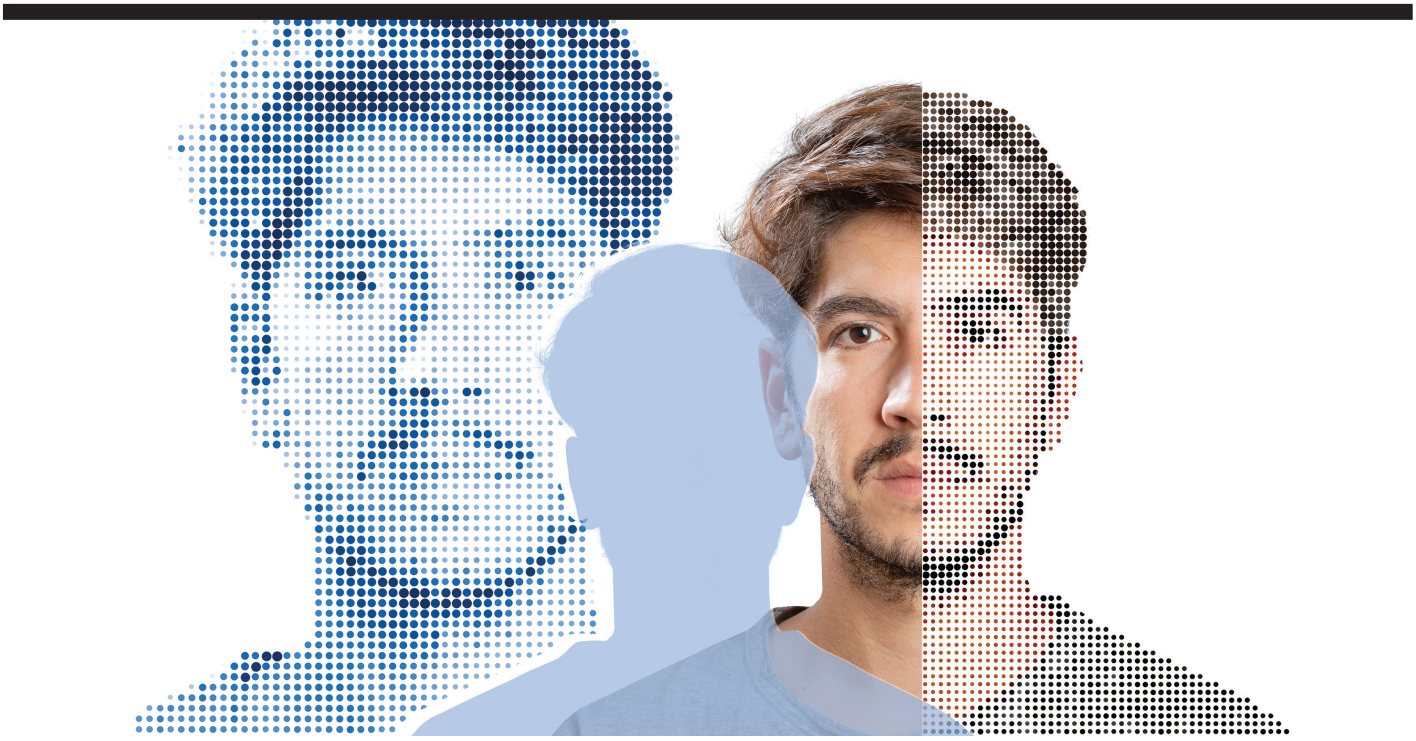
Yuxin (Myles) Liu, Habiba Farrukh, Ardalan Amiri Sani *University of California, Irvine*
Sharad Agarwal *Microsoft* **Gene Tsudik** *University of California, Irvine*



Scoop: Mitigation of Recapture Attacks on Provenance-Based Media Authentication

Excerpted from "Scoop: mitigation of recapture attacks on provenance-based media authentication," from *SEC '25: Proceedings of the 34th USENIX Conference on Security Symposium* with permission. https://dl.acm.org/doi/10.5555/3766078.3766299 ©ACM 2025

Today, digital media is constantly produced and consumed in enormous volumes. We rely heavily on smartphone images and videos from daily social sharing and entertainment to critical tasks, such as verifying a new Uber driver's identity, online banking operations, or providing evidence in legal proceedings. However, continuous advances in digital media manipulation, especially with the introduction of generative AI, yield increasingly sophisticated deepfakes [14]. This poses a massive threat to society, facilitating the spread of fake news, misinformation, and personal slander that greatly endanger our perception of reality.

