Private Eye: On the Limits of Textual Screen Peeking via Eyeglass Reflections in Video Conferencing

Yan Long*, Chen Yan†, Shilin Xiao†, Shivan Prasad*, Wenyuan Xu†, and Kevin Fu*
*Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA
† College of Electrical Engineering, Zhejiang University, Hangzhou, China
{yanlong, shprasad, kevinfu}@umich.edu, {yanchen, xshilin, wyxu}@zju.edu.cn

Abstract*—Personal video conferencing has become a new norm after COVID-19 caused a seismic shift from in-person meetings and phone calls to video conferencing for daily communications and sensitive business. Video leaks participants' on-screen information because eyeglasses and other reflective objects unwittingly expose partial screen contents. Using mathematical modeling and human subjects experiments, this research explores the extent to which emerging webcams might leak recognizable textual and graphical information gleaming from eyeglass reflections captured by webcams. The primary goal of our work is to measure, compute, and predict the factors, limits, and thresholds of recognizability as webcam technology evolves in the future. Our work explores and characterizes the viable threat models based on optical attacks using multi-frame super resolution techniques on sequences of video frames. Our models and experimental results in a controlled lab setting show it is possible to reconstruct and recognize with over 75% accuracy on-screen texts that have heights as small as 10 mm with a 720p webcam. We further apply this threat model to web textual contents with varying attacker capabilities to find thresholds at which text becomes recognizable. Our user study with 20 participants suggests present-day 720p webcams are sufficient for adversaries to reconstruct textual content on big-font websites. Our models further show that the evolution towards 4K cameras will tip the threshold of text leakage to reconstruction of most header texts on popular websites. Besides textual targets, a case study on recognizing a closed-world dataset of Alexa top 100 websites with 720p webcams shows a maximum recognition accuracy of 94% with 10 participants even without using machine-learning models. Our research proposes near-term mitigations including a software prototype that users can use to blur the eyeglass areas of their video streams. For possible long-term defenses, we advocate an individual reflection testing procedure to assess threats under various settings, and justify the importance of following the principle of least privilege for privacy-sensitive scenarios.

I. Introduction

Online video calls have become ubiquitous as a remote communication method, especially since the recent COVID-19 pandemic that caused almost universal work-from-home policies in major countries [24], [27], [31] and made video conference a norm for companies and schools to accommodate interpersonal communications even after the pandemic [6], [15], [43], [51].

While video conferencing provides people with the convenience and immersion of visual interactions, it unwittingly reveals sensitive textual information that could be exploited by a malicious party acting as a participant. Each video participant's screen could contain private information. The participant's own webcam could capture this information when it is reflected by the participant's eyeglasses and unwittingly provide the information to the adversary (Figure 1). We refer to this attack as a *webcam peeking attack*. Furthermore, adversary capabilities will only continue to increase with improvements to resolution, frame rate, and more. It is thus important to understand the consequences and limits of webcam peeking attacks in present-day and possible future settings.

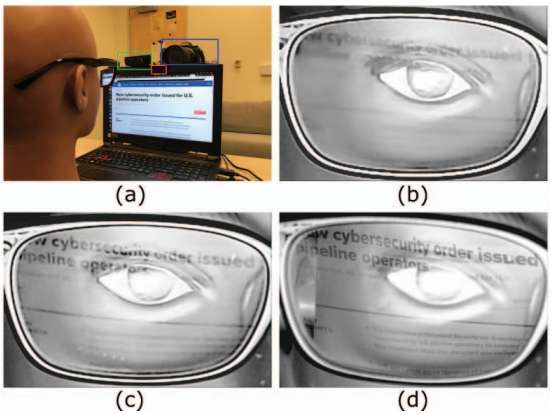

Fig. 1. The optical emanations of the victim's screen are reflected by eyeglasses, captured by the victim's webcam, and streamed to the adversary, which can then be used to reconstruct the screen contents. The experimental setup (a) with a laptop built-in webcam (b) (red box, 720p), an external Logitech webcam (c) (green box, 1080p), and a Nikon DSLR (d) (blue box, 4K) helps us predict the future fidelity of the attacks as video conferencing technologies evolve.

Previous work shows that similar attacks exploiting optical reflection off nearby objects in controlled setups are feasible, such as observing teapots on a desk with high-end digital single-lens reflex (DSLR) cameras and telescopes at a distance [25], [26]. The challenge and characterization of peeking using the more ubiquitous webcams, however, are qualitatively different due to the lower-quality images of present-day webcams. The lower-quality webcam images are caused by unique types of distortions, namely the shot and ISO noise due to insufficient light reception, and call for new image-enhancing techniques. In addition, new mathematical models and analysis frameworks are needed to understand

the threat model of webcam peeking attacks. Finally, this new threat model requires a dedicated evaluation to clarify the potential threats and mitigations to the average video conference user.

There are many types of media that can leak over optical reflections, including text and graphics. We focus on textual leakage in this work as it's a natural starting point for measurable recognizability and modeling of the fundamental baseline of information leakage, but also provides insights into the leakage of non-textual information such as inferring displayed websites through recognizing graphical contents on the screen. We seek to answer the following three major questions: $Q1$: What are the primary factors affecting the capability of the webcam peeking adversary? $Q2$: What are the physical limits of the adversary's capability in the present day and the predictable future, and how can adversaries possibly extend the limits? $Q3$: What are the corresponding threats of webcam peeking against cyberspace targets and the possible mitigations against the threats?

To answer $Q1$, we propose a simplified yet reasonably accurate mathematical model for reflection pixel size. The model includes factors such as camera resolution and glass-screen distance and enables the prediction of webcam peeking limits as camera and video technology evolve. By using the complex-wavelet structural similarity index as an objective metric for reflection recognizability, we also provide semi-quantitative analysis for other physical factors including environmental light intensity that affect the signal-to-noise ratio of reflections.

To answer $Q2$, we analyze the distortions in the webcam images and propose multi-frame super resolution reconstruction for effective image enhancement to extend the limits. We then gather eyeglass reflection data in optimized lab environments and evaluate the recognizability limits of the reflections through both crowdsourcing workers on Amazon Mechanical Turk and optical character recognition models. The evaluation shows over 75% accuracy on recognizing texts that have a physical height of 10 mm with a 720p webcam.

To answer $Q3$, we focus on web textual targets to build a benchmark that enables meaningful comparisons between present-day and future webcam peeking threats. We first map the limits derived from the model and evaluations to web textual content by surveying previous reports on web text size and manually inspecting fonts in 117 big-font websites. Then, we conduct a user study with 20 participants and play a challenge-response game where one author acts as an adversary to infer HTML contents created by other authors. Results of the user study suggest that present-day 720p webcams can peek texts in the 117 big-font websites and future 4K webcams are predicted to pose threats to header texts from popular websites. We investigated the underlying factors enabling easier webcam peeking in the user study by analyzing the correlation between adversary recognition accuracy and multiple factors. We found, for example, user-specific parameters including browser zoom ratio play a more important role than the glass-screen distance. Besides texts, we also explored the feasibility of recognizing websites through graphical content with 10 participants and observed accuracies as high as 94% on recognizing a closed-world dataset of Alexa top 100 websites.

Finally, we discuss possible near-term mitigations including adjusting environmental lighting and blurring the glass area in software. We also envision long-term solutions following an individual reflection assessment procedure and a principle of least privilege. In summary, the goal of this work is to provide a theoretical foundation and benchmark for the study of emerging webcam peeking threats with evolving webcam technologies and the development of securer video conferencing infrastructures. We summarize our main contributions:

- Our work quantifies the limits and primary factors that predict the degree of information leakage from webcam peeking by using theoretical modeling and experimentation. This characterization helps predict future unknown vulnerabilities tied to the limits of evolving webcam technologies that do not yet exist.
- A benchmark centering on web textual targets that enables comparisons of webcam peeking threats. Our benchmarking methodology builds upon web text design conventions and a 20-participant user study on present-day cameras such that the benchmark can be applied to both hypothetical and emerging cameras in the coming years.
- Analysis on near-term mitigations including using software-based blurring filters and changing physical setups as well as possible long-term defenses by proactive testing and following a principle of least privilege. Our analysis investigates the potential effectiveness and implementation methods of different protections.

II. Threat Model & Background

A. Threat Model

In this work, we study the webcam peeking attack during online video conferences, where the adversary and the victim are both participants. We assume the device the victim uses to join the video conference consists of a display screen and either a built-in or an external webcam that is mounted on the top of the screen as in most cases, and the victims wear glasses with a reflectance larger than 0, i.e., at least a portion of the light emanated by the monitor screen can be reflected from the glasses to the webcams. We do not enforce constraints on the devices used by the adversary. When the adversary launches the attack, we assume the victim is facing the screen and webcam in the way that the screen emanated light has a single-reflection optical path into the webcam through the eyeglass lens's outer surface. We do not assume the adversary has any control or information on the victim's device.

We assume that the victim's up-link video stream is enabled during the attack, and the adversary can acquire the down-link video stream of the victim. The adversary can achieve that by either directly intercepting the down-link video stream data, or recording the victim's video with the video conferencing platform being used or even third-party screen recording services. Since the webcam peeking attack does not require active interaction between the victim and the adversary, the

adversary does not need to attempt a real-time attack but can store the video recording and analyze the videos offline.

B. Glasses

The most common types of glasses that people wear in a video conferencing setting are prescription glasses [40] and blue-light blocking (BLB) glasses [11], [50]. BLB glasses can either have prescriptions with BLB coating or be non-prescription (flat). The reflectance and curvature of glass lenses are the two most important characteristics in the process of reflecting screen optical emanations.

Reflectance. Reflectance of a lens surface is the ratio between the light energy reflected and the total energy incident on a surface [5]. Reflectance is wavelength-dependent. The higher the reflectance, the more light can be reflected to and captured by a webcam.

Curvature. Curvature of a lens surface represents how much it deviates from a plane. The concepts of curvature, radius, and focal length of an eyeglass lens are used interchangeably on different occasions and are related by: $Curvature = 1/Radius = 2/FocalLength$. Smaller curvature leads to larger-size reflections. Both the outer and inner surfaces of a lens can reflect, but the outer surface often has smaller curvature and thus produce better quality reflections (Appendix A). This paper refers to the eyeglass lens curvature/radius/focal length as that of the outer surface if not specified otherwise.

C. Digital Camera Imaging System

Digital cameras have sensing units uniformly distributed on the sensor plane, each of which is a Charge-coupled Device (CCD) or Complementary Metal-oxide-semiconductor (CMOS) circuit unit that converts the energy of the photons it receives within a certain period of time, i.e., the exposure time, to an amplitude-modulated electric signal. Each sensing unit then corresponds to a "pixel" in the digital domain. The quality of a digital image to human perception is mainly determined by its pixel resolution, color representation, the amount of received light that is of our interest, and various imaging noise. The 2 key imaging parameters that are closely related to webcam peeking attacks are described below.

Exposure Time. Theoretically, the longer the exposure time, the more photons will hit the imaging sensors, and thus there can be potentially more light of interest captured. The images with a longer exposure time will generally be brighter. The downside of having a longer exposure time is the aggravated motion blur when imaging a moving object.

ISO Value. The ISO value represents the amplification factor of the photon-induced electrical signals. In darker conditions, the user can often make the images brighter by increasing the ISO value. The downside of having a higher ISO is the simultaneous amplification of various imaging noises.

D. Text Size Representations

It is important to select proper representations of text size in both digital and physical domains since the size of the smallest recognizable texts is the key metric for webcam peeking limits. When texts are digital, i.e., in the victim's software such as browsers and in the webcam image acquired by the adversary, we use point size and pixel size to represent the text size respectively. In the physical domain, i.e., when the texts are displayed on users' screens as physical objects, we use the cap height of the fonts and the physical unit mm to represent the size as it is invariant across different computer displays and enable quantitative analysis of the threats. Cap height is the uniform height of capitalized letters when font style and size are specified and is thus usually used as a convenient representation of physical text size and the base for other font parameters [22], [23].

