Audio-Visual Autoencoding for Privacy-Preserving Video Streaming

Honghui Xu, Zhipeng Cai, *Senior Member, IEEE,* Daniel Takabi, and Wei Li, *Member, IEEE*

***Abstract*—The demand of sharing video streaming extremely increases due to the proliferation of Internet of Things (IoT) devices in recent years, and the explosive development of Artificial Intelligent (AI) detection techniques have made visual privacy protection more urgent and difficult than ever before. Although a number of approaches have been proposed, their essential drawbacks limit the effect of visual privacy protection in real applications. In this paper, we propose a cycle Vector Quantized Variational Autoencoder (cycle-VQ-VAE) framework to encode and decode the video with its extracted audio, which takes advantage of multiple heterogeneous data sources in the video itself to protect individuals' privacy. In our cycle-VQ-VAE framework, a fusion mechanism is designed to integrate the video and its extracted audio. Particularly, the extracted audio works as the random noise with a non-patterned distribution, which outperforms the noise that follows a patterned distribution for hiding visual information in the video. Under this framework, we design two models, including frame-to-frame (F2F) model and video-to-video (V2V) model, to obtain privacy-preserving video streaming. In F2F, the video is processed as a sequence of frames; while, in V2V, the relations between frames are utilized to deal with the video, greatly improving the performance of privacy protection, video compression, and video reconstruction. Moreover, the video streaming is compressed in our encoding process, which can resist side-channel inference attacks during video transmission and reduce video transmission time. Through the real-data experiments, we validate the superiority of our models (F2F and V2V) over the existing methods in visual privacy protection, visual quality preservation, and video transmission efficiency. Our codes and more data are now available in https://github.com/ahahnut/cycle-VQ-VAE.**

***Index Terms*—Audio-Visual, Privacy, Video Streaming, VQ-VAE**

I. Introduction

RECENTLY, sharing video streaming has been becoming increasingly popular with the wide applications of Internet of Things (IoT) devices [1]–[4], the number of which is predicted to reach about 45 billion by 2022 [5]. In transmission process, however, the video streaming may be maliciously intercepted by attackers who intend to infer individuals' private information from the videos using detection/prediction approaches [6]–[13]. Meanwhile, recent breakthroughs in deep learning accelerate the development of machine learning-based detection techniques, such as face detection [14]–[18] and semantic segmentation [19]–[23], which greatly increases the risk of privacy leakage in the video streaming. For example, from the video, attackers are able to use these advanced machine learning models to accomplish speech recognition [24], [25], action recognition [26], [27], and other activity detection. According to the latest Cost of a Data Breach Report proposed by IBM and the Ponemon Institute, privacy leakage causes property loss of millions dollars every year for individuals or companies concerned [28]. In addition, privacy protection has been regulated by law – on May 25th, 2018, the European Union's new General Data Protection Regulation (GDPR) came into force, requiring that people should have more control over their personal data. To this end, *privacy protection is deemed to be an indispensable component for video sharing.*

So far, a lot of research has been conducted to protect visual privacy in various ways. Some works aim to hide (partial) visual information for privacy protection [29]–[34], some approaches achieve anonymity through disturbing the original visual information [35]–[39], some methods protect privacy by changing the visual style of original information [1], [40], and some studies apply encryption methods to protect privacy in video [41]–[45]. However, *the existing works still have their limitations, which also challenges the design of effective protection for visual privacy*: (i) random noise is added to disturb the visual information in noise-based models, but the added noise usually follows some patterned distributions (*e.g.*, normal distribution), which can be utilized as prior knowledge in attackers' detection models to infer private information; (ii) some noise-based models are just trained to fool a certain kind of discriminative model, which cannot be used to defend general detection models in real applications; (iii) all the existing models, even the encryption-based ones, do not fully consider leakage of side-channel information (*e.g.*, traffic size) during video transmission, leading to vulnerability to side-channel inference attack; and (iv) these previous privacy-preserving models only focus on visual privacy in separated video frames but overlook the temporal information (*i.e.* the relations between frames) in video streaming, resulting in the low effect of privacy protection.

To overcome the above challenges, in this paper, *we propose to encode and decode video streaming with its extracted audio to achieve visual privacy protection while maintaining the expected visual quality and enhancing video transmission efficiency*. The extracted audio is a kind of random noise without any patterned distribution, which can better disturb the visual information as well as reduce the accuracy of malicious detection, compared with the noise that follows patterned distributions. For any video, its extracted audio cannot be generated or manipulated easily by attackers without any prior

H. Xu, Z. Cai, D. Takabi and W. Li were with the Department of Computer Science, Georgia State University, Atlanta, GA, USA (email: *hxu16@student.gsu.edu,* {*zcai, takabi, wli28*}*@gsu.edu*.)

Corresponding Author: Wei Li

knowledge, which ensures that the encoded video can only be decoded by the receivers who obtain the extracted audio. In other words, we aim to fuse multiple heterogeneous data sources (*i.e.*, the video and its extracted audio in this paper) to hide private visual information to defend detection attack and side-channel inference attack simultaneously during video transmission, which has not been addressed in literature.

To realize our proposed design, we develop a cycle-VQ-VAE framework to accomplish the fusion of heterogeneous data sources by employing the idea of codebook. Our framework consists of two VQ-VAE components with one working as the encoder and the other working as the decoder. Considering a pair of sender and receiver in video sharing applications, this kind of cycle framework can guarantee that the encoded video frame can be properly encoded at the sender and decoded at the receiver with high visual quality. To fuse different data sources, we map both the video and its extracted audio into an appropriate low-dimension space such that the codes of audio can disturb the codebook of video and the video information can be compressed effectively in the encoder. This encoding process that has not been presented in previous works makes sure that our framework can also be used to defend side-channel inference attack because it changes the traffic pattern of video streaming. Correspondingly, in the decoder, the same audio can be used to decode the encoded video by removing the extra codes of the audio from the disturbed codebook. Under this cycle-VQ-VAE framework, we develop two different models, including Frame-to-Frame (F2F) and Video-to-Video (V2V) models. In F2F model, we divide the video into a series of frames and reconstruct the images in a frame by frame manner. In V2V model, we treat the video as time-series data to perform image reconstruction taking into account the temporal information in video. Finally, we use the AVE dataset [46], two AI detection models, and one side-channel inference attack model to evaluate the superiority of our proposed F2F and V2V models over the state-of-the-art schemes in terms of visual privacy protection, visual quality preservation, and video transmission efficiency. In the following, the contributions of this paper are summarized.

- To the best of our knowledge, this is the first work to study the fusion of multiple heterogeneous data sources in video streaming for privacy protection.
- The extracted audio used in the cycle-VQ-VAE framework does not follow any patterned distribution and thus outperforms the works using the noise that follows some patterned distributions (*e.g.*, normal distribution).
- A novel cycle-VQ-VAE framework is developed to process video streaming, where the video and its extracted audio can be fused properly for protecting visual privacy, preserving visual quality, and compressing video information simultaneously.
- The integration of video compression and encoding is proposed to defend side-channel inference attack and reduce video transmission overhead.
- F2F and V2V models are designed under the cycle-VQ-VAE framework to achieve the goal of privacy protection; especially, V2V model exploits the temporal information for performance enhancement in privacy protection, video compression, and video reconstruction.
- The real-data experiment results confirm the effectiveness and the advantages of our proposed models compared with the state-of-the-art.

The rest of this paper is organized as follows. Related works are briefly summarized in Section II. After introducing preliminaries in Section III, we detail our proposed models in Section IV. In Section V, comprehensive experiments are conducted and analyzed. Finally, Section VI concludes this paper and discusses our future work.

II. Related Works

The state-of-the-art about visual privacy protection is summarized in this section.

A. Noise-based Privacy-Preserving Models

In the existing works, the methods of protecting visual privacy via adding noise can be classified into three main categories: (i) applying noise to disturb the feature attributes in order to decrease the accuracy of recognition results [29]–[31]; (ii) using steganography algorithms to generate the stego images to protect privacy [35]–[37]; and (iii) changing the image styles to hide original visual information for privacy preservation [1], [40].

