

Learnable Cube-based Video Encryption for Privacy-Preserving Action Recognition

Yuchi Ishikawa, Masayoshi Kondo, Hirokatsu Kataoka

LY Corporation*, Tokyo, Japan

{yuchi.ishikawa, masayoshi.kondo, jpz4219}@lycorp.co.jp

Abstract

With the development of cloud services and machine learning, there has been an inevitable need to enhance privacy and security when serving video recognition models. Although existing image encryption methods can be used to address this issue, applying them frame by frame to videos is insufficient in two respects: model performance degradation and security strength. In this paper, we propose a novel encryption approach for privacy-preserving action recognition. It consists of two encrypting operations; Learnable Cube-based Video Encryption (LCVE) and ViT Scrambling. LCVE is video encryption based on spatio-temporal cubes, which has a large key space and can provide robust privacy protection. ViT Scrambling encrypts the Vision Transformer (ViT) model, which enables it to recognize the encrypted videos in the same manner as unencrypted videos without modifying the model architecture or fine-tuning on the encrypted data. We evaluate our method in an action recognition task with seven datasets containing a variety of action classes as well as motion and visual patterns. Empirical results demonstrate that LCVE combined with ViT Scrambling can preserve video privacy while recognizing action in encrypted videos as well as unencrypted videos. As a result, our approach outperforms existing privacy-preserving action recognition methods.

1. Introduction

With the spread of web cloud services and machine learning, it is becoming increasingly popular to develop and serve computer vision applications on cloud servers (e.g. human behavior recognition using surveillance cameras / work recognition in manufacturing plants). Although using web cloud services provides reasonable prices and stable operation, analyzing real-world videos often requires the preservation of personal information. A malicious at-

*LINE Corp underwent a merger and has been renamed to LY Corporation as of October 2023.

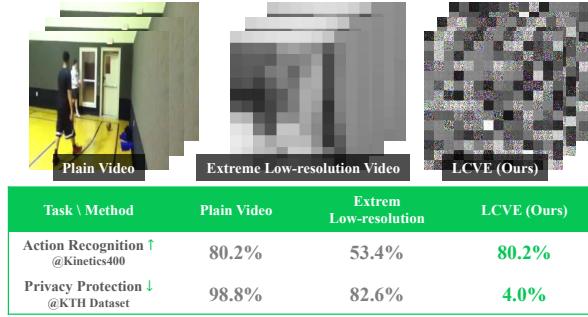


Figure 1. **Efficacy of our proposed method.** Our proposed framework demonstrates high accuracy while protecting privacy in the action recognition task in comparison with existing methods.

tacker may obtain these videos in transmission or at rest on cloud servers, allowing them to analyze private information. Recent reports state that training data may be reconstructed from the leak of information such as trained models [19] or model confidence [16]. Therefore preventing the leak of models trained with confidential data is also essential.

To mitigate these security concerns, some studies have been conducted on privacy-preserving image and video recognition. Most propose ideas for removing privacy information from images and videos [31, 46, 58]. However, these approaches are accompanied by degradation of model performance because the spatial features of objects and persons are lost as well. Some studies [37, 55] have proposed image encryption methods; however, the model structure needs to be modified to recognize the encrypted images. Therefore, it is difficult to leverage existing large-scale pretrained models that are optimized for recognition tasks. In addition, most approaches for privacy-preserving video recognition have only been evaluated on small datasets (to the best of our knowledge, except for [10]). Therefore, it is unclear whether these approaches scale to larger video datasets with more varying action classes, motion patterns, and people appearances.

To address these problems, we propose a novel encryption approach for privacy-preserving action recognition. It consists of two encrypting operations; **Learnable Cube-**

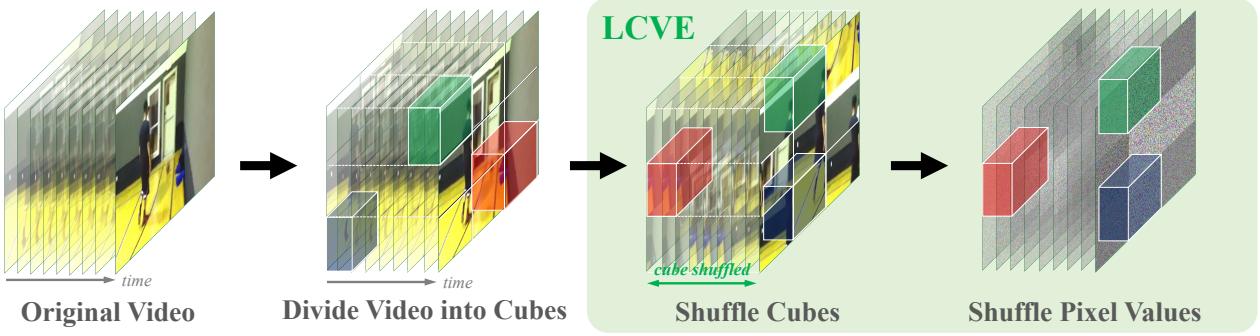


Figure 2. **Overview of LCVE.** LCVE consists of two shuffling operations; cube shuffling and pixel shuffling, using a single security key that determines the way of shuffling. LCVE has a larger key space and can provide more secure encryption than image-level encryption. The combination of LCVE and ViT Scrambling (Sec 3.3) enables ViT to recognize actions in encrypted videos.

based Video Encryption (LCVE) and ViT Scrambling. LCVE performs shuffling operations on the spatio-temporal cubes of a video instead of applying existing image encryption methods to each frame (Sec. 3.2). ViT Scrambling shuffles part of the Vision Transformer (ViT) [11]’s parameters in correspondence with LCVE for generating invariant outputs (Sec. 3.3). This enables the model to recognize the encrypted videos without additional training or modification of the model architecture. The combination of LCVE and ViT Scrambling can protect private information strongly while recognizing action in encrypted videos (Fig. 1). The contributions of this study are threefold.