Restoring trust in visual content has immense societal benefits, ensuring that organizations, institutions, and individuals can once again safely rely on the digital media they consume, restoring the principle of "seeing is believing." A good solution must provide a reliable way to verify where, when, and how a piece of media was created, rather than relying solely on deepfake detection algorithms, which is unfortunately shaping up to be a never-ending arms race.

To this end, an alternative promising direction (and the line of work we have pursued) is provenance assertion [2, 3]. This approach blends hardware-based secure camera designs with cryptography to authenticate the source of visual content and any post-processing (e.g., filters) applied to it. By leveraging secure components like a Trusted Execution Environment (TEE) [4] and specialized hardware like a secure camera module [5], provenance-based techniques aim to provide strong security guarantee for both generation and subsequent benign modification of visual content. Many prominent industry players from diverse sectors, including Canon, Adobe, BBC, and Microsoft, have already committed to or commercialized provenance-based media authentication techniques [2]. For example, Truepic teamed up with Qualcomm to provide provenance-

Illustration, istockphoto.com

based media authentication for smartphones with Qualcomm Snapdragon SoCs [1], and Adobe recently integrated content authenticity into its creative tools [2].

However, this promising defense has a critical, physical blind spot: recapture attack. In a recapture attack, an adversary simply displays maliciously fabricated content on a digital screen (or prints it) and takes a picture or video of that physical medium using a secure, provenance-asserting camera.

This attack is incredibly hard to defend against because existing provenance techniques only protect the pipeline inside the device, from the camera sensor to the consumption display. The camera faithfully authenticates exactly what its sensor "sees," while being completely unaware that it is looking at a high-definition TV rather than a real-world scene. Figure 1 demonstrates the effectiveness of this attack, showing how a fake, manipulated photo displayed on a TV and recaptured by a smartphone easily passes as real.

To address this critical vulnerability, we introduce *Scoop*, a systematic solution designed to mitigate recapture attacks. Scoop bridges this physical blind spot by leveraging state-of-the-art hardware depth sensing technologies alongside learning-based depth estimation to detect misleading recaptures. While a camera's traditional RGB sensor might be easily fooled by a 2D screen displaying a 3D scene, a Time-of-Flight (ToF) depth sensor measures the true, flat physical distance between the camera and the objects in front of it. By comparing the real hardware-generated depth map against an AI-generated perceived depth map, Scoop can effectively detect when a supposedly 3D scene is actually being displayed on a flat surface, highlighting the discrepancy for the viewer.

FIGURE 1. Demonstration of effectiveness of recapture attacks. One of these photos is real (i.e., captured by a camera in a real environment) and one is a recapture attack (i.e., captured by a camera pointed at a screen showing a modified photo). Even though we have used a mid-range TV, detecting the attack is very challenging (especially if the user is not suspecting an attack). *See answer at the end of this article.*

FIGURE 2. An example of a recapture attack on a provenance-based secure camera.

THE THREAT OF RECAPTURE ATTACKS

While recapture attacks are not a new concept and have historically targeted specific systems like face authentication [8, 9, 10, 11], they are now a potent threat against all modern provenance-based techniques. In such an attack, the adversary prepares some fake content, which it then either displays on a screen (e.g., TV, monitor, or projector) or prints/replicates on a physical surface (e.g., paper, cardboard, or canvas). Finally, the adversary captures this fake content with a camera that uses a provenance-based technique. Figure 2 demonstrates how a recapture attack is formed on a provenance-based secure camera.

The adversary's goal in a recapture attack is to present all captured content as real, including the fake components. Since a provenance-based camera cannot distinguish real content from the one that is displayed, painted, or printed, the entirety of captured content attains credibility as having been generated by a secure camera.