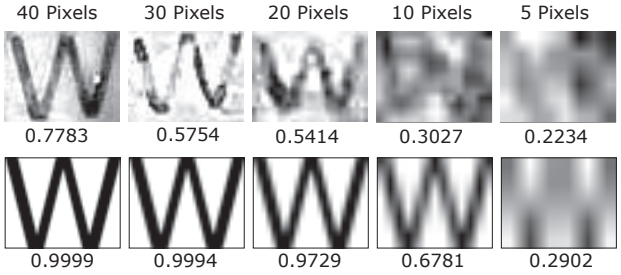

Fig. 2. (Upper) The captured images of the reflections. Compared with the ideal reflections, additional distortions exist that undermine image recognizability. (Lower) The estimated ideal reflections in the feasibility test corresponding to letters with a height of 80, 60, 40, 20, 10 mm respectively. The images are subjected to aliasing when enlarged.

III. Webcam Peeking Through Glasses

In this section, we start with a feasibility test that reveals the 3 key building blocks of the webcam peeking threat model, namely (1) reflection pixel size, (2) viewing angle, and (3) light signal-to-noise ratio (SNR). For the first two building blocks, we develop a mathematical model that quantifies the related impact factors. For light SNR, we analyze one major factor it encompasses, i.e., image distortions caused by shot noise, and investigate using multi-frame super resolution (MFSR) to enhance reflection images. We will analyze other physical factors that affect light SNR in Section IV-D. Experiments are conducted with the Acer laptop with its built-in 720p webcam, the pair of BLB glasses, and the pair of prescription glasses described in Appendix A.

A. Feasibility Test

We conduct a feasibility test of recognizing single alphabet letters with a similar setup as in Figure 1. A mannequin wears the BLB glasses with a glass-screen distance of 30 cm. Capital letters with different cap heights (80, 60, 40, 20, 10 mm) are displayed and captured by the webcam. Figure 2 (upper) shows the captured reflections. We find that the 5 different cap heights resulted in letters with heights of 40, 30, 20, 10, and 5 pixels in the captured images. As expected, texts represented by fewer pixels are harder to recognize. The reflection pixel size acquired by adversaries is thus one key building block of the characteristics of webcam peeking

TABLE I
PARAMETERS FOR MODELING REFLECTION PIXEL SIZE

Notation	Parameter
h_o	Physical size (cap height) of the object on the screen
h_s	Physical size of the object's projection on the sensor
s_p	Pixel size of the imaged object
h_i	Physical size of the object's virtual image
P	Physical size of a single imaging sensor pixel
N	Number of pixels the camera has in the dimension
W	Physical size of the imaging sensor in the dimension
f	Camera focal length
d_o	Distance between screen and glasses
d_i	Distance between glasses and virtual image
f_g	Focal length of the glasses convex outer surface

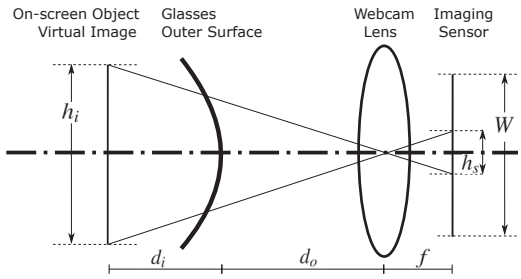

Fig. 3. The model of reflection pixel size. To better depict the objects, the sizes are not drawn up to scale. The screen overlaps with the webcam lens and is omitted in the figure.

attack that we need to model. In addition, Figure 2 (lower) shows the ideal reflections with these pixel sizes by resampling the template image. Comparing the two, we notice small-size texts are subjected to additional distortions besides the issue of small pixel resolution and noise caused by the face background, resulting in a bad signal-to-noise ratio (SNR) of the textual signals.

To quantify the differences using objective metrics, we embody the notion of reflection quality in the similarity between the reflected texts and the original templates. We compared multiple widely-used image structural and textural similarity indexes including structural similarity Index (SSIM) [56], complex-wavelet SSIM (CWSSIM) [53], feature similarity (FSIM) [59], deep image structure and texture similarity (DISTS) [32] as well as self-built indexes based on scale-invariant feature transform (SIFT) features [49]. Overall, we found CWSSIM which spans the interval $[0, 1]$ with larger numbers representing higher reflection quality produces the best match with human perception results. Figure 2 shows the CWSSIM scores under each image.

The differences show that the SNR of reflected light corresponding to the textual targets is another key building block we need to characterize. Finally, we notice that when we rotate the mannequin with an angle exceeding a certain threshold, the webcam images do not contain the displayed letters on the screen anymore. It suggests that the viewing angle is another critical building block of the webcam peeking threat model which acts as an on/off function for successful recognition of screen contents. In the following sections, we seek to characterize these three building blocks.

B. Reflection Pixel Size

In the attack, the embodiment of textual targets undergoes a 2-stage conversion process: digital (victim software) → physical (victim screen) → digital (adversary camera). In the first stage, texts specified usually in point size in software by the user or web designers are rendered on the victim screen with corresponding physical cap heights. In the second stage, the on-screen texts get reflected by the glass, captured by the camera, digitized, and transferred to the adversary's software as an image with certain pixel sizes. Generally, more usable pixels representing the texts enable adversaries to recognize texts more easily. The key is thus to understand the mechanism of point size → cap height → pixel size conversion.

Point Size → Cap Height. Mapping between digital point size and physical cap height is not unique but dependent on user-specific factors and software. The conversion formula for most web browsers can be summarized as follows:

$$h_o = \frac{4}{3} p_t \cdot \frac{H_{scr}}{N_{os}} \cdot s_{os} \cdot s_b \cdot r_{cap} \tag{1}$$

where h_o is the physical cap height of the text, $\frac{4}{3}p_t$ is the number of display hardware pixels most web browsers use to render the text given a point size p_t, H_{scr} is the physical height of the screen, N_{os} is the screen resolution on the height dimension set in the OS which can be equal to or smaller than the maximum supported resolution, s_{os} and s_b are the OS and browser zoom/scaling ratios respectively, and r_{cap} is the ratio between the cap height and the physical point size which is on average $\frac{2}{3}$ [22], [23].

Cap Height → Pixel Size. We would like to remind the readers that we only use pixel size to represent the size of texts living in the images acquired by the adversary[1]. Figure 3 shows the model for this conversion process. To simplify the model, we assume the glasses lens, screen contents, and webcam are aligned on the same line with the same angle. The result of this approximation is the loss of projective transformation information, which only causes small inaccuracies for reflection pixel size estimation in most webcam peeking scenarios. Figure 3 only depicts one dimension out of the horizontal and vertical dimensions of the optical system but can be used for both dimensions. In this work we focus on the vertical dimension for analysis, i.e., the reflection pixel size we discuss is the height of the captured reflections in pixels. We summarize the parameters of this optical imaging system model in Table I. Through trigonometry, we know

$$\begin{cases} \frac{h_s}{f} = \frac{h_i}{d_o + d_i} \\ h_s = s_p P \\ P = \frac{W}{N} \end{cases} \quad \Rightarrow \quad s_p = \frac{h_i}{d_o + d_i} \cdot \frac{f}{W} \cdot N \tag{2}$$

[1]Since web/software designers sometimes also directly specify text size in pixel size ($\frac{4}{3}P_t$ in Equation 1), the two pixel sizes can be easily confused without explanation.

TABLE II
THE PREDICTED FEASIBLE ATTACK RANGES FOR THE VIEWING ANGLE.

Type	Theoretical	Measurement
Pres: All Page + Horizontal	±15°	±17°
Pres: Center + Horizontal	±5°	±8°
Pres: All Page + Vertical	±9°	±13°
Pres: Center + Vertical	±3°	±5°
BLB: All Page + Horizontal	±20°	±25°
BLB: Center + Horizontal	±10°	±13°
BLB: All Page + Vertical	±14°	±19°
BLB: Center + Vertical	±8°	±10°

As pointed out in Section II-B, the reflective outer surface of glasses is mostly convex mirrors which shrink the size of the imaginary object h_i and also decrease d_i compared to an ideal flat mirror. To calculate the reflection pixel size s_p in this case, we can use the convex mirror equations [38]

$$\begin{cases} \frac{1}{(-f_g)} = \frac{1}{d_o} + \frac{1}{(-d_i)} \\ \frac{h_i}{h_o} = \frac{d_i}{d_o} \end{cases}$$

where f_g is the focal length of the convex mirror which is half of the radius of the glasses lens and is defined to be positive. Plugging the above equations into Equation 2 we can then get

$$s_p = \frac{h_o f_g}{d_o^2 + 2 d_o f_g} \cdot \frac{f}{W} \cdot N, \tag{3}$$

The term $\frac{f}{W}$ of typical laptop webcams can be estimated to be in the range $1.1 - 1.4$ (see Appendix A). With the Acer laptop and BLB glasses and a glass-screen distance of $d_o = 30$ cm, the estimated vertical pixel size of a 20 mm-tall object displayed on the screen is in the range of $9.2 - 11.7$ pixels, which matches with the 10 pixels found in the feasibility test and verifies the accuracy of the model despite the approximation we made.

C. Viewing Angle

To model the effect of viewing angle and the range of angles that enables webcam peeking attack, we model the lens as spherical with a radius $2f_g$. A detailed derivation of the viewing angle model can be found in Appendix B. We consider two cases of successful peeking with a rotation of the glass lens. The first case All Page claims success as long as there exists a point on the screen whose emitted light ray can reach the camera. The second case Center claims success only if the contents at the center of the screen have emitted lights that can be reflected to camera. Table II summarizes the calculated theoretical angle ranges and the measured values. Both the theoretical model and measurements show that the webcam peeking attack is relatively robust to human positioning with different head viewing angles, which is validated later by the user study results (Section V-B).

D. Image Distortion Characterization

Generally, the possible distortions are composed of imaging systems' inherent distortions and other external distortions. Inherent distortions mainly include out-of-focus blur and various imaging noises introduced by non-ideal camera circuits. Such inherent distortions exist in camera outputs even when no user interacts with the camera. External distortions, on the other hand, mainly include factors like motion blur caused by the movement of active webcam users.

User Movement-caused Motion Blur. When users move in front of their webcams, reflections from their glasses move accordingly which can cause blurs in the camera images. User motions can be decomposed into two components, namely involuntary periodic small-amplitude tremors that are always present [33], and intentional non-periodic large-amplitude movements that are occasionally caused by random events such as a user moving its head to look aside. By approximating user motions as displacements of h_o and utilizing Equation 3, the number of blurred pixels δ_p can be estimated by[2]:

$$\delta_p = \frac{\delta^T f_g}{d_o^2 + 2 d_o f_g} \cdot \frac{f}{W} \cdot N$$

where δ^T is the motion displacement amplitude within the exposure time of a frame.

For tremor-based motion, existing research suggests the mean displacement amplitude of dystonia patients' head tremors is under 4 mm with a maximum frequency of about 6 Hz [34]. Since dystonia patients have stronger tremors than healthy people, this provides an estimation of the tremor amplitude upper bound. With the example glass in Section III-B and a 30 fps camera, the estimated pixel blur is under 1 pixel. Such a motion blur is likely to affect the recognition of extremely small reflections. Intentional motion is not a focus of this work due to its random, occasional, and individual-specific characteristics. We will experimentally involve the impacts of intentional user motions in the user study by letting users behave normally.