- We propose a novel encryption approach, called Learnable Cube-based Video Encryption (LCVE) and ViT Scrambling, for privacy-preserving action recognition. The encrypted model can recognize the encrypted videos without additional training or modification of the model architecture (Sec. 3.2 and 3.3).
- Our LCVE can provide greater privacy protection because it has a larger key space and is more secure than existing image encryption methods (Sec. 3.4).
- We evaluate the effectiveness of our approach on seven datasets, including both large and small datasets, in an action-recognition task. We demonstrate that our approach can strictly protect private information without performance degradation on benchmarks with a wide variety of action classes, motion patterns, and the way people appear in videos (Sec. 5).

2. Related Work

Privacy protection has become a growing interest in recent years. In this section, we give an overview of privacy-preserving methods in computer vision systems.

2.1. Federated Learning

Federated learning [27, 33, 38, 66] enables a single shared model to be learned on individual user devices. Because

federated learning does not require sending raw data to one server, there is less data breach risk. However, several studies have shown that the training data can be reconstructed from the leakage of information, such as the model confidence [16], the difference between a model before and after being updated [49], the gradient [22], and trained models [19]. Therefore, for privacy protection, it is also essential to train and infer by using anonymized data instead of raw data. In this work, we focus on encrypting videos to hide private information.

2.2. Privacy-preserving Image Recognition

For privacy-preserving vision systems, there are early image anonymization approaches using blur [1], edge motion history images [8], and pixelation [32]. There are also several studies that focus on privacy protection at the hardware level [23, 43, 44, 61], but these are outside the scope of this paper. In recent years, image encoding approaches have also been proposed [20, 25, 26, 53, 65]. For example, there are several works on image transformation operations to remove visual information from images using adversarial training [26, 53]. However, these approaches lose spatial features in an image in the process of anonymization, which causes performance degradation. Also, some works reported that encoded images by InstaHide [25] and NeuraCrypt [65] can be reconstructed [6, 7].

As another way of protecting private information, several image encryption approaches have also been proposed. In particular, image scrambling [40, 41, 67] is one of the popular methods for image encryption. Image scrambling is to change the order of pixels and the bit values in an image with a secret key, making the image unrecognizable to humans. This process is invertible and can be reconstructed using the secret key. Recently, several studies proposed image scrambling methods that can be learned by neural networks but unrecognizable by humans [2, 37, 45, 52, 55]. For example, [55] proposed Learnable Encryption (LE), which

encrypts an image by dividing it into blocks, shuffling the pixels in the blocks, and subsequently changing its pixel values. However, these methods require modification of the model structure to recognize encrypted images.

On the other hand, [2, 45] proposed image scrambling methods that are easily applicable to Vision Transformer (ViT) architectures. In [45], an image is divided into patches and shuffled; then, its patches are further divided into sub-patches, and the pixels in its sub-patches are shuffled. The encryption method takes advantage of ViT’s ability to learn the relationship between shuffled pixels in a sub-patch and the relationship between reordered patches. So it does not require a dedicated model adapted for the encrypted image and provides high recognition accuracy. However, these encryption methods have a smaller key space than [37, 55], so there is a trade-off between model performance and confidentiality. Because of these issues, we do not apply image-level encryption methods to each video frame for privacy-preserving video understanding.

2.3. Privacy-preserving Action Recognition

With the emergence of large-scale datasets [18, 28, 29, 39], action recognition has become an active topic in computer vision. The performance of action recognition has been greatly improved by sophisticated architectures such as CNN-based methods [13, 15, 21, 51, 57, 59, 63] and ViT-based methods [3, 5, 12, 17, 35, 42, 64]. However, when considering the real-world application of these techniques, privacy protection should also be addressed.

For privacy-preserving action recognition, there are existing works using extreme low-resolution videos [9, 24, 47, 48]. Downsampling makes it more difficult to recognize private information in videos, but these methods can work effectively only for videos in which people are clearly captured. On the other hand, there are video encoding methods based on adversarial training to remove private information from videos [10, 31, 46, 58, 62]. The authors of [10, 58, 62] proposed an adversarial training framework for privacy-preserving action recognition, which learns an anonymization function that removes personal information. [31] proposed a BDQ encoder that removes privacy information through three modules: Blur, Difference, and Quantization. However, when these methods remove private information from videos, the spatio-temporal features that are necessary for action recognition can also be removed. This degrades the performance of the models.

