VVSec: Securing Volumetric Video Streaming via Benign Use of Adversarial Perturbation

Zhongze Tang
Rutgers University
Piscataway, NJ, USA
zhongze.tang@rutgers.edu

Xianglong Feng
Rutgers University
Piscataway, NJ, USA
xf56@scarletmail.rutgers.edu

Yi Xie
Rutgers University
Piscataway, NJ, USA
yi.xie@rutgers.edu

Huy Phan
Rutgers University
Piscataway, NJ, USA
huy.phan@rutgers.edu

Tian Guo
Worcester Polytechnic Institute
Worcester, MA, USA
tian@wpi.edu

Bo Yuan
Rutgers University
Piscataway, NJ, USA
bo.yuan@soe.rutgers.edu

Sheng Wei
Rutgers University
Piscataway, NJ, USA
sheng.wei@rutgers.edu

ABSTRACT

Volumetric video (VV) streaming has drawn an increasing amount of interests recently with the rapid advancements in consumer VR/AR devices and the relevant multimedia and graphics research. While the resource and performance challenges in volumetric video streaming have been actively investigated by the multimedia community, the potential security and privacy concerns with this new type of multimedia have not been studied. We for the first time identify an effective threat model that extracts 3D face models from volumetric videos and compromises face ID-based authentications. To defend against such attack, we develop a novel volumetric video security mechanism, namely *VVSec*, which makes benign use of adversarial perturbations to obfuscate the security and privacy-sensitive 3D face models. Such obfuscation ensures that the 3D models cannot be exploited to bypass deep learning-based face authentications. Meanwhile, the injected perturbations are not perceivable by the end-users, maintaining the original quality of experience in volumetric video streaming. We evaluate *VVSec* using two datasets, including a set of frames extracted from an empirical volumetric video and a public RGB-D face image dataset. Our evaluation results demonstrate the effectiveness of both the proposed attack and defense mechanisms in volumetric video streaming.

CCS CONCEPTS

• **Security and privacy → Systems security**; • **Information systems → Multimedia streaming**;

KEYWORDS

Volumetric video; video streaming; face authentication; adversarial perturbation

ACM Reference Format:
Zhongze Tang, Xianglong Feng, Yi Xie, Huy Phan, Tian Guo, Bo Yuan, and Sheng Wei. 2020. VVSec: Securing Volumetric Video Streaming via Benign Use of Adversarial Perturbation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413639

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413639

1 INTRODUCTION

Volumetric video (VV) is an emerging type of rich multimedia that records objects and space in three dimensions (3D) with six degrees of freedom (6-DOF), providing the users with fully immersive virtual reality (VR) or augmented reality (AR) experiences [12, 32]. It used to be depicted only in science fiction in the past decades [12]. However, with the recent developments in computer graphics and high-performance VR/AR devices, volumetric video has witnessed a gradual commercial development and deployment in the consumer market [14, 27]. It has been regarded as the next generation of video type after the traditional 2D video and the recently deployed 360-degree video [66], and the global volumetric video market is estimated to grow from $1.4 billion in 2020 to $5.8 billion by 2025 [22]. Different from the pixel-based 2D and 360-degree videos, volumetric video captures 3D objects, represented by 3D meshes [57] or point clouds [58], with significantly higher amount of data and computation involved for capturing, storage, transmission, and rendering. Consequently, it poses significant challenges to the traditional video processing and streaming technologies.

To date, the state-of-the-art research efforts have all been focusing on addressing various resource and performance challenges in volumetric video capturing [14, 27, 29], encoding [55, 63, 70], and streaming [36, 64, 76], to make it deployable under the existing network and video processing/streaming infrastructures. Although the existing efforts are still in the early stage, the challenges and solutions under exploration resemble the community's past experiences

with 2D and 360-degree videos where there was a similarly large gap between the capacity of the network/computation and the demand. Given the past success in addressing these similar challenges, it is foreseeable that the resource and performance challenges of volumetric video streaming can be well addressed eventually, especially with the rapid advancements in 5G wireless networks [53] and high-performance computing software/hardware stacks for video processing [9, 18].

However, the community has not explored the potential *security and privacy* vulnerabilities of volumetric video caused by its unique characteristics, e.g., with 6-DOF 3D objects, which did not exist in the traditional 2D or 360-degree videos. *First*, from the economic and business point of view, the 3D objects precisely presented in volumetric videos are significantly more valuable assets than 2D or 360-degree video content, often involving copyrights or intellectual properties that must be protected. While digital rights management (DRM) mechanisms have been well studied and widely deployed for 2D videos [4, 5, 23], the solution for protecting the even more valuable volumetric video is desirable before they can be widely deployed for consumer-facing applications. *Second*, in addition to its economic values, the 3D models in volumetric videos, if leaked, may trigger significant security/privacy concerns as they may involve rich information of human faces or other privacy-sensitive objects [60] or lead to biometrics-based security exploits. The potential security and privacy issues may become a critical roadblock for the future deployment of volumetric video streaming, even after all the current resource and performance challenges have been effectively addressed. In this paper, we aim to address this brand new dimension of security and privacy challenges posed by volumetric video streaming.

The most straightforward solution to the aforementioned security and privacy issues is to conduct end-to-end encryption of the volumetric video content coupled with a secure licence management mechanism, similar to those adopted in the DRM for traditional 2D videos [4, 5, 23]. However, under the context of volumetric video streaming, the traditional encryption-based approach is subject to the following two limitations. *First*, the performance and power overhead of end-to-end encryption would increase considerably in volumetric video given the significantly increased data volume. Such overhead could become even worse considering that the primary use case of volumetric video streaming is towards mobile VR/AR devices with limited computation and power resources. In addition, the highly interactive nature of the 6-DOF immersive experience makes volumetric video very sensitive to any increase of end-to-end transmission or processing delay. *Second*, end-to-end encryption can still be subject to potential security vulnerabilities even if a secure key management scheme is adopted. This is because the video content must be eventually decrypted before showing and, therefore, it could leave a traceable moment for the adversary to retrieve the decrypted volumetric video content via either reverse engineering based on memory access patterns [79] or screen recording of the displayed content.

To address the aforementioned limitations of end-to-end encryption, we develop *VVSec*, the first multimedia security framework aiming to protect volumetric videos with a focus on the 3D face models presented in the video content. The key idea of *VVSec* is to *obfuscate* the volumetric video via a benign use of adversarial examples [34, 73], which are small human non-perceivable perturbations added to the original video frames to mislead the deep learning-based face recognition/authentication systems [78, 86]. In particular, we propose to inject adversarial perturbations in the target volumetric video, so that the adversary would fail to impersonate the victim in face authentication by leveraging the extracted 3D face models. Additionally, we control the amount of introduced perturbation for effective defense without impacting the quality of experience (QoE) perceived by human users. While developing *VVSec* and demonstrating its effectiveness, we make the following major research contributions primarily targeting 3D face models presented in volumetric videos.