Raval *et. al* [47] designed a perturbation mechanism that can obtain the trade-off between privacy and utility to protect visual secrets based on denoising autoencoder through the adversarial training. Brkić *et. al* [29] proposed to hide some biometric attributes with noise to reduce the accuracy of face recognition. They also proposed a Conditional Generative Adversarial Network (CGAN) to generate a human image of full body while offering a solid level of identity protection in [48]. Uittenbogaard *et. al* [30] designed a framework based on Generative Adversarial Network (GAN) to achieve the goal of detecting, removing, and inpainting moving objects in multi-view imagery while removing private regions that users care about. Meng *et. al* [35] proposed a steganography algorithm based on image-to-image translation using cycle-GAN to obtain the stego images for the purse of concealment and security in the transmission process. Tang *et. al* [36] developed an automatic stegangraphic distortion framework using GAN (named ASDL-GAN), which can be applied to images for the enhancement of privacy preservation. Kim and Yang [37] proposed a privacy-preserving adversarial protector network (termed PPAPNet), where a noise amplifier was used to optimize noise for effective image anonymization. Wu *et. al* [1] designed a method to keep video transmission secure by using a two-dimensional noise matrix as the 4-th channel of image combining with a 3-channel RGB image, in which a video frame was transformed from one style to another based on the architecture of cycle-GAN. Chen *et. al* [40] also proposed to transfer the realistic images into cartoon images based on GAN to protect privacy to a certain extent.

B. Encryption-based Privacy-Preserving Models

Besides, encryption-based methods are proposed to hide the private visual information in video.

Paruchuri *et. al* [41] encrypted foreground video bit-stream to hide the private information in surveillance systems. Liu *et. al* [42] obscured the human face region in real time by encrypting the spatial chaotic map of face. Zhang *et. al* [44] generated a key through a cryptographic MAC function by using the information of the head contour in the video frame, and the key is used in a stream cipher to lock the head information detected pedestrians for privacy preservation. Chu *et. al* [45] proposed a fast homomorphic encryption method to encrypt the video frames for secure video transmission.

C. Limitations of Existing Works

In the existing noise-based models, the used noise follows the normal distribution, which, however, can be utilized as prior knowledge by attackers to mitigate the impact of noise in their detection models and enhance the accuracy of information prediction. Even for the encryption-based models, all of the current works fail to fully consider privacy leakage in the video transmission process and thus may be vulnerable to the side-channel inference attack where attackers are able to infer private information by analyzing the users' traffic data [10]. What's worse, recent advanced machine learning models can achieve action recognition and activity detection in video by exploiting the temporal information (*i.e.* the relations between frames) [11], [26], [27], which has not been taken into account for privacy preservation yet. Due to the aforementioned limitations, these existing works may not be adequate to effectively accomplish the task of protecting visual privacy in video.

In this paper, to improve the performance of visual privacy protection, we propose F2F and V2V models based on cycle-VQ-VAE to encode and decode the video by employing the video's extracted audio and temporal information. The technical advantages and innovations of our models lie in several aspects. (i) The audio of a video is extracted as the noise whose distribution is random and unknown. Thus, *applying such extracted audio can disturb the visual information more effectively, compared with using the noise following patterned distribution (e.g., normal distribution)*. (ii) Different from the noise that follows patterned distribution, the extracted audio is unique and meaningful for its corresponding video, so that *it guarantees that the noise cannot be generated or manipulated easily and can be used to decode the encoded video only by the receivers who have the audio*. (iii) The process of video compression is incorporated into our cycle-VQ-VAE framework, *improving the resistance to side-channel inference attack during transmission and reducing the video transmission time*. (iv) The relations between frames are utilized in V2V by integrating cycle-VQ-VAE with the RNN layers, *making privacy protection, video compression, and video reconstruction more efficient.*

III. Preliminaries

Vector Quantized Variational AutoEncoder (VQ-VAE) is a state-of-the-art image generation model with convolutional layers' architecture, in which all features of video frames are mapped into the codebook [49]. With the help of codebook, high-dimension data can be mapped into a low-dimension space and also can be reconstructed from the mapped low-dimension space.

VQ-VAE model consists of one encoder E and one decoder D, in which E and D share a common codebook c. The encoder is used to embed the original observations x into feature maps that should be close to the codebook vector c, and the decoder is used to recover the original observations $\|x-D(c)\|_2^2$ using the codebook vector c. During this process, performance loss includes: (i) the codebook loss, which is the distance between the selected codebook c and the outputs of encoder and is computed by $\|sg[E(x)]-c\|_2^2$ with the codebook variables, and (ii) the communication loss, which is the distance between the outputs of encoder and the selected codebook c and is calculated via $\|sg[c]-E(x)\|_2^2$ with the encoder weights, where $E(x)$ is the output of the encoder, sg is the stop-gradient to learn the code mappings for the codebook generation, and β is a hyperparameter to control the reluctance to change the codebook c to the encoder output. The objective function of VQ-VAE is expressed in Eq. (1).

$$L = \|x-D(c)\|_2^2+\|sg[E(x)]-c\|_2^2+\beta\|sg[c]-E(x)\|_2^2. \quad (1)$$

All abbreviations used in this paper are listed in Table I.

TABLE I
ABBREVIATIONS

Abbreviation	Full Name
AE	Autoencoder
VQ-VAE	Vector Quantized Variational Autoencoder
cycle-VQ-VAE	cycle Vector Quantized Variational Autoencoder
F2F	Frame-to-Frame
V2V	Video-to-Video
MAC	Message Authentication Code
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
sg	stop-gradient

IV. Methodology

In this section, we propose a cycle Vector Quantized Variational AutoEncoder (cycle-VQ-VAE) framework, based on which we design two novel models to generate privacy-preserving video.

A. Cycle-VQ-VAE Framework

The architecture of our cycle-VQ-VAE framework is shown in Fig. 1. This framework consists of one encoder and one decoder, where the encoder is designed to generate the encoded video frames for privacy protection, the decoder is designed to recover the encoded video frames, and the process of mapping video is based on VQ-VAE.

In the encoder of our cycle-VQ-VAE framework, the video frames and its extracted audio that are of high-dimension data

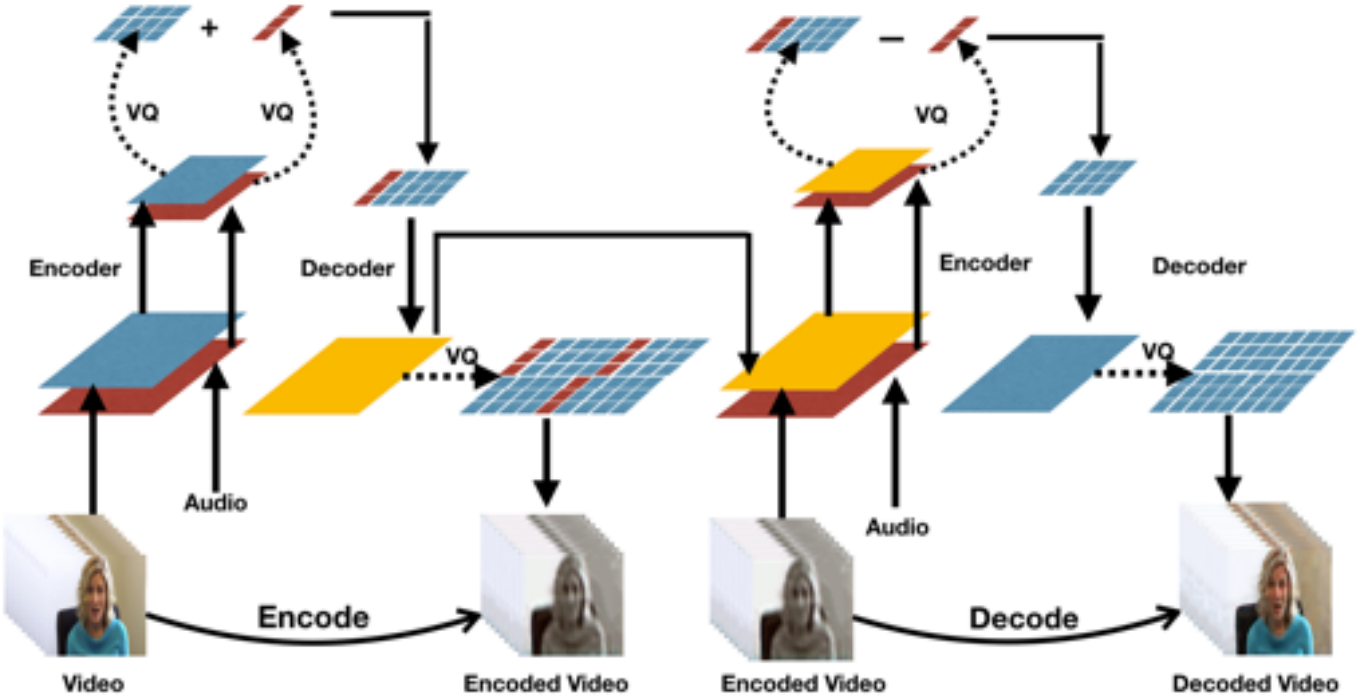


Fig. 1. The architecture of our cycle-VQ-VAE model

are mapped into a low-dimension space. The low-dimension representations of the audio are treated as the extra codes and added into the original codebook of video frames. Then, the disturbed codebook is used to generate the encoded video for privacy-preserving transmission. In the decoder, the low-dimension representations of the audio are removed from the disturbed codebook, and the original video frames can be reconstructed from the clean codebook.