We distinguish among two types of recapture attacks:

- **Full-recapture attacks:** All contents (i.e., pixels) captured with the camera are recaptured, such as a camera pointing directly at a TV screen that captures a part of the screen, while the TV displays fake content
- **Partial-recapture attacks:** Captured contents contain both genuine and recaptured components, such as a camera pointing at a TV while a real person stands in front of the screen.

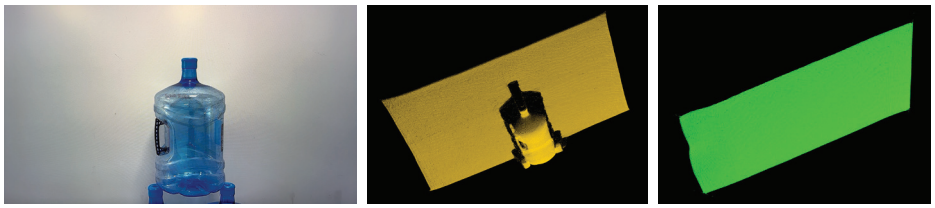
Figure 3 illustrates full-recapture and partial-recapture attacks. This could create a false sense of the person being physically present at the displayed location.

Furthermore, the presence of a display medium in a photo or video does not constitute an attack if the viewer can clearly recognize it (e.g., by seeing the physical bezels of a TV). When the display medium is not visually identifiable, we refer to the recapture as *misleading*, since the viewer might not realize the content is recaptured. For a misleading recapture to count as an attack, the content shown on the display medium must be crafted in order to fool viewers. Figure 4 illustrates this terminology: Malicious Recapture is a subset of Misleading Recapture, which is a subset of all Recaptures.

To demonstrate how incredibly effective these attacks are, we conducted a user study to see if human viewers could detect misleading recaptures. We showed 16 photos (8 original and 8 recaptured) to 43 adult participants in a random order. The results showed that original and recaptured photos are perceptually indistinguishable. The participants'

Full-Recapture Partial-Recapture

FIGURE 3. Illustration of (a) full-recapture and (b) partial-recapture attacks.



RGB Photo | Learning-Based Point Cloud | ToF-Based Point Cloud

FIGURE 5. Illustration of a learning-based depth estimation model got tricked by a photo displayed on a TV, while a ToF sensor captured depth map correctly represents the depth of the real scene.

correct classification accuracy was 50.15%, which is close to pure chance at 50%. Also, most participants stated they did not have any confidence when making their decisions.

Scoop: USING DEPTH TO DEFEAT RECAPTURE ATTACKS

To mitigate recapture attacks, we introduce Scoop, which leverages state-of-the-art depth sensing technologies as well as learning-based depth estimation to detect misleading recaptures, i.e., a recaptured photo or video where the presence of a display medium is not visually identifiable.

Our first key idea to detect recaptures is to use a Time-of-Flight (ToF) depth sensor, available in some modern smartphones (e.g., newer iPhones and some Android devices) and cameras, to capture depth information of the scene. A ToF sensor measures the physical distance between the camera lens and objects in front of it, providing a precise depth map. Because a ToF sensor detects the light it emits, it works well under low-light conditions.

There are two main types of ToF sensors [15]:

- **Direct Time-of-Flight (dToF):** Works by directly counting the time difference between the time the sensor emits light and the time it receives that light (e.g., Apple's LiDAR sensor).
- **Indirect Time-of-Flight (iToF):** Works by comparing the phase of emitted light and received light to derive the distance value (e.g., Samsung Galaxy S20 Plus).

However, mere use of depth information to mitigate recapture attacks faces an important problem: the existence of many naturally flat surfaces, such as walls. Scoop's goal is to detect display mediums, and not such naturally flat surfaces.

To address this challenge, Scoop computes a *learning-based depth map* of the scene, which yields the perceived visual depth in the same photo. Monocular depth estimation [6, 7] uses machine learning to determine the distance to each pixel in a photo using only a single RGB image. Recent advances allow achieving centimeter-level accuracy.

However, just like human beings, these models can be tricked easily by recapture attacks. When a 2D RGB photo is displayed on a TV screen without the screen frame being visible, state-of-the-art depth estimation models calculate depth based on the content displayed on the screen, creating an illusion of 3D depth. While monocular depth estimation models fail to recognize a recapture attack, the hardware ToF depth sensor recognizes that there is no depth in the scene, because it measures the actual physical distance between the camera and each point in the scene.