Distortion Analysis. To observe and analyze the dominant types of distortions, we recorded videos with the laptop webcam and a Nikon Z7 DSLR [17] representing a higher-quality imaging system. The setup is the same as the feasibility test except that we tested with both the still mannequin and a human to analyze the effects of human tremor. Figure 14 (a) shows the comparison between the ideal reflection capture and the actual captures in three consecutive video frames of the webcam (1st row) and Nikon Z7 (2nd row) when the human wears the glasses. Empirically, we observed the following three key features of the video frames in this setup with both the mannequin and human (see Appendix D for details):

- Out-of-focus blur and tremor-caused motion blur are generally negligible when the reflected texts are recognizable.
- Inter-frame variance: The distortions at the same position of each frame are different, generating different noise patterns for each frame.
- Intra-frame variance: Even in a single frame, the distortion patterns are spatially non-uniform.

[2]We mainly consider motions that are parallel to the screen because generally, they cause larger blurs than other types of motions

One key observation is that the captured texts are subjected to occlusions (the missing or faded parts) caused by shot noise [19] when there is an insufficient number of photons hitting the sensors. This can be easily reasoned in light of the short exposure time and small text pixel size causing reduced photons emitted and received. In addition, other common imaging noise such as Gaussian noise gets visually amplified by relatively higher ISO values due to the bad light sensitivity of the webcam sensors. We call such noise ISO noise. Both two types of distortions have the potential to cause intra-frame and inter-frame variance. The shot and ISO noise in the webcam peeking attack plays on a see-saw with an equilibrium point posed by the quality of the camera imaging sensors. It suggests that the threat level will further increase (see the comparison between the webcam and Nikon Z7's images in Figure 14) as future webcams get equipped with better-quality sensors at lower costs.

E. Image Enhancing with MFSR.

The analysis of distortions calls for an image reconstruction scheme that can reduce multiple types of distortions and tolerate inter-frame and intra-frame variance. One possible method is to reconstruct a better-quality image from multiple low-quality frames. Such reconstruction problem is usually defined as multi-frame super resolution (MFSR) [58]. The basic idea is to combine non-redundant information in multiple frames to generate a better-quality frame.

We tested 3 common light-weight MFSR approaches that do not require a training phase, including cubic spline interpolation [58], fast and robust MFSR [36], and adaptive kernel regression (AKR) based MFSR [41]. Test results on the reflection images show that the AKR-based approach generally yields better results than the other two approaches in our specific application and setup. All three approaches outperform a simple averaging plus upsampling of the frames after frame registration, which may be viewed as a degraded form of MFSR. An example of the comparison between the different methods and the original 8 frames used for MFSR is shown in Figure 4 (a). We thus use the AKR-based approach for the following discussions.

One parameter to decide for the use of webcam peeking is the number of frames used to reconstruct the high-quality image. Figure 4 (b) shows the CWSSIM score improvement of the reconstructed image over the original frames with different numbers of frames used for MFSR when a human wears the glasses to generate the reflections. Note that increasing the number of frames do not monotonically increase the image quality since live users' occasional intentional movements can degrade image registration effectiveness in the MFSR process and thus undermine the reconstruction quality. Based on the results, we empirically choose to use 8 frames for the following evaluations. In addition, the improvement in CWSSIM scores also validates that MFSR-resulted images have better quality than most of the original frames. We thus only consider evaluation using the MFSR images in the following sections.

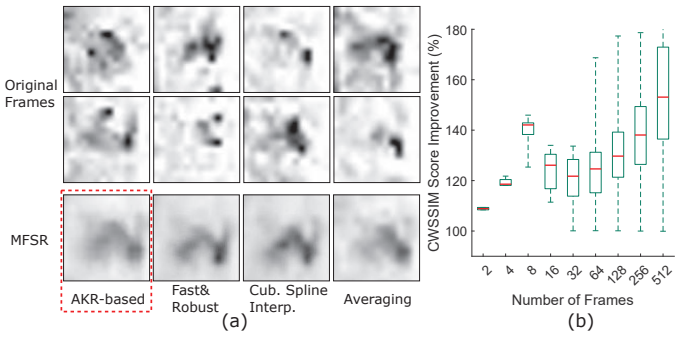

Fig. 4. (a) Comparison between single frames and the MFSR-reconstructed images with 4 different MFSR approaches. The MFSR images are reconstructed with the 8 frames shown at the top. The AKR-based approach generally produces the best reconstruction results in our task of reflection image reconstruction. (b) The improvement of reflection reconstruction quality as the number of frames used for MFSR increases.

IV. Reflection Recognizability & Factors

In this section, we evaluate the recognizability limits of reflected texts enhanced by the MFSR method given a specific set of webcams, glasses, and advantageous environmental conditions. We then investigate the impact of the most significant factors. The evaluations in this section are performed in a controlled lab environment and serve as the foundation for the analysis in Section V.

A. Experimental Setup

Equipment. We collected all data with the aforementioned Acer laptop as the victim device, and another Samsung laptop [18] as the adversary's device. The two laptops were in a lab environment with WiFi network connection. The victim laptop was measured to have an internet download speed of 246 Mbps and upload speed of 137 Mbps while those for the adversary laptop were 144 Mbps and 133 Mbps respectively. We used two pairs of glasses, i.e., the pair of BLB glasses and prescription glasses.

Data Collection. We asked a person to wear the glasses and sit in front of the victim's laptop. The glass-screen distance was chosen to be 40 cm which was also found to be close to the average distance in the user study (see Figure 9 (b)). The screen brightness was 100%. To estimate the limits of recognition, we used an environmental light intensity of 100 lux to generate the best reflections. We then displayed single capital letters (26 letters) on the victim screen with different heights ranging from 20 mm to 7 mm. The victim and adversary laptops had a Zoom [21] session with a video resolution of 1280×720. For each display of the letters, we recorded a 3s video of the victim's images on the adversary's laptop. We then used 8 consecutive frames starting from 1s for MFSR reconstruction and generated one corresponding image for each video. We generated 208 images in total for the 2 glasses each with 4 different sizes.

Recognizability Evaluation. In order to evaluate the recognizability of the reconstructed single-letter images and avoid potential bias introduced by the authors' prior knowledge

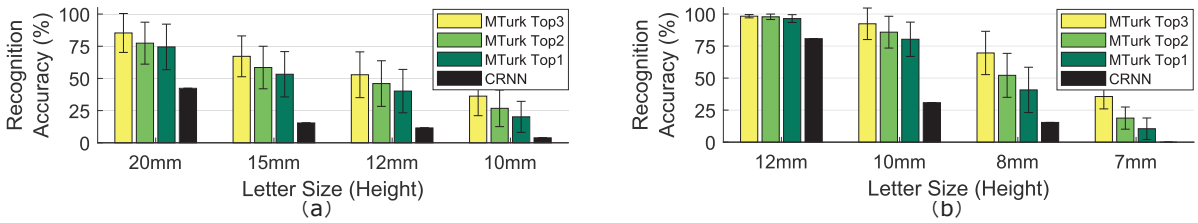

Fig. 5. The recognition accuracy of letters in different sizes with (a) the BLB glasses and (b) the prescription glasses. Although the pair of BLB glasses have higher reflectance than the prescription glasses, the prescription glasses enable reading smaller on-screen texts because of their smaller curvature leading to larger reflection pixel size. Note that the conclusion is device-specific and cannot be applied to general BLB-prescription glass comparison. Humans are found more capable of recognizing the reflected texts than SOTA OCR models.

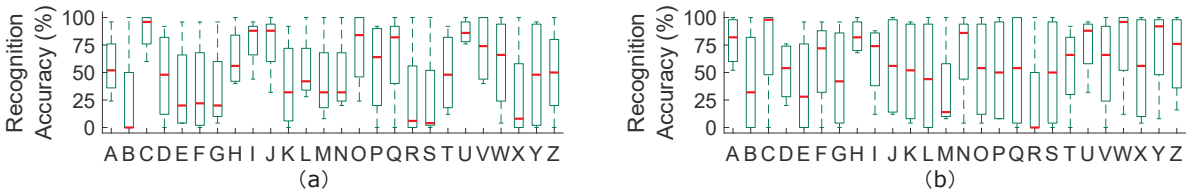

Fig. 6. The human recognition accuracy of different letters with (a) the BLB glasses and (b) the prescription glasses. Letters such as "R" have been found the most difficult to read in the reflections while letters such as "C" and "U" have high recognizability. The difference is mostly due to the simplicity and symmetry in the letters' structures which lead to smaller degradation of recognizability when the reflections are subject to distortions.

of the reflections, we acquired recognition accuracy by (1) using multiple SOTA pre-trained deep-learning OCR models including Google Tesseract and Keras CRNN, and (2) conducting a survey (Section VII-D) on Amazon Mechanical Turk (AMT) [9]. For the AMT study, we collected answers from 25 crowdsourcing workers for each reconstructed image and thus collected 5200 answers in total. We showed to the workers all reconstructed images in a randomized manner without providing them with any information on the original letters on the screen. We asked the workers to provide 3 best guesses of the single letter in each reconstructed image. They were allowed to input the same answer for multiple guesses if they feel confident in a guess, or if they have no clue about making subsequent guesses. The recognizability of the texts in the reconstructed images is then represented by the recognition accuracy, i.e., correctly recognized number of letters over the total number of letters in each case.

B. Recognizability vs. Size & Letter

Figure 5 shows the recognition accuracy with the BLB and prescription glasses respectively with different letter sizes. The AMT accuracy for each letter size is calculated by including all 25 answers for all 26 letters, i.e., with a denominator of $25 \times 26 = 650$. We picked 4 representative letter sizes for each pair of glasses respectively, and show the top 1, 2, and 3 recognition accuracy. we also use error bars to show the standard deviations. The SOTA OCR models performed considerably worse than AMT workers. We believe the main reason is that data distribution in the models' training sets is very different from the actual data in webcam peeking. After testing different image data on the models, we found the two main causes for their bad performance are (1) significantly lower contrast, (2) occlusions caused by insufficient photons. Surprisingly, we also found the models sensitive to how we crop the images, which suggests the convolutional layer features and potential data augmentation schemes employed by these models are not dealing well with our data's distribution. We think future researchers can potentially utilize these pretrained models and collect their own webcam peeking dataset to fine-tune the model weights to better adapt machine learning recognition models to this scenario.

The prescription glasses generally yield better results for the webcam peeking attack, showing that 10 mm texts can be recognized in the reconstructed images with over 75% accuracy. Although not as good as the prescription glasses, the recognition accuracy with the BLB glasses is also high enough to support efficient peeking attacks against texts of 10-20 mm. Despite the better reflective characteristics of the BLB glasses, the prescription glasses still generate better results due to their smaller curvature, highlighting the risks of the peeking attack even without highly reflective glasses.

Intuitively, different letters in the alphabet would be recognized with different levels of hardships due to their structural characteristics (see Figure 6). For instance, the letters "R" and "B" have been found the hardest to recognize in both cases of the two pairs of glasses. On the other hand, letters such as "C", "U", "I", and "O" have generally the highest recognizability in all the sizes, which we suspect is due to their simple or highly symmetric structures that prevent the recognizability of such letters from dropping too seriously when the texts are down-sampled and occluded. Furthermore, we found letters having similar structures are confused with each other more easily in the recognition. For instance, "J" and "L" are mostly recognized as "I" when the letter size gets small because the distortions to the bottom part of "J" and "L" makes them just like "I" in the reflection images.