It is worth mentioning that mapping high-dimension data into a low-dimension space is not a trivial issue. If the information in the codebook of video frames is much more than that in the codebook of audio in the low-dimension space, the codes of audio are not enough to disturb the codebook of video frames; if the information in the codebook of video frames is much less than that in the codebook of audio in the low-dimension space, it will be hard to extract the extra codes from the disturbed codebook of video frames to reconstruct the original video frames. That is, it is necessary to explore an appropriate low-dimension space, in which the codebook of video frames can be effectively disturbed using the codebook of its extracted audio. In this paper, we do comprehensive experiments by adjusting the dimension of codebook in the training process until we find a proper low-dimension space such that the encoded video frame reconstructed by the disturbed codebook is hardly detected by AI detection models, and the decoded video frame reconstructed by the clean codebook is similar to the original video frame.

Under our proposed cycle-VQ-VAE framework, a frame-to-frame (F2F) model and a video-to-video (V2V) model are developed. Especially, by utilizing the relations between video frames, V2V obtains an enhanced performance of privacy protection, video compression, and video reconstruction. The details of F2F and V2V models are demonstrated in Section IV-B and Section IV-C, respectively.

B. Frame-to-Frame (F2F) Model

1) Encoder: The encoder in F2F model includes one encoder module, one decoder module, and one codebook c_a^v as shown in Fig. 2. We encode the video frames with its extracted audio a to generate the encoded video v_a for protecting visual privacy. In other words, we use the low-dimension representations of audio as the extra codes c_a to disturb the codebook of the video frames c_v.

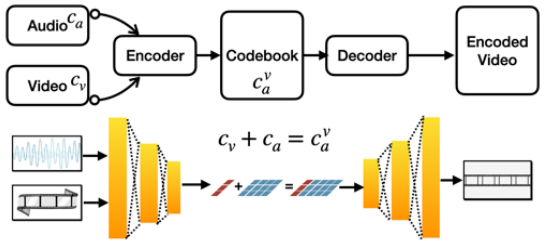


Fig. 2. The process of encoding (adding c_a into codebook c_a^v)

In the encoder module, we map both the video frames v and the audio a into the low-dimension space represented by codebook c_a^v, which is performed by using the stop-gradient sg [49]. Let $E(v_a^v|(v,a))$ be the expectancy of obtaining the encoded video with the video frames and the audio as inputs. According to the VQ-VAE mechanism, we can compute the codebook loss in Eq. (2) and the commitment loss in Eq. (3).

$$L_{E1} = \|sg[E(v_a|(v,a))] - c_a^v\|_2^2. \quad (2)$$

$$L_{E2} = \|sg[c_a^v] - E(v_a|(v,a))\|_2^2, \tag{3}$$

where $||\cdot||_2^2$ denotes the squared L2-norm.

In the decoder module, we generate the encoded video frames v_a from the disturbed codebook c_a^v, in which the reconstruction loss is computed by Eq. (4).

$$L_{D1} = \|v_a - D(c_a^v)\|_2^2. \tag{4}$$

To sum up, the loss function of the encoder in F2F model can be expressed in Eq. (5).

$$L_{Total1} = L_{E1} + \beta_e L_{E2} + L_{D1}, \tag{5}$$

where β_e is a hyperparameter to control the reluctance to change the codebook c_a^v to the encoded video v_a.

2) Decoder: At the side of receivers, the encoded video and the audio are high-dimension data. In order to obtain the original video, we first map the received data into the low-dimension space so as to clean the disturbed codebook of encoded video in the low-dimension space. Then, we reconstruct the decoded video in the high-dimension space.

Accordingly, the decoder in F2F model also has three components, including one encoder module, one decoder module, and one codebook c_v as shown in Fig. 3. In the decoder, we use the same audio a to decode the encoded video frames v_a with an aim that the decoded video frames should be similar to the original video frames v. To this end, we remove the extra codes c_a from the disturbed codebook c_a^v to obtain the clean codebook c_v of video frames.

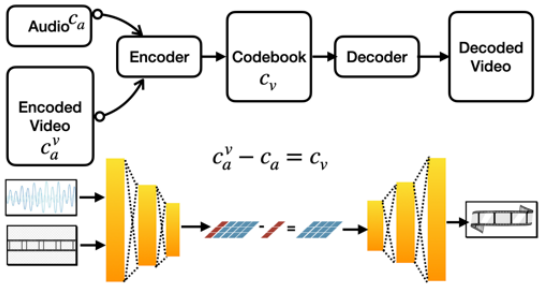


Fig. 3. The process of decoding (removing c_a from codebook c_a^v)

In the encoder module, we map both the encoded video frames v_a and the audio a into low-dimension space and learn the mappings through the stop-gradient sg operation. Let $(c_a^v|a)$ denote the disturbed codebook c_a^v, in which the codes of audio a are removed and $E(v|(v_a,a))$ denote the expectancy of obtaining decoded video frames with the encoded video frames and the audio being the inputs. The codebook loss and the commitment loss in this VQ-VAE are calculated by Eq. (6) and Eq. (7), respectively.

$$L_{E3} = \|sg[E(v|(v_a,a))] - (c_a^v|a)\|_2^2. \tag{6}$$

$$L_{E4} = \|sg[c_a^v|a] - E(v|(v_a,a))\|_2^2. \tag{7}$$

In the decoder module, we produce the decoded video frames from the clean codebook such that the decoded video frames are similar to the original video frames v. We remove the codes of audio c_a from the disturbed codebook c_a^v. The reconstruction loss is shown below.

$$L_{D2} = \|v - D(c_a^v|a)\|_2^2. \tag{8}$$

The loss function of the decoder in F2F model can be calculated by Eq. (9).

$$L_{Total2} = L_{E3} + \beta_d L_{E4} + L_{D2}, \tag{9}$$

where β_d is a hyperparameter to control the reluctance to change the clean codebook c_v to the original video v.

In summary, the loss function of our proposed F2F model is as follows,

$$L_{Total} = L_{Total1} + L_{Total2}. \tag{10}$$

We aim to minimize Eq. (10) in the training process, where L_{Total1} is minimized to obtain the encoded video frames using its extracted audio and L_{Total2} is minimized to decode the encoded video frames using the same audio such that the decoded video is similar to the original video.

C. Video-to-Video (V2V) Model

In F2F model, we divide the video into a series of frames and reconstruct the images in a frame by frame manner without considering the relations between frames. Motivated by the idea of video reconstruction in [50]–[53], we propose V2V model with the help of RNN layers, in which the temporal information (*i.e.* the relations between frames) in video is used for performance improvement in protection visual privacy, compressing video, and reconstructing video.

The architectures of encoder and decoder in V2V model are presented in Fig. 4(a) and Fig. 4(b), respectively. The difference between our F2F and V2V models is that we deploy a recurrent layer after each Convolutional Neural Network (CNN) block. A hidden state h in each recurrent layer (denoted by function f) is an output from the previous time step, i.e., for i-th CNN block, the output is $o_i = h_i = f(v, h_{i-1})$, where h_i is the hidden state in the i-th CNN block, and h_{i-1} is the hidden state in the $(i-1)$-th CNN block.