- We for the first time develop an effective security threat model exploiting volumetric video, in which the adversary extracts 3D face models from the video and uses them to impersonate the victim in deep learning-based face authentications;
- We for the first time develop an effective countermeasure to defend against the potential volumetric video attack, which makes benign use of adversarial perturbations to evade from the face authentication attack while still maintaining the original quality of experience to the end user; and
- We evaluate *VVSec* using a set of frames extracted from an empirical volumetric video, as well as a public RGB-D face dataset. Our experimental results demonstrate the success of both the proposed attack and defense mechanisms.

The remainder of the paper is organized as follows. Section 2 introduces the background information of volumetric video streaming and face authentication system that serve as the basis of the target problem. Section 3 describes the proposed face authentication attack using the facial information extracted from a volumetric video. Section 4 presents the proposed defense mechanism via the benign use of adversarial perturbation. Section 5 includes the experimental results for both the attack and defense mechanisms. Section 6 summarizes the closely related works to *VVSec*. Section 7 discusses the limitations of *VVSec* that we plan to explore and address in future work. Finally, Section 8 concludes the paper.

2 BACKGROUND

2.1 Volumetric Video Streaming

Several formats for storing and presenting volumetric video content have been developed recently, such as point cloud-based volumetric video [6, 51] and depth image (i.e., RGB-D)-based volumetric video [15, 77]. Different vendors like Microsoft [7], 8i [6], Depthkit [15], and several other companies [16, 31] have their unique ways of capturing, rendering, and delivering the volumetric videos. However, almost all of them take advantage of depth cameras, such as Microsoft Kinect [2, 3] and Intel RealSense depth cameras [20], to capture color and depth information at the same time. Without loss of generality, we use the RGB-D based video format from Depthkit [15] in our study of volumetric videos, as it can be processed with the off-the-shelf video coding techniques and is compatible with the widely adopted video streaming standard, e.g., Dynamic Adaptive Streaming over HTTP (DASH) [71].

Also, several research efforts have been focusing on DASH-based volumetric video streaming [39, 76]. Figure 1 shows a representative end-to-end volumetric video streaming system following the

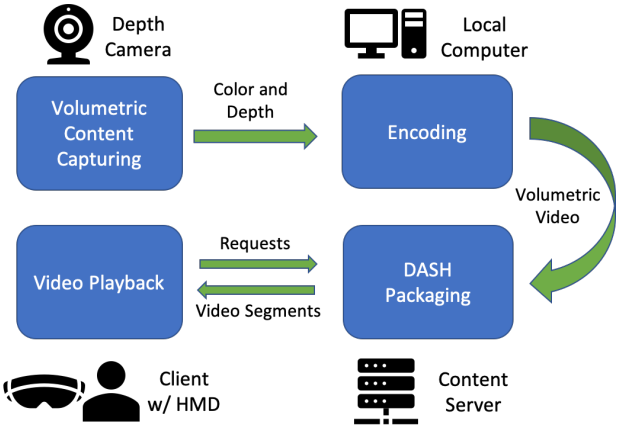

Figure 1: A representative end-to-end volumetric video streaming system.

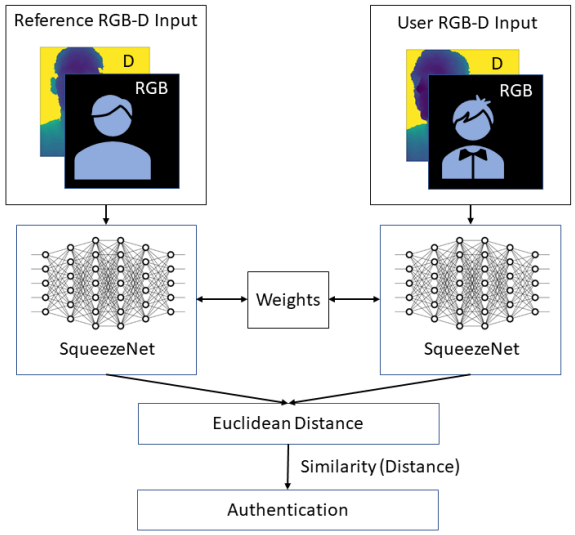

Figure 2: Architecture of the Siamese network-based face authentication system [61].

DASH standard, which delivers the volumetric content from the depth camera to the head-mounted display (HMD). The volumetric content, which consists of both color and depth information, is first captured by a depth camera and then encoded by a local computer. Next, the encoded video is sent to the content server, where the DASH packager generates the DASH-compatible video segments for further distribution. Finally, the HMD can request and receive the video segments from the content server via HTTP and play them back, providing the user with an immersive viewing experience.

2.2 Face Authentication System

A face authentication system leverages human face as the biometric to authenticate users by employing techniques like deep neural networks [78, 86]. It typically accomplishes the authentication task by comparing the user's facial information with the pre-recorded reference face model. Conventional face authentication systems rely on 2D images for the sake of convenience, but these systems are subject to straightforward attacks using simple 2D pictures to impersonate the legitimate users [11, 30, 80]. To counter the 2D image-based attacks, state-of-the-art face authentication mechanisms, such as Apple Face ID [72] and many other commercial or non-commercial systems [61], employ 3D imaging that takes into account the depth information to authenticate the user.

Without loss of generality, we adopt an open-source 3D image-based face authentication system [61] to validate the attack and defense mechanisms presented in this work. As shown in Figure 2, the face authentication system is based on the Siamese network [46], which involves a pair of deep neural networks (i.e., SqueezeNet [41]) to infer the similarity (i.e., euclidean distance) between the user RGB-D image and the reference RGB-D image based on a pre-trained model. In particular, each SqueezeNet is able to map an input image to a 128-dimensional array, and the euclidean distance between the two input images can be calculated to represent the similarity score needed for making the authentication decision. To pass the face authentication, the similarity score between the two input images must be under a certain threshold (e.g., 0.4) defined based on the target security sensitivity of the authentication. Furthermore, the Siamese network is trained with a contrastive loss [65] to minimize the similarity between the two input images for the same user (i.e., reduce false negatives) and maximize that for different users (i.e., reduce false positives).

3 THREAT MODEL: VOLUMETRIC VIDEO-ENABLED ATTACK

3.1 Attack Procedure

Our proposed threat model targets the scenario of a volumetric video streaming session between two users, namely Alice and Malice, for an immersive video streaming session, similar to the demonstration presented by Microsoft [67, 68]. In this scenario, Alice's 3D model is captured by one or more 3D cameras, encoded into an RGB-D video, and delivered to Malice's HMD (e.g., Microsoft Hololens [10] or HTC Vive [21]) to display a volumetric, 6-DOF view of Alice. Despite the premium immersive experience, we argue that the streaming of Alice's 3D model without any protection would raise potential security concerns due to the security-sensitive biometric information contained in the 3D model, such as the 3D facial information as a face ID.