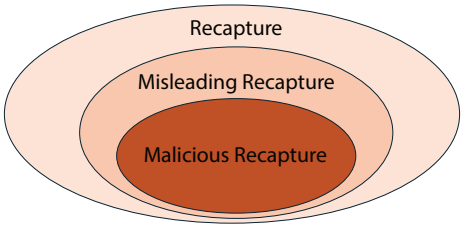


FIGURE 4. Terminology of recaptured visual content.

Figure 5 illustrates this discrepancy with depth maps generated using a state-of-the-art monocular depth estimation model and a ToF sensor. By comparing the depth information generated by the learning-based method with the ToF sensor, Scoop detects content that has visual depth information while being actually shown on a flat display medium.

SYSTEM DESIGN

Scoop is designed to be seamlessly integrated with both current camera pipelines and provenance-based media authentication systems. The camera captures depth information with its ToF sensor alongside ordinary RGB information and embeds it as part of the final captured content's provenance information.

Figure 6 shows the entire workflow of the Scoop viewer. During content playback on a Scoop-compatible viewer, the system generates two point clouds for each photo or video frame: a ground truth point cloud based on the ToF sensor and a learning-based point cloud based on the estimation model using RGB information.

Comparing these two clouds presents a significant technical hurdle. Although ToF sensors produce highly precise absolute depth maps, their resolution is relatively low, and they are prone to noise. Conversely, although learning-based point clouds have high resolution, their absolute precision is relatively low compared to ToF sensors. Because of this, Scoop does not blindly

compare absolute depth values across all pixels. Instead, it compares them by regions.

After applying noise reduction, Scoop uses a region growing algorithm to perform cluster segmentation on the ground truth point cloud to separate different surfaces. It then correlates each ground truth region with its pixel-level counterpart in the learning-based point cloud. Scoop employs three progressive comparison techniques to find mismatches:

1. **Direct comparison:** Scoop performs cluster segmentation on the learning-based counterpart to check if it yields only one region. If so, it means that the learning-based point cloud agrees with the ground truth.
2. **Deviation correction:** If a direct comparison fails, Scoop relaxes the standard to examine if the biggest sub-region covers more than a certain threshold (e.g., 90%) of the entire region being checked.
3. **Transformation:** As a final attempt, Scoop uses the iterative closest point (ICP) [13] and clustered viewpoint feature histogram (CVFH) [12] algorithms. ICP tries to register one point cloud in another by transforming depth points, while CVFH compares feature descriptors of the point clouds' geometric properties to ensure ICP does not get stuck in local minima.

If any mismatch is found between the two point clouds after applying these techniques, the content viewer flags the region as suspicious and highlights the mismatch to warn the users. Figure 7 illustrates this process with an example.

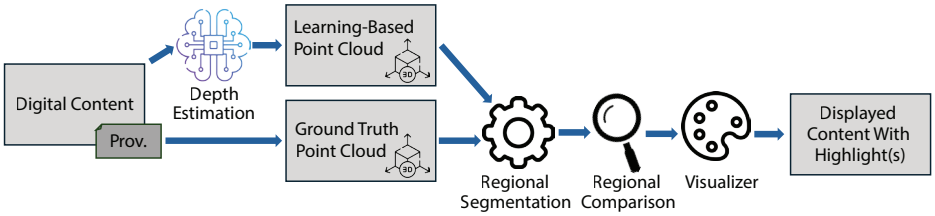


FIGURE 6. Scoop viewer's workflow.