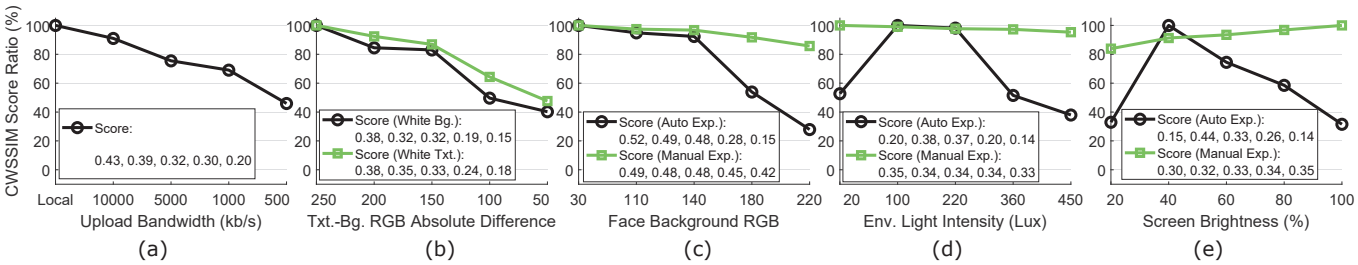

Fig. 7. Effects of impact factors evaluated by CWSSIM scores. The original score numbers are displayed along with the legend at the bottom, and we plot the ratio between each score and the highest score in each case as a percentage. Visualizations of the effects can be found in the appendix.

C. Network Influence

Video conferencing platforms like Zoom cause different levels of distortions in the images through video encoding and decoding under various network bandwidths. To analyze the impact, we compared the quality of the reconstructed images under different network bandwidths to that when the video is recorded by the victim's local device without going through Zoom. A visual demonstration of the effect is shown in Figure 17 which is quantified with CWSSIM scores and shown in Figure 7 (a). We found that when the upload bandwidth is larger than 10 Mbps, the quality of the reconstructed images generally remains the same, and is close to the locally-captured and reconstructed images with a minor degree of added distortions. An upload bandwidth smaller than 10 Mbps starts to undermine the reconstructed image quality over Zoom. When the bandwidth is smaller than 1000 kpbs, the letters get hard to recognize. It's almost unrecognizable with a bandwidth smaller than 500 kbps. When the bandwidth was larger than 1500 kbps, Zoom was generally able to maintain a 720p video resolution with a frame rate close to 30 fps (Appendix C).

D. Physical Factors

The recognizability of the reflections is a highly complex multi-variate function over many physical factors. We categorize the factors into 2 groups, namely those mainly affecting the reflection pixel size (Section III-B) and those affecting the light SNR. Comprehensive quantitative modeling of light SNR is very challenging due to the need for accurate imaging sensor models. Nevertheless, we provide qualitative analysis and quantify representative cases by calculating changes in CWSSIM scores (Figure 7).

In light SNR, the signal portion comes from the light emanating from the screen, reflected by the glasses, and then captured by the imaging sensors corresponding to the area of the screen. Other light captured by sensors in this area can be treated as noise. Counter-intuitively, more reflected light does not always lead to higher reflection recognizability as we will discuss next. Figure 7 (b-e) show the factors that can change light SNR most significantly. (c-e) also inspect how auto exposure and manual (fixed) exposure can affect the light SNR-recognizability relationships in surprisingly different ways by using the laptop built-in webcam and the configurable Nikon Z7 respectively.

Text Color Contrast. Different colors of texts can affect the reflection recognizability because the texts and screen background colors produce a certain contrast. We found that chroma has smaller effects than luma and show how luma affects reflection quality in Figure 7 (b) (visualization in Figure 17 (b)) by using the absolute difference in RGB values of gray-scale text and background colors to represent the contrast. As expected, lower contrast (smaller RGB difference) undermines the reflection recognizability.

Face Background Reflectance. Face background reflectance is decided by sub-factors such as skin color. We tested different background reflectance by pasting the inner side of the glasses with papers of different gray-scale colors that have the same values for RGB. When the background has a higher reflectance (larger RGB values), more light from the environment as well as the screen will be reflected by it, increasing the noise portion of the light SNR and thus undermining the recognizability of the reflections as shown in Figure 7 (c) (visualization in Figure 17 (c)).

Environment Light Intensity. A decrease in the environmental light intensity causes a smaller degree of noise and thus increases the light SNR. This increase, however, does not necessarily lead to better recognizability in the case of webcams which often have auto-exposure control to adjust the overall brightness of the videos they take. When the overall environment is too dark, the webcam's firmware automatically increases the exposure time trying to compensate for the dark environment. This increase in the exposure time can cause an over-exposure for the reflected contents on the glasses which could have much higher light intensity than the environment, leading to smaller contrast and thus harder-to-read images. Such over-exposure is found in multiple participants' videos in the user study (Section V-B). On the other hand, the recognizability monotonically increases in the case of manual-exposure cameras such as the Nikon Z7 in manual mode. Figure 7 (d) (visualization in Figure 17 (d)) shows the different behaviors of auto and manual exposure.

Screen Brightness. Screen brightness is the opposite of environmental light intensity in terms of its impact on the reflection recognizability. When the screen is brighter, the signal portion in the light SNR increases and can lead to more readable reflections for manual-exposure cameras. However, auto-exposure of most webcams can again negatively affect

recognizability. Specifically, if the screen gets too bright compared to the environmental lighting condition, the webcams will often adjust their exposure time and ISO based on the dominant environmental light condition, and thus cause over-exposure to the screen reflections. Figure 7 (e) (visualization in Figure 17 (e)) shows the effects.

Summary. The results show that variations in physical conditions can change the actual limits of the attack dramatically. The fact that reflection recognizability does not change monotonically with some factors like environmental light intensity and screen brightness further challenges the attack by making it more difficult to predict the possible outcomes in uncontrolled settings.

E. Eyeglass Lens

The difference in recognition accuracies between the pair of BLB and prescription glasses (Figure 5) suggests parameters of different eyeglass lenses will influence the performance of webcam peeking. To examine the impact, we analyzed 16 pairs of eyeglasses by inspecting the correlation between their reflection quality quantified by CWSSIM scores and several lens factors. The CWSSIM scores are acquired with the 16 glasses when all other factors are kept the same.

The results suggest lens focal length, which determines the pixel size of reflections (Equation 3), has the strongest influence on the reflections with a correlation score of 0.56. The minimum, mean, and maximum focal length of the 16 pairs of glasses are 10, 268, and 110 cm respectively. With a correlation score of 0.42, the second strongest factor is found to be prescription strength (lens power) as lens power usually has a positive correlation with focal length following design conventions (see Appendix A for explanation). Lens reflectance and surface coating conditions that mainly affect reflection light SNR produce correlation scores of 0.32 and 0.31 respectively. We empirically defined and added the factor of lens coating condition that gauges how much the lens coatings have worn off with higher values representing more intact coating. The motivation is our observation that damage in lens coating reduces the recognizability of reflections (see Figure 11). We also estimated lens reflection spectrum by calculating the ratio between RGB values of the reflections in the image but only found correlation scores lower than 0.15. This suggests the glass type (e.g., BLB or non-BLB) does not have a strong influence on reflection quality. Finally, we expect the parameters analyzed above have certain relationships with lens and coating materials, which require specialized optical equipment to measure and determine.

V. Cyberspace Textual Target Susceptibility

The evaluations so far are based on the text's physical size and carried out in controlled environments to better characterize user-independent components of the reflection model as well as the range of theoretical limits for webcam peeking. In this section, we start by mapping the limits to common cyberspace objects in order to understand the potential susceptible targets. We then conduct a 20-participant user study with both local and Zoom recordings to investigate the feasibility and challenges of peeking these targets and various factors' impact.

A. Mapping Theoretical Limits to Targets

We use web texts as an enlightening example of cyberspace textual targets considering their wide use and the relatively mature conventions of HTML and CSS. The discussion is based upon (1) a previous report [48] scraping the most popular 1000 websites on Alex web ranking [8], and (2) a manual inspection of 117 big-font websites archived on SiteInspire [10]. We further divide the inspected web texts into 3 groups ($\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, see Appendix E and Table III) in order to discuss separately how the webcam peeking attack with current and future cameras could have effects on them. As pointed out in Section III-B, the conversion between digital point size and physical cap height is dependent on specific user settings such as browser zoom ratio. The cap height values in Table III are thus measured with the Acer laptop with default OS and browser settings as a case study.

Based on the results in Figure 5, we hypothesize that the smallest cap heights adversaries can peek using mainstream 720p cameras is 7-10 mm. We then calculate the corresponding limits with 1080p and 4K cameras with Equation 3 and show them in the Theoretical column of Table III. Considering participants are most likely to use 720p cameras, we then choose point sizes S1-S6 in Table III for evaluations.

B. User Study

The user study (Section VII-D) is designed in the following challenge-response way: An author generates HTML files each with one randomly selected headline sentence containing 7-9 words [3] from the widely-used "A Million News Headlines" dataset [46]. Only each word's first letter is capitalized. The participants display the HTML page in their browsers when they are recorded, and another author acting as the adversary tries to recognize the words from the videos containing the 20 participants' reflections without knowing the HTML contents by using the same techniques as in Section IV. We then calculate the percentage of correctly recognized words.

Data Collection. Each participant was given 6 HTML files of increasing point sizes from S1 to S6 as shown in Table III. Note that the 6 sizes are specified in point size in HTML so that user-dependent factors such as screen size and browser zoom ratio can be studied (Equation 1). The participants display each HTML file on their own computer display in their accustomed rooms and behave normally as in video conferences. We allow participants to choose their preferred environmental lighting condition except asking them to avoid other close light sources besides the screen in front of their face. The reason is that we found a close frontal light source can seriously decrease light SNR, which can potentially be used as a physical mitigation against this attack but prevents us from examining the impact of all the other

[3] Uniform lengths (e.g., all 8 words) are avoided to prevent the adversary from guessing the words by knowing how long the sentences are.

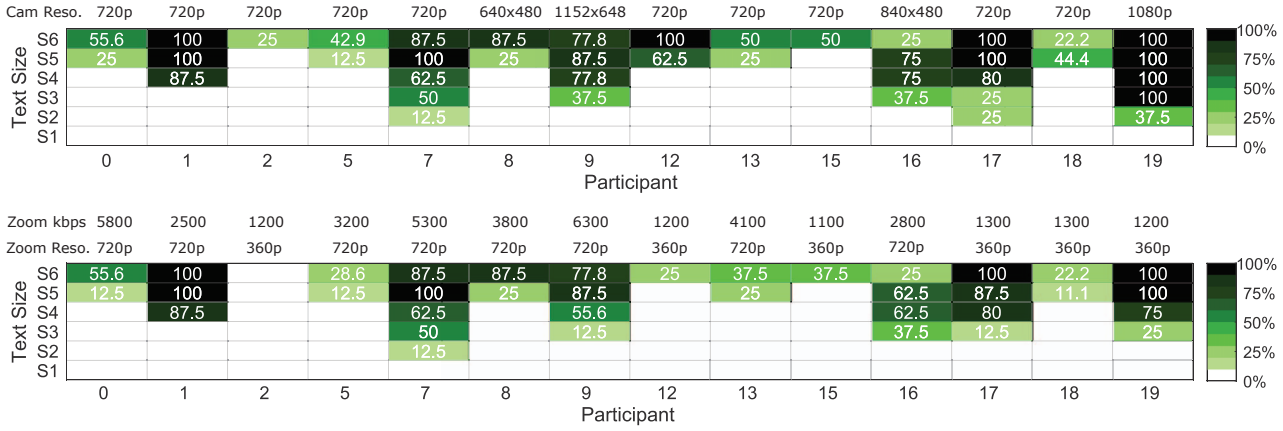

Fig. 8. The recognition results of textual reflections collected with local and Zoom-based remote video recordings from 20 user study participants. Participants 4, 14, and 3, 6, 10, 11 did not generate glass reflections that allow successful recognition due to problems of out-of-range viewing angles and very low light SNR respectively and are thus omitted from the figure.