In summary, existing works have a trade-off between the model performance and privacy protection, and require extra cost for a model to recognize anonymized videos. In addition, their efficacy is limited to small-scale and less diverse datasets. To address these issues, we propose a novel video encryption method for privacy-preserving video understanding. The video encrypted by our method can be

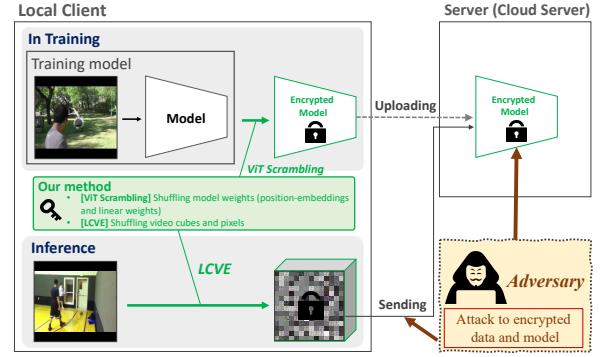


Figure 3. **Overview of our scenario.** On the client side, the data and model are encrypted by a pre-defined security key. Thus, the data and model sent by the client are securely encrypted. Even when there is leakage or interception by attackers, our method ensures that the key space is sufficient and private information in the videos is concealed.

recognized by ViT without extra cost or performance degradation, while private information is protected. Our method is also applicable to videos with various action classes, motion patterns, and the manner in which a person appears.

3. Proposed Method

In this section, we describe our approach, Learnable Cube-based Video Encryption (LCVE) and ViT Scrambling. First, we define the scenario of our privacy-preserving action recognition (Sec. 3.1). Next, we explain the algorithm of our video encryption method and how to recognize the encrypted videos (Sec. 3.2 and 3.3). Finally, we provide theoretical support for its security strength (Sec 3.4).

3.1. Overview and Scenarios

Figure 3 shows the relationship between the encryption process and the attacker when the system of action recognition is being served in the cloud. In this scenario, it is assumed that the video data is encrypted using a client-generated encryption key and is sent to the cloud server.

Our method determines a unique shuffling of the parameters and weights given an encryption key such that the encrypted video is recognized from the encrypted model equivalently to that of unencrypted videos with an unencrypted model. This retains the same performance as standard action recognition models. Apart from existing approaches, our method does not require re-training on the encrypted video data. We refer to this process of encryption as ViT Scrambling in our work. (See Sec. 3.3 for details).

Since the encryption key is created and used only on the client side, only the encrypted model and encrypted video data are sent to the cloud server. Thus, even if attackers

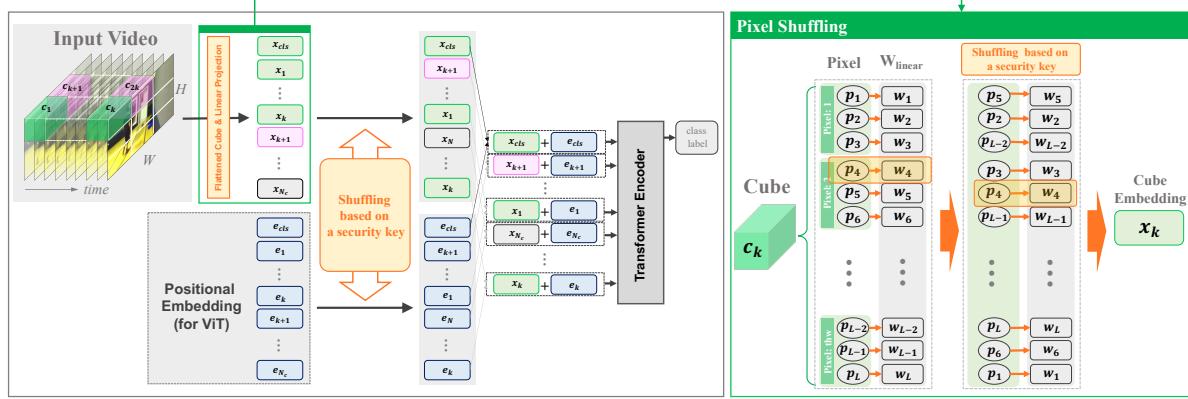


Figure 4. **Overview of ViT Scrambling.** ViT Scrambling consists of two shuffling operations on model parameters. The first operation shuffles the linear weights that are multiplied to the flattened pixel sequence of the spatio-temporal cube to calculate the cube embedding. The second shuffles the order of the positional encodings which correspond to the sequence of the cube embeddings. This way, calculated cube embeddings x_k are invariant, and the inputs to the transformer encoder only change order (except for the class token x_{cls} and its positional encoding e_{cls}).

can intercept videos from the cloud server or during data transmission, it will be difficult for the attackers to analyze or use them. Here, we evaluate the difficulty of recovering encrypted data and models by the size of the key space.

3.2. Learnable Cube-based Video Encryption

Next, we explain a novel video encryption method, called Learnable Cube-based Video Encryption (LCVE). Our approach is inspired by learnable image scrambling methods [37, 45, 55]. Image scrambling is comprised of two steps: confusion and diffusion [40]. The confusion step involves changing the order of pixels in an image. For example, [45] shuffles patches in an image as well as the pixel positions in sub-patches. The diffusion step changes the bit values to make them imperceptible to humans. For instance, LE [55] and ELE [37] apply a negative-positive transform. This transform reverses the intensities of pixels randomly in a 6-channel image created from the upper 4-bit and the lower 4-bit images. These two steps are processed using a security key, and the images are then encrypted. Image scrambling is invertible and encrypted images can be restored using the same key.

