EM Eye: Characterizing Electromagnetic Side-channel Eavesdropping on Embedded Cameras

Yan Long*§, Qinhong Jiang†§, Chen Yan†, Tobias Alam*, Xiaoyu Ji†, Wenyuan Xu†, Kevin Fu‡
*University of Michigan, †Zhejiang University, ‡Northeastern University
{yanlong, tobiasal}@umich.edu, {qhjiang, yanchen, xji, wyxu}@zju.edu.cn, k.fu@northeastern.edu

***Abstract*—IoT devices and other embedded systems are increasingly equipped with cameras that can sense critical information in private spaces. The data security of these cameras, however, has hardly been scrutinized from the hardware design perspective. Our paper presents the first attempt to analyze the attack surface of physical-channel eavesdropping on embedded cameras. We characterize EM Eye—a vulnerability in the digital image data transmission interface that allows adversaries to reconstruct high-quality image streams from the cameras' unintentional electromagnetic emissions, even from over 2 meters away in many cases. Our evaluations of 4 popular IoT camera development platforms and 12 commercial off-the-shelf devices with cameras show that EM Eye poses threats to a wide range of devices, from smartphones to dash cams and home security cameras. By exploiting this vulnerability, adversaries may be able to visually spy on private activities in an enclosed room from the other side of a wall. We provide root cause analysis and modeling that enable system defenders to identify and simulate mitigation against this vulnerability, such as improving embedded cameras' data transmission protocols with minimum costs. We further discuss EM Eye's relationship with known computer display eavesdropping attacks to reveal the gaps that need to be addressed to protect the data confidentiality of sensing systems.**

I. Introduction

Cameras, being one of the highest-entropy sensors, are becoming omnipresent even in private spaces. Recent advances in the miniaturization of semiconductor electronics have spurred the wide integration of cameras into various embedded and mobile systems ranging from smartphones to IoT gadgets such as smart locks and home monitors. For smart home security cameras alone, the number of families owning such devices is predicted to grow from 99 million to 180 million between 2023 and 2027 [37]. Given the near-universal adoption of embedded cameras and the critical information they could capture such as the private activities and personnel information in offices and households, it is imperative to prevent unauthorized access to camera data. While previous research examined the data eavesdropping vulnerabilities in networked IP cameras' software stack [3], [15], [23], [38], the hardware design of these embedded camera devices has not been scrutinized yet. To understand the threats more thoroughly, our work investigates a new dimension of the problem by asking *how may adversaries eavesdrop on camera data by exploiting the side-channel byproducts generated by the cameras' physical operations?*

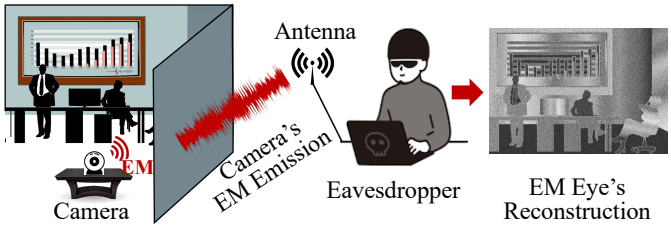


Fig. 1: Embedded cameras leak EM signals in operation, allowing eavesdroppers to visually spy on private spaces by reconstructing camera images.

Our work draws inspiration from recent works showing that embedded cameras' electromagnetic (EM) emissions allow people to detect the presence of cameras [25], [34], [41]. While these works simply use the existence of EM emissions as an on/off indicator of camera operations, essentially extracting a single bit of information, our work further investigates how much information of camera data is leaked from such EM emissions[1], and how adversaries can eavesdrop on the camera image streams by reconstructing synthesized images from the EM signals. Through experiments with the open-source Raspberry Pi camera, one of the most used embedded camera prototyping platforms, we observe highly predictable correlations between the EM emission patterns and the camera image contents. Nevertheless, mapping the 1D EM signals to 2D images is conceptually challenging without further knowledge of the EM generation process. Our investigations unveil that the primary EM leakage source is the digital image data transmission interface between the image sensor chips and the downstream image processing components. We carry out a detailed analysis of the physical layer of the embedded camera's data transmission interface. We find that RAW sensor data represented in bits are transmitted in a deterministic way following a frame-by-frame, row-by-row, and column-by-column order. By understanding the serialized data transmission scheme and reverse-engineering the transmission parameters, adversaries can directly generate eavesdropped image streams in real-time using portable equipment including an antenna, a software-defined radio receiver, and a laptop.

Despite the ability of direct image reconstructions, our experiments reveal additional challenges that limit adversaries'

§ Yan Long and Qinhong Jiang are co-first authors.

[1] Demo and tutorial are available at https://emeyeattack.github.io/Website/

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.24552
www.ndss-symposium.org

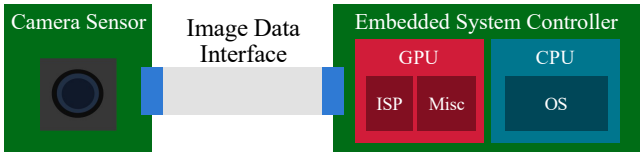


Fig. 2: The typical architecture of embedded camera systems.

capability of retrieving intelligible information from the reconstructions. For example, the eavesdropped images suffer from loss of colors and incorrect gray-scale values, as well as significant noise that causes degradation of image quality. We thus develop a model to characterize the physical leakage process of digital image transmission and analyze the root cause of these distortions. Our analysis shows that the limited EM signal bandwidths that could be afforded by adversaries in practical settings cause irreversible loss of image data structures in the EM signals, which then manifests itself as very structured distortions in the reconstructions. An adversary aiming to get a high-quality image then faces the challenge of partially recovering the data structures by leveraging their prior knowledge of the physical leakage process. To explore to what extent an adversary can achieve this, we develop an enhanced eavesdropping pipeline to strategically combine available EM signals and infer high-quality images using a supervised image-to-image translation network that learns the structured mappings between original and distorted images. We find the pipeline capable of removing most distortions, recovering authentic gray-scale images, and even producing colored images that well-approximate the camera scenes.

To examine the scope of EM Eye's risks, we conducted experiments with popular IoT camera development platforms including Raspberry Pi 3B+/4B, Nvidia Jetson Nano, Asus Tinkerboard 2S, and 12 commercial-off-of-the-shelf (COTS) devices with embedded cameras. With middle-end EM receiving equipment, our evaluations show that smartphone camera EM emissions could be received from up to 30 cm away, allowing adversaries to install low-profile hidden antennas to eavesdrop on smartphone photography. Dash cams and smart home cameras could be eavesdropped on from up to 5 m away, allowing adversaries to spy on physically-isolated spaces such as the interiors of cars, households, and offices through doors and walls as shown in Fig. 1. Our investigation of camera EM side-channel further uncovers the underlying physical vulnerability of unprotected image data baseband transmission. We note that this vulnerability is also shared by the well-known TEMPEST and acoustic side-channel eavesdropping attacks against computer displays. Despite the past 40 years of computer display eavesdropping research, our work shows that there still exists a semantic gap between the understanding of TEMPEST vulnerabilities and how modern sensors process and transmit data. Finally, we analyze how to protect embedded cameras by improving the data transmission protocols and discuss how future adversaries may apply the same eavesdropping methodology to other types of sensor data. We summarize our main contributions as follows:

- The characterization and modeling of the electromagnetic side-channel eavesdropping on embedded cameras. Our investigation bridges the gap between image data eavesdropping vulnerabilities and emerging sensor data transmission mechanisms.
- The analysis of the image distortion problems rooted in the physics of digital image transmissions and the design of an image reconstruction pipeline for improving image quality. The analysis and design methodology is reusable by side-channel image eavesdropping research.
- The evaluation results on 12 COTS devices as well as the lessons for mitigations gleaned from our evaluations. Our results aim to motivate researchers and manufacturers to systematically examine the side-channel eavesdropping risk on a wider range of sensor systems.

II. Background

A. Prior Work

This work builds upon the main hypothesis that the EM leakage of cameras is correlated with camera contents and can be used to infer or even reconstruct camera outputs. This hypothesis is motivated by recent research discoveries of the EM characteristics of embedded cameras. Several works have shown that smartphone cameras and hidden spy cameras produce EM emissions when they are turned on, allowing people to detect forbidden malicious operations of these cameras [25], [34], [41]. Essentially, these works only extract a single bit of entropy (on/off) from camera EM emissions. It also remains unclear how the EM emissions are generated by cameras. In the opposite direction, Jiang et al. [19] demonstrate the feasibility of injecting EM interference to partially control CMOS camera's outputs with an image row-level granularity; Köhler et al. [20] demonstrate a pixel-level injection granularity with Charge-Coupled Device (CCD) cameras which are less common in modern consumer electronics. Their results suggest there is significantly more entropy embedded in camera EM characteristics that can be harvested. Building upon these insights, our work seeks to characterize the feasibility, causality, and limits of eavesdropping on pixel-level information from the EM leakage of cameras in embedded systems.

B. Embedded Cameras

Embedded system devices are increasingly equipped with camera peripherals. Compared to traditional cameras such as digital single-lens reflex (DSLR) cameras, embedded cameras often feature open-standard designs that allow them to interface with a wide range of controllers. Fig. 2 shows the architecture of a typical embedded camera system. The camera's semiconductor image sensors convert photons hitting the semiconductors into proportional electrical signals. Each image sensor contains millions of sensing units corresponding to "pixels" in the digital image domain. The electrical signals are amplified, conditioned, digitized by analog-to-digital converters (ADCs), and transmitted to the computation units such as the image signal processor (ISP) in GPUs. The GPU then produces the final images after debayering (also known as demosaicing), image corrections, and miscellaneous post-processing. Like most sensor peripherals, embedded camera modules are often supplied by third parties and integrated by consumer electronics manufacturers.

RAW Images and Debayering. RAW images refer to the unprocessed data generated by image sensors. Since each semiconductor sensing unit only captures a single channel of RGB color that is selected by a color filter array, each pixel only has one color in the RAW images. To get a normal color image that users are familiar with which has all three RGB color channels, the ISP needs to perform a debayering step to interpolate the missing RGB channels for each pixel based on available colors from its neighbors [13].

Pixel Data Transmission. Image sensors and ISPs are connected by a pixel data transmission interface that transmits the RAW pixel data. Some examples of such interfaces include the High-speed Serial Pixel Interface (HiSPi) [5], the Digital Video Port (DVP) [16], the Low-voltage Differential Signaling (LVDS) [40], and the MIPI Camera Serial Interface 2 (MIPI CSI-2) [28]. MIPI CSI-2 has been widely adopted for its good usability, dedicated EM anti-interference designs, and capacity to support a variety of camera applications. It has become the de-facto standard for embedded cameras due to the rising demand for higher throughput and compatibility between hardware and software from different vendors. Same as most digital image transmission interfaces, MIPI CSI-2 transmits videos frame by frame. For each frame which is a 2D matrix, the camera transmits each row sequentially from top to bottom; for each row, each column (pixel) is also transmitted sequentially from left to right as shown by Fig. 3 (a). There often exists blanking between the transmission of consecutive frames and rows where the data transmission interface stays in an idle state without active transmissions. On the physical layer, MIPI CSI-2 uses high-speed differential signaling wires with up to four data lanes and a shared clock lane. Fig. 3 (b) demonstrates a MIPI CSI-2 interface with two data lanes.

III. Threat Model

We characterize the threat of passive eavesdropping on the confidential camera data of embedded systems by exploiting the unintentional EM emissions from camera sensors, the image data transmission interfaces, and image signal processors. The goal of the adversary is to reconstruct an image stream that approximates the authentic camera output as closely as possible. We assume the adversary uses a set of readily available commercial hardware equipment that is able to collect the EM emissions generated by the cameras. This often includes an antenna, a low-noise amplifier (LNA), a software-defined radio (SDR) device such as a USRP [10], and a laptop that runs the image reconstruction algorithms. We consider various camera-antenna distances and two corresponding categories of eavesdropping scenarios, namely the hidden-antenna (HA) and physical-isolation (PI) scenarios. In the former scenario, we assume the adversary manages to install a low-profile antenna near the target camera to receive stronger EM emissions. In the latter scenario, we assume the camera is located in a physically isolated space such as a private room and the adversary's antenna can only be placed outside the room to receive EM emissions through walls or doors. In both cases, the camera scenes contain private information that is supposed to be visible only to the legitimate camera owner.