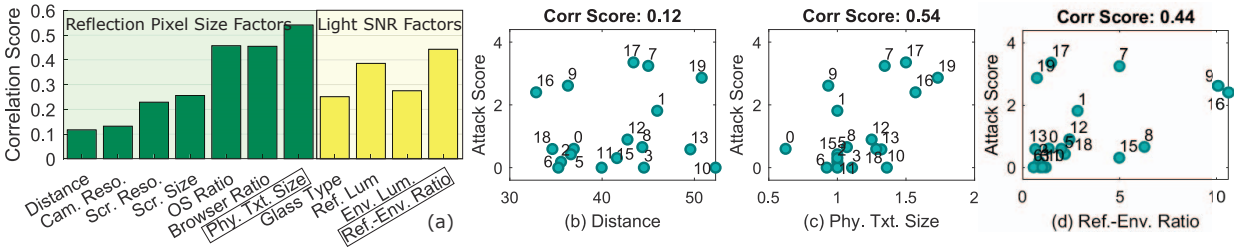

Fig. 9. (a) The degree of influence of different factors on the reflection recognition performance evaluated by the correlation scores. Factors highlighted with boxes are computed with other raw factors according to our model. (b-d) The joint distribution of three factors and the recognition results.

factors. We did not tell the participants to stay stationary and let them behave normally as in browsing screen contents. Their webcams record their image for 30 seconds for each HTML.

Network bandwidth and resulted video quality are artifacts of video conferencing platforms that improve in a rapid way [4] compared to other user-dependent physical factors. To study the present-day and possible future impact of video conferencing platforms, we record the 20 participants' videos both locally and remotely through Zoom. Our experiments focused on Zoom since it is the most used platform and also provides the most detailed video and network statistics.

We asked the participants to report their user-dependent parameters including screen resolution (N_{os}), screen physical size (H_{sr}), OS and browser zoom ratio (s_{os}, s_b) webcam resolution in Equation 1, webcam resolution (N) in Equation 3, and the type of their glasses. Some other physical factors including environmental light intensity, screen brightness, glass-screen distance, and the physical size of displayed texts are difficult to be measured by the participants themselves and are not reported. We thus estimated the values of these factors by utilizing their videos.

General Adversary Recognition Results. The recognition results achieved by the adversary with local and remote recordings are shown in Figure 8 (upper and lower respectively). Two participants (4 and 14) did not generate glass reflections of their screens in the video recordings due to the problem of out-of-range vertical viewing angles as predicted in Section III-B. Four participants (3, 6, 10, 11) yield 0% textual recognition accuracy due to a very low light SNR.

With local video recordings, the percentage out of the 20 participants that are subjected to non-zero recognition accuracy against S6-S1 are 70%, 60%, 30%, 25%, 15%, and 0% respectively. Videos of participants 7 and 17 using 720p cameras allowed the adversary to achieve 12.5% and 25% accuracies on recognizing S2. Videos of participant 16 using a 480p camera allowed the adversary to achieve an 37.5% accuracy on recognizing S3. These results translate to the predicted susceptible targets with cameras of different resolutions as listed in the User column of Table III, where 720p webcams pose threats to large-font webs ($\mathcal{G}_3$) and future 4K cameras pose threats to various header texts on popular websites ($\mathcal{G}_1$ and $\mathcal{G}_2$). As expected, this result is worse than the theoretical limits in the table that are derived with prescription glass data in the controlled lab setting (Section IV). Our observations suggest the main reasons include: (1) The environmental lighting conditions of the users are more diverse and less advantageous to screen peeking than the lab setup, generating reflections with worse light SNR. (2) Texts in the user study are mostly lower-case and have thus smaller physical sizes than the upper-case letters used in Section IV. (3) The prescription glasses used in Section IV have a larger focal length than the average user's glasses. (4) More intentional movements exist

in the user study leading to more motion blur.

With Zoom-based remote recordings, the percentage of participants with non-zero recognition accuracy against S6-S1 degraded to 65%, 55%, 30%, 25%, 5%, and 0% respectively. We logged the video network bandwidth and resolution reported by Zoom as shown in Figure 8. The correlation between Zoom bandwidth, resolution, and their impact on video quality agrees with the observations in Section IV-C. Generally, bandwidths smaller than 1500 kbps led to 360p resolutions for most of the time and decreased the recognizable text size by 1 level. Zoom's 720p videos also caused degradation in recognition accuracy but mostly kept the recognizable text size to the same level as the local recordings, suggesting the same predictions of susceptible text sizes and corresponding cyberspace targets.

Besides the mostly used platform Zoom, we also acquired remote recordings of participant 19 with Skype and Google Meet. The adversary achieved better results with Skype than Zoom by recognizing S3 and S2 with 89% and 25% accuracies respectively, which is likely due to Skype's capability of maintaining better-quality video streams with a 1200 kbps bandwidth. The web-based Google Meet platform provided the lowest quality videos and only allowed the adversary to achieve 22% accuracy on recognizing S4.

Underlying Reasons. To find out the dominant reasons enabling easier webcam peeking by analyzing the correlation between the recognition results and different factors, we turn each participant's results (6 sizes) into a single *attack score* that is a rectified weighted sum of the recognition accuracy of the six text sizes tested. Figure 9 (a) shows correlation scores with 11 factors that affect reflection pixel size (left) and light SNR (right) respectively when $w = 1.5$. The glass type includes prescription (15/20) and prescription with BLB coatings (5/20). The physical text size and reflection-environment light ratio highlighted in the boxes are two composite factors. In short, the physical text size represents the ratio between the actual physical size of texts displayed on each participant's screen and the case study values in Table III and is calculated with Equation 1 with other raw factors such as browser zoom ratios. The reflection-environment light ratio represents how strong the screen brightness is compared to the environmental light intensity and is calculated by dividing glass luminance by environmental luminance. Basically, these two composite factors represent our model's prediction of reflection pixel size and light SNR and are found to generate higher correlation scores than the other raw factors, which validates the effectiveness of our models. Figure 9 (b-d) further show the joint distribution of the attack score and three representative factors. It can be seen from (b) that the 40 mm screen-glass distance used in the evaluation of Section IV is about the average of the participants' values, and distances of these participants actually only have a very weak correlation with the easiness of webcam peeking attack. Figure 9 (d) suggests that when the screen brightness-environmental light intensity ratio gets lower than a certain threshold, the likelihood of preventing adversaries from peeking is very high, which may be considered as a temporary mitigation.

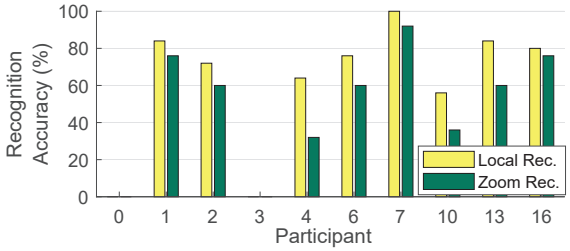

Fig. 10. Accuracy of recognizing Alexa top 100 websites from eyeglass reflections. Each participant browsed 25 websites. Participant 0 and 4 did not yield recognizable reflections due to bad light SNR and viewing angles.

VI. Website Recognition

The results so far suggest it may still be challenging for present-day webcam peeking adversaries with mainstream 720p cameras to eavesdrop on common textual contents displayed on user's screens. During our experimentation, we observed that recognizing graphical contents such as shapes and layouts on the screen is generally easier than reading texts. Although shapes and layouts contain more coarse-grained information compared to texts, a webcam peeking adversary may still pose non-trivial threats by correlating such graphical information with privacy-sensitive contexts. This work further explored to which degree can a webcam peeking adversary recognize on-screen websites by utilizing non-textual graphical information.

Data Collection. 10 out of the 20 participants in the user study participated in the website recognition evaluation. Following a similar methodology as in [42], we used the Alexa top 100 websites as a closed-world dataset. We only investigate the recognition of the home page of each website in this work. [42] shows that other pages of a website can also lead to the recognition of the website. We believe the easiness of recognizing a website using different pages is worth exploring in future works. The experiment followed a similar procedure as the textual recognition experiment in Section V. For each participant, one author generates a unique random sequence of 25 websites for the participant to browse (10 seconds for each website) while another author acts as the adversary that analyzes the video recordings. Both local and Zoom-based remote recordings were obtained and recognized by the adversary. The adversary was given the whole recording and was asked to match each segment of the video to a specific website out of the 100 websites in the correct order. A random guess naive adversary is supposed to have a success rate of about 1%. Note that some participants changed their environment and ambient lighting compared to the previous textual recognition experiment since the two experiments were conducted five months apart.

Recognition Results. Figure 10 shows the percentage of websites (out of 25) correctly recognized by the adversary. Participants 0 and 4 did not yield recognizable reflections due to bad light SNR and viewing angles respectively. This ratio of zero recognition (2 out of 10) agrees with that in the textual recognition test (6 out of 20), suggesting that webcam peeking

may be impossible in 20-30% video conferencing occasions due to extreme user environment configurations.

As expected, participants with higher textual recognition accuracies such as participant 7 generally yield higher website recognition accuracies too. In addition, we observe that website recognition is more robust to various lighting conditions in the participants' ambient environment. For example, we found participant 10 who had 0% textual recognition accuracy due to bad light SNR produced 56% (local) and 36% (remote) accuracies in website recognition with the same environment and lighting. The reasons are two-fold. First, solid graphical contents such as color blocks commonly found on web pages occupy larger areas than the body of texts and are thus much easier to identify in low-quality videos. Second, compared to black texts on white backgrounds which only have two different colors, the overall web pages with multiple graphical contents have more colors and contrast, leading to better robustness against over- and under-exposure of the usable screen contents in the webcam videos.

Recognition Easiness and Web Characteristics. Compared to texts, websites feature more abundant and diverse characteristics. We conducted qualitative and quantitative analyses to identify the characteristics that make certain websites more susceptible to webcam peeking. To that end, we ranked the 100 websites by their easiness of recognition utilizing recognition accuracies. Figure 16 shows rotated screenshots of the websites that rank the top and bottom 15 by their recognition easiness. Visual inspections suggest websites with higher contrast, larger color blocks, and more salient relative positions between different color blocks are easier to recognize. Websites that are mostly white with sparse textual and graphical components on them are the hardest to recognize. We calculated the correlation scores between the rank of each website and the average as well as the standard deviation of the websites' pixel values. Generally, a higher average means the website is closer to a pure white screen; a higher standard deviation means the website has more abundant high-contrast textures. The correlation scores obtained are -0.33 and 0.45.

VII. Discussion

A. Proposed Near-Term Mitigations

Given the threats, it is worthwhile exploring feasible mitigations that can be applied immediately. A straightforward approach involves users modifying the dominant physical factors identified in this work to reduce reflections' light SNR, e.g., by placing a lamp facing their face whose light increases the noise portion of light SNR. For software mitigations, we notice Zoom provides virtual filters of non-transparent cartoon glasses that can completely block the eye areas and thus eliminate reflections. Such features are not found in Skype or Google Meet. Other software-based approaches that support better usability involve fine-tuned blurring of the glass area. Although none of the platforms supports it now, we have implemented a real-time eyeglass blurring prototype that can inject a modified video stream into the video conferencing software. The prototype program[4] locates the eyeglass area and apply a Gaussian filter to blur the area. Figure 15 demonstrates the effect of using different strengths of Gaussian filtering by tuning the σ parameter. Stronger filtering (higher σ) reduces reflection quality more but also undermines usability and user experience to a larger degree as it makes the users' eye areas look more unnatural. We believe the usable strength also depends on the characteristics of specific glasses. For example, Figure 15 shows three pairs of glasses with increasing reflectance. Since glasses with higher reflectance (e.g., the 3rd row) may already have produced screen reflections that occupy and distort images of users' eye areas, applying stronger filtering may cause less degradation in user experience in this case. On the other hand, lower-reflectance (e.g., the 1st row) glasses may require weaker filtering to maintain the same degree of usability. In general, we believe it is a good idea for future platforms incorporating this protection mechanism to allow users to adjust filtering strength by themselves.