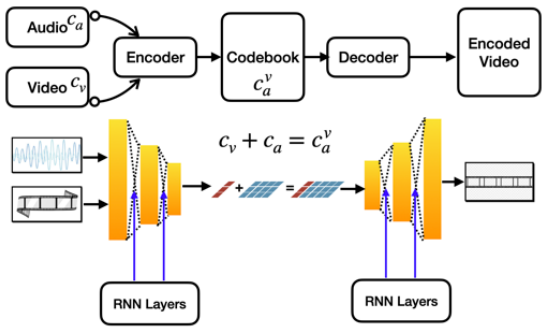
V. Experiments

In order to validate the effectiveness of our F2F and V2V models, extensive experiments are conducted to qualitatively and quantitatively evaluate the results of video encoding/decoding, the performance of privacy protection, and the efficiency of video transmission.

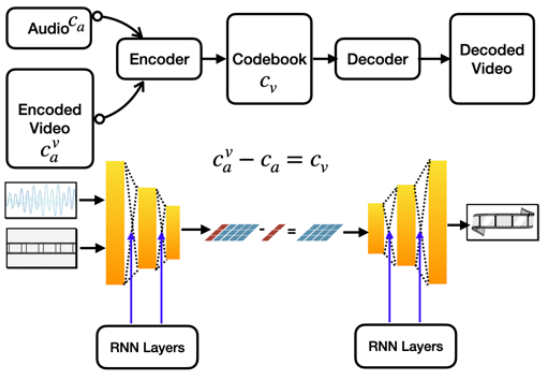
A. Experiment Settings

1) Dataset: In our experiments, we extract the video frames and the audio from 200 videos in the AVE dataset [46] to form the video dataset and audio dataset.

2) AI Detection Models for Video Frames: To illustrate that in our F2F and V2V models, the encoded video frames can resist AI detection and the decoded video frames can maintain visual quality, we adopt two AI detection models that have been widely used in real applications with mature technology. One is a face detection model that can detect the human face with a rectangle [14], and the other is the semantic segmentation model that can segment the human body with a pink color [19].



(a) The encoder architecture of V2V



(b) The decoder architecture of V2V

Fig. 4. V2V model

3) Side-Channel Inference Attack Model for Video Streaming: In real applications, a video can be typically encoded via a standard encoding method H264 [54] and then encrypted by TLS/SSL using 128-bit AES [55] for secure transmission. Nevertheless, the traffic pattern can be still utilized as side-channel information to infer individuals' activities in video streaming as the data traffic size can indicate the existence/type of an activity, resulting in privacy leakage. In our experiments, the attack approach of [10] is adopted, in which the traffic streaming is firstly divided into separate parts and then statistical coefficients (including mean, variance, skewness and kurtosis) of each separated traffic data are used as features to do activity recognition by using k-NN classification algorithm.

4) Two Baselines: We compare our proposed F2F and V2V models with two baselines. (1) AE based model: it is based on autoencoder (AE) architecture and adds the noise generated from the normal distribution into images [47] for privacy protection. (2) Style Translator based model: it changes the style of video frames to hide visual information based on cycle-GAN architecture [1].

All the experiment results are analyzed in Subsections V-B, V-C, V-D, and V-E. In this paper, video frames are presented to illustrate the effectiveness of our F2F and V2V models. More results of video and video frames can be found in https://github.com/ahahnut/cycle-VQ-VAE, and you can also create your own datasets for training using our open-source codes.

B. Qualitative Evaluation

There are original video frames, encoded video frames, and decoded video frames in the whole process of our cycle-VQ-VAE framework.

1) Video Frames of F2F and V2V: In Fig. 5, we show video frames in different phases in F2F and V2V models for performance comparison.

For the encoded/decoded video frames generated by F2F and V2V models, the results of face detection are presented in Fig. 5(a) and Fig. 5(b), and the results of semantic segmentation are presented in Fig. 5(d) and Fig. 5(e). Compared with the original video frames, we can draw a conclusion that in F2F and V2V models, the encoded video frames lose sufficient visual information to resist detection while the decoded video frames can recover the lost visual information effectively for the detection task.

From Fig. 5(c) and Fig. 5(f), one can see that by utilizing the relations between frames for video processing, V2V model outperforms F2F model in terms of video compression and video reconstruction. In Fig. 5(c), the encoded video frame of V2V is harder to be recognized, and the decoded frame of V2V is clearer for face detection. In Fig. 5(f), the encoded video frame of V2V loses more visual information causing worse semantic segmentation performance, and the decoded video frame of V2V has a higher visual quality for better semantic segmentation.

2) Encoded Video Frames: In Fig. 6(a), the encoded video frames in F2F and V2V cannot be detected by the face detector with a rectangle, but those of the AE based model and the Style Translator based model can be detected by the face detector. From Fig. 6(c), one can see that in our F2F and V2V models, human cannot be segmented by the semantic segmentation model from the encoded video frames, but in the AE based model and the Style Translator based model, human body can be segmented correctly. The main reason why our two models perform better is that the noise (*i.e.*, the extracted audio) of F2F and V2V does not follow any patterned distribution, greatly disturbs the visual information, and reduces the detection accuracy. Besides, V2V outperforms F2F in the video compression process due to consideration of the relations between frames even if they are both trained by our proposed cycle-VQ-VAE framework.

Moreover, since the noise can be filtered from real data by analyzing energy distribution [56], the energy distribution of encoded video frames is drawn in Fig. 8 for performance comparison. From Fig. 8(a), we observe that the energy distribution of original frames looks like a valley. Similarly, in Fig. 8(c) and Fig. 8(d), the energy distribution of the encoded frames of the two baselines only has one valley, which indicates that it is possible to recover the original frames from the encoded ones by removing the patterned noise in real applications. Differently, in Fig. 8(b) and Fig. 8(e), the energy distribution of encoded video frames of F2F and V2V contain several valleys, which means that our extracted audio can disturb the video information in a proper low-dimensional space where the audio energy can effectively influence the energy distribution of video frames. As a result, it becomes harder to recover the original frames from our

(a) Face Detection on F2F Video Frames (b) Face Detection on V2V Video Frames (c) Face Detection Comparison

(d) Semantic Segmentation on F2F Video Frames (e) Semantic Segmentation on V2V Video Frames (f) Semantic Segmentation Comparison

Fig. 5. The Results of Face Detection and Semantic Segmentation in F2F and V2V

encoded video frames just by removing the noise, which is consistent with the results of Fig. 5. Particularly, when comparing Fig. 8(b) with Fig. 8(e), we can find out that the energy distribution of encoded video frame in V2V is more irregular than that of encoded video frame in F2F because V2V achieves a better video compression performance by taking the relations between frames into consideration, leading to a larger difficulty in removing the noise for recovery.

3) Decoded Video Frames: As shown in Fig. 6(b), the decoded video frames of the four models can be observed. However, only the decoded video frames of our F2F and V2V models can be detected by the face detection model with a rectangle. Similarly, in Fig. 6(d), only the decoded video frames of F2F and V2V models can be segmented with a pink color through the semantic segmentation model. It is worth mentioning that the decoded video frames should have satisfied visual quality for observation/detection in real applications. From Fig. 6(b) and Fig. 6(d), we can see that our models can make the decoded video frames maintain the expected visual quality but the two baselines fail to make it, indicating that our models outperform the two baselines. In addition, in Fig. 6, compared with the decoded video frames in F2F, the decoded video frames in V2V can be better reconstructed when considering the relations between frames with respect to the video reconstruction task.

In Fig. 7, one more same video frame is chosen to compare our models with two baselines qualitatively for better illustrating the superiority of our models, especially V2V model. From Fig. 7(a), we observe that the encoded video frames in AE based and Style Translator based models can be detected by the face detection model, but the encoded video frames in F2F and V2V models cannot be detected, which means that our models outperform the two baselines. Especially, the encoded video frames in F2F model, AE based model, and Style Translator based model can be more or less segmented by the semantic segmentation model, but the encoded video frame in V2V model can not be segmented, indicating that V2V has the best performance of video compression and privacy protection. The results of Fig. 7(b) show that the decoded video frames in the four models can be detected by the face detection model and the semantic segmentation model, which means that F2F and V2V models can be used in video reconstruction. However, the decoded video frame in V2V has the highest visual quality, illustrating the advantage of V2V model in video reconstruction.