For video encryption, we extend these image scrambling approaches to videos. Our approach uses spatio-temporal cubes in a video instead of the spatial patches in an image. By expanding them in the time dimension, the encryption method becomes more secure than image-level encryption applied to each frame in a video. Given a video $V \in \mathbb{R}^{T \times H \times W \times C}$ ¹, the algorithm for the proposed LCVE is as follows:

1. Define a security key. Using the security key, deter-

¹ T is the number of frames in a video. H and W are the width and the height of each video frame. C is the number of the channel.

mine the random order of the cubes K_1 and the random order of RGB values K_2 .

2. Divide a video V into cubes with a size of $t \times h \times w$ as $x = \{x_1, x_2, \dots, x_{N_c}\}$, where N_c is the number of cubes and is calculated as $N_c = \frac{T \cdot H \cdot W}{t \cdot h \cdot w}$.
3. Shuffle these cubes x in the order $K_1 = [a_1, \dots, a_i, a_j, \dots, a_{N_c}]$, where $a_i \in \{1, \dots, N_c\}$ and $a_i \neq a_j$ if $i \neq j$. We call this procedure Cube Shuffling.
4. Flatten each cube into a vector $y \in \mathbb{R}^L$, where $L = t \cdot h \cdot w \cdot C$.
5. Shuffle RGB values in each cube in the order $K_2 = [b_1, \dots, b_i, b_j, \dots, b_L]$, where $b_i \in \{1, \dots, L\}$ and $b_i \neq b_j$ if $i \neq j$. We refer to this operation as Pixel Shuffling. Note that the way of shuffling pixels is the same over all cubes.
6. Reshape shuffled vectors into cubes with a size $t \times h \times w$
7. Concatenate all cubes to generate the encrypted video.

Figure 2 shows the example process of our encryption method. Our LCVE encrypts videos so that humans cannot perceive visual information. For an action recognition model to recognize actions in the encrypted video, the model architecture must be modified or fine-tuned to learn spatio-temporal patterns. On the other hand, our encryption method works as long as the model has a standard transformer architecture, so the encrypted videos can be recognized with a simple operation to the ViT weights.

3.3. ViT Scrambling

In this section, we first give an overview of Vision Transformer (ViT), and next describe our core method, ViT Scrambling. ViT for action recognition divides a video V

into spatio-temporal cubes and reshapes the flattened cubes $\mathbf{p} \in \mathbb{R}^{N_c \times L}$, where $L = t \cdot h \cdot w \cdot C$. Then each cube is fed into a linear projection $\mathbf{W} \in \mathbb{R}^{(t \cdot h \cdot w \cdot C) \times D}$ and mapped to D -dimension cube embeddings $\mathbf{x} \in \mathbb{R}^{N_c \times D}$. After concatenating the class token to cube embeddings, cube embeddings are added to positional encodings $\mathbf{e} \in \mathbb{R}^{(N_c+1) \times D}$. This is an input to a transformer encoder.

To recognize LCVE-encrypted videos without extra cost, we perform a simple operation to ViT, called ViT Scrambling. Using the same security key used in LCVE, ViT Scrambling rearranges both the weights of the linear projection to generate cube embeddings and positional encodings. This allows the ViT to treat the encrypted videos in the same way as plain videos. Figure 4 shows the overview of ViT Scrambling. ViT Scrambling consists of two steps; (1) Rearranging the positional encoding \mathbf{e} in the order K_1 except for the positional encoding for the class token, (2) Rearranging the linear projection \mathbf{W} in the order K_2 .

Next, we explain why this operation works well. We leverage the property that the output for the class token in a standard transformer encoder is not influenced by the order of input cube embeddings, as long as the correspondence between shuffled cubes and positional encoding is the same as that between original cubes and positional encoding. In this work, our target task is action recognition. Therefore, by rearranging positional encoding in the order K_1 , we can obtain the same output from shuffled cubes as the original video. In addition, since the linear projection of ViT takes as input each cube, shuffling pixels in a cube affects only the cube embeddings. Hence, rearranging the weight of the linear projection in the order K_2 enables ViT to generate the same cube embeddings as the original embeddings. Note that this process only needs to be applied to the model once when the encryption key is determined, so the key is not stored in the cloud server.

3.4. Security Strength

Here, we explain the security strength of LCVE, or the size of the key space, which refers to the theoretical set of all possible permutations of encryption. In our algorithm, when we shuffle N_c cubes and $t \cdot h \cdot w \cdot C$ values in each cube, the key space of LCVE is defined as below;

$$S = N_c! \cdot (thwC)! \quad (1)$$

When we assume that the video size is $16 \times 224 \times 224$ and the cube size is $2 \times 16 \times 16$, N_c and $thwC$ are calculated as $N_c = \frac{16 \times 224 \times 224}{2 \times 16 \times 16} = 1568$, $thwC = 2 \times 16 \times 16 \times 3 = 1536$. Therefore, the size of the key space S is;

$$S = 1568! \times 1536! \approx 3.5 \times 10^{8560} \quad (2)$$

Table 1 shows the comparison of the key space for each encryption method. LCVE has a larger key space than existing image encryption methods such as LE [55] and [45]. We

Table 1. Key space comparison for image and video encryption methods. Note that when we use our LCVE as image-level encryption, we shuffle patches in an image instead of cubes. We also extend existing encryption methods [45, 55] to a temporal dimension. In this case, we apply the patch-wise processes to the spatiotemporal cubes.