B. Improve Video-conferencing Infrastructure

Individual Reflection Assessment Procedure. Our analysis and evaluation reveal that different individuals face varying degrees of potential information leakage when subjected to webcam peeking. Specifically, various factors of software settings, hardware devices, and environmental conditions affect the quality of reflections. Even for the same user, the potential level of threats varies when the user joins video conferences from different places or at different times of the day. These factors make it infeasible to recommend or implement a single set of protection settings (e.g., what glasses/cameras/filter strength to use) before the actual user settings are known.

Providing usable security requires an understanding of how serious the problem is before trying to eliminate the problem. In light of this, we advocate an individual reflection assessment procedure that can potentially be provided by future video conferencing platforms. The testing procedure can be made optional to users after notifying them of the potential risk of webcam peeking. The procedure may follow a similar methodology as the one used in this work by (1) displaying test patterns such as texts and graphics, (2) collecting webcam videos for a certain period of time, (3) comparing reflection quality in the video with test patterns to estimate the level of threats of webcam peeking. With the estimated level of threats, the platform can then notify the user of the types of on-screen content that might be affected and offers options for protection such as filtering or entering the meeting with the PoLP principle that will be discussed below.

Principle of Least Pixels. Cameras are getting more capable than what average users can understand—unwittingly exposing information beyond what users intend to share. The fundamental privacy design challenge with webcam technology is "oversensing" [28] where overly-capable sensors can provide too much information to downstream processing—more data

[4]Details and open-source code of this prototype implementation can be found at https://github.com/longyan97/EyeglassFilter.

than is needed to complete a function, such as a meaningful face-to-face conversation. This oversensing leads to a violation of the sensor equivalent to the classic *Principle of Least Privilege (PoLP)* [52]. We believe long-term protection of users ought to follow a PoLP (perhaps a Principle of Least Pixels) as webcam hardware and computer vision algorithms continue to improve. Thus, we recommend that future infrastructure and privacy-enhancing modules follow the PoLP not just for software, but for the camera data streams themselves. In sensitive conversations, the infrastructure could provide only the minimal amount of information needed and allow users to incrementally grant higher access privileges to the other parties. For example, PoLP blurring techniques might blur all objects in the video meeting at the beginning and then intelligently unblur what is absolutely necessary to hold natural conversations.

C. User Opinion Survey

We collected opinions on our findings of webcam peeking risks and expectations of protections from 60 people including the 20 people who participated in the user study and 40 people who did not. We did not find apparent differences between the two group's opinions. The overall opinions are reported below.

Textual Recognition. For the discovered risk of textual recognition, 40% of the interviewees found it a larger risk than what they expected; 48.3% thought it was almost the same as their expectation; 11.7% expected worse consequences than what we found. In addition, 76.7% of the interviewees think this problem needs to be addressed while 23.3% think they can tolerate this level of privacy leakage.

Website Recognition. 61.7% of the interviewees found it a larger risk than what they expected; 30% thought it was almost the same as their expectation; 8.3% expected worse consequences than what we found. In addition, 86.7% of the interviewees think this problem needs to be addressed while 13.3% think they can tolerate this level of privacy leakage.

Reflection Assessment. Regarding the proposed idea of reflection assessment procedures that may be provided by video conferencing platforms in the future, 95% of the interviewees said they would like to use it; 85%, 68.3%, 45%, and 20% of the 60 interviewees would like to use it when meeting with strangers, colleagues, classes, and family/friends respectively.

Glass-blur Filters. Regarding the possible protection of using filters to blur the glass area, 83.3% of the interviewees said they would like to use it; 78.3%, 51.7%, 43.3%, and 11.7% of the 60 interviewees would like to use it when meeting with strangers, colleagues, classes, and family/friends respectively.

D. Ethical Considerations

The AMT and user opinion survey received IRB waiver (No.HUM00208544) from the authors' institutes. The downloaded results are de-anonymized by only keeping their answers and deleting all other identifiable information including worker IDs. The results on the AMT and survey websites are deleted. We provided compensation of $18/h for the workers.

The textual and website recognition user studies are IRB-approved (No.ZDSYHS-2022-5). We ensured that participants and others who might have been affected by the experiments were treated ethically and with respect and anonymized participants with random orders. No personal information other than the videos and questionnaires was collected. The HTML files they used were created randomly by the authors and do not involve the participants' private information or contain any unethical or disrespectful information. The securely stored videos were used only for this research and not disclosed to third parties or used for other purposes.

E. Limitation & Future Work

This work used human-based recognition to evaluate the performance limits of reflection recognition. In future scenarios such as forensic investigations carried out by specialized institutions, we believe trained expert humans or machine learning methods may be employed to further increase the accuracy of reflection recognition. Compared to machine learning-based recognition, human-based recognition helps us understand the threats posed by a wide range of adversarial parties including even common users of video conferencing, and thus provides an estimate of the lower bound of the limits posed by camera hardware and other factors. We believe it is always possible to improve the attack performance by designing a more sophisticated machine learning model with more parameters, increasing the size and diversity of the training dataset, etc. Further, machine learning recognition is likely to face over-fitting and generalizability problems in webcam peaking due to highly varying personal environment conditions. Thus, we believe limits posed by a machine learning recognition back end are subjected to very large variances and require dedicated future works to quantify

Certain levels of biases have been introduced in the user study by informing the participants of the study's purpose. We envision that a future study may conduct a real-world validation of this attack by performing it without participants' awareness while carefully following ethical regulations. Alternatively, public videos on social media may be analyzed to investigate how often such information leakage happens. A future study could also systematically interview professionals in different types of businesses and explore information leakage conditions, frequencies, and concerns. Contextual factors and user attitudes in real-world situations are complementary to this work's focus and worth investigating in future research.

VIII. Related Work

The problem of screen reconstruction is a long-studied challenging problem. In this section, we analyze the past works that served as the foundations for our thinking in the context of video conferencing today and in the predicted future.

Screen Peeking Using Cameras. Screen-peeking with cameras through optical emanation reflections has been explored in previous works. In 2008, Backes et al. [26] showed that adversaries can use off-the-shelf telescopes and DSLR cameras to spy victims' LCD monitor screen contents from up to 30m

away by utilizing the reflective objects that can be commonly found next to the monitor screen such as teapots placed on a desk. In 2009, the authors [25] took the attack to the next level by addressing the challenges of motion blur and out-of-focus blur by performing deconvolution on the photos with Point Spread Functions (PSF). Our work differs from these previous works by exploiting the victims' own webcams in video conferences for a remote attack. Such changes call for different imaging enhancing techniques due to the different types of image distortions. In addition, reflective objects on the desks and human eyes cannot be easily utilized due to very large curvatures. We thus exploit the glasses people wear to video conferences as a modern attack vector. [57] proposed a relevant idea of using adversary-controlled webcams to detect changes in webpage links' colors for inferring visited websites. It requires the adversary to take control over the victim's webcam with malicious web modules and exploits coarse-grain color variations, while our work studies more natural attack vectors in video conferencing and investigate the limits of textual reconstruction.

Screen Content Reconstruction With Other Emanations. Besides the direct optical emanations from the screen that we exploit in this work, previous works also explored other channels such as electromagnetic radiation [44], [45], [55] and acoustic emanations [37]. Reconstructing screen contents with such emanations usually requires using additional eavesdropping hardware that is placed close to the victims by the adversary. On the other hand, our work exploits the victim's own webcams, making the attack more accessible.

Remote Eavesdropping Via Audio/Video Calls. Similar to our work, such attacks assume the adversary and victim are both participants of an audio/video conference, and the adversary can eavesdrop on privacy-sensitive information by analyzing the audio/video channels. For example, Voice-over-IP attacks for keystroke inference eavesdrop on the victim's keyboard inputs by utilizing timing and/or spectrum information embedded in the keystroke acoustic emanations [29], [30], [35], [54]. Recently, Sabra et al. [51] proposed works solving the problem of inferring keystrokes by analyzing the dynamic body movements embedded in the videos during a video call. Hilgefort et al. [39] spies victims' nearby objects through virtual backgrounds in video calls by carrying out foreground-background analysis and accumulating background pixels. In contrast, our work explores the related problem of content reconstruction using only the optical reflections from participants' glasses embedded in the videos.

IX. Conclusion

In this work, we characterized the threat model of the webcam peeking attack in video conferencing settings. We developed mathematical models that describe the relationship between the attack limits and different user-dependent factors. The analysis enables the prediction of future threats as webcam technology evolves. We conducted experiments both in controlled lab settings and with a user study. Results showed that present-day 720p cameras pose threats to the contents on users' screens when users browse certain big-font websites. Future 4K cameras are predicted to allow adversaries to reconstruct various header texts on popular websites. We also found adversaries can recognize the website users are browsing through webcam peeking with 720 cameras. We analyzed both short-term mitigations and long-term defenses and collected user opinions on the possible protections.

Acknowledgement

This work is supported by a gift from Analog Devices Inc. and China NSFC Grant 61925109 and 62201503. We thank our reviewers for their insightful comments that helped us improve this manuscript; we thank Dr. Cheng Yang for volunteering to wear eyeglasses in the lab experiments.

References

[1] Converting diagonal field of view and aspect ratio to horizontal and vertical field of view. http://vrguy.blogspot.com/2013/04/converting-diagonal-field-of-view-and.html, 2013.
[2] Webcam Field of View . https://www.telehealth.org.nz/assets/Uploads/1511-webcam-field-of-view.pdf, 2015.
[3] Approximate Focal Length for Webcams and Cell Phone Cameras. https://learnopencv.com/approximate-focal-length-for-webcams-and-cell-phone-cameras/, 2016.
[4] Cisco Annual Internet Report (2018–2023) White Paper. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, 2020.
[5] Schott AG: Transmittance of optical glass. https://www.schott.com/d/advanced_optics/5b1f5065-0587-4b3f-8fc7-e508b5348012/, 2020.
[6] The most maddening part about working from home: video conferences. . https://www.washingtonpost.com/technology/2020/03/16/remote-work-video-conference-coronavirus/, 2020.
[7] Acer Predator 15. https://www.acer.com/ac/en/IN/content/predator-model/NH.Q1YSI.001, 2021.
[8] Alexa SEO and Competitive Analysis Software. https://www.alexa.com/, 2021.
[9] Amazon Mechanical Turk. https://www.mturk.com/, 2021.
[10] Big Type Websites. https://www.siteinspire.com/websites?categories=22, 2021.
[11] Blue Light Blocking Glasses Market Size 2021 with a CAGR of 7.7% , Research by Business Opportunities, Top Companies data report covers, globally Market Key Facts and Forecast to 2025. https://www.wboc.com/story/43536337/blue-light, 2021.
[12] Blue Light Blocking Glasses on Amazon. https://www.amazon.com/gp/product/B07VBFSY33/, 2021.
[13] Cheese. https://wiki.gnome.org/Apps/Cheese, 2021.
[14] Default style sheet for HTML 4. https://www.w3.org/TR/CSS2/sample.html, 2021.
[15] For better or worse, working from home is here to stay. https://www.cnbc.com/2021/03/11/one-year-into-covid-working-from-home-is-here-to-stay.html, 2021.
[16] Let's Talk About Base Curves. https://opticianworks.com/lesson/lets-talk-base-curves/, 2021.
[17] Nikon Z7. https://www.nikonusa.com/en/nikon-products/product/mirrorless-cameras/z-7.html, 2021.
[18] Samsung Notebook 9. https://www.samsung.com/hk/pc/notebook-9-np900x5m-k03/, 2021.
[19] Shot Noise. https://en.wikipedia.org/wiki/Shot_noise, 2021.
[20] Web Style Sheets CSS tips & tricks: EM. https://www.w3.org/Style/Examples/007/units.en.html#units, 2021.
[21] Zoom. https://zoom.us/, 2021.
[22] Aries Arditi. Adjustable typography: an approach to enhancing low vision text accessibility. *Ergonomics*, 47(5):469–482, 2004.
[23] Aries Arditi and Jianna Cho. Serifs and font legibility. *Vision research*, 45(23):2926–2933, 2005.
[24] Melanie Arntz, Sarra Ben Yahmed, and Francesco Berlingieri. Working from home and covid-19: The chances and risks for gender gaps. *Intereconomics*, 55(6):381–386, 2020.