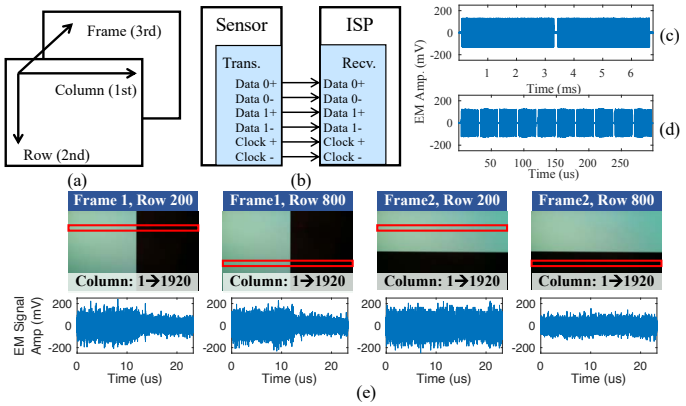


Fig. 3: How embedded cameras' operations generate EM signals that leak camera image information. (a) Each video frame is transmitted row by row and column by column. (b) The MIPI CSI-2 interface transmits image data with multiple lanes of differential data wires and clock wires, all generating EM leakage. (c) EM signals of two consecutive frames. (d) EM signals of ten consecutive rows. (e) EM signals of transmitting different frames, rows, and columns, showing clear correlations with the image contents.

IV. Optical EM Side Channels

Adversaries are able to eavesdrop on the camera images by analyzing the electromagnetic signals that are converted from the optical signals captured by the camera's image sensor. This section investigates the feasibility, model, and characteristics of these optical EM side channels.

A. Feasibility

We use a Raspberry Pi camera V1 (RPi V1) to record a computer monitor displaying two simplified black/white scenes. Fig. 16 shows the experiment setup of our feasibility tests. The top row of Fig. 3 (e) shows the two scenes recorded by the camera. Meanwhile, we collect the EM signals around the camera using a near-field EM probe connected to an oscilloscope. The camera records with a frame rate of 30 fps. At various center frequencies including different multiples of 51 MHz, we receive periodic signals at 30 Hz matching the camera frame rate. Fig. 3 (c) shows such signals at 204 MHz with two consecutive frames and blanking between them. We have confirmed that the received signals are from the camera instead of the computer monitor which has a refresh rate of 120 Hz. When zooming in, we can also see the transmission of different rows with blanking in between, as shown by Fig. 3 (d). Inspecting the EM signals corresponding to different frames, rows, and columns, we found obvious correlations between the shape of the EM signals and the pixel values of the camera image, as shown in Fig. 3 (e).

EM Leakage Source. To determine where the EM leakage comes from, we use a tiny near-field magnetic probe to collect the EM emissions from each component of the camera device while shielding the other components. We find that the EM signals have significantly better signal-to-noise ratios (SNRs) when the probe is placed near the image data transmission cable that connects the image sensor and downstream image

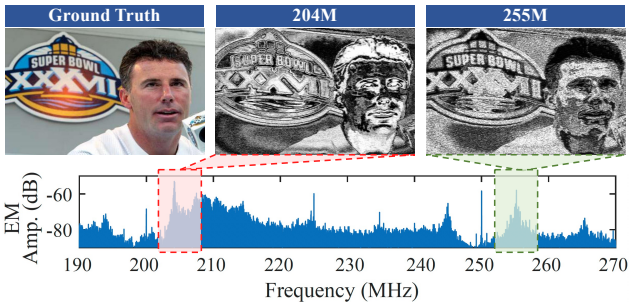


Fig. 4: Illustrations of EM emission's spectrum and two reconstructed images using signals around 204 and 255 MHz.

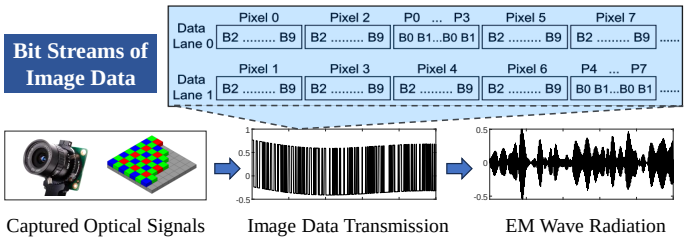


Fig. 5: The information flow of camera EM leakage. Optical signals captured by image sensors are converted to bit streams shown on the top. The transmission cables act as unintentional antennas that convert the bits into radiated EM waves.

processing components. We thus conclude that the cable for image data transmission is the main EM leakage source.

Basic Image Reconstruction. To reconstruct an image, the adversary needs to map the one-dimension EM signals received by the antenna within a certain frequency band back to a two-dimension matrix by associating each segment of the EM signals to specific pixels of the image. This requires the adversary to model key parameters including the pixel transmission rate, row transmission rate, image height and width, blanking periods, etc. The adversary then needs to convert 1D vectors of EM signals to scalar pixel values of the reconstructed image, essentially demodulating the EM signals that are modulated by the image contents. Since the EM emission process is an unintentional communication channel, we believe simpler modulation schemes such as amplitude and frequency modulation are more appropriate than other sophisticated man-made schemes. A closer look at the temporal-spectral variations of the EM signals reveals that only very wide-band and rapid variations exist in the frequency components of the emissions, which could require a GHz-level sampling bandwidth to provide sufficient coverage and is thus not feasible. We thus hypothesize that amplitude demodulation is the most appropriate method based on our observations in Fig. 3 and use the amplitudes of EM signals as the gray-scale values of the pixels. We denote this reconstruction process as $\mathcal{R}_{base}$ and provide further details in Section V-A. With $\mathcal{R}_{base}$, we are able to reconstruct images that share very similar structures as the camera ground truths in real time. Fig. 4 provides an example of such reconstructed images and the spectrum of the corresponding EM signals.

B. Digital Image Transmission Leakage Model

To understand why the reconstruction method above can recover an image similar to the camera ground truth and the potential ways to further improve the reconstruction performance, we analyze the fundamental information leakage model that unpins the optical EM side channels in embedded cameras. We use one of the most popular data transmission protocols, MIPI CSI-2 with RAW10 image data format and two data lanes, as an example for developing the model. This protocol is also used by RPi V1. Nevertheless, we note that the modeling and analysis methodology also applies to other digital image transmission interfaces.

1) Fundamental Principle: Fig. 5 demonstrates how the optical information received by a camera sensor is transformed into EM signals that adversaries can capture. The process can be divided into two stages. In the first stage, the camera sensor transmits image data represented by digital bits row by row. The alternating currents/voltages caused by bit flips produce EM waves in the camera environment according to Maxwell's equation. In the second stage, the cable between the image sensor and ISP acts as an unintentional transmission antenna and propagates the EM waves to the adversary's receiving antenna. The EM signals are subjected to various environmental noises. With an off-of-the-shelf USRP device, the adversary can then sample the EM signals in specific frequency bands.

Fig. 5 also demonstrates how MIPI CSI-2 of image sensors transmits RAW10 images in the form of digital bits with two data lanes. Each pixel/column is represented by 10 ordered bits B0 to B9 (least significant bit (LSB) to most significant bit (MSB)) with the least significant bits transmitted first. The sensors treat a byte as a transmission element, although there is often no blanking between bytes during transmission. Since each pixel has 10 bits, RAW10 has to pack four consecutive pixels into a unit of five bytes where the two LSBs of the four pixels are packed into the last byte. Two units (8 pixels) are further grouped together. Using the dual data rate (DDR) technique, the clock f_{clk} frequency is twice the frequency of transmitting a bit f_b. For RPi V1, f_{clk} is measured to be 204 MHz, which means the byte transmission frequency is 51 MHz. When more than one data lane is used, consecutive byes are distributed to the lanes sequentially. It is worth pointing out that each wire of the transmission system, including the data and clock wires generate its own EM signals and the final signal the adversary receives is a mixture of them.

2) Modeling: Based on the understanding of the leakage process, we develop a mathematical model that can explain and simulate the physical leakage process's key characteristics. Assume the adversary tries to reconstruct an image that approximates the ground-truth camera image I_{GT} from the EM signals in the frequency band $[f_{lo}, f_{hi}]$ with a function $\mathcal{R}_{base}\{\cdot\}$, the EM reconstruction image can be calculated by

$$I_{EM}^{[l,h]} = \mathcal{R}_{base}\Big\{ z + b_{clk} + \mathcal{F}_{filt}\left[l, h, \mathcal{F}_{data}(I_{GT})\right] \Big\}, \quad (1)$$

where z represents the noise, b_{clk} represents a constant signal offset produced by the clock wire's emissions given that clock amplitudes are stable, $\mathcal{F}_{filt}[l, h, \cdot]$ represents the EM energy transfer function in the frequency band $[l, h]$, and

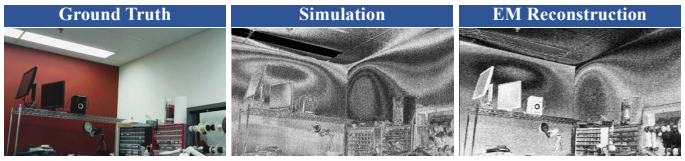


Fig. 6: The camera ground truth, simulated, and actual EM reconstruction. Distortions such as the amplification of light gradients and high-frequency noises appear.

$\mathcal{F}_{data}(\cdot)$ is the digital data transmission function that maps a 2D ground-truth image to a 1D bit stream transmitted by the data wires (Appendix B). Although theoretically, all the non-deterministic functions and variables in Eq. (1) are dependent on the environment and challenging to measure and model accurately, we found that simplified approximations (e.g., setting $\mathcal{F}_{filt}$ to a constant in the sampled frequency range) can produce simulated images that have very close quality and characteristics to the actual EM reconstructions. Fig. 6 provides some examples of the simulated and actual reconstructions using $\mathcal{R}_{base}$.

3) Key Characteristics: Based on the model, we then investigate several key observations of the eavesdropped images and analyze their causalities.

Baseband Leakage Frequency Dependency. The emitted EM signals are baseband signals of the digital bits instead of narrow-band signals that are modulated onto certain carriers such as clock frequencies of the system, which are more common for intentional communication systems. Since the baseband signal is wideband, every frequency band can contain different information about the ground-truth image. For example, Fig. 4 shows how 204 MHz and 255 MHz better capture the edge and gray-scale of the ground truth respectively. In practice, the adversary can only sample a subset of the digital wide-band information at a time. Advanced adversaries may thus need to combine information from different frequency bands. Besides the different information contained, each frequency band also has its unique EM wave propagation efficiency (transfer function) that leads to different SNRs for the adversary's received signals. We find that frequency components near the fundamental and harmonic frequencies of the digital transmission byte frequency (51 MHz) have the strongest signal strengths and lead to the best-quality reconstructions. This is because of the strong periodicity of transmitted bytes, leading to high EM amplitudes at these frequencies that can tolerate environmental noise better.

Multi-wire Signal Polarity Inversion. Another key phenomenon is that at certain frequency bands that contain f_{clk} and its harmonics, the amplitude of the EM signals could be inverted when the antenna moves relative to the cameras, leading to inversion of the reconstructed image's grayscale polarity (Fig. 17). Based on this observation, we hypothesize that the inversion of polarity is caused by the superposition of EM signals emitted by the data and clock wires. We then verified our hypothesis by measuring emissions from the clock and data lines separately (see Appendix C for details). Essentially, the clock emissions can interfere with the EM emissions from data wires. When the antenna is placed at a position that receives EM signals as a mixture of the data clock wire signals, the two signals can cancel each other out, producing an image that approximates a white image subtracted by the data line image. This image thus has an inverted polarity compared to the data line-only reconstructions.

Practical Sampling Distortion. We observe well-structured distortion patterns in all reconstructions, including:

- Loss of color information. Only gray-scale information remains in the reconstructions.
- Shuffled gray-scale mapping. The original and reconstructed images have different but correlated gray scales.
- Light gradient & high-frequency noise. Light gradients result in ellipse/contour-like shapes that are not visible in the original camera images, e.g., in Fig. 6. The reconstructions also have additional high-frequency noise.

