier resolution of authenticity disputes, and greater attention to the risks created by synthetic media [3, 6, 7, 11, 12]. That literature establishes the institutional problem, but it does not provide the empirical infrastructure needed to measure how authentication systems perform on photographic evidence of the kind courts may encounter. Our corpus addresses this missing empirical layer. It does not claim that detection alone can solve the legal problem, but it gives researchers a benchmark for evaluating the reliability, failure modes, and limits of detection systems in a setting aligned with evidentiary practice.

3 Resource Design and Construction

Our proposed corpus is designed for evidentiary image authentication rather than general synthetic-image detection. This setting requires a benchmark that represents the kinds of photographs submitted in legal and institutional contexts, includes manipulations that can change the evidentiary meaning of an image, reflects the consumer-tool threat model through which non-experts can now produce plausible edits, and provides metadata that supports controlled analysis beyond aggregate binary classification.

We operationalize these requirements along four axes. **First**, the corpus covers three evidentiary image families, namely surveillance, dashcam imagery, and consumer photographs. These families differ in viewpoint, compression, resolution, scene structure, and evidentiary function, so treating them as explicit benchmark dimensions allows detector performance to be measured under more realistic conditions. **Second**, the corpus distinguishes three manipulation tiers. Tier 1 changes scene conditions such as time of day, weather, lighting, or crowd density while preserving the main scene content. Tier 2 inserts, removes, or changes an evidentiary salient element such as a person, vehicle, sign, hazard, or visible damage. Tier 3 fabricates a complete image from a text prompt with no source photograph. Figure 1 shows representative before-and-after pairs: a Tier-1 weather change to a dashcam frame and a Tier-2 insertion of a person into a surveillance scene. **Third**, the corpus uses consumer-accessible generative systems from multiple organizations so that the benchmark does not depend on the artifacts of a single model or provider. **Fourth**, each artifact is released with metadata describing its source, evidentiary family, manipulation tier and subtype, generator, prompt, and scene attributes.

(a) Authentic source (b) T1: weather edit

(c) Authentic source (d) T2: person added

Figure 1: Example manipulations. *Top:* Tier-1 scene-condition edit adds snow to a dashcam frame (a→b). *Bottom:* Tier-2 localized edit inserts an additional person into a surveillance frame (c→d).

Table 2: Performance of off-the-shelf detectors.

Detector	Acc.	Prec.	Rec.	FPR	ROC-AUC	PR-AUC
D3	0.616	0.696	0.469	0.224	0.674	0.684
UNIVFD	0.573	0.686	0.336	0.168	0.643	0.633
DRCT	0.535	0.647	0.238	0.141	0.621	0.642
CNNSPOT	0.478	0.500	0.003	0.003	0.639	0.587

For constructing the corpus, we first defined a fixed set of prompt templates, each tied to a family, tier, and subtype. The templates were written to model a non-expert using a consumer image generation interface. Each prompt requests a single scoped change, asks the model to preserve all other scene content, and aims to produce a plausible evidentiary image rather than an arbitrary edit. The final prompt set contains 27 templates spanning the three families and three tiers.

We then constructed a source pool from public datasets matched to the evidentiary families. Surveillance images are drawn from VIRAT [18] and the normal-activity subset of UCF-Crime [21], dashcam stills from BDD100K [23], and consumer photographs from CASIA v2.0 [9]. Surveillance imagery is intentionally drawn from two sources so that detectors cannot rely on the visual signature of a single dataset as a proxy for authenticity. Tier 1 and Tier 2 items require authentic source frames; Tier 3 items are generated directly from prompts. For the source-based tiers, we computed quotas from the prompt templates and generation matrix, broken down by family, pool, tier, and required scene attributes. Candidate frames were then streamed from the source datasets. Up to three deterministic timestamps were sampled from each VIRAT and UCF-Crime video, one keyframe was sampled from each BDD100K clip, and each CASIA image was treated as one candidate frame. Each candidate was tagged with a local vision model that recorded general attributes such as time of day, weather, lighting, and number of people, together with family-specific attributes such as signage, traffic-light state, vehicle presence, hazards, and visible damage. A frame was retained only when its tags matched an unfilled quota. No source frame is reused anywhere in the corpus.

Manipulated items were generated by applying the appropriate prompt template to the assigned source frame for Tier 1 and Tier 2, or from the prompt alone for Tier 3. The main corpus uses GPT Image 2, Gemini 3.1 with Nano Banana 2, and Firefly 5. GPT Image 2 and Gemini 3.1 were accessed through APIs, while Firefly 5 was accessed through its consumer interface. The corpus also includes a smaller *Hero* set produced through iterative human-in-the-loop refinement in Photoshop and ChatGPT. The main variants model ordinary single-pass manipulation, while the Hero set models a higher-effort adversary willing to manually refine an image.