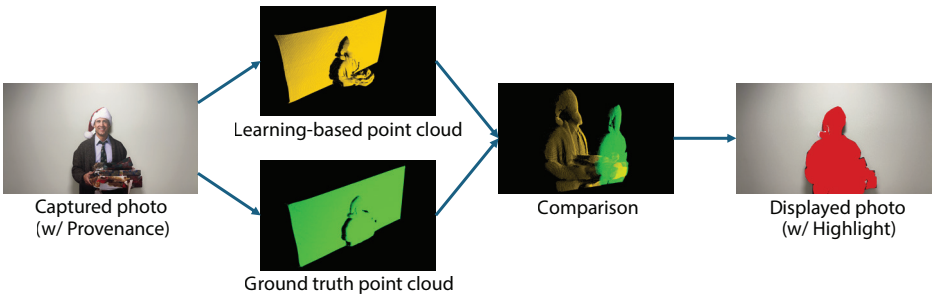


FIGURE 7. Demonstration of Scoop viewer prototype's workflow with a sample photo containing a life-size cardboard cutout of a celebrity.

REAL-WORLD EVALUATION

Prototype & Dataset

We implemented a complete prototype of Scoop. The producer (camera) side runs on both iOS (Apple iPhone 14 Pro with dToF) and Android (Samsung Galaxy S20 Plus with iToF) platforms. The consumer (viewer) side runs on a Linux desktop, utilizing the state-of-the-art *depth-pro* [6] model with PyTorch and CUDA for learning-based depth estimation.

To thoroughly evaluate the efficacy and practicality of Scoop, we constructed a first-of-its-kind dataset containing 122 unique data points, including 78 recapture scenarios and 44 benign, ordinary daily scenes. This dataset covers diverse content categories including background complexity (plain, textured, complex), objects (items, humans, animals), and lighting (very dark to well lit). To test the system to its limits, the recapture scenarios utilized diverse display mediums, including an 85-inch LCD TV, a 77-inch OLED TV, a 65-inch Mini-LED TV, an Epson laser projector, and life-size cardboard cutouts. Figure 8 demonstrates three example photos for each flat surface category.

Real-World Effectiveness

Evaluation results show that the iPhone 14 Pro (dToF) prototype achieves exceptional overall results with a 94.81% true positive rate (TPR) and only 0.02% false positive rate (FPR). The Galaxy S20 Plus (iToF) prototype achieves a solid 74.03% TPR and 17.78% FPR.

We observed that the Android prototype struggles more in complex scenes, particularly those with reflective surfaces (e.g., LG OLED TV screen) or scenarios that allow light to bounce back and forth in different paths. This aligns with industry understanding that dToF sensors generally perform better under complex lighting and long distances, whereas iToF sensors tend to suffer more from multi-path interference noise [15]. Comparing these results to our user study, where humans achieved only 44.19% TPR and 56.1% FPR on a comparable subset, Scoop proves to be much more capable of recognizing misleading recaptures than human beings.

Overheads

Scoop introduces computational overhead in three main places: camera capture, learning-based depth estimation, and viewer analysis.

- **Capture Performance:** At camera capture time, no human-noticeable runtime overhead was observed. Using the Android Power Profiler, we found that capturing photos with Scoop incurred an energy consumption overhead of 56.2%, and 38.07% during video capture. Storage overhead for appending the depth map alongside the RGB photo was approximately 648 KB (about 26.9%) for the iOS prototype, and 125 KB (about 10.4%) for Android.
- **Viewer Performance:** Depth estimation on the viewer side takes on average 0.28 seconds for the iOS captured photo. The viewer analysis itself currently introduces the most runtime overhead (e.g., averaging 69.38 seconds for iOS), as our region-by-region clustering algorithms are not yet optimized for performance (i.e., currently only using a single CPU thread).

(1) Print (2) Screen (3) Projection

FIGURE 8. Examples from our dataset showing different display mediums used.

FUTURE CONSIDERATIONS AND CONCLUSION

While Scoop systematically addresses flat display-based recapture attacks, it opens new avenues for future research. For instance, attacks using 3D displays or life-size custom 3D-printed models remain out of scope for now, although Scoop's geometric transformation limits (Comparison technique 3) can provide a baseline defense. Future hardware improvements could incorporate binocular vision technology or mobile Forward Looking Infrared (FLIR) cameras to further augment depth and liveness detection beyond standard ToF limitations. Furthermore, expanding Scoop's support for videos by leveraging consecutive frame motion and hardware gyroscopes represents an important next step.