Such distortion patterns are caused by the imperfect sampling of the EM leakage signals that adversaries could achieve in practice. The imperfection is two-fold. First, adversaries often can only sample an EM signal bandwidth on the order of 10 MHz with common USRPs and laptops while digital image transmissions have bandwidths on the order of 1 GHz. This causes the loss of a significant amount of information. Second, even if a hypothetical adversary can sample the whole bandwidth, e.g., by using multiple USRPs or sampling multiple times, it is still impractical for them to recover the original bit stream transmitted because of the added noise during EM propagation and the requirement of perfect synchronization for determining which bit is being transmitted. With these problems in mind, we can analyze the causality of the distortions above.

To recover the RGB colors of images using debayering, the adversary needs to know the original gray-scale value of each pixel precisely which requires perfect sampling of the digital bits and is thus impractical. In the original image, the gray-scale values represent an ordered array of bits; in the reconstructions, the gray-scale values represent the EM signal amplitudes which approximately correspond to the numbers of bit flips in the array. As a result, gray-scale values of the camera outputs are mapped to different values in the reconstructions in a shuffled but deterministic way. For example, bright lights and windows in the original images are often mapped to dark polygons in the EM reconstructions (see the first two columns of Fig. 6 for example). This is because the saturated bright pixels in the original image are mapped to constant ones in the transmitted digital data and cause significantly lower EM amplitudes due to the few bit flips.

The high-frequency noise exists everywhere in the reconstructed images while the light gradient distortions appear mostly on single-color surfaces in the scene. The culprit of light gradient and high-frequency noise is the loss of data structure due to imperfect sampling. Specifically, it is because *the EM emissions of different bits get combined without correct bit ordering.* In the original digital transmission protocol of cameras, each bit has its own weight and the ground-truth pixel value is calculated by $v_{GT} = \sum_{i=0}^{9} 2^i \boldsymbol{Bi}$. The adversary, however, can only calculate the pixel values while losing bit-ordering information in practice, because it is challenging to

determine the current bit being transmitted. Practically, all bits are considered equivalent whose EM emissions are added up without weights assigned. Conceptually, this can be modeled as $v_{EM} = \sum_{i=0}^{9} Bi$ which amplifies the light intensity variations and high-frequency noises that are often embedded in the least significant bits.

4) Insights: Our investigation reveals several challenges and opportunities for adversaries to reconstruct higher-quality images compared to the basic reconstructions presented above. (a) The frequency dependency problem calls for a method for integrating information in different frequency bands in order to harvest more entropy from the original camera outputs. (b) Although the multi-wire signal polarity inversion does not affect human visual perception significantly, it can cause additional noise to automated data processing and pattern recognition pipelines and thus needs to be mitigated to improve the eavesdropping performance. (c) The practical sampling distortions cause obvious degradation of the images' visual quality and intelligibility. As a result, the adversaries need to employ additional techniques for correcting these distortions. We will introduce the improved eavesdropping design that supports adversaries to extend their performance limits in the next section.

C. Relationship with Computer Display Eavesdropping

We discover that the eavesdropping vulnerability of embedded cameras shares the same physical principle as previous computer display eavesdropping attacks (Section VII) where the transmitted plain digital image data leaks in the form of EM waves. Furthermore, we confirm that all the key phenomena above are also observable when we replicate computer display eavesdropping attacks following previous research. However, many of these phenomena such as light gradient amplification and polarity inversions have not been reported and analyzed before. We believe this is because computer display eavesdropping only investigated simple screen contents of uniform texts on uniform backgrounds (e.g., no light gradients), which do not suffer significantly from the practical sampling distortions. In contrast, camera image scenes have more complex and diverse structures and textures, posing greater challenges for adversaries to reconstruct intelligible images. In addition, our survey shows that amplitude demodulation has also been the state-of-the-art method for mapping 1D EM signals to scalar pixel values in display eavesdropping attacks, which confirms our design choice in Section IV-A.

V. Eavesdropping System Design

To support the evaluation of eavesdropping limits and factors, we design a system that employs the signal processing pipeline shown in Fig. 7. The adversary first finds at least one frequency band that contains the EM leakage of transmitted digital image data. For each frequency band, the adversary reconstructs a single-band EM image from the received EM signals in this band. The adversary then strategically combines the images from different available frequency bands using a distortion-guided combination algorithm. The output of this algorithm, i.e., the multi-band EM image, is then input into an image-to-image translation network to acquire a final reconstructed image.

A. Single-band Image Reconstruction

The single-band image reconstruction process $\mathcal{R}_{base}$ on each frame can be formulated as

$$\begin{cases} I_{EM}^{[l,h]}[i_r, i_c] = \frac{1}{n_{samp}} \sum_{n=n_1}^{n_2} a[n] \\ n_{samp} = n_2 - n_1 + 1, a[n] = \mathcal{F}_{amd}[m[n]] \\ n_1 = \lfloor f_s(i_f T_f + i_r T_r + i_c T_c) \rfloor \\ n_2 = \lfloor f_s(i_f T_f + i_r T_r + (i_c + 1) T_c) \rfloor, \end{cases} \tag{2}$$

where i_f, i_r, i_c are the frame, row, and column indexes, T_f, T_r, T_c are the frame, row, and column transmission duration that needs to be estimated by the adversary through EM measurements, $m[n]$ is the discrete IQ measurements output of USRP with a sampling rate f_s, and $\mathcal{F}_{amd}[\cdot]$ is the amplitude demodulation function. Apparently, when f_s is on the order of 10 MHz in practical settings, n_1 and n_2 will be the same which is also the same for multiple consecutive i_c. This means the actual column resolution W_{EM} of the reconstructed image is smaller than the transmitted image and is determined by $W_{EM} = f_s T_{fd} / H_{EM}$ where H_{EM} is the row resolution that remains the same as the original transmitted image and T_{fd} is the actual frame data transmission duration excluding inter-frame blanking. As a result, T_c degrades to T_r / W_{EM} in most cases and does not need to be estimated separately. Fig. 18 provides more details on how to find the parameters. To improve the signal quality, we also perform frame averaging on the consecutive frames of camera outputs, which aims to mitigate the random noise in the EM wave propagation process and help the useful signals stand out. It is worth noting that this reconstruction process is also the current state-of-the-art (SOTA) used in computer display eavesdropping attacks, which we use as a building block as well as a baseline for our enhanced image reconstruction pipeline. We conduct an additional polarity-correction step that compares single-band reconstructions with data wire-only simulations and inverts the polarity if inversion is detected. We then apply histogram equalization to the image to further reduce the impact of clock signal offset b_{clk} (Eq. (1)) on image contrast.

B. Distortion-guided Multi-band Combination

We design a combination criterion based on the heuristic that the best combination can mitigate the light gradient distortions on single-color surfaces to the largest degree. As Section IV-B3 points out, the light gradient distortions arise because the bit-ordering information is lost. For example, both B2 and B6 have a periodicity of 8-bit cycles in RAW10 (Fig. 5), producing the same EM frequency that cannot be separated apart. Nevertheless, we observe that different frequency bands could still contain some inter-bit information. For example, if the 8-bit cycle frequency is a Hz, then the frequency of $2a$ Hz embeds the variation between B2 and B6. Similarly, we know that *different frequency bands embed different inter-bit information.* As a result, we propose that an adversary who can perform multi-band combination effectively should be able to minimize the light gradient distortions to restore the single-color surfaces. In our experiments, we

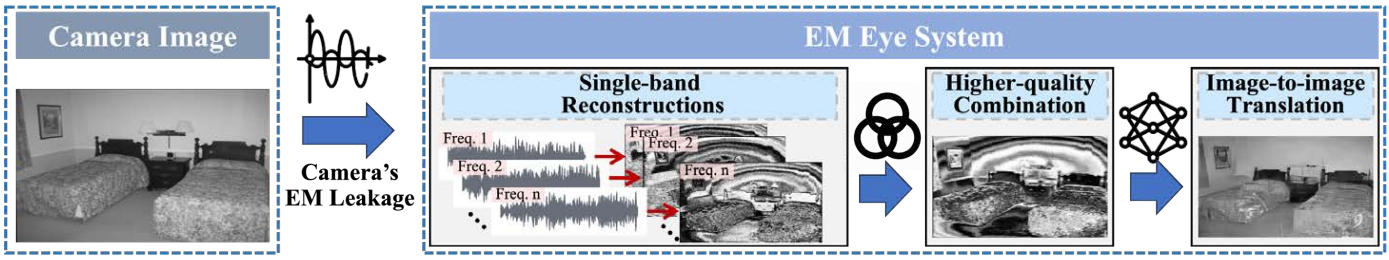


Fig. 7: The image eavesdropping pipeline of EM Eye.

empirically formulated this as

$$\hat{I}_{EM} = \sum_{i=0}^{N} w_i \cdot I_{EM}^{[l_i,h_i]}, \;\; s.t. \;\; [w_i] = \min_{[w_i]} ||c - S(\hat{I}_{EM})||, \quad (3)$$

where N is the number of available bands, w_i is the weight of band i, $S[\cdot]$ is a segmentation function that allows the adversary to manually select a subarea of the image that is likely a single-color surface, and c is a constant that the adversary can select to represent the color (gray scale) of the surface. Note that such an operation is possible because the single-band EM images often contain important structural information about the scene and experienced adversaries are able to hypothesize some key objects in the scene such as the walls of a room (see Fig. 7 for example). When selecting the frequency bands to combine, we also employ a thresholding criterion similar to [8] in order to remove components that are too noisy. Fig. 7 shows an example of this process. Typical values of N are in the range of 1-3 in our evaluations.

C. Image-to-image Translation

To further mitigate the remaining image distortions, we employ a supervised image-to-image translation process. This is inspired by our observation that additional semantic information in the image domain can be utilized to reconstruct images that are closer to the ground-truth image. For example, when observing the remaining light gradient distortion patterns, experienced human adversaries are able to understand that these distorted areas are likely to be single-color surfaces (which have the strongest light gradients) in the original camera output and thus manually correct the images. Another example is that the dark polygons in the EM reconstructions often map to the bright lights and windows in the original images. Given the very structured mappings, we hypothesize that it is possible to automate this process of correcting structured distortions in the EM reconstructions using machine learning-based approaches.

To verify this hypothesis, we formulate the task as an image-to-image translation problem from the EM-reconstructed image space to the original camera output space. We adopt pix2pix [17], an aligned image translation model based on a conditional generative adversarial network (GAN) to reconstruct a higher-quality image I_{EM} from $\hat{I}_{EM}$. Fig. 7 demonstrates an example of the translated reconstruction image in comparison with the gray-scale ground truth. We find the translation process capable of removing almost all remaining distortions when the testing images are within a reasonable range of variation compared to the training images. Although the generative model can also recover similar colors (see Appendix G), color information is often less useful for image pattern recognition tasks. In addition, the color recovery problem only relies on image semantic information and is completely detached from the EM leakage physics. We thus focus on gray-scale images in our following evaluations.

VI. Evaluation

A. Overview

Our evaluation seeks to measure the limits of the embedded camera eavesdropping risks under various camera designs and environmental conditions.

Experimental Setup. To provide reproducibility and scalability over multiple devices, we use the same setup as Section IV-A where images of different scenes are displayed by a monitor screen and recorded by the cameras under test, as shown in Fig. 8. We utilize two existing datasets to cover the common camera scenes pertinent to the threat model. The first dataset is a subset of the Face Detection Data Set and Benchmark [18] and has 3000 randomly selected images, each containing at least one person in the scene. The second dataset is a subset of the MIT Indoor Scenes Benchmark [31] that also has 3000 randomly selected images. Since the supervised image-to-image translation requires a training phase, we use 2700 images' corresponding $\hat{I}_{EM}$ from each dataset for training. In Section VI-B, we calculate the quantitative metrics over all 600 test images to evaluate the performance of the eavesdropping pipeline. To support scalable tests with fine-grained variations in the evaluation of factors and COTS devices, we also use a randomly-selected test subset of 35 images for each dataset which provides a confidence level of 90% at a resolution of 0.5 times the standard deviation of the test set population's scores [1]. For the training of the image-to-image network, we use the default hyper-parameters of the model [17] with 100 training epochs. We then use the last epoch's model as the final network. By default, we use the same model trained on a base case (Section VI-B) to test various test sets to examine the generalizability of this supervised network over different factors. The only exception is the evaluation of different camera sensors and controllers (Section VI-B) where we also train models using their own EM reconstructions as a comparison to investigate the improvement of dedicated image translation models. In total, we have collected 32400 training images and 10460 test images. We use an EM sampling rate (f_s) of 8 MHz in all experiments.