More concretely, if Malice retrieves the 3D model from the HMD during streaming, she can obtain and exploit Alice's facial information to impersonate Alice in a face ID-based authentication. Figure 3 demonstrates the proposed procedure that Malice can adopt to launch such an attack. First, the volumetric video containing Alice's 3D facial information is streamed to Malice in the RGB-D format. Then, Malice retrieves the 3D face model residing in the memory of the HMD before rendering, which is feasible even in the scenario of end-to-end encryption, as the video content must be decrypted prior to rendering. Lastly, with the Alice's RGB-D face model, Malice can impersonate Alice with face authentication

using either a virtual camera or a printed 3D mask [43]. The key challenge in the attack procedure is for Malice to extract the RGB-D face model from the volumetric video, which we discuss next.

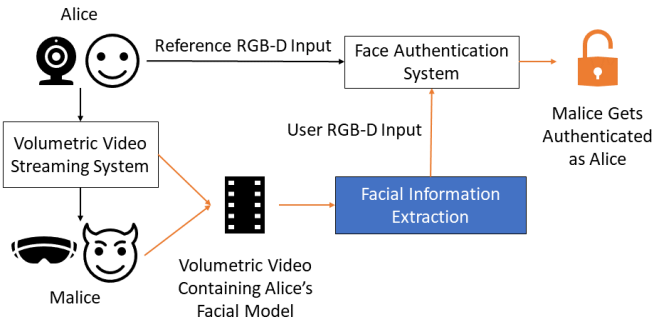

Figure 3: The proposed volumetric video-enabled face authentication attack.

3.2 Facial Information Extraction

We develop a *Facial Information Extractor* to accomplish the key step of extracting the 3D facial information from the target volumetric video, as shown in Figure 3. Also, Figure 4 shows the procedure of facial information extraction. The upper half of the frame is the RGB portion of the 3D model, and the lower part is the depth portion. The depth portion uses RGB color to present the depth information, where the scale of the hue value of the corresponding pixel follows the scale of the depth. Malice first cuts a frame from the volumetric video and, then, the *Facial Information Extractor* generates the RGB image together with the depth image by cropping, rotating, and expanding both the RGB and depth portions of the frame. Moreover, the extractor translates the depth image to quantitative depth data as the input to the face authentication system. In particular, it reads the value of pixels in the HSV (*hue, saturation, and value*) color space and maps the value of *hue* to the corresponding *depth* value. Equation (1) reveals the relationship between the *hue* and the *depth*, where D_{max} and D_{min} are the maximum and minimum distances between the camera and the user, respectively, and e is the rescaling factor that ensures a reasonable *depth* value for the face authentication system.

$$depth = \left((D_{\max} - D_{\min}) \times hue + D_{\min}\right) \times \frac{e}{D_{\min}} \quad (1)$$

4 PROPOSED DEFENSE: VVSEC

4.1 Challenges of Protecting Volumetric Video

Generally speaking, protecting the confidentiality of data is a well studied and addressed problem in the security community, especially with the state-of-the-art hardware isolation-based trusted execution environments (TEEs) [25] and the cryptographic algorithms [59]. However, both categories of defense mechanisms have their limitations in the specific scenario of volumetric video streaming. *First,* hardware-based TEEs (e,g., ARM TrustZone and Intel SGX) [1, 25] suffer from several vulnerabilities related to side channel attacks and hardware physical attacks [74, 75], which would

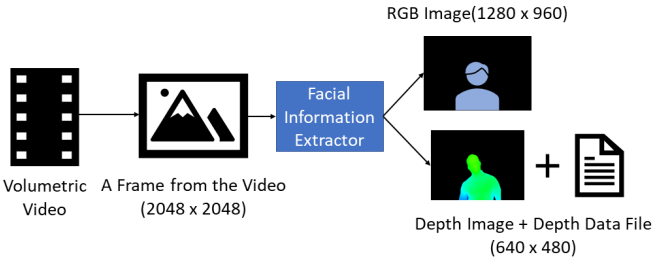

Figure 4: Our proposed procedure for facial information extraction from volumetric video.

lead to the leakage of sensitive data. *Second,* no matter what kind of end-to-end encryption strategies are employed, eventually the video must be decrypted and stored in certain location of the memory in plaintext, which can be exposed to the attackers [47, 54]. Furthermore, encryption may worsen the performance (e.g., end-to-end latency) of volumetric video streaming and raise the already high power consumption of the computation-intensive multimedia application dealing with 3D models.

4.2 Solution: Benign Use of Adversarial Attack

From the defense perspective, our observation is that the volumetric video streaming is a very unique use case where the sensitive data (i.e., Alice's 3D face model retrieved by Malice) must not pass the deep learning-based authentication, while it must be perceivable by human users as per the requirement of the video streaming application. Such a requirement for *defense* is essentially a close match with the state-of-the-art adversarial *attacks* where an adversary adds deliberately-designed perturbations to the original benign inputs of a deep neural network. Such adversarial perturbations are imperceptible to humans but would cause significant degradations in the accuracy of the neural networks, leading to incorrect inference results [34, 73].

Inspired by the nature of the adversarial attacks, we propose a novel defense mechanisms, *VVSec*, to protect the confidentiality of volumetric video. In a nutshell, *VVSec* adds adversarial perturbations at the sender (i.e., Alice) side of the volumetric video streaming, so that even if Malice could extract the RGB-D facial information in plaintext, the face authentication would fail due to the effect of the "adversarial" perturbation on the deep neural network. On the other hand, the original functionality of volumetric streaming especially the perceivable quality of experience to human users is unchanged, as ensured by the design principle of adversarial perturbations [34, 73].

In order to fully understand the mechanism of adversarial attacks, to date, numerous attack methods have been extensively investigated. Goodfellow et al. [34] introduced the fast gradient sign method (FGSM), which is a simple but effective technique to quickly produce adversarial examples. The key idea of FGSM is to utilize the gradients of the loss function with respect to inputs to craft the adversarial perturbations in a single step. Inspired by FGSM, researchers have proposed to take multiple steps of FGSM (I-FGSM) [50, 56] in an iterative manner to achieve stronger attack performance while keeping smaller perturbations. Moreover,

C&W [26] attack is an optimization-based method, which can generate high-robustness adversarial examples that break many state-of-the-art defense methods [28, 83]. Different from the original intent of the existing adversarial attacks, in *VVSec*, we make benign use of adversarial attack to protect the 3D facial information in the volumetric video, as described next.

4.3 Algorithm Design: Content-Aware Adversarial Perturbation Generation

To clearly present the steps of our perturbation generation algorithm, we define the notations for the rest of the paper. Recall that the goal of the face authentication system is to determine whether the input face image belongs to the enrolled user (i.e., the user represented by the reference RGB-D input). The input image x is denoted as $h \times w \times c$ where h, w, c represent the height, width, and number of channels, respectively. In particular, $c = 4$ where the first three channels represent RGB, and the last one is the depth (D) channel. Next, we model the face authentication system as a function $F(x, y)$, which takes a face image x and the stored face image of the enrolled legitimate user y as inputs and outputs the similarity score S. Typically, if S is under a certain predefined threshold τ, the input image x is considered passing the authentication. In this work, to prevent the extracted frames of the volumetric video from passing the face authentication, we aim to find an imperceptible perturbation δ that could achieve $x' = x + \delta$ such that $S' = F(x', y) > \tau$.