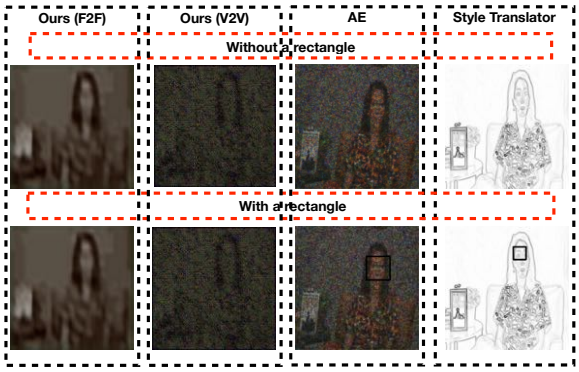
C. Quantitative Evaluation

We evaluate the quantitative performance of F2F and V2V models in terms of the average accuracies of face detection and semantic segmentation, and present the results in Table II and Table III.

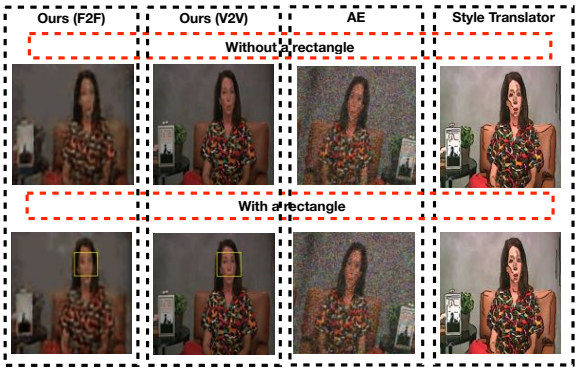
TABLE II
ACCURACY OF FACE DETECTION

	Ours(F2F)	Ours(V2V)	AE	Style Translator
Original	96.67%	96.67%	96.67%	96.67%
Encoded	6.00%	0.00%	26.67%	36.67%
Decoded	80.00%	96.67%	46.67%	63.33%

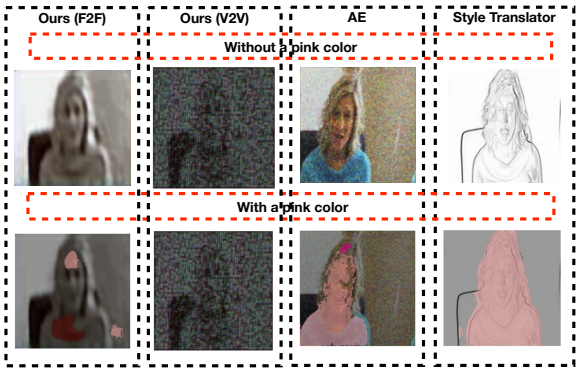
1) Video Frames of F2F and V2V: Compared with the average accuracy of face detection on the original video frames (*i.e.*, 96.67% in Table II), this accuracy is only 6.00% for the encoded video and can reach 80.00% for the decoded video in F2F model, and this accuracy decreases to 0.00% for the



(a) Face Detection on Encoded Video Frames: Ours *v.s.* Others



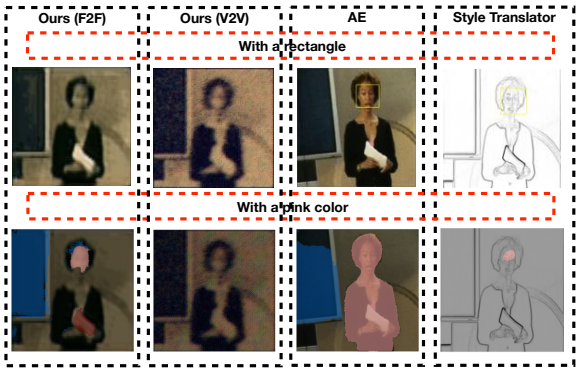
(b) Face Detection on Decoded Video Frames: Ours *v.s.* Others



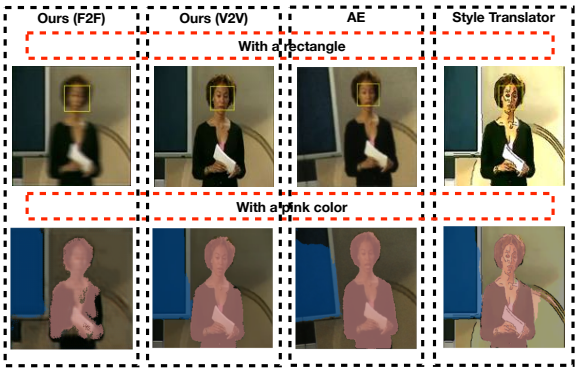
(c) Semantic Segmentation on Encoded Video Frames: Ours *v.s.* Others

(d) Semantic Segmentation on Dencoded Video Frames: Ours *v.s.* Others

Fig. 6. Performance Comparison for Face Detection and Semantic Segmentation

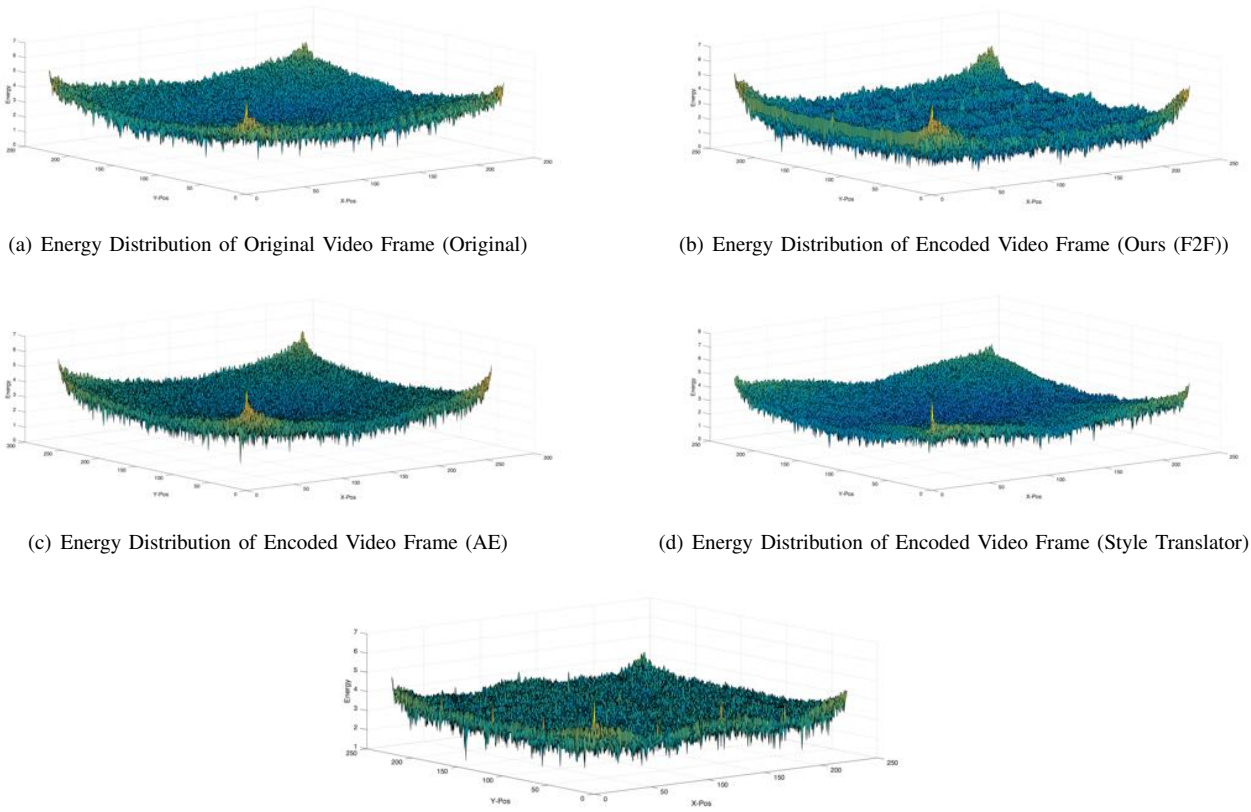


(a) Face Detection (Top) and Semantic Segmentation (Bottom) on Encoded Frames: Ours *v.s.* Others



(b) Face Detection (Top) and Semantic Segmentation (Bottom) on Dencoded Frames: Ours *v.s.* Others

Fig. 7. Recognition Comparison for Encoded Video Frames and Decoded Video Frames



(a) Energy Distribution of Original Video Frame (Original)

(b) Energy Distribution of Encoded Video Frame (Ours (F2F))

(c) Energy Distribution of Encoded Video Frame (AE)

(d) Energy Distribution of Encoded Video Frame (Style Translator)

(e) Energy Distribution of Encoded Video Frame (Ours (V2V))

Fig. 8. Energy Distribution of Encoded Video Frame

TABLE III
ACCURACY OF SEMANTIC SEGMENTATION

	Ours(F2F)	Ours(V2V)	AE	Style Translator
Original	93.30%	93.30%	93.30%	93.30%
Encoded	6.70%	0.00%	20.00%	36.67%
Decoded	73.33%	93.30%	43.30%	60.00%

encoded video and can be recovered back to 96.67% for the decoded video in V2V model. As shown in Table III, the average accuracy of semantic segmentation on original video frames is 93.30%; by using F2F model, the accuracy decreases to 6.70% on the encoded video frames and achieves 73.33% on the encoded video frames; and by using V2V model, this accuracy is only 0.00% on the encoded video and can reach 93.30% on the decoded video. These results illustrate that our F2F and V2V models can reduce the risk of privacy leakage in the encoded video frames while successfully recovering the lost visual information in the decoded video frames for real applications. In other words, our models are effective for privacy preservation in video streaming.