Quantitative Metrics. To quantify the impact of different factors on the eavesdropped information on both the EM signal and the image perception levels, we use the following metrics:

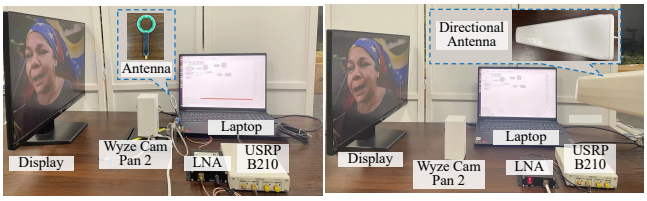
1) Unintentional signal-to-noise ratio (USNR) calculates the ratio of the unintentional EM emission power to the background noise power [8].
2) Structural similarity index measure (SSIM) measures the similarity between the eavesdropped and ground-truth camera images.
3) Face detection rate (Fdetect) calculates the ratio between the number of faces detected in the eavesdropped and ground-truth face dataset images.
4) Indoor scene captioning rate (Icaption) calculates the ratio of the longest common subsequence between the descriptional caption texts generated from the eavesdropped and ground-truth indoor dataset images.

SSIM, Fdetect, and Icaption range from 0 to 1, with larger values representing closer replicates of the ground truth. Apparently, the meanings of the absolute values are less intuitive. We thus also show example images corresponding to different values in our evaluations. Nevertheless, the variations of these metrics can still inform us of how different factors affect the quality of reconstructed images. Different from previous computer display eavesdropping research whose targets are simple texts, Fdetect and Icaption are specifically designed by us to measure how machines/humans perceive the complex visual information in camera scenes. Appendix E explains how we calculate these metrics.

B. Sensor and Controller

As pointed out in Section II-B, the camera data transmission interface can connect various camera sensors and controllers from different manufacturers. Given that different models of sensors and controllers could change the image data processed and transmitted, we first evaluate the impact of them on EM Eye's performance (Table I). We employ Raspberry Pi 3B+ and Cam V1 (#1) as the base case for collecting $\hat{I}_{EM}$ to train a base model (TrainA). We then change the sensors and controllers and collect corresponding $\hat{I}_{EM}$ to train their own models (TrainB).

The TrainA results in Table I suggest that sensors have a larger impact on the EM reconstructions than controllers. When the sensors change (e.g., [#1, #3, #6]), we observe larger degrees of variations in the image quality than when the controllers change (e.g., [#1, #2] and [#3, #4, #5]). This can be explained by the fact that it is often the sensors that decide the image data's format, amount, transmission speed, etc. The signals that EM Eye eavesdrops on are all produced by the camera sensors while the downstream processors mostly perform post-processing of the image data. Besides sensor hardware that determines the maximum supported image capacity, each camera sensor can also be configured to have various software/firmware settings such as resolution, frame rate, and sensor mode. Our tests show that setting the camera resolution does not change the transmitted data and EM emissions because the sensor always transmits the full resolution supported by a certain sensor mode and lets ISPs to down-sample the images in software. A different frame rate will change the number of frames transmitted per second and require the adversary to adjust the eavesdropping frame rate setting accordingly. Different sensor modes [30], which are combinations of camera firmware settings that decide the actual resolutions used by the sensor chips, will change the width and height of transmitted images and require the change of eavesdropping parameters.



(a) (b)

Fig. 8: Experiment setups of using (a) a near-field probe within 10 cm and (b) a directional antenna beyond 10 cm.

Fig. 9 compares some examples of direct EM reconstructions using state-of-the-art (SOTA) techniques and enhanced reconstructions using the EM Eye pipeline. Overall, obvious improvements in the visual quality are observed. The only caveat is that the image-to-image translation network can sometimes distort certain details of the images such as small textual objects. In these cases, the adversary may refer to the untranslated images $\hat{I}_{EM}$ to capture such information. Table I show the percentage of improvement in the quantitative image quality metrics compared to the SOTA results. On average, we observe 166.5%, 72.2%, and 52.2% increases in the SSIM, Fdetect, and Icaption scores for TrainA. The average values increase to 256.2%, 143.7%, and 88.7% for TrainB. The comparison between the metrics in TrainA and TrainB also shows that dedicated image translation models trained with each sensor-controller combination's EM data can indeed improve the quality of the eavesdropped images. The EM emissions of RPi 4B with Cam V1 are the most similar to the base case while those of Nvidia Jetson Nano with Cam V2 are the most dissimilar. The non-trivial metrics of all cases show that the base case model has a reasonable level of generalizability to process data from various sensors and controllers.

Summary. Different sensors and controllers can affect the EM signals while the EM Eye pipeline is able to reconstruct images with various sensor and controller settings. It also provides sufficient generalizability to allow the reconstructed images to outperform the SOTA results of direct EM reconstructions in most cases. In addition, resourceful adversaries may train dedicated models on each target camera system to further improve the eavesdropping performance.

C. Transmission Cable & Environmental Factors

Next, we measure the limits of EM Eye under various physical factors of the transmission cable and environment.

Cable EM Shielding. EM shielding uses special cable shield materials to block or reduce the propagation of EM waves. We evaluate its impact using 15 cm cables in three forms, namely the default cable of Raspberry Pi cameras without shielding, a cable shielded with conductive fabric, and one with aluminium foil. We use a near-field antenna to capture the EM emissions at a distance of 1 cm, and compare the values of USNR, SSIM, Fdetect, and Icaption for each cable with the same experimental setup. We depict the

TABLE I: Evaluation Results of EM Eye on 6 Sets of Sensor and Controller.

#	Sensor Module (Reconstruction Parameters)†	Controller Module	USNR (dB)	$W_{EM} \times H_{EM}$	TrainA(Improvement)* SSIM	TrainA(Improvement)* Fdetect	TrainA(Improvement)* Icaption	TrainB(Improvement)* SSIM	TrainB(Improvement)* Fdetect	TrainB(Improvement)* Icaption
1	Cam V1: OV5647	Raspberry 3B+ (Base)	39.68	186 × 1080	0.58(↑235.0%)	0.80(↑78.2%)	0.33(↑21.6%)	N/A	N/A	N/A
2	(T_f: 33.31 ms, T_r: 29.58 us)	Raspberry 4B	40.30	186 × 1080	0.45(↑221.8%)	0.75(↑50.7%)	0.29(↑19.1%)	0.55(↑298.8%)	0.78(↑57.7%)	0.32(↑30.9%)
3	Cam V2: IMX219	Raspberry 3B+	41.34	84 × 1290	0.29(↑186.9%)	0.51(↑95.5%)	0.23(↑80.8%)	0.45(↑349.4%)	0.70(↑168.3%)	0.27(↑115.5%)
4	(T_f: 33.84 ms, T_r: 18.90 us)	Nvidia Jetson Nano	42.51	84 × 1080	0.30(↑132.4%)	0.35(↑102.4%)	0.21(↑71.5%)	0.43(↑240.1%)	0.69(↑298.2%)	0.27(↑117.5%)
5		Asus Tinkerboard 2S	40.47	144 × 2466	0.39(↑112.5%)	0.60(↑79.5%)	0.26(↑49.6%)	0.53(↑193.0%)	0.76(↑129.6%)	0.31(↑78.7%)
6	Cam V3: IMX708 (T_f: 33.24 ms, T_r: 26.72 us)	Raspberry 3B+	43.54	104 × 1080	0.34(↑110.4%)	0.52(↑27.1%)	0.20(↑70.5%)	0.48(↑199.6%)	0.68(↑65.0%)	0.24(↑100.7%)

† The frame duration T_f and row duration T_r need to be estimated to decode the eavesdropped EM emission to reconstruct the images (Appendix D).
* EM Eye is evaluated on TrainA (base model) and TrainB (retrained model), and the percentage represents the improvement over the SOTA approach.

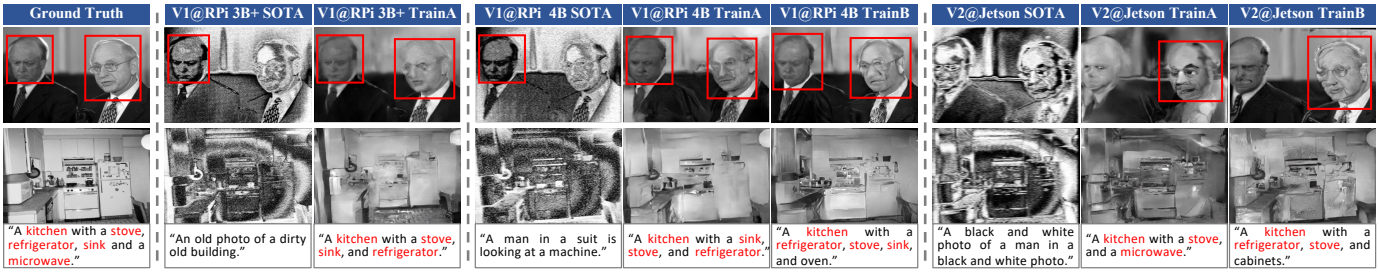


Fig. 9: Examples of eavesdropped images from three camera-controller systems using the SOTA and EM Eye pipelines, where A is the camera and B is the controller in A@B. Training dedicated models for each camera-controller combination (TrainB) provides better results than the base case model (TrainA). The detected faces of the face dataset images and the generated captions of the indoor dataset images are shown.

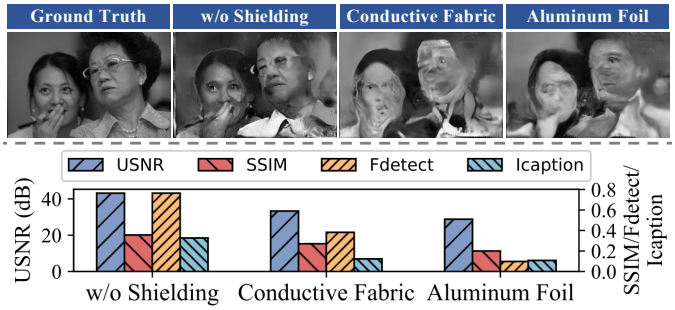


Fig. 10: Illustrations of (bottom) the impact of different cable EMI shielding, and (top) the same image reconstructed with different cable EMI shielding.

results in Fig. 10 (bottom). The cable shielded with conductive fabric and aluminium foil material significantly reduces the intensity of the EM emission radiated by the cable. We observe 9.84 dB and 14.33 dB decrease in USNR values respectively. Nevertheless, it is still possible to reconstruct images with acceptable SSIM and Fdetect values on these two shielded cables as shown in Fig. 10 (top).

Antenna-camera Distance and Cable Length. With the transmission cable acting as an unintentional antenna, the strength of EM emission attenuates with the antenna-camera distance. To quantify the impact, we measure the metrics under different distances with five typical cable lengths, namely 3, 10, 15, 30, and 50 cm. We use a near-field probe in Fig. 8(a) and a directional antenna in Fig. 8(b) with the same experimental setup to capture the EM Emission within and beyond 10 cm. The results are shown in Fig. 11. Notably, USNR, SSIM, Fdetect, and Icaption values gradually decrease with increasing distances, and longer cables often have higher values for these metrics at the same distance in our experiments. As shown in Fig. 11 (top), we observe almost unanimously better-quality images with longer cables. This is because the gains of different cable lengths vary, and longer cables provide a larger effective area, resulting in greater efficiency in radiating EM waves [6]. The maximum distances we could achieve for 3 cm, 10 cm, standard 15 cm, 35 cm and 50 cm cables are 50 cm, 200 cm, 270 cm, 400 cm, and 450 cm respectively. We note that the distance can be further increased by employing a professional antenna with superior directionality and gain.