Method	Level	Key Space	
		Order	Our setting
Pixel Shuffle	frame	$(hwC)!$	1.8×10^{1884}
Patch Shuffle	frame	$N_p!$	5.1×10^{365}
LE [55]	frame	$(2hwC)! \cdot 2^{2hwC}$	3.9×10^{4691}
Z. Qi et al. [45]	frame	$(hw/4)! \cdot N_p!$	6.4×10^{454}
LCVE (Ours)	frame	$(hwC)! \cdot N_p!$	9.3×10^{2249}
Pixel Shuffle	video	$(thwC)!$	1.6×10^{4229}
Cube Shuffle	video	$N_c!$	2.1×10^{4331}
LE [55]	video	$(2thwC)! \cdot 2^{2thwC}$	1.3×10^{10306}
Z. Qi et al. [45]	video	$(hwt/8)! \cdot N_c!$	2.7×10^{4420}
LCVE (Ours)	video	$(thwC)! \cdot N_c!$	3.5×10^{8560}

Table 2. Dataset details.

Dataset	#class	#video	#train / #valid / #test
Kinetics400	400	306,245	241,181 / 19,877 / 38,671
SSV2	174	220,847	168,913 / 24,777 / 27,157
Diving48	48	18,404	15,026 / - / 1,970
UCF-101	101	13,320	9,536 / - / 3,783
HMDB51	51	6,766	3,570 / 1,666 / 1,530
IPN	13	4218	3117 / - / 1101
KTU	6	2390	1791 / - / 599

also compare the temporally extended versions of existing image encryption methods. As seen in Table 1, the video-extended LE is more secure than ours. However, it highly degrades the model performance in action recognition because its complex diffusion process makes it more difficult to recognize the encrypted videos (Sec. 5.1). On the other hand, our approach consists of only confusion steps, and combining with ViT Scrambling can achieve high model performance (Sec. 5.3), while providing much more secure encryption than video-extended [45].

4. Experiments

In this work, through three experiments, we demonstrate the efficacy of our approach. First, to evaluate how well the model understands the content from encrypted videos, we evaluate the accuracy of action recognition tasks using encrypted videos. In addition to the popular datasets used in previous studies, we evaluate our approach on large datasets of action recognition to demonstrate its generality (Sec. 5.1). Second, for evaluating how strictly privacy is protected using our encryption, following [31], we conduct privacy label prediction tasks (Sec. 5.2). Note here that we

aim to achieve lower prediction accuracy for concealing private attributes that are portrayed in videos. Third, we compare several settings in our approach as an ablation study to validate the impact of each component of our approach (Sec. 5.3).

4.1. Datasets and Evaluation Metrics

Action Recognition: Our goal is to accurately recognize actions in videos while preserving private information. To evaluate the performance in an action recognition task, we use a total of seven datasets with details in Table 2. Of these seven, Kinetics400 (K400) [29], Something-Something V2 (SSV2) [18] and Diving48 [34] are large-scale datasets that are often used as benchmarks for action recognition. UCF-101 [54], HMDB51 [30], IPN Hand Dataset [4] and KTH Dataset [50] are relatively small datasets and are often used to evaluate privacy-preserving action recognition. Some datasets do not provide validation sets. In that case, we re-split the original training split into training and validation sets by a ratio of 9 to 1. We report *top1* accuracy as in existing works.

Privacy Label Prediction: To evaluate the performance on privacy protection, we compare our method with existing works in a privacy label prediction task. Our goal is to prevent attackers to predict private attributes from encrypted videos. Therefore, the lower the accuracy on the privacy label prediction task is, the more strictly private information is preserved. For this experiment, we use IPN Hand Dataset and KTH Dataset. As privacy labels, We use gender labels for IPN Hand Dataset and actor labels for KTH Dataset aligned with [31]. We report *top1* accuracy for the privacy prediction tasks as in [31].

4.2. Comparative Methods

As comparative encryption methods, we use two image encryption methods proposed in [55] and [45]. LE [55] divides an 8-bit RGB image into $h \times w$ patches and splits each patch to the upper 4-bit and the lower 4-bit patches, making 6-channel image patches. Then, it applies to reverse intensities of randomly selected pixel positions and shuffling pixels in each patch. On the other hand, the method proposed in [45] divides an image into $h \times w$ patches and shuffles them. In addition, it splits each patch into $\frac{h}{2} \times \frac{w}{2}$ sub-patches and randomly shuffles pixel positions in a sub-patch.

We also implement temporally extended versions of these two methods by applying each process to a spatio-temporal cube instead of a spatial patch. In our experiments, we compare our LCVE with four encryption methods in total.

4.3. Implementation

On data processing, as a video, we take as input 16 frames of size 224×224 (i.e. $16 \times 224 \times 224$). The size of the cube in the LCVE is $2 \times 16 \times 16$. Therefore, the number of cubes N_c is $8 \times 14 \times 14$. As an action recognition model, we use a vanilla ViT backbone [11] pretrained on Kinetics400 using VideoMAE [56], one of the state-of-the-art methods. ViT uses joint space-time attention [3, 35] alongside joint space-time cube embedding [3, 12, 35], which treats a cube of size $2 \times 16 \times 16$ as a single token embedding. Note that this is the same size as the LCVE cube. In training, we used AdamW [36] as an optimizer. Except when we apply ViT Scrambling to a model, we finetune a model with encrypted videos. During the evaluation, we follow the method of multi-view testing [15, 60] used in [14, 56]. We create 5 clips from each video and generate 3 spatial views for each clip. The final prediction result is the average of these results obtained from all inputs. For privacy label prediction, we use the same model and settings as in action recognition. We experimented on 8 A100 Nvidia GPUs.