2) Encoded Video Frames: With respect to face detection on the encoded video frames, the average accuracies in our F2F model, our V2V model, the AE based model, and the Style Translator based model are 6.00%, 0.00%, 26.67%, and 36.67%, respectively (see Table II). In addition, for semantic segmentation on the encoded video frames, the average accuracies in our F2F model, our V2V model, the AE based model, and the Style Translator based model reach 6.70%, 0.00%, 20.00%, and 36.67%, respectively (see Table III). From the above comparison, one can see that our F2F and V2V models can lower detection accuracy on the encoded video frames in face detection and semantic segmentation and thus perform better than the two baselines in protecting visual privacy. This is because for the video, our models utilize the extracted audio that is a type of random and non-patterned distributed noise to blur the visual information while the two baselines use patterned distributed noise. What's more, V2V can obtain a lower detection accuracy than F2F in face detection and semantic segmentation on the encoded video frames since more visual information is lost in the V2V's encoding process when taking the relations between frames into account.

3) Decoded Video Frames: The decoded video frames are expected to recover the lost visual information as much as possible for further utilization. From Table II and Table III, one can see that a higher average accuracy of face detection/semantic segmentation on the decoded video frames is achieved by our F2F and V2V models, which means our models outperform the two baselines in terms of the visual quality of decoded video frames. In addition, by comparing F2F and V2V, the decoded video frames in V2V can better be applied in face detection and semantic segmentation tasks, which means that considering the relations between frames in V2V is helpful for reconstructing a high-quality video.

D. Security Analysis

In our F2F and V2V models, we can encode the video frames with its extracted audio and decode the encoded video frames with the same audio. The encoded video frames can (i) defend against the detection attacks using face detection and semantic segmentation during the transmission process, (ii) defend against side-channel inference attack, and (iii) only be decoded with the same audio received by the authorized receivers, which is deeply analyzed as follows.

1) Defense against Detection Attacks: We use two mainstream detection models to validate that our encoded video frames can prevent the visual information from being accurately detected. As shown in Table II and Table III, compared with the two baselines, our F2F and V2V models obtain a lower average accuracy in both face detection and semantic segmentation for the encoded video frames. The main reason lies in the method noise generation: in our encoded video frames, the noise (*i.e.*, the extracted audio) is extracted from the video so that it owns non-patterned distribution and sufficient randomness to help improve the performance of protecting visual information; while in the two baselines, the noise is generated from patterned distribution (*i.e.* normal distribution), which can be used as prior knowledge for information detection. Moreover, compared with F2F, V2V can obtain a lower accuracy and even decrease the detection accuracy to 0.00% in both face detection and semantic segmentation for the encoded video frames. This is because considering the relations between frames is effective to encode the visual information of video frames.

2) Defense against Side-Channel Inference Attack: The prior work [10] reveals that the traffic pattern of video streaming can be used as side-channel information to infer human's activities during the transmission even if the video streaming is encrypted by TLS/SSL. Fig. 9 shows that the traffic pattern of original video streaming and that of the encrypted original video streaming have a pretty high similarity.

To investigate the performance of video encoding methods in resisting the side-channel inference attack, the encoded video streaming is generated using the encoded video frames. The traffic size of the original video streaming, the encoded video streaming of F2F, the encoded video streaming of V2V, and the encoded video steaming of two baselines are presented in Fig. 10(a). Then, we use the side-channel inference method in [10] to calculate the accuracy of activity inference in video streaming and report the results in Table IV, where the average accuracy of activity inference is 95.8% in the original video streaming. The average accuracy of activity inference is 95.60% in AE encoded video streaming and 94.50% in Style Translator encoded video streaming, indicating that these two encoding methods cannot prevent side-channel information leakage. Notably, the average accuracy of activity inference is reduced to 42.86% in the encoded video streaming of F2F and even reduced to 0.00% in the encoded video streaming of V2V. The reason is that the encoding process of our F2F and V2V model can effectively smooth the traffic pattern. In particular, the relations between frames are exploited for video compression in V2V, further increasing the difficulty of traffic analysis during transmission. Thus, we can conclude that our F2F and V2V models can effectively resist side-channel inference attack.

Moreover, experiments are conducted to compare our F2F and V2V models with two baseline models after using TLS/SSL (AES 128 bit) encryption method for video transmission, traffic size are shown in Fig. 10(b), and results of activity inference are presented in Table IV. In Fig. 10(b), the traffic pattern of video streaming seems almost unchanged after video encryption. In Table IV, the average accuracy of activity inference is 94.80% in AE encrypted encoded video streaming, 93.70% in Style Translator encrypted encoded video streaming, 41.98% in F2F encrypted encoded video streaming, and still 0.00% in V2V encrypted encoded video streaming. These results indicate that the encoding methods of AE and Style Translator cannot prevent the side-channel attack even if the encryption method is used during video transmission. On the contrary, our encoding models outperform these two baselines and can prevent the side-channel attack effectively.

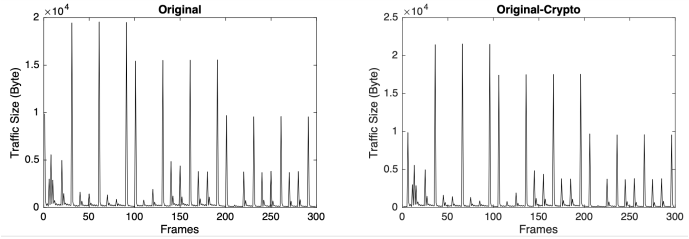


Fig. 9. Traffic Size of Original Video Streaming before and after Encryption

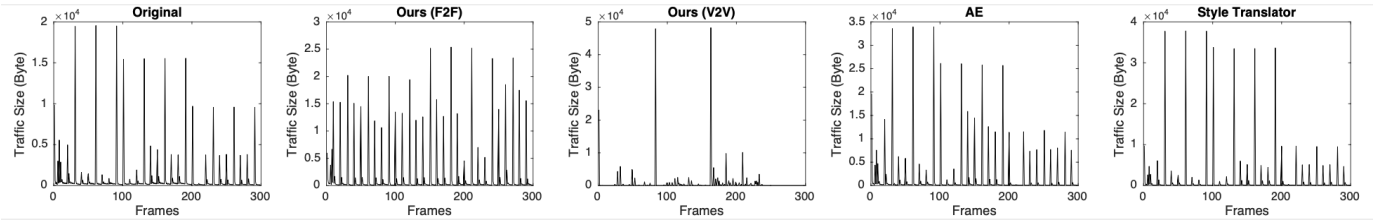
TABLE IV
RESULTS OF ACTIVITY INFERENCE

	Accuracy		Accuracy
Original	95.80%	Original-Crypto	94.90%
Ours (F2F)	42.86%	Ours (F2F)-Crypto	41.98%
Ours (V2V)	0.00%	Ours (V2V)-Crypto	0.00%
AE	95.60%	AE-Crypto	94.80%
Style Translator	94.50%	Style Translator-Crypto	93.70%

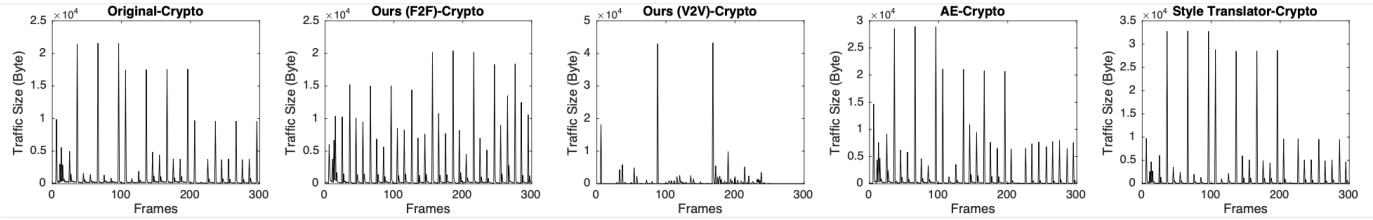
3) Defense against Un-authorization: In our F2F and V2V models, we train the same audio to encode the video frames and decode the encoded video frames. Different from the noise that follows certain distributions (*e.g.*, normal distribution), the audio extracted from its corresponding video is unique and cannot be easily generated or manipulated. Therefore, the video streaming can only be recovered by the authorized receivers who have the extracted audio.