A volumetric video can contain one or more 3D objects, and the extra space besides the objects is considered as background. If the perturbation is added to the background portion of the volumetric video, it can be obviously perceived by the user and significantly impact the quality of experience. Therefore, in *VVSec* we develop a content-aware perturbation generation algorithm to add the perturbation only to the 3D objects instead of the background. In a nutshell, our algorithm takes advantage of the RGB-D data, where the information of the object location could be approximately inferred by the depth channel. Specifically, we first generate a boolean mask α ($\alpha \in \{0, 1\}$) with the same size of the input image x,

$$\alpha(n, p, q) = \begin{cases} 1; & x(n, p, 4) \geq t \text{ and } q \leq 3, \\ 0; & \text{otherwise}, \end{cases} \tag{2}$$

where t is the predefined threshold. With the guidance of such derived mask α, our desired content-aware adversarial perturbation is calculated as:

$$x' = Clip(x + \delta \cdot \alpha, -\epsilon, +\epsilon), \tag{3}$$

where $Clip(\cdot)$ denotes removing the values under certain noise level ϵ to constrain the perturbations using the L_1 distance metric. Since we keep all values for the last channel of α as 0, no pixels would be changed on the depth channel of the original input x.

Next, we utilize an optimization-based attack method inspired by *C&W* [26] to craft our content-aware adversarial examples x'. In particular, we formulate the adversarial perturbation generation as the following optimization problem:

$$\textbf{minimize } Loss = -F(x', y) + \beta \parallel \delta \parallel_2, \tag{4}$$

where β is a pre-chosen constant to control the magnitude of the perturbation. Specifically, the first term is the model prediction score S', and the second term penalizes the perturbation magnitude. Gradient descent is applied to find the optimal perturbation until S' is larger than the threshold τ. Given that the face authentication system mainly leverages the features extracted from the face model to calculate the similarity score, our content-aware perturbation generation targeting only the face portion of the frame would effectively alter the authentication results with minimum perturbation. The details of our adversarial perturbation generation algorithm are presented in Algorithm 1.

Algorithm 1: Content-Aware Adversarial Perturbation Generation

Input: Extracted facial images x, face authentication system $F(\cdot)$, enrolled face images y, identification threshold τ, attacking strength ϵ, penalty constant β.
Result: Adversarial perturbation δ.
Initialize $\delta \leftarrow 0$;
Compute boolean mask α following Equation (2);
$x' \leftarrow Clip(x + \delta \cdot \alpha, -\epsilon, +\epsilon)$;
$S' \leftarrow F(x', y)$;
while $S' < \tau$ **do**
    $Loss \leftarrow -F(x', y) + \beta \parallel \delta \parallel_2$;
    Minimize $Loss$ to update δ;
end

5 EXPERIMENTAL RESULTS

In this section, we first evaluate the volumetric video-enabled attack on the face authentication system. Then, we validate the effectiveness of *VVSec* in preventing the attack.

5.1 Experimental Setup

5.1.1 Volumetric video streaming system. In this work, we use a pre-recorded volumetric video [15] from the Depthkit. On the server side, we adopt the GPAC filter [35] as both the encoder and the DASH packager to generate the DASH segments from the source video. Moreover, we deploy a web server using Node.js [17] to serve the video segments. On the client side, we employ the Vimeo Depth Player [77], which is a browser-based volumetric video player, to process and play the video.

5.1.2 Datasets. We employ two datasets to evaluate the effectiveness and performance of *VVSec*, including a dataset containing frames extracted from volumetric video [15] (i.e., Dataset #1) and an RGB-D face dataset [38] (i.e., Dataset #2).

- Dataset #1 contains 11 RGB-D images of one user extracted from the volumetric video demo in the Depthkit [15], in which we use 1 image as the reference input and 10 images as the user inputs in our evaluation of face authentication.
- Dataset #2 [38] consists of 31 users with 17 different poses each, including 13 face orientations and 4 facial expressions (i.e., smiling, sad, yawn, and angry). For each pose of each user, 3 images are captured, making the total of 1581 RGB-D images in the dataset. All the images are collected by a Microsoft Kinect v1 device [2].

For both datasets, the RGB images are stored as a 32-bit bitmap with resolution $1280 \times 960px$. The depth images are stored as plain-text files, where a depth value represents the corresponding depth pixel, with resolution $640 \times 480px$. In our experiments, we crop and re-scale all the face images to $200 \times 200px$ to meet the requirement of the face authentication system [61]. In our evaluation of the attack, we use Dataset #2 to train and validate the face authentication model. While evaluating the proposed defense, we add the content aware adversarial perturbations to both datasets.

5.1.3 Parameter Settings. In the face authentication system, we use the same default parameter settings as in [61] for the face authentication system. Note that it is our intention to adopt the default settings, as the goal of this work is to demonstrate the identified new attack surface in typical and commonly used face authentication systems, which is well represented by the default settings in [61]. In particular, the Siamese network is trained with a batch size of 16 on Dataset #2 for 50 epochs. Also, we use the Adam optimizer [45] with the learning rate $\eta = 0.001$ and the momentum terms $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Moreover, in the content-aware adversarial perturbation described in Algorithm 1, we set the the predefined threshold t in Equation (2) as 0, the threshold of the similarity score τ as 0.4, attacking strength ϵ as 32/255, and penalty constant β as 0. For the facial information extractor in Section 3.2, we use Equation (1) to map the hue value to the depth value. In our experiments, $D_{max} = 2370$ and $D_{min} = 1270$ in millimeters and $e = 850$.

5.1.4 Evaluation Metrics. To evaluate the effectiveness of the defense, i.e., whether *VVSec* can successfully prevent the face authentication attack, we use the success rate of face authentication defined as follows:

$$success\ rate = \frac{\#\ of\ successful\ cases\ of\ defense}{\#\ of\ valid\ test\ cases} \tag{5}$$

In particular, if the original similarity score before adding the perturbation is smaller than 0.4, we consider it as a valid test case, i.e., the face authentication attack is successful. Then, for a valid test case, if the similarity score after adding the perturbation is greater than 0.4, we consider it as a successful case of defense.

Furthermore, we adopt the normalized L2 norm to quantify the perturbations added by *VVSec*, which is a commonly used metric in the adversarial attack research domain to evaluate the quality impact of adversarial perturbations [26, 37, 73]:

$$perturbation = \frac{\| x - x' \|_2}{\| x \|_2} \tag{6}$$

where x is the user input face image, and x' is the image with the generated adversarial perturbation.