4 Dataset Content and Benchmarking

To establish reference points for future work, we evaluate publicly available image-manipulation detectors on the corpus in a zero-shot setting. Each detector scores every image once using its released weights and default decision threshold, with no fine-tuning, and we report both threshold-free metrics (ROC-AUC, PR-AUC), which rank detectors independently of any operating point, and threshold-dependent metrics (accuracy, precision, recall, FPR) at each detector's shipped default. The evaluation uses the DetectZoo framework [10] and treats the task as binary authentication over all 1,505 items. The detectors span released synthetic-image detectors: a diffusion-reconstruction method (D3), a CLIP-feature classifier (UNIVFD), a diffusion-reconstruction contrastive detector (DRCT), and a CNN-based GAN-artifact detector (CNNSPOT).

Across all systems, performance is low: the best detector, D3, reaches only 0.674 ROC-AUC, and the rest sit near chance (Table 2). This is the central benchmarking result—off-the-shelf detectors, trained predominantly on fully synthesized GAN or diffusion imagery, transfer poorly to the localized edits and consumer-tool manipulations that characterize evidentiary content. Default operating points are also poorly aligned with evidentiary use: at its shipped threshold D3 catches fewer than half of all manipulations (47% recall) yet wrongly flags 22% of authentic images as manipulated, and falsely impugning genuine evidence is the more damaging error in court. A benchmark for evidentiary authentication must therefore report this error structure, not only aggregate accuracy.

A breakdown of the strongest detector, D3, illustrates the diagnostic value of the corpus design. Performance is uneven in ways that track evidentiary structure rather than overall difficulty. By family, dashcam imagery is markedly harder than surveillance or consumer photographs (AUC 0.55 versus 0.69 and 0.78). By generator, detectability varies sharply: items from one consumer model are far less detectable than another's (per-generator recall ranging from 0.46 to 0.71), confirming that the corpus exposes cross-generator generalization gaps. The high-effort *Hero* set is the most revealing case: manually refined Photoshop edits receive extremely low manipulation scores, the lowest scoring at 0.04, and slip past the detector entirely, direct evidence that a motivated, hands-on adversary can defeat current systems. These findings position the corpus as a diagnostic benchmark for studying not just whether authentication systems succeed, but where they fail and which failure modes matter for evidentiary use.

5 Concluding Remarks

This paper introduces the CIFAR Synthetic Evidence Corpus for Detecting AI-Manipulated Images, a benchmark for studying evidentiary image authentication in court and justice-system contexts. This open resource addresses a gap between existing image-forensics benchmarks and the forms of photographic evidence that legal institutions increasingly need to assess, including surveillance, dashcam, and consumer-photo imagery manipulated through contemporary generative tools. By organizing the corpus around evidentiary families, manipulation tiers, generators, prompt templates, and source provenance, the dataset supports diagnostic evaluation rather than only aggregate detection. The state-of-the-art baselines show that current detectors remain uneven in ways that matter for evidentiary use, particularly around localized edits and false-positive behavior on authentic images. By releasing the data, prompts, manifest, code, and benchmark scripts, we aim to provide a shared resource for research on visual evidence authentication, information integrity, and trustworthy AI in the justice system.

6 Acknowledgments

This work was supported by the *Safeguarding Courts from Synthetic AI Content* Solution Network, funded through the Canadian AI Safety Institute (CAISI) Research Program at CIFAR.

7 GenAI Usage Disclosure

The generative models central to this work (GPT Image 2 (OpenAI), Gemini 3.1 with Nano Banana 2 (Google), and Firefly 5 (Adobe)) were used to construct the manipulated portion of the corpus, as described in Section 3. This use is the subject of the resource itself rather than an authoring aid, and all generated artifacts are released with the provenance metadata documented above.

In preparing the manuscript, the authors used Claude to assist with improving the clarity of author-written prose and generating draft LaTeX for tables. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the content of the paper. No AI system is listed as an author.

References

[1] Sergi D Bray, Shane D Johnson, and Bennett Kleinberg. 2023. Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity* 9, 1 (Jan. 2023), tyad011. https://doi.org/10.1093/cybsec/tyad011

[2] Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. 2025. Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. https://doi.org/10.48550/arXiv.2503.02857 arXiv:2503.02857 [cs] version: 1.

[3] Robert Chesney and Danielle Keats Citron. 2018. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. https://doi.org/10.2139/ssrn.3213954

[4] Connor Heaton, Shay Cleary, and Michael Navin. 2026. AI-generated evidence is a threat to public trust in the courts | National Center for State Courts. *National Center for State Courts* (Feb. 2026). https://www.ncsc.org/resources-courts/ai-generated-evidence-threat-public-trust-courts

[5] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2022. On the detection of synthetic images generated by diffusion models. https://doi.org/10.48550/arXiv.2211.00680 arXiv:2211.00680 [cs].