Antenna-camera Angle. To examine the impact of camera-antenna angles, we change the angle from 0 °to 360 °with a step of 30 °(12 angles in total). The angle is defined as the angle between the centerline of the camera cable and the antenna. We conduct two sets of experiments using a near-field probe at a distance of 3 cm and a directional antenna at 40 cm respectively. Fig. 12 shows the impact of angles with the quantitative metrics. The angle has a small impact on EM Eye's performance at a close distance while some angles slightly outperform others. Due to the nature of the directional antenna, the angle has more impact on the eavesdropped images at a larger antenna-camera distance. As shown in Fig. 12, when the angle is between 90 °and 270 °at 40 cm, the values of these three metrics are significantly lower than when the angle is between 0 °to 90 °or 270 °to 360 °.

Interference from Electrical Devices and Background Noises. *(a) Electrical Devices.* The interference from displays of some electrical devices (such as TV, monitor, smartphone, etc.) is the most likely to affect EM Eye since the EM emission pattern of these displays is similar to that of the camera. However, modern displays offer refresh rates of 60, 120, or even 240 fps [36], whereas embedded cameras' frame rates are often limited to 30 fps. Therefore, adversaries can distinguish

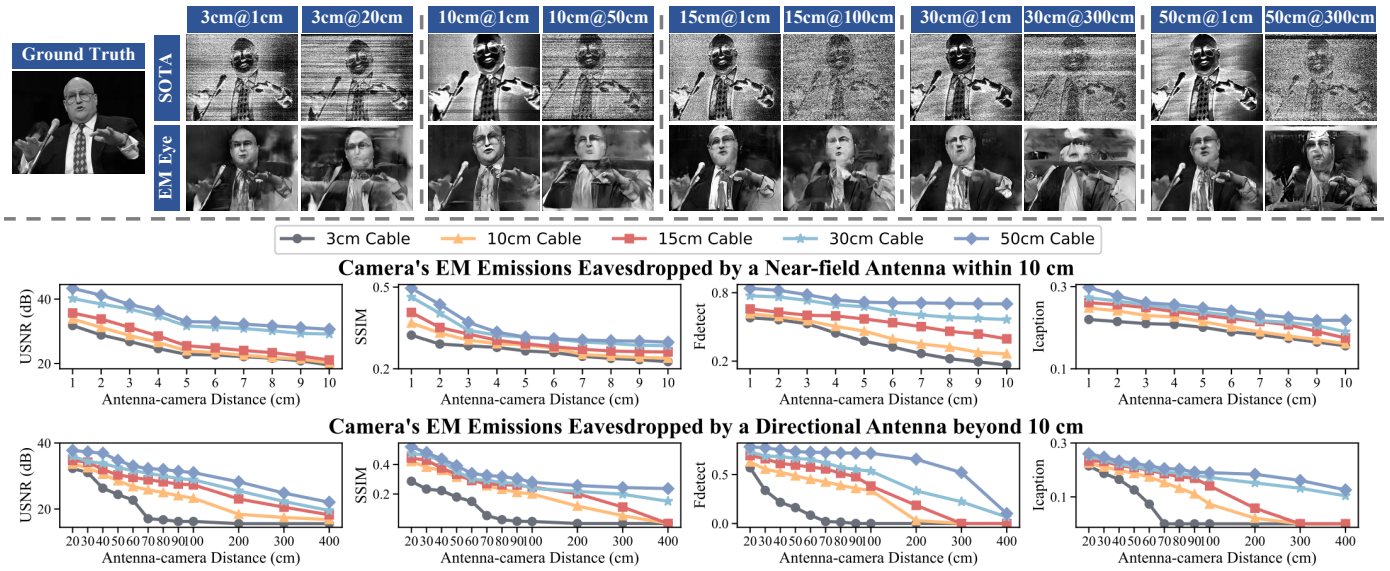


Fig. 11: Illustrations of (bottom) the impact of distances with different cable lengths, and (top) the same image reconstructed at different distances with different cable lengths, where A is the cable length and B is the distance in A@B.

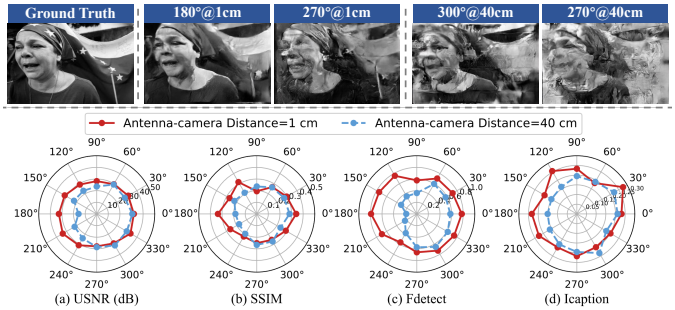


Fig. 12: Illustrations of (bottom) the impact of angles at 1 cm and 40 cm, and (top) the same image reconstructed at different angles at these two distances, where A is the antenna-camera angle and B is the distance in A@B.

camera emissions from the display's interference by setting the center frequency at those frequencies with no repetitions above 30 Hz to minimize the interference. We have verified this through experiments which are illustrated in Fig. 19. Besides, the EM emission pattern of cameras is very different from that of earbuds [8], recorders [42], wireless eavesdroppers [7], [35], etc. *(b) Background Noises.* Since EM Eye works at various frequencies, adversaries can improve image quality by avoiding selecting eavesdropping frequencies that conflict with common communication frequency bands (Table III). It is also effective to use analog filters to filter out background noises (see Fig. 20 for an example). Appendix F provides more discussions on the impact of interference.

D. COTS Camera Devices & Case Study

We have evaluated EM Eye on 12 commercial-off-the-shelf (COTS) camera devices from three different categories to investigate the common use cases of embedded cameras. These include 4 smartphones, 6 smart home cameras, and 2 dash cams. All of these devices are intact with their original packaging. Table II shows the specifications and eavesdropping parameters of these devices. Besides evaluating the eavesdropped image quality at 1 cm, we also measured the approximate maximum eavesdropping distance for each device at which we can still recover intelligible images. The maximum distances vary from 1 cm to 500 cm and with significant differences across devices. While all devices can be eavesdropped on in hidden-antenna scenarios where the antenna is close to the camera, we also observe that 8 out of the 12 devices allow adversaries to perform physical-isolation eavesdropping through windows, doors, and walls. We believe the variations in eavesdropping distances are mostly decided by the length and shielding materials used by these devices. For example, we found that smartphones often use short cables with better shielding designs to minimize the EM interference between the onboard components. Dash cams and home security cameras, on the other hand, tend to use cheap unshielded cables to reduce the manufacturing cost and longer cables to support different form factors of the mechanical structures. Despite the variations in these devices' designs, we note that the EM Eye vulnerability is a shared problem in common embedded camera devices, and we have reported our findings to the camera vendors (Appendix A). Based on the results above, we carry out case studies on three typical attack scenarios that we envision to be applicable to the threat model.

Smartphone Camera Eavesdropping. Since smartphone camera emissions only allow adversaries to eavesdrop from a close distance, we envision a hidden-antenna scenario where the antenna and EM signal receiver could be installed in modified power banks. Such power banks may either be tampered with from the supply chain as distributed products or provided by shared power bank rentals that are common in shopping malls. Existing COTS products of miniaturized low-cost SDR receivers such as the RTL-SDR dongles [33] suggest

TABLE II: Evaluation Results of EM Eye on 12 COTS Camera Devices.

#	COTS Camera Devices Manu. and Model	Year	Reconstruction Parameters† T_f (ms)	T_r (us)	Freq.	EM Eye Performance USNR	$W_{EM} \times H_{EM}$	SSIM	Fdetect	Icaption	Max. Dist.*	Scenarios HA	PI
1	Google Pixel 1	2013	33.45	21.49	600,1649 MHz	42.17 dB	168×1140	0.30	0.44	0.19	30 cm	✓	✗
2	Google Pixel 3	2018	33.27	10.89	515,680 MHz	41.81 dB	74×2840	0.24	0.36	0.19	2 cm	✓	✗
3	Samsung S6	2015	33.32	10.50	527,1054 MHz	38.92 dB	184×3000	0.31	0.70	0.19	5 cm	✓	✗
4	ZTE Z557	2019	41.70	17.00	522,1740 MHz	35.09 dB	310×1940	0.28	0.68	0.14	1 cm	✓	✗
5	Wyze Cam Pan 2	2019	49.98	29.63	890,1185 MHz	42.39 dB	164×1080	0.31	0.43	0.23	350 cm	✓	✓
6	Xiaomi Dafang	2019	66.66	29.63	322,890 MHz	39.06 dB	190×1080	0.35	0.67	0.17	500 cm	✓	✓
7	Baidu Xiaodu X9	2023	66.67	53.33	204,1470 MHz	35.86 dB	460×1080	0.24	0.23	0.15	200 cm	✓	✓
8	TeGongMao	2023	66.00	44.00	763,1144 MHz	40.58 dB	190×720	0.19	0.24	0.14	120 cm	✓	✓
9	Goov V9	2022	33.00	44.00	546,656 MHz	33.79 dB	190×720	0.31	0.32	0.18	70 cm	✓	✓
10	QiaoDu	2021	66.48	29.61	293,1191 MHz	38.79 dB	84×1080	0.22	0.38	0.17	50 cm	✓	✓
11	360 M320 Dashcam	2020	40.00	22.00	450,1261 MHz	39.71 dB	142×1440	0.29	0.17	0.22	250 cm	✓	✓
12	Blackview Dashcam	2022	33.22	27.78	155,1015 MHz	34.38 dB	190×1080	0.30	0.21	0.24	300 cm	✓	✓

† We only report two frequencies of the strongest emission.
* The maximum distance can be further increased by using higher-end EM receiving equipment such as professional direction antennas and analog filters.

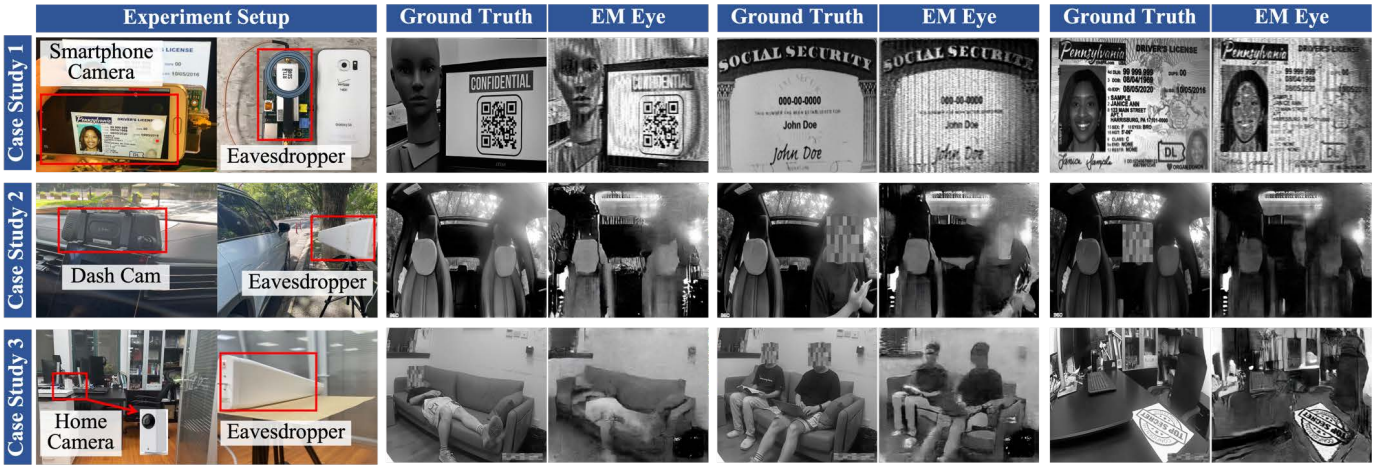


Fig. 13: Three case studies of how EM Eye poses eavesdropping threats against smartphones, dash cams, and home security cameras. For each case, the experimental setup and three examples of ground truths and eavesdropped images are shown.

the possibility of manufacturing such power banks. Fig. 13 (top) showcases an envisioned prototype and three sensitive images eavesdropped when the victim takes photos of private documents, including a QR code, a social security card, and a driver's license, with a Samsung S6 phone.