5.2 Effectiveness of the Attack

We evaluate the effectiveness of the proposed attack by feeding a pair of input images to the face authentication system. One is the pre-enrolled user face image, namely the reference input; and the other is the new face image for authentication, namely the user input. Each reference or user input consists of an RGB image and the corresponding depth image to represent the full RGB-D face data. The face authentication system would output the similarity score between the two input images, which we use as an indicator for the success of the attack. In particular, the attack is successful if the similarity score is less than 0.4, as discussed in Section 5.1.4.

Table 1: Quantitative evaluation results of both the attack and the defense using 10 test cases from Dataset #1.

Case	Similarity (Original)	Similarity (*VVSec*)	Perturbation (L2 norm)	Time (second)
1	0.157	0.402	0.042	4.674
2	0.149	0.404	0.042	4.656
3	0.225	0.401	0.092	7.182
4	0.209	0.404	0.038	4.274
5	0.095	0.402	0.113	8.475
6	0.053	0.401	0.136	9.443
7	0.046	0.400	0.141	10.591
8	0.064	0.402	0.076	6.096
9	0.154	0.404	0.056	5.203
10	0.136	0.404	0.057	5.134

Table 1 demonstrates 10 test cases of face authentication using Dataset #1. The *Similarity (Original)* column indicates the resulting similarity score under the attack scenario. We observe that all the original similarity scores are below the threshold 0.4, indicating that, without *VVSec*, the attacker is able to impersonate the legitimate user and successfully pass the face authentication in all the test cases.

5.3 Effectiveness of the Defense

We execute the adversarial perturbation generation algorithm (i.e., Algorithm 1) on both datasets to evaluate the effectiveness of the defense. Table 1 shows the results of 10 test cases from Dataset #1. *First*, we observe that with *VVSec* all the similarity scores, as shown in the *Similarity (VVSec)* column, are larger than the threshold 0.4, indicating that the RGB-D images containing the generated perturbations fail to pass the face authentication system and thus the effectiveness of the defense. *Second*, we also present the quantitative values of the perturbations in the *Perturbation (L2 norm)* column, which vary among different cases, as the perturbation generation algorithm is content-dependent. In addition, the time costs of the perturbation generation are shown in the *Time (second)* column, ranging from around 4 to 11 seconds, which are acceptable if *VVSec* is used offline to generate the protected volumetric video for the video-on-demand (VOD) streaming scenario.

We further evaluate the effectiveness and performance of *VVSec* on all the RGB-D images from Dataset #2. Among the 51 images (i.e., 17 poses and 3 images per pose) of each user, we use the one with front-facing pose as the reference input and the other 50 images as the user inputs for testing, which creates 1550 test cases in total. Among these test cases, there are 1529 cases that

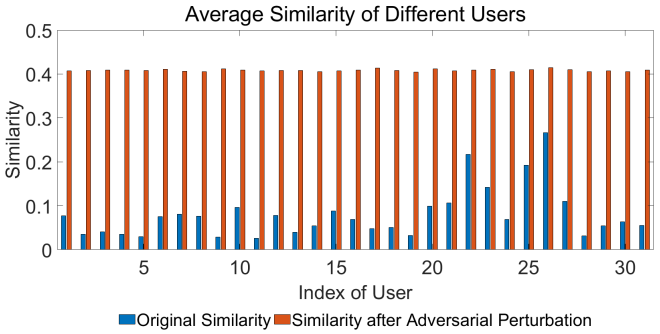

Figure 5: Average similarity scores output by the face authentication system for 31 users from Dataset #2 before and after *VVSec* adding the adversarial perturbation.

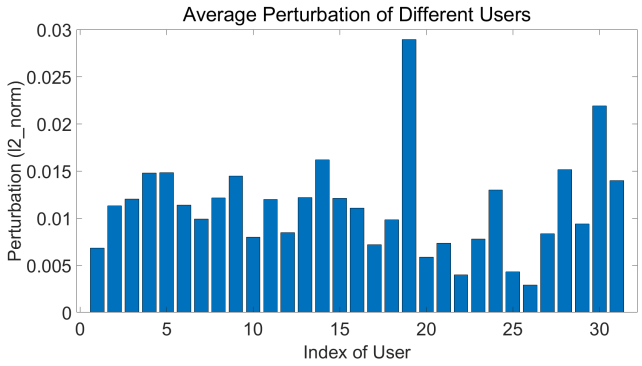

Figure 6: Average adversarial perturbations generated for 31 users from Dataset #2.

are valid ones based on the definition presented in Section 5.1.4 (i.e., the original similarity score before adding the perturbation is smaller than 0.4). Therefore, we conduct our experiments on Dataset #2 with these 1529 test cases and present the results in Figures 5 to 7. Figure 5 shows the average similarity scores with and without perturbation added by *VVSec* for different users. All the average similarities with perturbation are between 0.4 and 0.42, indicating the failure of face authentication and thus the success of defense accomplished by *VVSec*. Figure 6 reveals the average perturbations added to the images are in the range of 0.003 to 0.029, which vary among different users. Lastly, Figure 7 presents the average running time of perturbation generation, which ranges from around 3 seconds to around 9 seconds per image. The results confirm our observations in Table 1 with Dataset #1, indicating that *VVSec* can be applied to VOD volumetric video streaming.

Overall, combining our experiments with both Dataset #1 and Dataset #2, we have evaluated the defense mechanism provided by *VVSec* using 1560 test cases in total, 1539 of which are valid test cases based on the definition in Section 5.1.4. Our experimental results indicate that all the 1539 test cases successfully return a larger than 0.4 similarity score when *VVSec* is applied, achieving a 100% success rate of the defense.

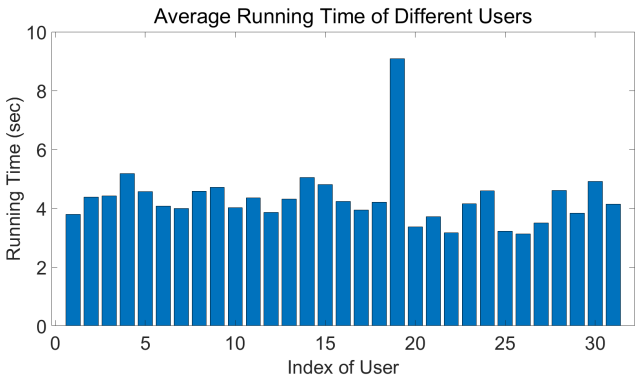

Figure 7: Average running time of our content-aware adversarial perturbation generation for 31 users from Dataset #2.

6 RELATED WORK

Benign Use of Adversarial Attack. As discussed in Section 4.2, many research works have been focusing on adversarial perturbation generation [34, 73]. Also, some researchers have utilized adversarial attacks for benign use cases as a means of obfuscation. For example, Yu et al. [87] developed an adversarial example generation algorithm to protect mobile voice data from being eavesdropped using automatic speech recognition. Xu et al. [85] developed a framework called HAMPER to protect leaked images and voices from malicious speech and image recognition by using adversarial perturbations. Our *VVSec* is inspired by these existing works on making benign use of adversarial attacks for security protections, but we target a brand new and significantly more challenging scenario of volumetric video streaming.