Visual content is an essential form of information consumption, and the stakes for visual authenticity have never been higher. We are moving toward a future where society must implicitly trust the video feeds from police body cameras for legal accountability, C-SPAN broadcasts for political transparency, and the continuous streams of visual data guiding autonomous robots and cars. If an adversary successfully executes a recapture attack against the camera feeding an autonomous vehicle, the consequences can extend far beyond digital misinformation, where they become immediate physical safety threats.

Fortunately, the industry is working towards standardizing trust. Initiatives like the Coalition for Content Provenance and Authenticity (C2PA) [3] are establishing robust cryptographic pipelines from the moment light hits the sensor to the moment it renders on a display. Yet, as the internal pipeline solidifies, the attack surface naturally shifts to the physical world directly in front of the lens. Existing provenance-based techniques completely fail to protect us here.

Scoop provides a practical means of closing this physical blind spot by blending hardware depth sensors with AI-driven depth perception. To further achieve mass adoption, the mobile systems community must push for the ubiquitous integration of affordable, high-resolution Time-of-Flight (ToF) sensors across all device tiers, standardize depth metadata within existing frameworks like C2PA, and optimize analysis pipelines for real-time mobile processing. As hardware improves and multi-modal AI becomes more sophisticated, systems like Scoop will evolve into indispensable tools for maintaining trust in our digital and physical realities. ■

Answer to The Figure: In Figure 1, the one on the right is a real photo. The left photo was displayed on our TCL Mini-LED TV, with the person in the photo digitally erased.

Yuxin (Myles) Liu is a PhD candidate at the University of California, Irvine. His research focuses on digital provenance, embedded and mobile systems, operating systems, with a particular emphasis on system and hardware-assisted security.

Habiba Farrukh is an Assistant Professor at the University of California, Irvine. Her research interests involve security and privacy, and mobile computing with a focus on designing secure systems for emerging computing platforms.

Ardalan Amiri Sani is a Professor at the University of California, Irvine. His research involves building trustworthy systems, and are often at the intersection of mobile computing, security, and operating systems.

Sharad Agarwal is a Senior Principal Researcher at Microsoft. His research focus is on AI + ML systems that solve large-scale cloud challenges. His latest work includes building and deploying production AI systems for automated SaaS management and AI model routing.

Gene Tsudik is a Distinguished and ICS Alumni Professor of Computer Science at the University of California, Irvine. His research focuses on privacy, computer & network security, and applied cryptography.

REFERENCES

[1] Truepic. 2020. "Truepic Breakthrough Charts a Path for Restoring Trust in Photos and Videos at Internet Scale." https://www.prnewswire.com/news-releases/truepic-breakthrough-charts-a-path-for-restoring-trust-in-photos-and-videos-at-internet-scale-301152998.html

[2] Content Authenticity Initiative. 2019. https://contentauthenticity.org/

[3] Coalition for Content Provenance and Authenticity. 2021. https://c2pa.org/

[4] Y. Liu, et al. 2022. Vronicle: Verifiable provenance for videos from mobile devices. *ACM MobiSys.*

[5] Y. Liu, et al. 2024. Provcam: A camera module with self-contained tcb for producing verifiable videos. *ACM MobiCom.*

[6] A. Bochkovskii, et al. 2024. Depth pro: Sharp monocular metric depth in less than a second. *arXiv.*

[7] M. Hu, et al. 2024. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv.*

[8] Z. Boulkenafet, et al. 2017. Oulu-npu: A mobile face presentation attack database with real-world variations. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG).*

[9] H. Farid. 2009. Image forgery detection. *IEEE Signal Processing Magazine.*

[10] H. Cao, et al. 2010. Identification of recaptured photographs on LED screens. 2010 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[11] H. Farrukh, et al. 2020. Facerevelio: a face liveness detection system for smartphones with a single front camera. *ACM MobiCom.*

[12] X. Han, et al. 2023. 3d point cloud descriptors: state-of-the-art. *Artificial Intelligence Review.*

[13] Iterative Closest Point Algorithm. 2024. https://cs.gmu.edu/~kosecka/cs685/cs685-icp.pdf

[14] Y. Li, et al. 2019. Exposing deepfake videos by detecting face warping artifacts. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[15] Comparison between dToF and iToF sensors. 2021. https://faster-than-light.net/TOFSystem_C4