E. Transmission Efficiency Analysis

Notice that the efficiency of video transmission has not yet been incorporated into visual privacy protection by the existing works, but the consideration of transmission efficiency is a necessary component for IoT devices and applications. One major advantage of our cycle-VQ-VAE framework over the state-of-the-art is that it can achieve effective visual privacy protection and efficient video transmission simultaneously. The main reason is that the encoder component in our cycle-VQ-VAE framework leverages the extracted audio to encode



(a) Traffic Size of Video Streaming before Encryption: Ours *v.s.* Others



(b) Traffic Size of Video Streaming after Encryption: Ours *v.s.* Others

Fig. 10. Traffic Size of Encoded Video Frames

the corresponding video, in which the video actually is compressed to a reduced size, and the transmission time can be reduced as well. On the contrary, the previous visual privacy-preserving models (such as AE based and Style Translator based model) exploit the noise to hide the original visual content, where the additional noise increases the video size, and the transmission time is increased. Furthermore, we do real-data experiments and use the transmission time as a performance metric to illustrate the transmission efficiency of our models during video streaming transmission in real applications. In Table V, we list the transmission time of uploading 10-second video streaming to an edge server at different network bandwidths. Compared with the original video, the transmission time is averagely decreased by 16.2% in our F2F model due to the video compression in the encoding process. Even better, the transmission time is averagely reduced by 53.4% in our V2V model as a better video performance can be achieved by considering the relations between frames. But the transmission time is averagely increased by 43.8% in AE based model and 9.1% in Style Translator based model, in which noise is added to disturb the original visual information without compression.

VI. Conclusion

In this paper, we propose an audio-visual autoencoder framework, named cycle-VQ-VAE. To the best of our knowledge, this is the first work to use multi-source information to generate privacy-preserving video streaming; especially, the audio is extracted from its corresponding video and used as the random noise to disturb the visual information. Since the extracted audio is unique and meaningful, it cannot be generated or manipulated easily and thus can be used by the authorized receivers to decode the encoded video. In addition, we develop F2F and V2V models under cycle-VQ-VAE framework. The entire encoded video streaming of our models has a more smooth traffic pattern, which can prevent the side-channel inference attacks using traffic size analysis. Besides, with video compression in our encoding process, the time of video transmission can be greatly decreased. Via extensive experiments, we demonstrate that our F2F model can preserve the expected visual quality, reduce the risk of visual privacy leakage, and improve the efficiency of video transmission; especially, V2V model outperforms F2F model in all evaluation metrics owing to the consideration of the relations between frames for video compression and reconstruction.

Acknowledgment

This work was partly supported by the National Science Foundation of U.S. (1741277, 1829674, 1704287, 1912753, and 2011845) and the Microsoft Investigator Fellowship.

References

[1] H. Wu, J. Feng, X. Tian, F. Xu, Y. Liu, X. Wang, and S. Zhong, "secgan: A cycle-consistent gan for securely-recoverable video transformation," in *Proceedings of the 2019 Workshop on Hot Topics in Video Analytics and Intelligent Edges*. ACM, 2019, pp. 33–38.

[2] S. K. Verma, "Method and system for sharing an output device between multimedia devices to transmit and receive data," Mar. 3 2020, uS Patent 10,581,933.

[3] T. Onohara, R. Ueda, K. Daini, T. Yoshio, Y. Kawabe, S. Iwayagano, H. Takuma, and E. Sakai, "Information-sharing device, method, and terminal device for sharing application information," May 7 2019, uS Patent 10,282,316.

[4] Y. P. Ong, C. M. Tan, and C. J. Y. Siau, "Systems and methods for processing a video stream during live video sharing," Sep. 10 2019, uS Patent 10,412,318.

[5] L. Capital, "billion cameras by 2022 fuel business opportunities," *Tech reports*, 2017.

[6] C. Li, Q. Zhong, D. Xie, and S. Pu, "Collaborative spatiotemporal feature learning for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7872–7881.

[7] O. Ulutan, S. Rallapalli, M. Srivatsa, C. Torres, and B. Manjunath, "Actor conditioned attention maps for video action detection," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 527–536.

[8] X. Liu, J.-Y. Lee, and H. Jin, "Learning video representations from correspondence proposals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4273–4281.

[9] H. Wu, Z.-J. Zha, X. Wen, Z. Chen, D. Liu, and X. Chen, "Cross-fiber spatial-temporal co-enhanced networks for video action recognition," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 620–628.

TABLE V
TRANSMISSION TIME AT DIFFERENT BANDWIDTHS (OURS (F2F) V.S. OTHERS)

	Original	Ours (F2F)	Ours (V2V)	AE	Style Translator
0.5MB/s	3.84s	3.24s(↓ 15.6%)	1.75s(↓ 54.4%)	5.6s(↑ 45.8%)	4.2s(↑ 9.3%)
1MB/s	1.87s	1.57s(↓ 16.1%)	0.87s(↓ 53.1%)	2.68s(↑ 43.3%)	2.05s(↑ 9.6%)
2MB/s	0.94s	0.78s(↓ 17.1%)	0.44s(↓ 52.7%)	1.34s(↑ 42.5%)	1.02s(↑ 8.5%)
Average		↓ 16.2%	↓ 53.4%	↑ 43.8%	↑ 9.1%

[10] H. Li, Y. He, L. Sun, X. Cheng, and J. Yu, "Side-channel information leakage of encrypted video stream in video surveillance systems," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 2016, pp. 1–9.

[11] R. Jiang, C. Qu, J. Wang, C. Wang, and Y. Zheng, "Towards extracting highlights from recorded live videos: An implicit crowdsourcing approach," in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE, 2020, pp. 1810–1813.

[12] M. R. Anderson, M. Cafarella, G. Ros, and T. F. Wenisch, "Physical representation-based predicate optimization for a visual analytics database," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019, pp. 1466–1477.

[13] Z. Huang, L. Wang, H. T. Shen, J. Shao, and X. Zhou, "Online near-duplicate video clip detection and retrieval: An accurate and fast system," in *2009 IEEE 25th International Conference on Data Engineering*. IEEE, 2009, pp. 1511–1514.

[14] A. F. A. Nasir, A. S. A. Ghani, M. A. Zakaria, A. P. A. Majeed, and A. N. Ibrahim, "Automated face detection using skin color segmentation and viola-jones algorithm," *Mekatronika*, vol. 1, no. 1, pp. 58–63, 2019.

[15] Z. Zhang, W. Shen, S. Qiao, Y. Wang, B. Wang, and A. Yuille, "Robust face detection via learning small faces on hard images," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1361–1370.

[16] S. Zhang, L. Wen, H. Shi, Z. Lei, S. Lyu, and S. Z. Li, "Single-shot scale-aware network for real-time face detection," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 537–559, 2019.

[17] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "Dsfd: dual shot face detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5060–5069.

[18] C. Li, R. Wang, J. Li, and L. Fei, "Face detection based on yolov3," in *Recent Trends in Intelligent Computing, Communication and Devices*. Springer, 2020, pp. 277–284.

[19] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1529–1537.

[20] S. Ghanem, A. Imran, and V. Athitsos, "Analysis of hand segmentation on challenging hand over face scenario," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2019, pp. 236–242.

[21] T. Meenpal, A. Balakrishnan, and A. Verma, "Facial mask detection using semantic segmentation," in *2019 4th International Conference on Computing, Communications and Security (ICCCS)*. IEEE, 2019, pp. 1–5.