Adversarial Attacks on 3D Models. Adversarial attacks on 3D objects have recently been explored [40, 81, 82, 88]. For example, Xiang et al. [81] proposed several algorithms to generate adversarial examples targeting 3D point clouds. Xiao et al. [82] proposed to generate 3D meshes by manipulating objects with rich shape features but minimal textural variations. These 3D-based adversarial attack methods are closely related to our goal of injecting adversarial perturbations in volumetric videos. However, studies along this line are still at an early stage and cannot be directly applied to domain-specific applications such as volumetric video streaming.

Attacks on Face Authentication Systems. The most straightforward way to attack a face authentication system may be presenting a facial biometric artifact of the victim user to the authentication system. In such presentation attacks, a printed photo [11, 30, 80], a 3D face mask [8, 43], or an electronic display of a photo or video [84] have been exploited to successfully bypass the face authentication. Even modern commercial face recognition systems like Microsoft Windows Hello [13, 80], Apple's Face ID [8, 72], and payment authentication systems of Alipay and Wechat have been bypassed by such presentation attacks [43]. As countermeasures, face authentication systems have also been evolving by adding features like

depth matching and liveness detection [42, 48, 49]. These countermeasures could effectively increase the difficulty level of presentation attacks. However, even with the modern defense features, the volumetric video still has the strong potential to constitute a face authentication attack given the liveness and 3D features of its video content. In other words, there exists minimum difference between the 3D face model in volumetric video and the real human face to be distinguished by a face authentication system.

End-to-End Encryption and Video DRM. End-to-end encryption can be an effective approach to protect the confidentially of data in networked systems in general and video streaming in particular. In fact, state-of-the-art video DRM mechanisms [4, 5, 23] rely on end-to-end encryption to protect the commercial video streaming services (e.g., Netflix [44]) from piracy or illegal broadcasts. However, under the context of volumetric video streaming, the end-to-end encryption or video DRM mechanisms are not sufficient to address the face authentication attack targeted by *VVSec*, for the following reasons. *First,* the content of the video must be decrypted before being displayed to the end user, and it has been shown to be feasible for the adversaries to retrieve the decrypted content in memory by reverse-engineering the memory access pattern [79] or through potential vulnerabilities in modern processors such as Spectre [47], Meltdown [54], and ZombieLoad [69]. *Second*, even if the confidentially of the data in memory is not compromised, the nature of the video viewing experience determines that the video can still be re-captured by the attacker from the screen after it is displayed. The re-captured video can be exploited to bypass the face authentication even with lower resolution than the original video, as the deep learning-based face authentication is in general non-sensitive to the resolutions of the input images. Therefore, a content-based video obfuscation like *VVSec* is desirable to protect the volumetric video before it is ever exposed to the potential attack surfaces.

7 LIMITATIONS AND DISCUSSIONS

Despite the effectiveness of both the proposed attack and defense mechanisms, as supported by our experimental results, *VVSec* at the current stage still has a number of limitations that we would like to discuss and explore in our future work, including the consideration of the depth channel in perturbation generation, the timing overhead, and the QoE evaluation.

Adversarial Perturbation on the Depth Channel. Most of the adversarial attack studies to date focus on adding perturbation to the RGB domain, instead of the depth domain. The current version of *VVSec* also leverages only the RGB domain in the adversarial perturbation generation algorithm. The aforementioned recent research efforts on 3D adversarial attacks [40, 81, 82, 88] could provide us with a viable option to enhance *VVSec*, e.g., by leveraging the depth channel and further reducing the human perceivable perturbations in the RGB domain, which could potentially contribute to reducing the timing overhead as well.

Timing Overhead of Adversarial Perturbation Generation. An obvious limitation in the current *VVSec* system is that the timing overhead of generating the adversarial perturbation is relatively high, i.e., 3 to 9 seconds per image as shown in Figure 7. This restricts the applicability of *VVSec* to offline processing and VOD streaming only without the support of live streaming. The high timing overhead is caused by the iterative, optimization-based adversarial perturbation generation process, which we plan to improve with a brand new learning-based algorithm in the future to achieve real-time performance required by the live streaming use case. Given the recent advancements in real-time adversarial attacks [33, 52, 62], we believe that the objective of real-time *VVSec* is feasible with the unique challenges in the 3D domain that we aim to focus on.

QoE Evaluation. In this work, we adopt the L2 norm as the metric to quantify the impact of the perturbation posed to the quality of the volumetric video. Although the L2 norm is the most commonly used and the standard metric in the deep learning community to evaluate the quality impact of adversarial attacks [26, 73], it is still not a standard QoE metric for video streaming in general and volumetric video streaming in particular. As an alternative, we have explored the possibility of using other objective QoE metrics for volumetric video; however, at the time of writing this paper, there have been no effective QoE metrics developed in the field of volumetric video given that it is still an emerging research area. Such observation is also confirmed by other researchers in the area of volumetric video streaming [64] and in the related field of 360-degree video streaming [24]. In our future work, we plan to further explore the effective QoE metrics and/or conduct subjective user studies to improve the QoE evaluation of *VVSec*.

8 CONCLUSION

We for the first time investigated the security and privacy issues in volumetric video streaming originated from the rich user information involved in the 3D objects. Our exploration began with the development of a threat model, which compromises deep learning-based face authentication mechanisms through effectively extracting the 3D face models from the volumetric video. Then, we developed a countermeasure, namely *VVSec*, to secure the volumetric video via a benign use of adversarial perturbation generation. We showed that volumetric videos with perturbations generated by *VVSec* can effectively defend against the face authentication attack. Also, it poses no impact to the normal use case of volumetric video viewed by human end users thanks to the minimum and non-perceivable perturbations. We evaluated the effectiveness and performance of *VVSec* using an empirical volumetric video, as well as a large number of 3D face images with various poses obtained from a RGB-D face dataset. To motivate further volumetric video security research, we have open-sourced *VVSec* via a GitHub repository [19].

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their constructive feedback. This work was partially supported by the National Science Foundation under awards CNS-1912593, CNS-1815619, and CCF-1937403 and the Air Force Research Laboratory under Grant No. FA87501820058.