5. Result and Analysis

5.1. Privacy-Preserving Action Recognition

Comparison with Existing Methods: Table 3 shows the accuracy of LCVE + ViT Scrambling and previous methods on the encrypted action recognition task for three datasets; HMDB51, UCF-101, and Diving48. Therein, "Vision Transformer (unencrypted)" shows the performance on a standard action recognition task, and "Downsampling $N \times$ " shows the results when the frame size of the input video is reduced by the factor of N . Using our approach, inference on LCVE videos shows identical accuracy to the vanilla ViT when inferring on plain videos, therefore showing superior to all comparative methods. On the other hand, when ViT Scrambling is not applied for our method, in other words, when the security key is unknown, Table 3 shows that the performance of ViT trained on plain videos is significantly degraded. This result indicates how well visual information is concealed by encryption. When the security key is unknown, it is difficult to recognize actions in encrypted videos, preventing attackers to obtain information in case of a data breach.

Comparison with image encryption methods: We also compare the case when image encryption methods are applied to videos. Here, we encrypt videos in two ways; to patches within individual frames of the video, and to spatio-temporal cubes within the video as mentioned in Sec. 4.2. For extension of [55], the cube size is set to $2 \times 16 \times 16$. For that of [45], the cube size is set to $2 \times 16 \times 16$ and the sub-cube size to $1 \times 8 \times 8$.

Table 3 shows the results on three datasets; UCF-101, HMDB51, and Diving48. Although [45] performs

Table 3. Comparison with the existing privacy-preserving action recognition methods on HMDB51, UCF-101 and Diving48.

Method	Encryption Level	HMDB51	UCF-101	Diving48
Vision Transformer (unencrypted)				
- Plain Video	-	70.0	95.6	86.2
- Downsampling 2 \times	-	54.1	78.8	34.8
- Downsampling 4 \times	-	42.4	61.0	16.9
Extreme Low-resolution Video [48]	-	28.5	-	-
Extreme Low-resolution Video [47]	-	37.7	-	-
Extreme Low-resolution Video [24]	-	54.6	71.1	-
Adversarial Training [58]	-	42.3	62.1	-
Self-supervised Learning [10]	-	43.1	62.0	-
Z. Qi et al. [45]	frame	27.1	34.6	8.8
LE [55]	frame	7.3	4.4	5.0
Z. Qi et al. [45] extended to videos	video	40.7	59.8	14.5
LE [55] extended to videos	video	9.6	6.0	4.5
Inference on LCVE videos (Ours)				
- ViT trained on plain videos	video	1.9	1.0	3.7
- ViT finetuned on LCVE videos	video	19.0	33.0	7.9
- ViT trained on plain videos + ViT Scrambling	video	70.0	95.6	86.2

Table 4. Top1 Accuracy on large datasets; SSV2 and K400.

Method	SSV2	K400
Vision Transformer (unencrypted)		
- Plain Video	69.0	80.2
- Downsampling 2 \times	54.5	71.5
- Downsampling 4 \times	39.8	53.4
Inference on LCVE videos (Ours)		
- ViT trained on plain videos	0.4	0.3
- ViT finetuned on LCVE videos	19.2	59.5
- ViT trained on plain videos + ViT Scrambling	69.0	80.2

marginally better than LE [55] in the encrypted action recognition task, [45] has smaller key space, providing less encryption strength. Where model performance and encryption strength are often a trade-off, LCVE with ViT scrambling exhibits optimal performance while simultaneously having a large key space.

Evaluation on large-scale action recognition datasets:

We also evaluate LCVE on large-scale video datasets with more diverse action classes and visual appearance. Table 4 shows the results on two large-scale datasets; Something-Something V2 (SSV2) and Kinetics400 (K400). Since ViT Scrambling enables the model to treat the encrypted videos in the same manner as plain videos, the model can recognize action without performance degradation. As seen from the results, downsampling methods show a trade-off between downsampling rate and recognition performance.

Table 5. Comparing our proposed method in both action recognition and privacy label prediction tasks. Results with * are reported by the authors of [31].

Method	KTH		IPN	
	Action \uparrow	Actor \downarrow	Action \uparrow	Gender \downarrow
Plain Video	96.0	98.8	85.2	93.8
Downsample 2 \times *	91.6	91.8	82.3	80.0
Downsample 8 \times *	91.2	91.6	79.5	70.1
Downsample 32 \times *	85.6	82.6	53.0	63.3
H. Wang et al. [58]	85.9	19.3	76.0	65.0
BDQ [31]	91.1	7.2	65.0	59.0
Ours	96.0	4.0	85.2	61.3

5.2. Evaluation of Privacy Protection

Here we assume the scenario when an attacker, who does not know the encryption method, obtains the encrypted data and tries to predict private attributes using a classifier trained on unencrypted videos. In this experiment, we evaluate the encryption through two tasks; an action recognition task and a privacy label prediction task. Two datasets (the IPN dataset and the KTH dataset) are used, and the experimental settings follow [31]. Note that the privacy label prediction model is trained only on unencrypted videos, and encrypted videos are used only during inference. For the action recognition task, we evaluate the model with ViT scrambling applied, as the client has the key.