In-car Peeking. When victims park their cars with their interior dash cams on, an adversary may be able to peek at the inside of the cars using EM Eye eavesdropping from nearby. Fig. 13 (middle) shows an example setup with a 360 M320 dashcam [2] on the dash board of the car. The adversary sets up an antenna 50 cm away from the car (100 cm antenna-camera distance) to capture the EM emissions. Three eavesdropped images reveal no one in the car, one person in the driver's seat using his phone, and one in the back seat. When needed, the eavesdropping equipment can also be made portable as a suitcase, as has been demonstrated in previous research [14], to avoid further drawing the attention of the cars' owners.

Through-wall Room Spying. Another typical physical-isolation eavesdropping scenario involves an adversary spying on a private household or office room through the EM emissions of the IoT home security camera. The convention of installing such security cameras near the room's walls, windows, and doors could allow the adversary to receive the camera's EM emissions from only a few meters away. Fig. 13 (bottom) demonstrates a case where the antenna is placed 70 cm away outside an office room (150 cm antenna-camera distance). The adversary can see a person sleeping on a couch, two people sitting on the couch, and a confidential document on a desk by eavesdropping on a Xiaomi Dafang home camera [11].

VII. Related Work

A. Computer Display Side-channel Eavesdropping

It has been widely acknowledged that computer displays generate side-channel leakages in operation that allow adversaries to eavesdrop on the displayed contents. The most known research is TEMPEST attacks where EM leakage is used to reconstruct computer screens. Following the first work by Wim van Eck in 1985 [39] that proved the feasibility of reconstructing video display contents using non-military commercial-grade equipment, extensive research has been carried out over the last 40 years. Some notable works include Markus Kuhn's efforts to develop low-cost techniques to eavesdrop on analog CRT [21] and digital LCD flat panel displays [22]. While earlier works only investigated standalone computer display units which often generate stronger EM emissions, Hayashi et

al. showed it is possible to eavesdrop on smaller tablet and laptop screens from 2m away [14]. Recently, Liu et al. [24] extended this attack to smartphone displays. However, due to the very weak EM emissions generated by the small smartphone circuits, the researchers had to use machine learning classifiers to recognize the humanly-unintelligible reconstructions at a distance of 1 cm. Besides EM emissions, Genkin et al. [12] showed that acoustic side-channel signals generated by computer display circuits when processing different pixel data also allow adversaries to detect screen contents using machine learning classifiers. In all these works, texts on screens have been the sole target of eavesdropping.

Cameras work in similar ways as computer displays in that they both have to transmit streams of 2D images in a serialized manner. Our work shows that a more fundamental analysis framework for 2D digital image transmission leakage can be developed to model and generalize these attacks. Compared to previous works, our research bridges the gap between such information leakage mechanisms and a broad range of emerging sensor systems. From the standpoint of technical advances, this work shows that the camera image contents are significantly more complex and diverse than those of computer displays, causing new challenges such as light gradient distortions that increase the difficulty of using existing TEMPEST techniques to reconstruct high-quality recognizable images. We thus design and apply new computational techniques to address these unique challenges.

B. IP Camera Hijacking & Sniffing

With a similar purpose of accessing the outputs of unauthorized cameras, several works have found that networked IP cameras can be hijacked or sniffed by adversaries when there exist vulnerabilities in the network configurations. For example, Abdalla et al. showed that many cameras use default passwords and unencrypted communications [3]. Ling et al. demonstrated the feasibility of performing an online brute-force attack to uncover IP camera's password because many cameras only have only four-digits long passwords [23]. Herodotou et al. found that a generic camera module used by many spy camera manufacturers can be controlled by adversaries over the internet as long as the serial number of the camera is known [15]. Tekeoglu et al. successfully reconstructed 253 JPEG images from about 20 hours of video track by sniffing an IP camera's unencrypted network traffic [38]. While these works show the feasibility of eavesdropping on IP cameras when there exist software vulnerabilities, our work explores the complementary aspect of physical vulnerabilities of camera designs. This allows an adversary to eavesdrop on not only networked cameras but also locally-operated cameras as well as systems with strong software security such as smartphones and home security devices.

VIII. Discussion

A. Countermeasure

We analyze the possible countermeasures from the standpoint of camera and system designers.

EM Jamming. Jamming is a common technique used to disrupt intentional communication systems. However, we believe jamming is less suitable for mitigating camera eavesdropping given that the leaked signals are wide-band, requiring an expensive device to cover such a wide bandwidth. Furthermore, jamming can easily compromise the legitimate camera data stream itself as has been demonstrated by [19], [20]. Jamming devices can either be installed by camera manufacturers or users. The challenge is it needs to cover a large space as the EM field distribution can be unpredictable and varying. This is based on our observations that different probe positions and orientations will lead to very different results.

Shorter Cables & Better Shielding. Our evaluation shows that short cables often produce weaker EM emissions, especially in the far field. Device manufacturers are thus encouraged to employ shorter cables in their designs. However, we note that such changes may also require a complete redesign of the devices' mechanical structures since it requires the camera lens to be very close to the controller boards. Otherwise, the manufacturers can consider using better-shielded data transmission cables, which have been shown to be capable of reducing the EM signal strength by over 10 dB.

Increase and Randomize Transmission Blanking. With the same frame rate and resolution of the transmitted images, increasing the blanking between frames and rows will reduce the effective resolution of the eavesdropped images under a certain eavesdropping sampling rate (Fig. 18). This requires the transmission interface to have higher bit rates. Furthermore, adding intentional jitters to randomize the blanking duration can prevent adversaries from easily performing frame averaging and thus reduce the leakage USNR they receive.

Grouped Pixel Smoothing Protocol Improvement. We argue that the current image data transmission protocols are flawed and can be improved to mitigate EM leakage. Essentially, the EM emissions originate from the periodic bit flips. Ideally, the order of transmitted rows, columns, and even bits should be randomized, eliminating all the periodicity. However, we also realize such randomization requires a complete hardware redesign and could be expensive for manufacturers. We thus seek to improve the protocol by keeping the overall architecture but minimizing the number of periodic bit flips. We achieve this by simply rearranging the bits. Specifically, we observe that adjacent pixels (columns) have similar values in their bits, especially the MSBs. By putting the same bits from adjacent pixels in a byte as shown in Fig. 15, we can smooth out many bit transitions and reduce the EM emission amplitudes. In addition, the more adjacent pixels grouped together in this way, the fewer emissions there will be. Fig. 14 demonstrates the EM emission spectrum calculated by the simulation model (Eq. (1)) when there is no such defense and when 8 and 128 pixels are grouped together for smoothing, respectively. Most of the strong emission peaks at the multiples of the byte frequency (51 MHz) are mitigated by over 10 dB. Note that the original protocol already groups 8 pixels together in transmission, so supposedly 8-pixel smoothing requires minimal modifications to the interface designs.

B. Other Sensing Devices

We believe the threat of EM side-channel eavesdropping may be further extended to other sensing devices.

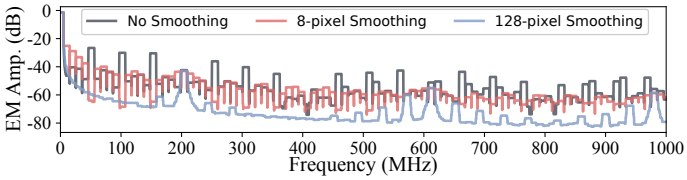


Fig. 14: The simulated EM emission strengths with no defense and with the proposed grouped pixel smoothing in the transmission protocol design.

Encoded Video Data Transmission. Although embedded systems widely use open-standard image data transmission interfaces that send uncompressed RAW data, many traditional camera devices such as USB webcams still use proprietary interfaces that send encoded (e.g., h264) video data. With such devices, the adversary cannot use the eavesdropping method of this work to directly reconstruct images. However, it is possible to use machine-learning-based classifiers to recognize human-unintelligible EM signals because image patterns can be recognized as long as the corresponding EM signals have sufficient separability. We experimented with a Logitech C920x HD Pro webcam which transmits compressed video data. We tried to classify 100 different face images recorded by the webcam using its EM emissions. We simply use the EM signals' Fast Fourier Transform coefficients processed by Linear Discriminant Analysis and a self-built three-layer neural network classifier. Even with the crude features, we could achieve a test accuracy of 90.12% for the 100-class classification. This makes us believe the eavesdropping threat can affect a wider range of cameras even if the adversary doesn't understand how data is transmitted.

General Sensors. Every sensor peripheral has to transmit data to the central processors. Most sensors transmit unencoded plain data. Given that the data throughput of most sensors is much smaller than cameras, we believe the EM side-channel eavesdropping on other sensors could be achieved even with less sophisticated equipment and data reconstruction algorithms. In addition, eavesdropping on sensors used in industrial settings may not require the adversaries to be physically isolated from the sensors. For example, an employee trying to steal the secret specifications of a product that is being measured by a benchmarking device may physically approach the automated device to collect EM signals.

C. Limitation & Future Work

Eavesdropping Distance. While many camera devices we examined could be eavesdropped on from physically-isolated rooms, some of the cameras still have a limited range of feasible distances compared to previous computer display eavesdropping attacks. This is due to the shorter data transmission cables and the lower voltage swings of the cameras. For example, HDMI has 3.3 V differential signals while MIPI CSI-2 has less than 500 mV. At large distances where the reconstructed images get too distorted to be intelligible to humans, we believe machine learning-based methods can be used to directly recognize the EM signals, as has been demonstrated by previous research on smartphone display screen eavesdropping [24]. Furthermore, we note that dedicated analog band-pass filters could be utilized to further reduce noise and extend the eavesdropping distance [22] (Appendix H).

Automated Tests. In our evaluations, it takes 30-60 min to find the eavesdropping parameters of an unseen camera device. By manually analyzing 12 COTS devices, our work aims to present an initiative that motivates stakeholders to examine how wide the problem is in the real world more thoroughly. In future large-scale studies, automated test methods may need to be developed to enable the testing of more devices. We envision that the main challenge is for the automated algorithm to robustly determine whether there is an eavesdropped image in the reconstructions while efficiently sweeping through a wide range of parameters.

IX. Conclusion

This work investigated EM Eye, an EM leakage vulnerability of embedded camera systems that allow eavesdroppers to reconstruct camera image streams from camera EM emissions. We have identified the cause of this vulnerability to be the unprotected deterministic digital image data transmissions between the image sensor and downstream image processing components. After developing an eavesdropping signal processing pipeline, we verified the impact of this vulnerability on 4 IoT development platforms and 12 COTS camera devices including smartphones, smart home cameras, and dash cams. For defenses, we found that dedicated cable shielding, shorter cables, and proposed improved data transmission protocols can effectively reduce camera EM leakage. Finally, we pointed out that EM Eye shares the same underlying physical principle with computer display eavesdropping attacks, which we believe could be further generalized to other sensors.

Acknowledgments

We thank our reviewers for their valuable comments and suggestions. We also thank Benjamin Cyr's insights that helped us identify the cause of the multi-wire signal polarity inversion problem. This work was supported by the NSF Phase I IUCRC grant from the Center for Hardware and Embedded System Security and Trust (CHEST), China NSFC Grant 62201503, 61925109, 62222114, and 62071428, and the Fundamental Research Funds for the Central Universities 226-2022-00223.

References

[1] Engineering Statistics Handbook: Sample Sizes Required. https://www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm, 2012.

[2] 360. 360 m320 dashcam. https://shopee.sg/-Local-Seller-Brand-360-M301-M320-M320C-HD-Car-Dash-Camera-(Front-Rear)-(SD-Card-included)-i.585005157.12446270348, 2022. [Online; accessed 27-June-2023].

[3] Peshraw Ahmed Abdalla and Cihan Varol. Testing IoT security: The Case Study of an IP Camera. In *Proceedings of the 2020 8th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE.

[4] MIPI Alliance. Mipi alliance specification for camera serial interface 2 (csi-2), 2005.

[5] Aptina. High-speed serial pixel (hispi) interface protocol. https://files.niemo.de/aptina_pdfs/High-Speed_Serial_Pixel_%28HiSPi%29_Interface_Specification.pdf, 2011. [Online; accessed 12-June-2023].

[6] Constantine A Balanis. *Antenna theory: analysis and design*. John wiley & sons, 2016.

[7] Anadi Chaman, Jiaming Wang, Jiachen Sun, Haitham Hassanieh, and Romit Roy Choudhury. Ghostbuster: Detecting the presence of hidden eavesdroppers. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018.