REFERENCES

[1] 2005. ARM Security Technology: Building a Secure System using TrustZone Technology.
[2] 2015. Kinect Docs. https://github.com/Kinect/Docs.
[3] 2015. Kinect for Windows. https://developer.microsoft.com/en-us/windows/kinect.
[4] 2016. Adobe Primetime DRM. https://www.adobe.com/content/dam/acom/en/marketing-cloud/primetime/pdf/54658.en.primetime.datasheet.drm-factsheet.pdf.
[5] 2016. Microsoft PlayReady. https://www.microsoft.com/playready/overview/.
[6] 2017. 8i Labs. https://www.8i.com
[7] 2017. Bring life to mixed reality at Mixed Reality Capture Studios. https://www.microsoft.com/en-us/mixed-reality/capture-studios
[8] 2017. Cyber Security Firm Uses a 3D Printed Mask to Fool iPhone X's Facial Recognition Software. https://3dprint.com/194079/3d-printed-mask-iphone-x-face-id/
[9] 2019. Alveo U200 and U250 Data Center Accelerator Cards Data Sheet. https://www.xilinx.com/support/documentation/data_sheets/ds962-u200-u250.pdf.
[10] 2019. Microsoft HoloLens | Mixed Reality Technology for Business. https://www.microsoft.com/en-us/hololens
[11] 2019. A photo will unlock many Android phones using facial recognition. https://nakedsecurity.sophos.com/2019/01/08/facial-recognition-on-42-android-phones-beaten-by-photo-test/.
[12] 2019. What is Volumetric Video? https://volumetric-video.com/what-is-volumetric-video/.
[13] 2020. Biometric Facial Recognition – Windows Hello - Microsoft. https://www.microsoft.com/en-us/windows/windows-hello
[14] 2020. Create holograms from real life. https://www.microsoft.com/en-us/mixed-reality/capture-studios.
[15] 2020. DepthKit. https://depthkit.tv.
[16] 2020. EF EVE Volumetric Video Capture Software. https://ef-eve.com/volumetric-capture-software/.
[17] 2020. Node.js. https://nodejs.org/en/
[18] 2020. Nouveau: Accelerated Open Source driver for nVidia cards. https://nouveau.freedesktop.org.
[19] 2020. Project VVSec. https://github.com/hwsel/VVSec.
[20] 2020. Stereo Depth – Intel RealSense Depth and Tracking Cameras. https://www.intelrealsense.com/stereo-depth/
[21] 2020. VIVE | Discover Virtual Reality Beyond Imagination. https://www.vive.com/us/
[22] 2020. Volumetric Video market 2020-2025 forecast. https://volumetric-video.com/volumetric-video-market-2020-2025-forecast/.
[23] 2020. Widevine DRM. https://www.widevine.com/solutions/widevine-drm.
[24] Shivang Aggarwal, Sibendu Paul, Pranab Dash, Nuka Saranya Illa, Y. Charlie Hu, Dimitrios Koutsonikolas, and Zhisheng Yan. 2020. How to Evaluate Mobile 360° Video Streaming Systems?. In *International Workshop on Mobile Computing Systems and Applications (HotMobile)*. 68–73.
[25] Ittai Anati, Shay Gueron, Simon Johnson, and Vincent Scarlata. 2013. Innovative technology for CPU-based attestation and sealing. In *International Workshop on Hardware and Architectural Support for Security and Privacy (HASP)*.
[26] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*. 39–57.
[27] Eugene d'Eon, Bob Harrison, Taos Myers, and Philip A. Chou. 2019. JPEG Pleno Database: 8i Voxelized Full Bodies (8iVFB v2) - A Dynamic Voxelized Point Cloud Dataset. https://jpeg.org/plenodb/pc/8ilabs/.
[28] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. 2018. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442* (2018).
[29] Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017. Motion2fusion: Real-Time Volumetric Performance Capture. *ACM Transactions on Graphics* 36, 6, Article Article 246 (2017), 16 pages.
[30] Nguyen Minh Duc and Bui Quang Minh. 2009. Your face is not your password face authentication bypassing lenovo–asus–toshiba. *Black Hat Briefings* 4 (2009), 158.
[31] EYEBEAM. 2012. RGBDToolkit Workshop. https://www.eyebeam.org/events/rgbdtoolkit-workshop/
[32] James George. 2017. The Brief History of Volumetric Filmmaking. https://medium.com/volumetric-filmmaking/the-brief-history-of-volumetric-filmmaking-32b3569c6831.
[33] Yuan Gong, Boyang Li, Christian Poellabauer, and Yiyu Shi. 2019. Real-time adversarial attacks. *arXiv preprint arXiv:1905.13399* (2019).
[34] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).
[35] GPAC. 2019. Towards GPAC 0.9.0. https://gpac.wp.imt.fr/2019/06/28/towards-gpac-0-9-0/
[36] Serhan Gül, Dimitri Podborski, Thomas Buchholz, Thomas Schierl, and Cornelius Hellge. 2020. Low-Latency Cloud-Based Volumetric Video Streaming Using Head Motion Prediction. In *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*. 27–33.
[37] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. 2017. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117* (2017).
[38] RI Hg, Petr Jasek, Clement Rofidal, Kamal Nasrollahi, Thomas B Moeslund, and Gabrielle Tranchet. 2012. An RGB-D database using Microsoft's Kinect for Windows for Face Detection. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*. 42–46.
[39] Mohammad Hosseini and Christian Timmerer. 2018. Dynamic Adaptive Point Cloud Streaming. In *Packet Video Workshop (PV)*. 25–30.
[40] Wenlong Huang, Brian Lai, Weijian Xu, and Zhuowen Tu. 2019. 3d volumetric modeling with introspective neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8481–8488.
[41] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv preprint arXiv:1602.07360* (2016).
[42] Hyung-Keun Jee, Sung-Uk Jung, and Jang-Hee Yoo. 2006. Liveness detection for embedded face recognition system. *International Journal of Biological and Medical Sciences* 1, 4 (2006), 235–238.
[43] Roberts Jeff. 2019. Look how easy it is to fool facial recognition – even at the airport. https://fortune.com/2019/12/12/airport-bank-facial-recognition-systems-fooled/.
[44] Daniel Kim. 2019. How Netflix protects its content – Part 1. https://medium.com/pallycon/how-netflix-protects-contents-part-1-a40508ed0001/.
[45] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[46] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, Vol. 2.
[47] Paul Kocher, Jann Horn, Anders Fogh, Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, et al. 2019. Spectre attacks: Exploiting speculative execution. In *IEEE Symposium on Security and Privacy (S&P)*. 1–19.
[48] Klaus Kollreider, Hartwig Fronthaler, and Josef Bigun. 2008. Verifying liveness by multiple experts in face biometrics. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–6.
[49] Klaus Kollreider, Hartwig Fronthaler, Maycel Isaac Faraj, and Josef Bigun. 2007. Real-time face detection and motion analysis with application in "liveness" assessment. *IEEE Transactions on Information Forensics and Security* 2, 3 (2007), 548–558.
[50] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533* (2016).
[51] Anna Qingfeng Li, William Cheung, Robert Kawiak, Dusty Robbins, Michael Chen, Pete Quesada, Tonaci Tran, Jason Juang, and Caoyang Jiang. 2019. An investigation of Volumetric VOD streaming compression techniques. https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/v2-volumetric-vod-streaming-whitepaper.pdf
[52] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy Chowdhury, and Ananthram Swami. 2018. Adversarial perturbations against real-time video classification systems. *arXiv preprint arXiv:1807.00458* (2018).
[53] Xingqin Lin, Jingya Li, Robert Baldemair, Jung-Fu Thomas Cheng, Stefan Parkvall, Daniel Chen Larsson, Havish Koorapaty, Mattias Frenne, Sorour Falahati, Asbjorn Grovlen, et al. 2019. 5G new radio: Unveiling the essentials of the next generation wireless access technology. *IEEE Communications Standards Magazine* 3, 3 (2019), 30–37.
[54] Moritz Lipp, Michael Schwarz, Daniel Gruss, Thomas Prescher, Werner Haas, Anders Fogh, Jann Horn, Stefan Mangard, Paul Kocher, Daniel Genkin, et al. 2018. Meltdown: Reading Kernel Memory from User Space. In *USENIX Security Symposium (Security)*. 973–990.
[55] J. Liu, J. Yao, J. Tu, and J. Cheng. 2019. Data-Adaptive Packing Method for Compression of Dynamic Point Cloud Sequences. In *IEEE International Conference on Multimedia and Expo (ICME)*. 904–909.
[56] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).
[57] Adrien Maglo, Guillaume Lavoué, Florent Dupont, and Céline Hudelot. 2015. 3D Mesh Compression: Survey, Comparisons, and Emerging Trends. *Comput. Surveys* 47, 3 (2015), 44:1–44:41.
[58] Rufael Mekuria, Kees Blom, and Pablo Cesar. 2017. Design, Implementation, and Evaluation of a Point Cloud Codec for Tele-Immersive Video. *IEEE Trans. Cir. and Sys. for Video Technol.* 27, 4 (2017), 828–842.
[59] Alfred J Menezes, Jonathan Katz, Paul C Van Oorschot, and Scott A Vanstone. 1996. *Handbook of applied cryptography*. CRC press.
[60] MIT Technology Review. 2019. Data leak exposes unchangeable biometric data of over 1 million people. https://www.technologyreview.com/f/614163/data-leak-exposes-unchangeable-biometric-data-of-over-1-million-people/.
[61] Norman Di Palo. 2018. GitHub Repository: How I implemented iPhone X's FaceID Using Deep Learning in Python. https://github.com/normandipalo/faceID_beta