Table 5 shows the results with existing methods. Our encryption method successfully protects privacy on the KTH dataset and also performs relatively well on the IPN dataset.

Table 6. Comparing the performance impact of each component in our proposed method.

Pixel Shuffling	Cube Shuffling	ViT Scrambling	HMDB51		UCF-101		Diving48	
			top1	top5	top1	top5	top1	top5
✓			70.0	90.9	95.6	99.4	86.2	98.5
	✓		32.6	67.7	52.8	82.1	9.3	38.3
		✓	54.4	82.1	80.4	96.6	35.5	78.4
✓	✓		1.5	9.9	1.0	4.5	4.0	12.9
✓	✓	✓	19.0	46.7	33.0	66.0	7.9	35.3
✓	✓	✓	70.0	90.9	95.6	99.4	86.2	98.5

As for action recognition on the two datasets, our method shows the best classification rate. On the IPN dataset, BDQ [31] shows lower prediction rates for private labels, but it requires training the image encoder by adversarial learning. In contrast, our encryption method comprises of simple shuffling of cubes and pixels within the video and does not require re-training the recognition model. Moreover, applying ViT Scramble shows the same performance on encrypted data as in unencrypted action recognition. From these perspectives, our proposed method shows efficacy in privacy-preserving action recognition.

5.3. Ablation Study

In the ablation tests, we examine the performance impact of each operation in our method. We test combinations of three operations; Pixel Shuffling, Cube Shuffling, and ViT Scrambling. For each of these settings, we evaluated the performance of the model based on the accuracy of action recognition tasks on three datasets; UCF-101, HMDB51, and Diving48. Table 6 shows the accuracies. First, we discuss the impact of data encryption, which are the Pixel Shuffling and Cube Shuffling operations. Applying either Pixel Shuffling or Cube Shuffling significantly reduces classification performance, and when both are applied together, performance degrades further. These demonstrate the model’s inability to deal with encrypted data. This result also implies that even when an attacker were to obtain the encrypted data, the content is incomprehensible. Next, we discuss the impact of ViT Scrambling, which encrypts the model. When only applying ViT Scrambling, model performance is significantly lower because the encrypted model could not recognize unencrypted data. Therefore, attackers would not be able to use the encrypted model when it is compromised. Finally, when all operations are applied, model performance is the same as that of unencrypted action recognition.

6. Discussion and Limitation

Through our experiments, we demonstrated that LCVE can strongly protect privacy. We also showed that the combination of LCVE and ViT Scrambling enables ViT to rec-

ognize actions in encrypted videos in the same way as unencrypted videos. While LCVE + ViT Scrambling is effective in privacy-preserving action recognition, there are limitations because we leverage two properties unique to the transformer encoder.

The first property is that the input sequence order only affects positional encodings. ViT Scrambling avoids performance issues by preserving the logical positions for the shuffled input sequence. On the other hand, ViT Scrambling might be difficult to apply to Convolutional Neural Network-based models like [13, 15, 21] because the convolution operation relies on the local features for the input sequential order.

The second property is that the first token, or class token, is not affected by the order of the input tokens in a vanilla transformer encoder. Since our target task is action recognition, we need only the first token. However, sequential labeling tasks generally require the correct ordering of the outputs from the transformer encoder. In such a case, because it is non-trivial to rearrange the ViT Scrambled-outputs, the model would not work effectively.

7. Conclusion

In this paper, we proposed a novel encryption approach for privacy-preserving action recognition; Learnable Cube-based Video Encryption (LCVE) and ViT Scrambling. LCVE is an encryption method for videos that are applied on spatio-temporal cubes, which can provide high encryption strength and privacy protection. ViT Scrambling scrambles model weights in a way that enables the recognition of encrypted videos in the same manner as non-encrypted videos. No modifications to the model architecture or additional training on the encrypted data is required. In the experiments on seven datasets, we demonstrated that LCVE preserves video privacy while ViT Scrambling aids the recognition of encrypted videos as accurately as non-encrypted videos. However, there still remain limitations such as applications to CNN architectures and sequence-labeling tasks. In future works, we will focus on solving these issues and developing advanced encryption methods.

References

[1] Prachi Agrawal and PJ Narayanan. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(3):299–310, 2011. 2

[2] MaungMaung AprilPyone and Hitoshi Kiya. Privacy-preserving image classification using an isotropic network. *IEEE MultiMedia*, 29(2):23–33, 2022. 2, 3

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3, 6

[4] Gibran Benitez-Garcia, Jesus Olivares-Mercado, Gabriel Sanchez-Perez, and Keiji Yanai. Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. In *2020 25th international conference on pattern recognition (ICPR)*, pages 4340–4347. IEEE, 2021. 6

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[6] Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. An attack on instahide: Is private learning possible with instance encoding. *arXiv preprint arXiv:2011.05315*, 1, 2020. 2

[7] Nicholas Carlini, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Florian Tramer. Neuracrypt is not private. *arXiv preprint arXiv:2108.07256*, 2021. 2

[8] Datong Chen, Yi Chang, Rong Yan, and Jie Yang. Tools for protecting the privacy of specific individuals in video. *EURASIP Journal on Advances in Signal Processing*, 2007:1–9, 2007. 2

[9] Ji Dai, Jonathan Wu, Behrouz Saghafi, Janusz Konrad, and Prakash Ishwar. Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 68–76, 2015. 3