[6] Abhishek Dalal, Chongyang Gao, Hon Paul Grimm, Maura Grossman, Daniel Linna Jr, Chiara Pulice, V. S. Subrahmanian, and Hon John Tunheim. 2025. Deepfakes in Court: How Judges Can Proactively Manage Alleged AI-Generated Material in National Security Cases. *University of Chicago Legal Forum* 2024, 1 (Jan. 2025). https://chicagounbound.uchicago.edu/uclf/vol2024/iss1/3

[7] Rebecca Delfino. 2023. Deepfakes on Trial: A Call To Expand the Trial Judge's Gatekeeping Role To Protect Legal Proceedings from Technological Fakery. *UC Law Journal* 74, 2 (Feb. 2023), 293. https://repository.uclawsf.edu/hastings_law_journal/vol74/iss2/3

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. https://doi.org/10.48550/arXiv.2006.07397 arXiv:2006.07397 [cs].

[9] Jing Dong, Wei Wang, and Tieniu Tan. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. 422–426. https://doi.org/10.1109/ChinaSIP.2013.6625374

[10] Sajad Ebrahimi, Nima Jamali, Bardia Shirsalimian, Kelly McConvey, Wentao Zhang, Jalehsadat Mahdavimoghaddam, Maksym Taranukhin, Maura Grossman, Vered Shwartz, Yuntian Deng, and Ebrahim Bagheri. 2026. DetectZoo: A Unified Toolkit for AI-Generated Content Detection Across Text, Audio, and Image Modalities. https://doi.org/10.48550/arXiv.2606.04205 arXiv:2606.04205 [cs.MM].

[11] Paul Grimm, Maura Grossman, and Gordon Cormack. 2021. Artificial Intelligence as Evidence. *Northwestern Journal of Technology and Intellectual Property* 19, 1 (Dec. 2021), 9. https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss1/2

[12] Maura Grossman and Paul Grimm. 2025. Judicial Approaches to Acknowledged and Unacknowledged AI-Generated Evidence. *Science and Technology Law Review* 26, 2 (May 2025). https://doi.org/10.52214/stlr.v26i2.13890

[13] Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy Yates, Andrew Delgado, Daniel F. Zhou, Timothée N. Kheyrkhah, Jeff Smith, and Jonathan G. Fiscus. 2019. MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation. *NIST* (Jan. 2019). https://www.nist.gov/publications/mfc-datasets-large-scale-benchmark-datasets-media-forensic-challenge-evaluation Last Modified: 2019-12-31T06:12-05:00.

[14] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. https://doi.org/10.48550/arXiv.2103.05630 arXiv:2103.05630 [cs].

[15] Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2021. OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Montreal, QC, 10097–10107. https://doi.org/10.1109/ICCV48922.2021.00996

[16] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. https://doi.org/10.48550/arXiv.1909.12962 arXiv:1909.12962 [cs].

[17] Jennifer Mnookin. 1998. The Image of Truth: Photographic Evidence and the Power of Analogy. https://papers.ssrn.com/abstract=135311

[18] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, J. K. Aggarwal, Hyungtae Lee, Larry Davis, Eran Swears, Xioyang Wang, Qiang Ji, Kishore Reddy, Mubarak Shah, Carl Vondrick, Hamed Pirsiavash, Deva Ramanan, Jenny Yuen, Antonio Torralba, Bi Song, Anesco Fong, Amit Roy-Chowdhury, and Mita Desai. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*. CVPR 2011, Colorado Springs, Colorado, 3153–3160. https://doi.org/10.1109/CVPR.2011.5995586 ISSN: 1063-6919.

[19] Thomas Roca, Anthony Cintron Roman, Jehú Torres Vega, Marcelo Duarte, Pengce Wang, Kevin White, Amit Misra, and Juan Lavista Ferres. 2025. How good are humans at detecting AI-generated images? Learnings from an experiment. https://doi.org/10.48550/arXiv.2507.18640 arXiv:2507.18640 [cs] version: 1.

[20] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. https://doi.org/10.48550/arXiv.1901.08971 arXiv:1901.08971 [cs].

[21] Waqas Sultani, Chen Chen, and Mubarak Shah. 2019. Real-world Anomaly Detection in Surveillance Videos. https://doi.org/10.48550/arXiv.1801.04264 arXiv:1801.04264 [cs].

[22] William A. Ryan and Allen Garrett. 2023. Practical Lessons from the Attorney AI Missteps in Mata v. Avianca. *Association of Corporate Counsel* (Aug. 2023). https://www.acc.com/resource-library/practical-lessons-attorney-ai-missteps-mata-v-avianca

[23] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. https://doi.org/10.48550/arXiv.1805.04687 arXiv:1805.04687 [cs].