[62] Huy Phan, Yi Xie, Siyu Liao, Jie Chen, and Bo Yuan. 2019. CAG: A Real-time Low-cost Enhanced-robustness High-transferability Content-aware Adversarial Attack Generator. *arXiv preprint arXiv:1912.07742* (2019).
[63] Cédric Portaneri, Pierre Alliez, Michael Hemmer, Lukas Birklein, and Elmar Schoemer. 2019. Cost-Driven Framework for Progressive Compression of Textured Meshes. In *ACM Multimedia Systems Conference (MMSys)*. 175–188.
[64] Feng Qian, Bo Han, Jarrell Pair, and Vijay Gopalakrishnan. 2019. Toward Practical Volumetric Video Streaming on Commodity Smartphones. In *International Workshop on Mobile Computing Systems and Applications (HotMobile 2019)*. 135–140.
[65] Rajalingappaa Shanmugamani. 2018. Contrastive loss - Deep Learning for Computer Vision [Book]. https://www.oreilly.com/library/view/deep-learning-for/9781788295628/0fe2ce8e-9141-4734-a311-41ff109b57c4.xhtml
[66] Sarah Redohl. 2019. Volumetric video is so much more than VR. https://www.immersiveshooter.com/2019/01/10/volumetric-video-means-so-much-more-than-vr/.
[67] Microsoft Research. 2016. Holoportation Project. https://www.microsoft.com/en-us/research/project/holoportation-3/
[68] Microsoft Research. 2016. Holoportation: virtual 3D teleportation in real-time. https://www.youtube.com/watch?v=7d59O6cfaM0
[69] Michael Schwarz, Moritz Lipp, Daniel Moghimi, Jo Van Bulck, Julian Stecklina, Thomas Prescher, and Daniel Gruss. 2019. ZombieLoad: Cross-privilege-boundary data sampling. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 753–768.
[70] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, and V. Zakharchenko. 2019. Emerging MPEG Standards for Point Cloud Compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 9, 1 (2019), 133–148.
[71] Iraj Sodagar. 2011. The MPEG-DASH Standard for Multimedia Streaming Over the Internet. *IEEE MultiMedia* 18, 4, 62–67.
[72] Apple Support. 2020. About Face ID advanced technology. https://support.apple.com/en-us/HT208108
[73] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
[74] Jo Van Bulck, Marina Minkin, Ofir Weisse, Daniel Genkin, Baris Kasikci, Frank Piessens, Mark Silberstein, Thomas F Wenisch, Yuval Yarom, and Raoul Strackx. 2018. Foreshadow: Extracting the Keys to the Intel SGX Kingdom with Transient Out-of-Order Execution. In *USENIX Security Symposium (Security)*. 991–1008.
[75] Jo Van Bulck, Daniel Moghimi, Michael Schwarz, Moritz Lipp, Marina Minkin, Daniel Genkin, Yarom Yuval, Berk Sunar, Daniel Gruss, and Frank Piessens. 2020. LVI: Hijacking transient execution through microarchitectural load value injection. (2020), 1399–1417.
[76] Jeroen van der Hooft, Tim Wauters, Filip De Turck, Christian Timmerer, and Hermann Hellwagner. 2019. Towards 6DoF HTTP Adaptive Streaming Through Point Cloud Compression. In *ACM Multimedia Conference*. 2405–2413.
[77] Vimeo. 2019. Vimeo Depth Viewer. https://github.com/vimeo/vimeo-depth-player.
[78] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5265–5274.
[79] Ruoyu Wang, Yan Shoshitaishvili, Christopher Kruegel, and Giovanni Vigna. 2013. Steal This Movie: Automatically Bypassing DRM Protection in Streaming Media Services. In *USENIX Security Symposium (Security)*. 687–702.
[80] Tom Warren. 2017. Windows 10's face authentication defeated with a picture. https://www.theverge.com/2017/12/21/16804992/microsoft-windows-10-windows-hello-bypass-security.
[81] Chong Xiang, Charles R. Qi, and Bo Li. 2019. Generating 3D Adversarial Point Clouds. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9136–9144.
[82] Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. 2019. MeshAdv: Adversarial Meshes for Visual Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6898–6907.
[83] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2017. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991* (2017).
[84] Yi Xu, True Price, Jan-Michael Frahm, and Fabian Monrose. 2016. Virtual u: Defeating face liveness detection by building virtual models from your public photos. In *USENIX Security Symposium (Security)*. 497–512.
[85] Zirui Xu, Fuxun Yu, Chenchen Liu, and Xiang Chen. 2019. HAMPER: high-performance adaptive mobile security enhancement against malicious speech and image recognition. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*. 512–517.
[86] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. 2017. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 4362–4371.
[87] Fuxun Yu, Zirui Xu, Chenchen Liu, and Xiang Chen. 2019. MASKER: Adaptive Mobile Security Enhancement against Automatic Speech Recognition in Eavesdropping. In *Proceedings of the 56th Annual Design Automation Conference 2019*. 1–6.
[88] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L. Yuille. 2019. Adversarial Attacks Beyond the Image Space. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4302–4311.