[22] S. Benini, K. Khan, R. Leonardi, M. Mauro, and P. Migliorati, "Face analysis through semantic face segmentation," *Signal Processing: Image Communication*, vol. 74, pp. 21–31, 2019.

[23] Y. Wang, B. Luo, J. Shen, and M. Pantic, "Face mask extraction in video sequence," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 625–641, 2019.

[24] H. Hung and S. O. Ba, "Speech/non-speech detection in meetings from automatically extracted low resolution visual features," Idiap, Tech. Rep., 2009.

[25] S. Chaudhuri, J. Roth, D. P. Ellis, A. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. G. Reid, K. Wilson *et al.*, "Ava-speech: A densely labeled dataset of speech activity in movies," *arXiv preprint arXiv:1808.00606*, 2018.

[26] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," *arXiv preprint arXiv:1912.04487*, 2019.

[27] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi *et al.*, "Ava-activespeaker: An audio-visual dataset for active speaker detection," *arXiv preprint arXiv:1901.01342*, 2019.

[28] IBM and the Ponemon Institute, "Cost of a data breach report highlights," https://www.ibm.com/security/data-breach, 2019.

[29] K. Brkić, T. Hrkać, Z. Kalafatić, and I. Sikirić, "Face, hairstyle and clothing colour de-identification in video sequences," *IET Signal Processing*, vol. 11, no. 9, pp. 1062–1068, 2017.

[30] R. Uittenbogaard, C. Sebastian, J. Vijverberg, B. Boom, D. M. Gavrila *et al.*, "Privacy protection in street-view panoramas using depth and multi-view imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 581–10 590.

[31] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 82–89.

[32] Z. Xiong, Z. Cai, Q. Han, A. Alrawais, and W. Li, "Adgan: Protect your location privacy in camera data of auto-driving vehicles," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, 2020.

[33] Z. Xiong, W. Li, Q. Han, and Z. Cai, "Privacy-preserving auto-driving: a gan-based approach to protect vehicular camera data," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 668–677.

[34] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey towards private and secure applications." *to appeawr in ACM Computing Surveys*, 2021.

[35] R. Meng, Q. Cui, Z. Zhou, Z. Fu, and X. Sun, "A steganography algorithm based on cyclegan for covert communication in the internet of things," *IEEE Access*, 2019.

[36] W. Tang, S. Tan, B. Li, and J. Huang, "Automatic steganographic distortion learning using a generative adversarial network," *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1547–1551, 2017.

[37] T. Kim and J. Yang, "Latent-space-level image anonymization with adversarial protector networks," *IEEE Access*, vol. 7, pp. 84 992–84 999, 2019.

[38] J. Wang, Z. Cai, and J. Yu, "Achieving personalized k-anonymity-based content privacy for autonomous vehicles in cps," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 4242–4251, 2019.

[39] Z. Cai, X. Zheng, and J. Yu, "A differential-private framework for urban traffic flows estimation via taxi companies," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 12, pp. 6492–6499, 2019.

[40] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9465–9474.

[41] J. Paruchuri, S.-c. Cheung, and M. Hail, "Video data hiding for managing privacy information in surveillance systems," *EURASIP Journal on Information Security*, vol. 2009, pp. 1–18, 2009.

[42] S. Liu and L. Kong, "Local chaotic encryption based on privacy protection on video surveillance," in *2018 8th International Conference on Social science and Education Research (SSER 2018)*. Atlantis Press, 2018.

[43] P. Zhang, T. Thomas, and S. Emmanuel, "Privacy enabled video surveillance using a two state markov tracking algorithm," *Multimedia systems*, vol. 18, no. 2, pp. 175–199, 2012.

[44] P. Zhang, T. Thomas, S. Emmanuel, and M. S. Kankanhalli, "Privacy preserving video surveillance using pedestrian tracking mechanism," in *Proceedings of the 2nd ACM workshop on Multimedia in forensics, security and intelligence*, 2010, pp. 31–36.

[45] K.-Y. Chu, Y.-H. Kuo, and W. H. Hsu, "Real-time privacy-preserving moving object detection in the cloud," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 597–600.

[46] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6047–6056.

[47] N. Raval, A. Machanavajjhala, and L. P. Cox, "Protecting visual secrets using adversarial nets," in *2017 IEEE Conference on Computer Vision*

and Pattern Recognition Workshops (CVPRW). IEEE, 2017, pp. 1329–1332.

[48] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic, "I know that person: Generative full body and face de-identification of people in images," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2017, pp. 1319–1328.

[49] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," *CoRR*, vol. abs/1906.00446, 2019.

[50] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 1144–1156.

[51] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," *arXiv preprint arXiv:1910.12713*, 2019.

[52] A. Mallya, T.-C. Wang, K. Sapra, and M.-Y. Liu, "World-consistent video-to-video synthesis," *arXiv preprint arXiv:2007.08509*, 2020.

[53] Y. Chen, Y. Pan, T. Yao, X. Tian, and T. Mei, "Mocycle-gan: Unpaired video-to-video translation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 647–655.

[54] C. Grecos and M. Y. Yang, "Fast inter mode prediction for p slices in the h264 video coding standard," *IEEE Transactions on Broadcasting*, vol. 51, pp. 256–263, 2005.

[55] H. K. Lee, T. Malkin, and E. Nahum, "Cryptographic strength of ssl/tls servers: Current and recent practices," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. ACM, 2007, pp. 83–92.

[56] A. K. Boyat and B. K. Joshi, "Image denoising using wavelet transform and wiener filter based on log energy distribution over poisson-gaussian noise model," in *2014 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2014, pp. 1–6.

Honghui Xu is a PhD student in Department of Computer Science at Georgia State University (GSU). He received a Bachelor's degree from University of Eletronic Science and Technology of China (UESTC). His research focuses on machine learning and deep learning, including the fundamental theory of machine learning, the applications of deep learning in computer vision field, and the topic about privacy-preserving machine learning.

Zhipeng Cai is currently an Associate Professor at Department of Computer Science, Georgia State University, USA. He received his PhD and M.S. degrees in the Department of Computing Science at University of Alberta, and B.S. degree from Beijing Institute of Technology. Prior to joining GSU, Dr. Cai was a research faculty in the School of Electrical and Computer Engineering at Georgia Institute of Technology. Dr. Cai's research areas focus on Internet of Things, Machine Learning, Cyber-Security, Privacy, Networking and Big data. Dr. Cai is the recipient of an NSF CAREER Award. He served as a Steering Committee Co-Chair and a Steering Committee Member for WASA and IPCCC. Dr. Cai also served as a Technical Program Committee Member for more than 20 conferences, including INFOCOM, ICDE, ICDCS. Dr. Cai has been serving as an Associate Editor-in-Chief for Elsevier High-Confidence Computing Journal (HCC), and an Associate Editor for more than 10 international journals, including IEEE Internet of Things Journal (IoT-J), IEEE Transactions on Knowledge and Data Engineering (TKDE), IEEE Transactions on Vehicular Technology (TVT).

Daniel Takabi received the B.S. degree in Computer Engineering from Amirkabir University of Technology in 2004, M.S. degree in Information Technology, Sharif University of Technology in 2007, and PhD degree in Information Science and Technology from University of Pittsburgh in 2013. He is currently an Associate Professor of computer science and the Next Generation Scholar with Georgia State University (GSU), Atlanta, GA, USA. He is also a Founding Director of the Information Security and Privacy: Interdisciplinary Research and Education (INSPIRE) Center which is designated as the National Center of Academic Excellence in Cyber Defense Research (CAE-R). His research interests include various aspects of cybersecurity and privacy, including privacy preserving machine learning, adversarial machine learning, advanced access control models, insider threats, and usable security and privacy. He is a member of IEEE and ACM.

Wei Li is currently an Assistant Professor in the Department of Computer Science at Georgia State University. Dr. Li received her Ph.D. degree in computer science, from The George Washington University, Washington DC, in 2016 and M.S. degree in Computer Science from Beijing University of Posts and Telecommunications, in 2011. She won the Best Paper Awards in ACM MobiCom Workshop CRAB 2013 and international conference WASA 2011, respectively. Her current research spans the areas of blockchain technology, security and privacy for the Internet of Things and Cyber-Physical Systems, secure and privacy-aware computing, Big Data, game theory, and algorithm design and analysis. She is a member of IEEE and a member of ACM.