[8] Jieun Choi, Hae-Yong Yang, and Dong-Ho Cho. Tempest Comeback: A Realistic Audio Eavesdropping Threat on Mixed-signal Socs. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[9] Pieterjan de Meulemeester, Bart Scheers, and Guy AE Vandenbosch. Eavesdropping a (Ultra-) High-definition Video Display from an 80 Meter Distance under Realistic Circumstances. In *Proceedings of the 2020 IEEE International Symposium on Electromagnetic Compatibility & Signal/Power Integrity (EMCSI)*.

[10] Ettus Research. Ettus research usrp products. https://www.ettus.com/products/, 2023. [Online; accessed 12-June-2023].

[11] Gadget Freakz. Xiaomi dafang 1080p smart monitor camera review. https://gadget-freakz.com/xiaomi-dafang-1080p-smart-monitor-camera-review/, 2018. [Online; accessed 27-June-2023].

[12] Daniel Genkin, Mihir Pattani, Roei Schuster, and Eran Tromer. Synesthesia: Detecting Screen Content via Remote Acoustic Side Channels. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*.

[13] Bahadir K Gunturk, John Glotzbach, Yucel Altunbasak, Ronald W Schafer, and Russel M Mersereau. Demosaicking: Color Filter Array Interpolation. *IEEE Signal processing magazine*, 22(1):44–54, 2005.

[14] Yuichi Hayashi, Naofumi Homma, Mamoru Miura, Takafumi Aoki, and Hideaki Sone. A Threat for Tablet Pcs in Public Space: Remote Visualization of Screen Images Using EM Emanation. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[15] Samuel Herodotou and Feng Hao. Spying on the spy: Security analysis of hidden cameras. *arXiv preprint arXiv:2306.00610*, 2023.

[16] Infinite Reality. Digital video port (dvp) specification. https://irix7.com/techpubs/007-3594-001.pdf, 2011. [Online; accessed 12-June-2023].

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[18] Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[19] Qinhong Jiang, Xiaoyu Ji, Chen Yan, Zhixin Xie, Haina Lou, and Wenyuan Xu. GlitchHiker: Uncovering Vulnerabilities of Image Signal Transmission with IEMI. In *Proceedings of the USENIX Security 23*, 2023.

[20] Sebastian Köhler, Richard Baker, and Ivan Martinovic. Signal injection attacks against ccd image sensors. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, 2022.

[21] Markus G Kuhn. Optical Time-domain Eavesdropping Risks of CRT Displays. In *Proceedings of the 2002 IEEE Symposium on Security and Privacy (SP)*.

[22] Markus G Kuhn. Electromagnetic Eavesdropping Risks of Flat-panel Displays. In *Proceedings of the International Workshop on Privacy Enhancing Technologies*. Springer, 2004.

[23] Zhen Ling, Kaizheng Liu, Yiling Xu, Yier Jin, and Xinwen Fu. An End-to-end View of IoT Security and Privacy. In *Proceedings of the 2017 IEEE Global Communications Conference (GLOBECOM)*.

[24] Zhuoran Liu, Niels Samwel, Léo Weissbart, Zhengyu Zhao, Dirk Lauret, Lejla Batina, and Martha Larson. Screen Gleaning: A Screen Reading TEMPEST Attack on Mobile Devices Exploiting an Electromagnetic Side Channel. In *Proceedings of the 28th Annual Network and Distributed System Security Symposium (NDSS)*, 2021.

[25] Ziwei Liu, Feng Lin, Chao Wang, Yijie Shen, Zhongjie Ba, Li Lu, Wenyao Xu, and Kui Ren. CamRadar: Hidden Camera Detection Leveraging Amplitude-modulated Sensor Images Embedded in Electromagnetic Emanations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(4):1–25, 2023.

[26] MathWorks. Face detection and tracking using the klt algorithm. https://www.mathworks.com/help/vision/ug/face-detection-and-tracking-using-the-klt-algorithm.html, 2023. [Online; accessed 20-June-2023].

[27] MathWorks. rougeevaluationscore. https://www.mathworks.com/help/textanalytics/ref/rougeevaluationscore.html, 2023. [Online; accessed 20-June-2023].

[28] MIPI Alliance. Mipi csi-2 specifications. https://www.mipi.org/specifications/csi-2, 2023. [Online; accessed 12-June-2023].

[29] NLP Connect. vit-gpt2-image-captioning. https://huggingface.co/nlpconnect/vit-gpt2-image-captioning, 2023. [Online; accessed 20-June-2023].

[30] picamera. Camera hardware: Sensor modes. https://picamera.readthedocs.io/en/latest/fov.html#sensor-modes, 2023. [Online; accessed 27-June-2023].

[31] Ariadna Quattoni and Antonio Torralba. Recognizing Indoor Scenes. In *Proceedings of the 2009 IEEE conference on computer vision and pattern recognition (CVPR)*.

[32] RF Bay. Rf bay, inc. lna-650. https://www.rfbayinc.com/upload/files/lna/lna-650.pdf, 2023. [Online; accessed 26-June-2023].

[33] RTL-SDR. Buy rtl-sdr dongles (rtl2832u). https://www.rtl-sdr.com/buy-rtl-sdr-dvb-t-dongles/, 2023. [Online; accessed 26-June-2023].

[34] Ariel Schwarz, Yosef Sanhedrai, and Zeev Zalevsky. Digital Camera Detection and Image Disruption Using Controlled Intentional Electromagnetic Interference. *IEEE transactions on electromagnetic compatibility*, 54(5):1048–1054, 2012.

[35] Cheng Shen and Jun Huang. Earfisher: Detecting wireless eavesdroppers by stimulating and sensing memory emr. In *Proceedings of the 18th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2021.

[36] Josef Spjut, Ben Boudaoud, Kamran Binaee, Jonghyun Kim, Alexander Majercik, Morgan McGuire, David Luebke, and Joohwan Kim. Latency of 30 ms benefits first person targeting tasks more than refresh rate above 60 hz. In *SIGGRAPH Asia 2019 Technical Briefs*, pages 110–113. 2019.

[37] Statista. Number of households with smart security cameras worldwide from 2016 to 2027. https://www.statista.com/forecasts/1301193/worldwide-smart-security-camera-homes, 2023. [Online; accessed 16-June-2023].

[38] Ali Tekeoglu and Ali Saman Tosun. Investigating Security and Privacy of a Cloud-based Wireless IP camera: NetCam. In *Proceedings of the 2015 24th International Conference on Computer Communication and Networks (ICCCN)*. IEEE.

[39] Wim Van Eck. Electromagnetic Radiation from Video Display Units: An Eavesdropping Risk? *Computers & Security*, 4(4):269–286, 1985.

[40] Wikipedia contributors. Low-voltage differential signaling. https://en.wikipedia.org/w/index.php?title=Low-voltage_differential_signaling&oldid=1021691966, 2021. [Online; accessed 12-June-2023].

[41] Baki Berkay Yilmaz, Elvan Mert Ugurlu, Milos Prvulovic, and Alenka Zajic. Detecting Cellphone Camera Status at Distance by Exploiting Electromagnetic Emanations. In *2019 IEEE Military Communications Conference (MILCOM)*, 2019.

[42] Ruochen Zhou, Xiaoyu Ji, Chen Yan, Yi-Chao Chen, Wenyuan Xu, and Chaohao Li. DeHiREC: Detecting Hidden Voice Recorders via ADC Electromagnetic Radiation. In *Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP)*.

Appendix A
Responsible Disclosure

We have reported our findings including a thorough description of the vulnerability and affected devices to the device vendors in Table II as well as MIPI Alliance. We provided detailed instructions[1] for replicating a simplified version of the attack and our suggestions for countermeasures.

Appendix B
Two-lane RAW10 Transmission Model

We provide the mathematical model for two-lane RAW10, one of the most common instances of $\mathcal{F}_{data}(\cdot)$ for converting images to bit streams in MIPI CSI-2. We also introduce an improved version that aims to mitigate the EM leakage problems.

The Original Protocol. The original protocol can be found in Figure 5 and [4]. Assume the transmission of the very first bit of a frame starts at time 0. Denote the time for transmitting a byte as T_B. At the time t, row i_r is being transmitted. Denote the current column and bit being transmitted by lane 0 and lane 1 as i_{c0}, i_{b0} and i_{c1}, i_{b1} respectively. With all indexes starting from 0, the protocol can be described as:

$$
\begin{cases}
\tilde{t} = t - T_r i_r \quad \text{current row working time} \\
s_c = 8 \cdot \lfloor \tilde{t}/5T_B \rfloor \quad \text{column offset of 8-column groups} \\
g_B = mod(\lfloor \tilde{t}/T_B \rfloor, 5) \quad \text{byte position in the 8-column group} \\
i_{c0} = s_c + \mathcal{I}_{\{g_B<2\}} \cdot 2g_B + \mathcal{I}_{\{g_B>2\}} \cdot (2g_B - 1) \\
\qquad\qquad + \mathcal{I}_{\{g_B==2\}} \cdot mod(\lfloor \tilde{t}/2T_{bs} \rfloor, 4) \\
i_{c1} = s_c + \mathcal{I}_{\{g_B<2\}} \cdot (2g_B + 1) + \mathcal{I}_{\{2\le g_B<4\}} \cdot 2g_B \\
\qquad\qquad + \mathcal{I}_{\{g_B==4\}} \cdot [4 + mod(\lfloor \tilde{t}/2T_{bs} \rfloor, 4)] \\
i_{b0} = \mathcal{I}_{\{g_B\neq 2\}} \cdot [2 + mod(\lfloor \tilde{t}/T_{bs} \rfloor, 8)] \\
\qquad\qquad + \mathcal{I}_{\{g_B==2\}} \cdot mod(\lfloor \tilde{t}/T_{bs} \rfloor, 2) \\
i_{b1} = \mathcal{I}_{\{g_B\neq 4\}} \cdot [2 + mod(\lfloor \tilde{t}/T_{bs} \rfloor, 8)] \\
\qquad\qquad + \mathcal{I}_{\{g_B==4\}} \cdot mod(\lfloor \tilde{t}/T_{bs} \rfloor, 2)
\end{cases}
$$

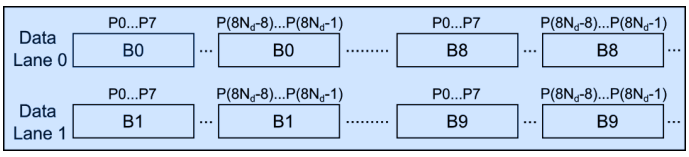


Fig. 15: The defense of grouped pixel smoothing where the same bits from $8N_d$ adjacent (similar) pixels are transmitted together to minimize bit flip-caused EM emissions.

Grouped Pixel Smoothing Protocol Improvement. Assume we group $N_d(\geq 1)$ 8-column groups together, the new smoothed protocol shown in Fig. 15 can be described as:

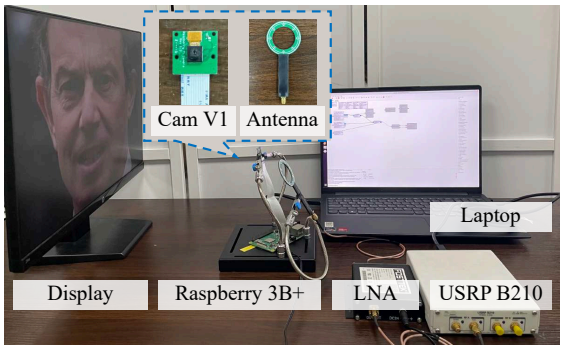


Fig. 16: The experiment setup for the feasibility tests.

$$
\begin{cases}
\tilde{t} = t - T_r i_r \quad \text{current row working time} \\
s_c = 8N_d \cdot \lfloor \tilde{t}/5N_d T_B \rfloor \quad \text{column offset} \\
g_B = mod(\lfloor \tilde{t}/T_B \rfloor, 5N_d) \text{ byte position in the group} \\
\hat{g}_B = \lfloor g_B/N_d \rfloor \\
i_{c0} = s_c + mod(\lfloor \tilde{t}/T_{bs} \rfloor, 8N_d) \\
i_{c1} = i_{c0} \\
i_{b0} = 2\hat{g}_B \\
i_{b1} = i_{b0} + 1
\end{cases}
$$

Appendix C
Multi-wire Polarity Inversion

Fig. 17 shows the setup and results for investigating our hypothesis of multi-wire polarity inversion. (a) shows the setup for measuring emissions from a data transmission clock and a data wire at the clock frequency. The green/left and yellow/right cables extend the data and clock wires respectively to allow the measurement of individual wire's emissions. (b) shows the first antenna position. (c) shows the second antenna position. (d) shows the EM reconstruction when only the data line is extended, with any antenna positions. (e) shows the reconstruction when only the clock line is extended, with any antenna positions. (f) shows the results of subtracting (d) from (e), demonstrating an inverted polarity. When we connect both the data and clock wires, (g) shows the reconstruction at the first antenna position and (h) shows the reconstruction at the second antenna position which is similar to (f). It demonstrates that the first antenna position receives most signals from the clock wire while the second position receives a mixture of the data and clock wires' signals.