[25] Michael Backes, Tongbo Chen, Markus Dürmuth, Hendrik PA Lensch, and Martin Welk. Tempest in a teapot: Compromising reflections revisited. In *2009 30th IEEE Symposium on Security and Privacy*, pages 315–327. IEEE, 2009.

[26] Michael Backes, Markus Dürmuth, and Dominique Unruh. Compromising reflections-or-how to read lcd monitors around the corner. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 158–169. IEEE, 2008.

[27] Alexander Bick, Adam Blandin, and Karel Mertens. Work from home after the covid-19 outbreak.*CEPR Discussion Paper*. 2020.

[28] Connor Bolton, Kevin Fu, Josiah Hester, and Jun Han. How to curtail oversensing in the home. *Communications of the ACM*, 63(6):20–24, 2020.

[29] Stefano Cecconello, Alberto Compagno, Mauro Conti, Daniele Lain, and Gene Tsudik. Skype & type: Keyboard eavesdropping in voice-over-ip. *ACM Transactions on Privacy and Security (TOPS)*, 22(4):1–34, 2019.

[30] Alberto Compagno, Mauro Conti, Daniele Lain, and Gene Tsudik. Don't skype & type! acoustic eavesdropping in voice-over-ip. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 703–715, 2017.

[31] Zechuan Deng, René Morissette, and Derek Messacar. Running the economy remotely: Potential for working from home during and after covid-19. *Statistics Canada*. 2020.

[32] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *arXiv preprint arXiv:2004.07728*, 2020.

[33] Rodger J Elble. Tremor. In *Neuro-geriatrics*, pages 311–326. Springer, 2017.

[34] Rodger J Elble, Helge Hellriegel, Jan Raethjen, and Günther Deuschl. Assessment of head tremor with accelerometers versus gyroscopic transducers. *Movement Disorders Clinical Practice*, 4(2):205–211, 2017.

[35] Fürkan Elibol, Uğur Sarac, and Işin Erer. Realistic eavesdropping attacks on computer displays with low-cost and mobile receiver system. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pages 1767–1771. IEEE, 2012.

[36] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.

[37] Daniel Genkin, Mihir Pattani, Roei Schuster, and Eran Tromer. Synesthesia: Detecting screen content via remote acoustic side channels. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 853–869. IEEE, 2019.

[38] Atsuki Higashiyama, Yoshikazu Yokoyama, and Koichi Shimono. Perceived distance of targets in convex mirrors. *Japanese Psychological Research*, 43(1):13–24, 2001.

[39] Jan Malte Hilgefort, Daniel Arp, and Konrad Rieck. Spying through virtual backgrounds of video calls. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pages 135–144, 2021.

[40] Brien A Holden, Timothy R Fricke, David A Wilson, Monica Jong, Kovin S Naidoo, Padmaja Sankaridurg, Tien Y Wong, Thomas J Naduvilath, and Serge Resnikoff. Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. *Ophthalmology*, 123(5):1036–1042, 2016.

[41] Mohammad Moinul Islam, Vijayan K Asari, Mohammed Nazrul Islam, and Mohammad A Karim. Video super-resolution by adaptive kernel regression. In *International Symposium on Visual Computing*, pages 799–806. Springer, 2009.

[42] Marc Juarez, Sadia Afroz, Gunes Acar, Claudia Diaz, and Rachel Greenstadt. A critical evaluation of website fingerprinting attacks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 263–274, 2014.

[43] Katherine A Karl, Joy V Peluchette, and Navid Aghakhani. Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly. *Small Group Research*, 53(3):343–365, 2022.

[44] Markus G Kuhn. Electromagnetic eavesdropping risks of flat-panel displays. In *International Workshop on Privacy Enhancing Technologies*, pages 88–107. Springer, 2004.

[45] Markus G Kuhn. Security limits for compromising emanations. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pages 265–279. Springer, 2005.

[46] Rohit Kulkarni. A Million News Headlines, 2018.

[47] Chao-Hsien Kuo and Zhen Ye. Sonic crystal lenses that obey the lensmaker's formula. *Journal of Physics D: Applied Physics*, 37(15):2155, 2004.

[48] Michael Li. I studied the fonts of the top 1000 websites. Here's what I learned. https://dribbble.com/stories/2021/04/26/web-design-data-fonts, 2021.

[49] Tony Lindeberg. Scale invariant feature transform. 2012.

[50] Tatsiana Palavets and Mark Rosenfield. Blue-blocking filters and digital eyestrain. *Optometry and Vision Science*, 96(1):48–54, 2019.

[51] Mohd Sabra, Anindya Maiti, and Murtuza Jadliwala. Zoom on the keystrokes: Exploiting video calls for keystroke inference attacks. *arXiv preprint arXiv:2010.12078*, 2020.

[52] Jerome H Saltzer and Michael D Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

[53] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009.

[54] Ilia Shumailov, Laurent Simon, Jeff Yan, and Ross Anderson. Hearing your touch: A new acoustic side channel on smartphones. *arXiv preprint arXiv:1903.11137*, 2019.

[55] Wim Van Eck. Electromagnetic radiation from video display units: An eavesdropping risk? *Computers & Security*, 4(4):269–286, 1985.

[56] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[57] Zachary Weinberg, Eric Y Chen, Pavithra Ramesh Jayaraman, and Collin Jackson. I still know what you visited last summer: Leaking browsing history via user interaction and side channel attacks. In *2011 IEEE Symposium on Security and Privacy*, pages 147–161. IEEE, 2011.

[58] Jianchao Yang and Thomas Huang. Image super-resolution: Historical overview and future challenges. In *Super-resolution imaging*, pages 1–34. CRC Press, 2017.

[59] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

Appendix A
Equipment Information

Lens Power & Focal Length. The power/Diopter of a lens is defined as the reciprocal of the lens' nominal focal length. Different from the f_g used before, this nominal focal length corresponds to the optical effect produced by the combination of the outer and inner surfaces of the lens, and is related to the radius of the outer and inner surfaces by the Lens Maker's Formula [47]:

$$D = \frac{1}{f} = (n-1)(\frac{1}{R_o} - \frac{1}{R_i})$$

where R_o and R_i are the radius of the outer inner surfaces respectively, and n is the refractive index of lens material. When the lens power and materials are set, R_o and R_i can both be adjusted to produce the desired power. However, flatter outer surfaces, as known as base curves, are often used for higher power lenses [16]. This is why we observe a positive correlation between f_g and the lens power in Section IV-E.

Webcam Parameter Estimation. Manufacturers of the laptop built-in webcams often do not share information about the webcam focal length f and imaging sensor physical size W. In this case, further estimation needs to be made. The term $\frac{f}{W}$ is a function of the vertical field-of-view (FoV) of the webcams. Specifically, the FoV angle α can be written as

$$\alpha = 2\tan^{-1}\frac{W}{2f}$$

Considering that typical webcams have a diagonal FoV of in the range $70-90°$, we can convert it to a typical vertical

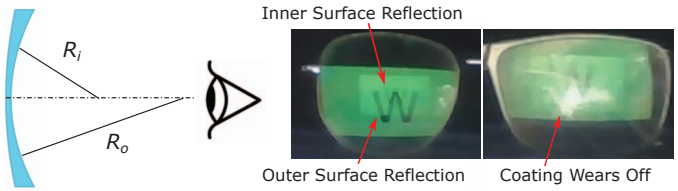

Fig. 11. Design conventions suggest that eyeglasses with higher prescription strength have smaller curvature (larger radius/focal length) on the lens outer surface, leading to larger-size reflections. Besides curvature and reflectance, lens coating conditions can also affect reflection quality.

FoV of about $40-50°$ for a 720p webcam and thus get $\frac{f}{W}$ approximately in the range of $1.1-1.4$ [1]–[3].

Lab Setting Experiment Equipment. The Acer laptop [7] has a screen width of 38 cm and height of 190 mm and a 720p built-in webcam. The OS is Ubuntu 20.04. The OS and browser zoom ratios are default (100%). All the photos and videos are collected with the Cheese [13] webcam application. The photos are in PNG format and the videos are in WEBM format. The Samsung laptop used as the attacker device has OS Windows 10 Pro. The recordings are collected with OBS Studio in MP4 format.

The pair of BLB glasses [12] has lenses with a horizontal and vertical chord length of 5 cm and 4 cm respectively, and a focal length (f_g) of 8 cm. The pair of prescription glasses [12] has lenses with a horizontal and vertical chord length of 6 cm and 5 cm respectively, and a focal length of 50 cm.

Nikon Z7: The photos are in JPEG format (highest quality) and the videos are in MP4 format. We compared these formats with the compression-less (raw) photo and video formats provided by Nikon Z7 but didn't find an obvious difference in the image quality.

Appendix B
Viewing Angle Model

Similar to the pixel size model, we only use 2D modeling (Figure 12) for simplicity which can represent either horizontal or vertical rotations, and we only consider one glass lens since the two lenses are symmetric. The lenses are further modeled as spherical with a radius $2f_g$. We set the origin O to the center of the head which is also treated as the rotation center, and assume the initial orientation without rotation is such that the center of the glass lens arc P_1 aligns with the rotation center and the laptop webcam P_4 on the X-axis. The distance between the glass lens center and the rotation center is s. To calculate the maximum feasible angles, we only need to consider the reflections from either one of the two boundary points of the glass lens since they are symmetric. We label the bottom boundary point as P_2. After a rotation of angle θ, P_1, P_2 are rotated to P_1', P_2' respectively, and the vector $\overrightarrow{P_1'P_2'}$ yields the normal $\vec{n}$ at the reflection point P_2'. P_3 denotes the point source on the screen whose light gets reflected to the camera with an incident angle β. With L_s being the length of the screen on the dimension, the camera should be able to

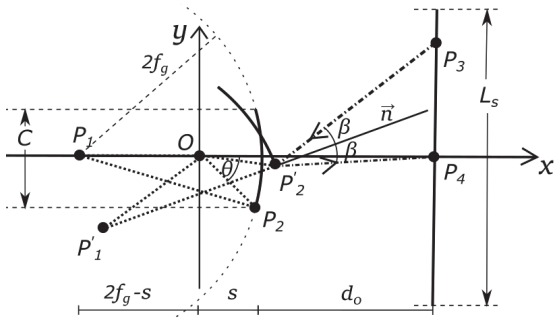

Fig. 12. The model of viewing angle.

peek reflections from the glass lens if P_3 falls in the range of the screen. C denotes the length of the glass lens chord.