[10] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20164–20173, 2022. 1, 3, 7

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6

[12] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3, 6

[13] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 3, 8

[14] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 6

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3, 6, 8

[16] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015. 1, 2

[17] Harshayu Girase, Nakul Agarwal, Chiho Choi, and Karttikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18759–18769, 2023. 3

[18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The “something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3, 6

[19] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. *arXiv preprint arXiv:2206.07758*, 2022. 1, 2

[20] Fusheng Hao, Fengxiang He, Yikai Wang, Fuxiang Wu, Jun Cheng, and Dacheng Tao. Privacy-preserving vision transformer on permutation-encrypted images, 2023. 2

[21] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 3, 8

[22] Ali Hatamizadeh, Hongxu Yin, Holger R Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10021–10030, 2022. 2

[23] Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Privhar: Recognizing human actions from privacy-preserving lens. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 314–332. Springer, 2022. 2

[24] Mingzheng Hou, Song Liu, Jiliu Zhou, Yi Zhang, and Ziliang Feng. Extreme low-resolution activity recognition using a super-resolution-oriented generative adversarial network. *Micromachines*, 12(6):670, 2021. 3, 7

[25] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. Instahide: Instance-hiding schemes for private distributed learning. In *International conference on machine learning*, pages 4507–4518. PMLR, 2020. 2

[26] Hiroki Ito, Yuma Kinoshita, Maungmaung Aprilpyone, and Hitoshi Kiya. Image to perturbation: An image transformation network for generating visually protected images for privacy-preserving deep neural networks. *IEEE Access*, 9:64629–64638, 2021. 2

[27] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. 2

[28] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 3

[29] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 6

[30] Hildegarde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 6

[31] Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 518–534. Springer, 2022. 1, 3, 5, 6, 7, 8

[32] Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 15(1):101–116, 2001. 2

[33] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 2

[34] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 6

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3, 6

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[37] Koki Madono, Masayuki Tanaka, Masaki Onishi, and Tetsuji Ogawa. Block-wise scrambled image recognition using adaptation network. *arXiv preprint arXiv:2001.07761*, 2020. 1, 2, 3, 4

[38] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 2

[39] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 3

[40] Bhaskar Mondal. Cryptographic image scrambling techniques. In *Cryptographic and Information Security*, pages 37–65. CRC Press, 2018. 2, 4

[41] Bhaskar Mondal and Tarni Mandal. A nobel chaos based secure image encryption algorithm. *International Journal of Applied Engineering Research*, 11(5):3120–3127, 2016. 2

[42] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3163–3172, 2021. 3

[43] Francesco Pittaluga and Sanjeev J Koppal. Privacy preserving optics for miniature vision sensors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 314–324, 2015. 2

[44] Francesco Pittaluga and Sanjeev Jagannatha Koppal. Pre-capture privacy for small vision sensors. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2215–2226, 2016. 2

[45] Zheng Qi, AprilPyone MaungMaung, Yuma Kinoshita, and Hitoshi Kiya. Privacy-preserving image classification using vision transformer. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 543–547. IEEE, 2022. 2, 3, 4, 5, 6, 7

[46] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018. 1, 3

[47] Michael Ryoo, Kiyo Kim, and Hyun Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3, 7

[48] Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 3, 7

[49] Ahmed Mohamed Gamal Salem, Apratim Bhattacharyya, Michael Backes, Mario Fritz, and Yang Zhang. Updates-leak: Data set inference and reconstruction attacks in online learning. In *29th USENIX Security Symposium*, pages 1291–1308. USENIX, 2020. 2

[50] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004. 6

[51] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 3

[52] Warit Sirichotendumrong, Yuma Kinoshita, and Hitoshi Kiya. Pixel-based image encryption without key management for privacy-preserving deep neural networks. *Ieee Access*, 7:177844–177855, 2019. 2

[53] Warit Sirichotendumrong and Hitoshi Kiya. A gan-based image transformation scheme for privacy-preserving deep neural networks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 745–749. IEEE, 2021. 2

[54] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[55] Masayuki Tanaka. Learnable image encryption. In *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 1–2, 2018. 1, 2, 3, 4, 5, 6, 7

[56] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 6

[57] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3

[58] Haotao Wang, Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Privacy-preserving deep visual recognition: An adversarial learning framework and a new dataset. *arXiv preprint arXiv:1906.05675*, 2, 2019. 1, 3, 7

[59] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 3

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 6

[61] Zihao W Wang, Vibhav Vineet, Francesco Pittaluga, Sudipta N Sinha, Oliver Cossairt, and Sing Bing Kang. Privacy-preserving action recognition using coded aperture videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[62] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European conference on computer vision (ECCV)*, pages 606–624, 2018. 3

[63] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 3

[64] Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Sfformer: Semi-supervised video transformer for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18816–18826, 2023. 3

[65] Adam Yala, Homa Esfahanizadeh, Rafael GL D’ Oliveira, Ken R Duffy, Manya Ghobadi, Tommi S Jaakkola, Vinod Vaikuntanathan, Regina Barzilay, and Muriel Medard. Neuracrypt: Hiding private health data via random neural networks for public training. *arXiv preprint arXiv:2106.02484*, 2021. 2

[66] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 2

[67] Guodong Ye. Image scrambling encryption algorithm of pixel bit based on chaos map. *Pattern Recognition Letters*, 31(5):347–354, 2010. 2