Appendix D
Reconstruction and Image Parameters

Fig. 18 (top) demonstrates how the blanking between frames and rows manifests itself in the domain of reconstructed images. It also depicts three boxes whose areas represent the duration of the frame transmission T_f, the actual frame data transmission time T_{fd}, and the row transmission time (T_r). Fig. 18 (bottom) shows a raw reconstruction that considers the blanking as part of the reconstructed image. As a result, the blanking areas corresponding to the background noise during the transmission idle time appear to be much darker than the

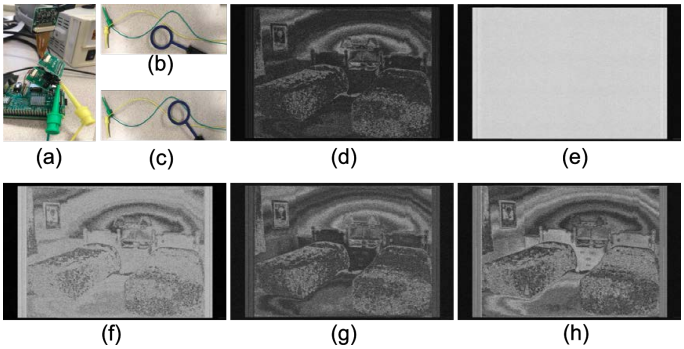


Fig. 17: Illustrations of (a) the setup for investigating the multi-wire polarity inversion problem, (b) the first antenna position, (c) the second antenna position, and the EM reconstructions when (d) only the data line is extended, and (e) only the clock line is extended. (f) illustrates the results of subtracting (d) from (e), (g) and (h) are images reconstructed at antenna position (b) and (c). To show the real EM amplitudes, histogram equalization is not applied.

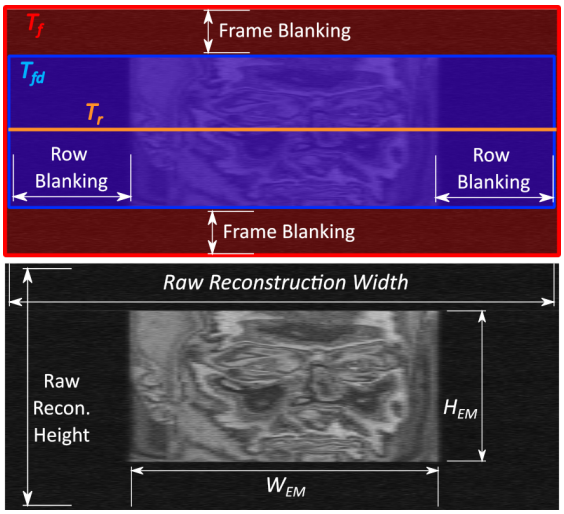


Fig. 18: A raw reconstruction that has both the eavesdropped image and blanking as well as the image reconstruction parameters shown.

eavesdropped image. In practice, it is often easier to utilize Eq. (2) to get the blanking-included raw reconstruction first, and then crop out the blanking areas to get I_{EM}. The procedure for the adversary to eavesdrop on images of an unseen camera is as follows. First, the adversary sets the USRP sampling rate f_s and then finds the T_f of the camera. The adversary gets $f_s \times T_f$ sample points for each frame that need to be mapped to the width and height of the raw reconstruction. The adversary then finds the height of the raw reconstruction by estimating T_r and calculating $raw\_height = T_f/T_r$. Accordingly, the maximum width (setting n_{samp} to 1 in Eq. (2)) is $f_s \times T_r$. Similar to the relationship between W_{EM} and H_{EM}, the width of the raw reconstruction is often much smaller than the height. Fig. 18 and all eavesdropped images that are shown in this paper resized the reconstructions through image processing by increasing the widths proportionally in order to get a more normal visualization of the eavesdropped image. Finally, H_{EM}, W_{EM}, T_{fd} can be estimated by comparing the cropped eavesdropped image with the raw reconstruction. In summary, the two most important that adversaries need to estimate from the EM signals are T_f and T_r. The estimation can be done using a combination of signal auto-correlation tests and trial-and-errors.

Appendix E
Metrics

Fdetect. The Fdetect rate is calculated by $N^{EM}_{face}/N^{GT}_{face}$ where the numerator and denominator are the numbers of detected faces in I_{EM} and I_{GT}. For face detection, we use the Cascade Object Detector provided in MATLAB [26] with the FrontalFaceCART model.

Icaption. The Fdetect rate is calculated with the rougeEvaluationScore function in MATLAB [27], where we run the ROUGE-L metric on the generated captions from I_{EM} and I_{GT}. We use an existing image captioning model developed by NLP Connect that uses transformers to generate the texts [29].

Appendix F
Interference from Electronic Devices and Background Noises

The interference from displays of some electrical devices is the most likely to interfere with EM Eye since the data transmission pattern of these displays is similar to that of the camera. However, the display emission patterns can still be effectively differentiated from camera emissions because of two reasons. First, the center frequencies of the best-quality signal bands of display emissions are very unlikely to overlap with those of cameras. Since the emissions cover a wide range of frequencies, the likelihood of getting overlaps in every frequency band is even lower. So the adversary can almost always find a camera eavesdropping frequency that does not overlap with the display emissions. Second, the eavesdropping parameters such as frame rate and image heights are also very different. As shown in Fig. 19, we can receive EM emissions from the display in the room around 888 MHz and 1037 MHz when eavesdropping on the Dafang camera. Since we use the eavesdropping parameters of the Dafang camera instead of the display, the reconstructed image shows these diagonal stripes, which can be easily distinguished from camera emissions.

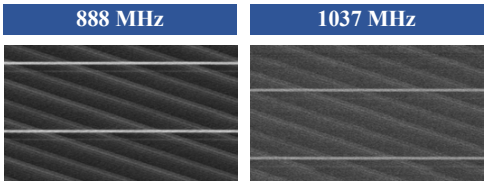


Fig. 19: The images show the display emissions received at two frequencies when eavesdropping on the Dafang camera. Since we use the eavesdropping parameters of the Dafang camera instead of the display, the reconstructed image shows these diagonal stripes, which can be easily distinguished from camera emissions.

The frequency bands of some common communication standards may overlap with the camera emissions (Table III), producing environmental background noise that can negatively affect the SNR of camera emissions. To address the problem, professional analog filters can be employed to filter out the interference of the background noise. As shown in Fig. 20, the analog filter has significantly improved the quality of reconstructed images.

TABLE III: Frequency Bands of Common Communication Standards.

Protocol	Frequency Band	Protocol	Frequency Band
GSM	880 - 960~MHz	Wi-Fi	2.4~GHz and 5~GHz
3G	800 - 2100~MHz	ZigBee	915~MHz and 2.4~GHz
LTE	700 - 2600~MHz	LoRa	868~MHz and 915~MHz
5G	850~MHz, 1900~MHz 1850 - 1990~MHz	NB-IoT	824 - 849~MHz, 869 - 894~MHz
Bluetooth	2.4~GHz	Z-Wave	868.42~MHz and 908.42~MHz

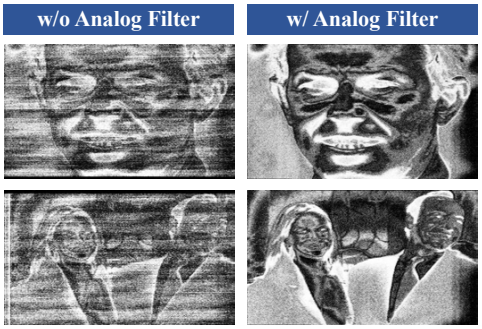


Fig. 20: Illustrations of filtering the background noises with an analog filter.

Appendix G
Colored Reconstruction

Fig. 21 shows some examples of the reconstructed color images compared with their ground truth and SOTA reconstructions. The data is from the base case in Section VI-B. As can be seen from the figures, the eavesdropping pipeline is able to infer colors in a reasonable deviation range. However, such recovered colors do not originate from the intrinsic characteristics of the leaked EM signals themselves. As explained in Section V-C, colors are inferred completely from the semantics of images. For example, when the image translation network detects a shape that looks like a tie, it can color the tie area based on the tie examples in the training set. This, however, can be different from the actual colors, as demonstrated by the 5th column of Fig. 21. Given these caveats, this work chooses to avoid the complications brought by colors to better investigate the fundamentals of the EM physical channel. Nevertheless, we believe the added color information could still be useful to adversaries in many cases. Future research could quantify the impact of colors on how adversaries perceive the eavesdropped images.

Appendix H
Eavesdropping Equipment

We employed the following equipment to build the middle-end EM eavesdropping devices in our evaluations. (1) Software

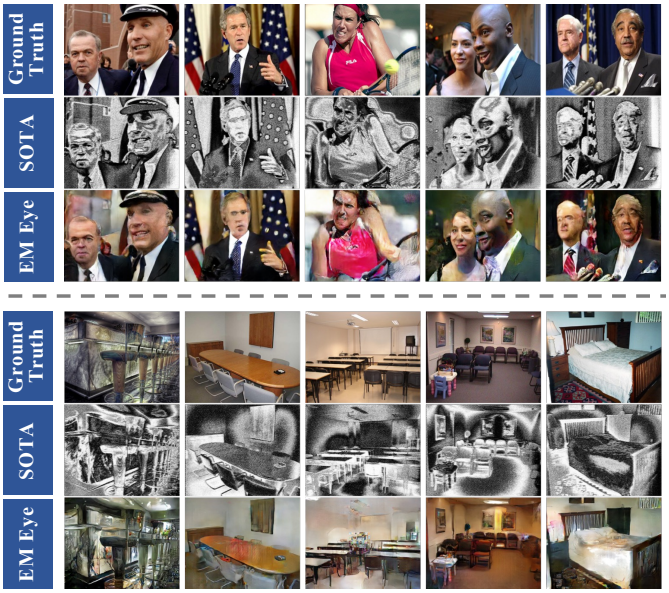


Fig. 21: Ground truth (top), SOTA reconstruction (middle), and colored EM Eye reconstruction (bottom).

Defined Ratio (SDR): Ettus USRP B210, a two-channel USRP device with continuous RF coverage from 70 MHz – 6 GHz with up to 56 MHz of real-time bandwidth that costs $2100. (2) Low Noise Amplifier (LNA): Foresight Intelligence FST-RFAMP06 LNA, which costs $207 and offers a frequency range of DC to 3.5 GHz with a gain of up to 40 dB. (3) Directional Antenna: A common outdoor Log-periodic directional antenna (LPDA), which costs $15 and offers a frequency range of 700 - 4900 MHz with a gain of up to 15 dBi.

Better Equipment. Our experiments only employed off-of-the-shelf middle-end eavesdropping equipment. It is possible to use more advanced devices to increase the eavesdropping distances and reconstruction quality further. For example, there are professional antennas with gains higher than 30 dBi. High-end LNAs can achieve gains up to 50 dB [32]. Furthermore, analog filters can significantly reduce noise and improve SNR, as shown in Appendix F. Resourceful adversaries can manufacture dedicated analog band-pass filters for each target camera or purchase expensive tunable filters. As an example, previous computer display research has shown that by employing a 45 dBi LPDA antenna, analog band-pass filters, and better software algorithms, it is possible to increase the maximum eavesdropping distance from 10 m to 80 m [9].