In order to find a mapping from the rotation angle θ to the light-emission point P_3 on the screen, the key is to find the slope of the line $P_2'P_3$ which intersects with the screen. Since $P_1'P_2'$ bisects $P_2'P_4$ and $P_2'P_3$, we denote the slope of these three lines as b_1, b_2, b_3 respectively, and have

$$b_3 = \frac{b_2 - 2b_1 - b_1^2 b_2}{b_1^2 - 2b_1 b_2 - 1}$$

To calculate b_1 and b_2, the coordinate of P_1' and P_2', P_4 can be denoted as,

$$\begin{cases} P_1' : \big((s-2f_g)cos\theta, (s-2f_g)sin\theta\big) \triangleq (C, D) \\ P_2' : \big(x_0 cos\theta - y_0 sin\theta, x_0 sin\theta + y_0 cos\theta\big) \triangleq (A, B) \\ P_2' : (s+d, 0) \triangleq (E, 0) \end{cases}$$

and thus

$$b_1 = \frac{B-D}{A-C}, \quad b_2 = \frac{B}{A-E}$$

The last missing piece is the coordinate of P_2, which is denoted as $P_2 : (x_0, y_0) = (r \times cos\alpha, r \times sin\alpha)$, where

$$\begin{cases} r = \sqrt{(\frac{C}{2})^2 + (\sqrt{R^2 - (\frac{C}{2})^2} - (R-s))^2} \\ \alpha = -arcsin(\frac{C}{2r}) \end{cases}$$

We note that the measured ranges in Table II are uniformly larger than the theoretical values, which could be caused by a coarse estimation of the distance s since the actual distance between the lens and the rotation center is hard to determine, and the fact that the model approximates the camera as a point instead of a surface.

Appendix C
Video Conferencing Platform Behaviors

Zoom Under Low Bandwidths. When network bandwidth got smaller than 4 Mbps, we found Zoom will first experience a short period of aggravated packet loss, and then rapidly decrease the video resolution to compensate for it. Video resolution will soon be increased again by sacrificing frame rate as well as compression loss. Zoom will still try to recover

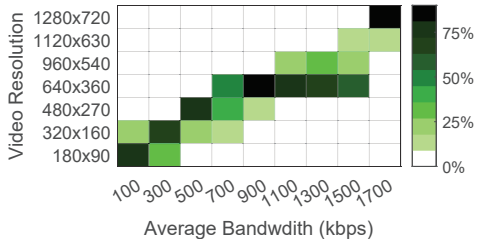

Fig. 13. Heat map of observed Zoom video resolutions under different low bandwidths that resulted in resolutions lower than 720p

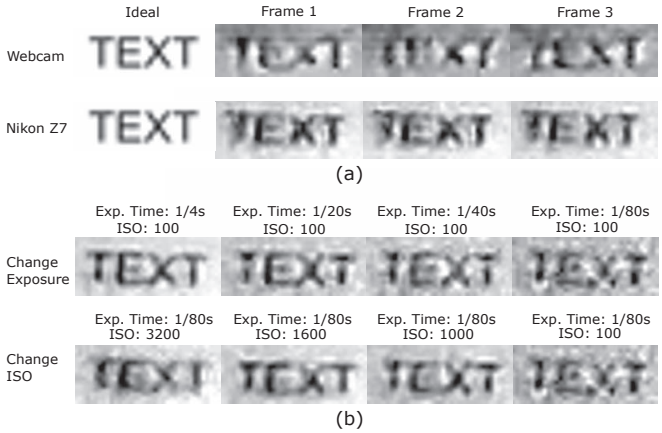

Fig. 14. (a) The ideal capture versus the actual captures in three consecutive frames by webcam (1st row) and Nikon Z7 (2nd row). The distortions feature occlusions with inter-frame and intra-frame variance. The webcam yields larger variances. (b) Photos captured by Nikon Z7 under different exposure times and ISO settings. Longer exposure time and medium ISO yield smaller distortions and increase SNR.

high frame rates later by further increasing the video compression loss. Through our experiments, we noticed that when the bandwidth was larger than 1500 kbps, Zoom was able to maintain a 1280*720 resolution with a frame rate very close to 30 fps. We observed lower resolutions when the bandwidth is lower than 1500 kbps, as shown in Figure 13. Skype and Google Meet do not provide statistics like resolution, frame rate, and bandwidth. But our visual inspection suggests they take a similar approach as Zoom to handle bandwidth issues.

Video Quality Control. Currently, Zoom and Skype do not provide an option for users to control video resolution or quality directly. Google Meet only allows users to switch between 720p and 360p send and receive video resolutions. However, users can limit their system or process bandwidths using software like NetLimiter to decrease video quality even without the conferencing platform offering such an option.

Appendix D
Distortion Analysis

We taped the inner surface of the glasses lens with black papers in order to eliminate the impact of the face background and better characterize the inherent distortions. Effects of the face background are discussed in Section IV-D. The webcam and Nikon Z7 were set to the same color temperature (3500 K) and frame rate (30 fps). For the highly configurable Nikon Z7, we set the ISO, aperture, and exposure time to 100, $F4$, and $\frac{1}{30}s$ respectively, disabled all active noise-reduction schemes including vibration reduction, and used manual focus mode. For both cases, we displayed the string "TEXT" and adjusted the size to make sure the captured text in both cameras' frames has a size of 10 pixels vertically.

Different from previous works [25], [26], motion blur and out-of-focus blurs that are theoretically uniform within a single frame is not the number one limiting factors in the webcam peeking threat model because of the relatively shorter exposure time and closer, more constant camera-object distance. Instead, distortions with intra-frame and inter-frame variance dominate which suggests the image quality cannot be easily improved with PSF deconvolution as in [25] and new image enhancing techniques are needed.

Figure 14 (b) taken with the configurable Nikon Z7 shows how these two forms of distortions (shot and ISO noise) affect the images. For the first set of images (1st row), we keep ISO at 100 and decrease the exposure time from $\frac{1}{4}s$ to $\frac{1}{80}s$ to show the effect of fewer photons hitting the image sensors which results in increased shot noise occlusions. For the second set of images (2nd row), we keep the exposure time at $\frac{1}{80}s$ while increasing ISO from 100 to 3200 to show the effect of increased ISO noise.

TABLE III
Text sizes of web contents

Target	Point Size	Cap Height (mm)
$\mathcal{G}_1$ P	12	2.1
$\mathcal{G}_1$ H3	14	2.5
$\mathcal{G}_1$ H2	18	3.2
$\mathcal{G}_1$ H1	24	4.3
$\mathcal{G}_2$ P	21	3.7
$\mathcal{G}_2$ H3	25	4.3
$\mathcal{G}_2$ H2	32	5.6
$\mathcal{G}_2$ H1 (S1)	42	7.4
$\mathcal{G}_3$ 0% (S2)	56	10
$\mathcal{G}_3$ 20% (S3)	80	14
$\mathcal{G}_3$ 40% (S4)	102	18
$\mathcal{G}_3$ 60%	136	24
$\mathcal{G}_3$ 80% (S5)	253	35
$\mathcal{G}_3$ 95% (S6)	340	60

Appendix E
Web Textual Targets

Web Text Design Conventions. Despite the fact that the default CSS font sizes are decided by web browser vendors separately, we find many of them follow the W3C recommendation [14], where H1, H2, H3 headers' font sizes are 2, 1.5, 1.17 em respectively. To briefly explain, a text size of x em means the size is x times the current body font size of the web page [20] which is usually the same as the font size of paragraph (P) elements. Nevertheless, we note that web design standards are lacking and designers have a large degree of freedom of choosing their own text designs. Sometimes bigger fonts are preferred in order to make the websites more stylish and eye-catching. In this section, we thus investigate both conventional and more stylish web text sizes.

Fig. 15. Different strengths of Gaussian filtering applied on three pairs of glasses. The reflected texts and their CWSSIM scores in each case are shown. Different glasses require different strengths of filters to reduce the reflection. We thus advocate an individual reflection testing procedure to determine protection scheme and settings.

Text Sizes. We summarize the text sizes investigated in Table III where The cap height values are measured with the Acer laptop and default OS and browser settings.

$\mathcal{G}_1$ and $\mathcal{G}_2$: The first group represents the median HTML P, H1, H2, H3 texts of the 1000 websites. [48] reports that the median size of the P elements is about 12 pt and H1, H2, H3 sizes are close to the 2, 1.5, 1.17 em ratios recommended [14]. We thus use these point sizes for $\mathcal{G}_1$ and specify the corresponding cap heights in Table III. The second group represents the largest HTML P, H1, H2, H3 texts of the 1000 websites in [48] with the same recommended em ratios for the headers. [48] finds that about 4% of the 1000 websites use a P size as large as 21 pt. This results in H1, H2, H3 sizes of 25, 32, and 45 pt respectively.

$\mathcal{G}_3$: The third group represents the 117 big-font websites' texts. We manually inspected all the 427 websites archived on SiteInspire [10]. The reason for manual analysis rather than scraping is that many large-font texts on the websites are embedded in the form of images instead of HTML text elements in order to create more flexible font styles. We then selected 117 of them based on the following criteria: (1) The webpage is still active. (2) The largest static texts that enable an adversary to identify the website through google search have a cap height of at least 10 mm when displayed on the Acer laptop. We show the different quantiles of the largest physical cap heights on the 117 websites and the converted point sizes in Table III. We find that most websites in $\mathcal{G}_3$ are related to art, design, and cinema industry which like to present their stylish design skills but unfortunately make the web peeking attack easier. About 1/3 of the websites are designers' or studios' websites that computer science/security researchers may overlook. Furthermore, 72 out of the 117 websites are ranked on Alexa from 38 to 8,851,402 with 5 websites among the top 10,000.

Fig. 16. A spectrum of Alexa top 100 websites that are found to be the easiest (upper) and hardest (lower) to recognize in our evaluation of website recognition under webcam peeking attacks. Screenshots of each website are rotated by 90 degrees and concatenated horizontally. Correlations scores between the rank of website recognition easiness and website pixel values' average and standard deviation are -0.33 and 0.45 respectively, suggesting darker websites with high-contrast graphical contents are easier to recognize.

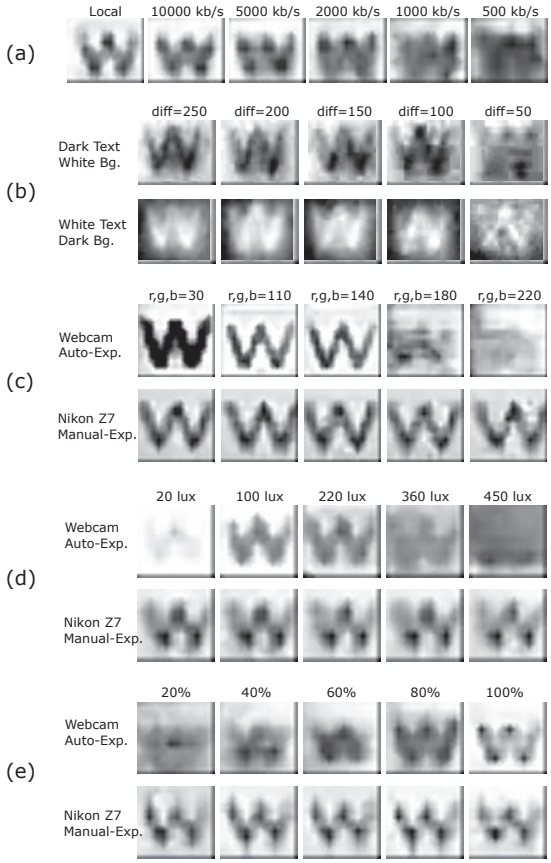

Fig. 17. (a) The comparison between reconstructed images when the video is recorded locally on the victim device and over Zoom with different network upload bandwidths. (b) Changes of reflection recognizability with different text-background color contrast. (c) Changes of reflection recognizability with different background colors (reflectance). We tested gray-scale colors with the same RGB values, which have relatively uniform reflectance on the visible light spectrum. (d) Changes of reflection recognizability under different environmental light intensities. (e) Changes in reflection recognizability with different screen brightness.