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

Supplementary Material

In this supplementary material, we provide more details on LCVE with respect to the following points:

- The datasets we use in our experiments (Sec. A).
- The implementation details (Sec. B).
- Quantitative comparison of image encryption methods and their temporal extensions (Sec. C).
- Visualization samples of LCVE (Sec. D).

A. Datasets

In our experiments, we use seven datasets to evaluate privacy-preserving action recognition methods. Most existing methods are just evaluated on small-scale datasets, and their effectiveness on large datasets with high diversity remains unclear. However, our proposed method works well on a variety of datasets including many types of action classes, motions, and visual patterns. Herein, we provide an overview of the datasets used in this study.

HMDB51 [11]: this dataset contains approximately 6k video data collected from movies as well as YouTube. The videos are annotated with 51 action classes, which consist of five types of action: (i) general facial actions (e.g. smile), (ii) facial actions with object manipulation (e.g. eat), (iii) general body movements (e.g. jump), (iv) body movements with object interaction (e.g. kick ball), (v) body movements for human interaction (e.g. punch). HMDB51 is often used for evaluation of action recognition methods in [4, 8, 17, 18, 24]

UCF-101 [20]: it contains about 13k videos with 101 action classes. Action classes can be divided into five types: (i) Human-Object Interaction (e.g. Juggling Balls), (ii) Body-Motion Only (e.g. Push Ups), (iii) Human-Human Interaction (e.g. Head Massage), (iv) Playing Musical Instruments (e.g. Drumming), (v) Sports (e.g. Archery). UCF-101 is also used to evaluate privacy-preserving action recognition methods [4, 8, 24].

KTH Dataset [19]: this has more than 2k action sequences, containing six human actions (walking, jogging, running, boxing, hand waving, and hand clapping). In this dataset, 25 people perform each action with four types of patterns; static homogeneous background, scale variations, different clothes, and lighting variations. Therefore, we use the 25 actor identities for the privacy label prediction task, as in [12]. Note that there are no details about how to split the data into train/test sets in [12], so we use videos recorded on the static homogeneous backgrounds as the test set.

IPN Hand Dataset [2]: it is a hand gesture dataset containing more than 4k videos with 13 static and dynamic hand gesture classes like "Throw up" and "Zoom in". This dataset is annotated with some metadata like gender, background, and so on. We use gender annotations for the privacy label prediction task as in [12].

Diving48 [13]: this is a competitive diving video dataset for fine-grained action recognition, consisting of approximately 18k videos and 48 different classes of dives. This dataset has a few noticeable biases for static representations. Therefore, it is used to evaluate whether the action recognition model can capture motion information. There is no existing privacy-preserving action recognition method that conducts experiments on Diving48.

Something-Something V2 [6]: it contains more than 220k videos. The videos are labeled with 174 action classes, composed of primitive actions with everyday objects. In this large dataset, action classes are defined as caption-templates such as "Moving something up" and "Covering something with something". Therefore, to predict these action classes, it is necessary for models to extract not only visual features but also motion features from videos. Something-Something V2 is not yet used in existing research in privacy-preserving action recognition tasks.

Kinetics400 [10]: it consists of about 300k videos with 400 human action labels such as playing musical instruments, shaking hands, or riding a bike. This is the largest dataset used in our experiments. The videos were collected from YouTube. The duration of each video clip is approximately 10 seconds. Kinetics400 is one of the most popular benchmarks for the action recognition task, but most existing works do not use it for evaluating privacy-preserving action recognition methods.

Table 1. **Finetuning setting for each dataset.**

configuration	Kinetics400	Something-Something V2	UCF-101 Diving48	Other Datasets
optimizer	AdamW [14]			
learning rate	1e-3	5e-4	5e-4	1e-3
weight decay	0.05			
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$			
batch size	128			
learning rate schedule	cosine decay			
warmup epochs	5			
epochs	75	30	100	100
repeated augmentation [7]	2			
flip augmentation	✓	-	✓	✓
RandAug [3]	(9, 0.5)			
label smoothing [21]	0.1			
mixup [28]	0.8			
cutmix [27]	1.0			
drop path [9]	0.1	0.1	0.2	0.2
dropout	0.0	0.0	0.5	0.5
layer-wise lr decay [1]	0.75			
sampling	dense sampling [5, 26]	uniform sampling [25]	dense sampling	dense sampling

B. Implementation

In our implementation, we use the PyTorch [15] framework. Our training and inference codes are based on the VideoMAE implementation[1]. In our experiments, we finetune VideoMAE with the ViT-Base backbone, which is pretrained on Kinetics400.

Details of the finetuning configuration for each dataset are listed in Table 1. We use the same training and inference schemes as in [23]. Therefore, our results should ideally be the same as those reported by [23]. However, our results are slightly worse. This may result from the differences in the number of GPUs used and batch sizes.

C. Qualitative Comparison with Image Encryption Method.

In our experiments, we use the two image encryption methods proposed in [16, 22]. LE [22] applies shuffling pixels in each patch and reverses the intensities of the randomly selected pixel positions. The latter process is called a negative-positive transform. In contrast, the method proposed by Qi et al. [16] executes shuffling patches and shuffles the pixel positions in a sub-patch.

In addition to applying them frame by frame to videos, we implement temporally extended versions of two methods by applying each process to a spatio-temporal cube. For example, the temporally extended version of LE divides an 8-bit RGB video into spatio-temporal cubes and splits each cube into the upper 4-bit and the lower 4-bit cubes, making 6-channel video cubes. Then, it applies the negative-positive transform and shuffles the pixels in each cube. A temporally extended method proposed by Qi et al. divides a video into cubes and shuffles them. Then, it splits each cube into sub-cubes and randomly shuffles the pixel positions in each sub-cube.

Figure 1 shows the qualitative comparison with LCVE and the two image encryption methods [16, 22]. For this visualization, we set the video size to $8 \times 224 \times 224$. The patch and cube sizes are set to 16×16 and $2 \times 16 \times 16$, respectively. Although the videos encrypted by the method of Qi et al. retain color information, The LCVE video contains little information for identifying the original video. LE also can completely make videos visually unreadable, but it is difficult for neural networks to recognize the action, as demonstrated in our experiments. On the other hand, LCVE videos can be recognized by ViT Scrambling in the same manner as non-encrypted videos.

One of the weaknesses in our method is that LCVE generates cube artifacts, which may allow an attacker to guess how the video is encrypted. In future work, we will focus on developing encryption methods that do not produce such artifacts.

D. Visualization

We show the LCVE visualization samples on HMDB51 (Figure 2), UCF-101 (Figure 3), KTH Dataset (Figure 4), IPN Hand Dataset (Figure 5), Diving48 (Figure 6), Something-Something V2 (Figure 7), and Kinetics400 (Figure 8). In these visualizations, we adopt the same setting as

[1] https://github.com/MCG-NJU/VideoMAE

in Figure 1. As shown in these figures, LCVE can hide private information from a variety of videos. Through our experiments, we also demonstrate that it is difficult for neural networks to recognize content without ViT Scrambling. Therefore, our proposed method strictly protects privacy.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] 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.

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[4] 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.

[5] 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.

[6] 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.

[7] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020.

[8] 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.

[9] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016.

[10] 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.

[11] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz 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.

[12] 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.

[13] 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.

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

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[16] 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.

[17] Michael Ryoo, Kiyoon 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.

[18] 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.

[19] 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.

[20] 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.

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[28] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

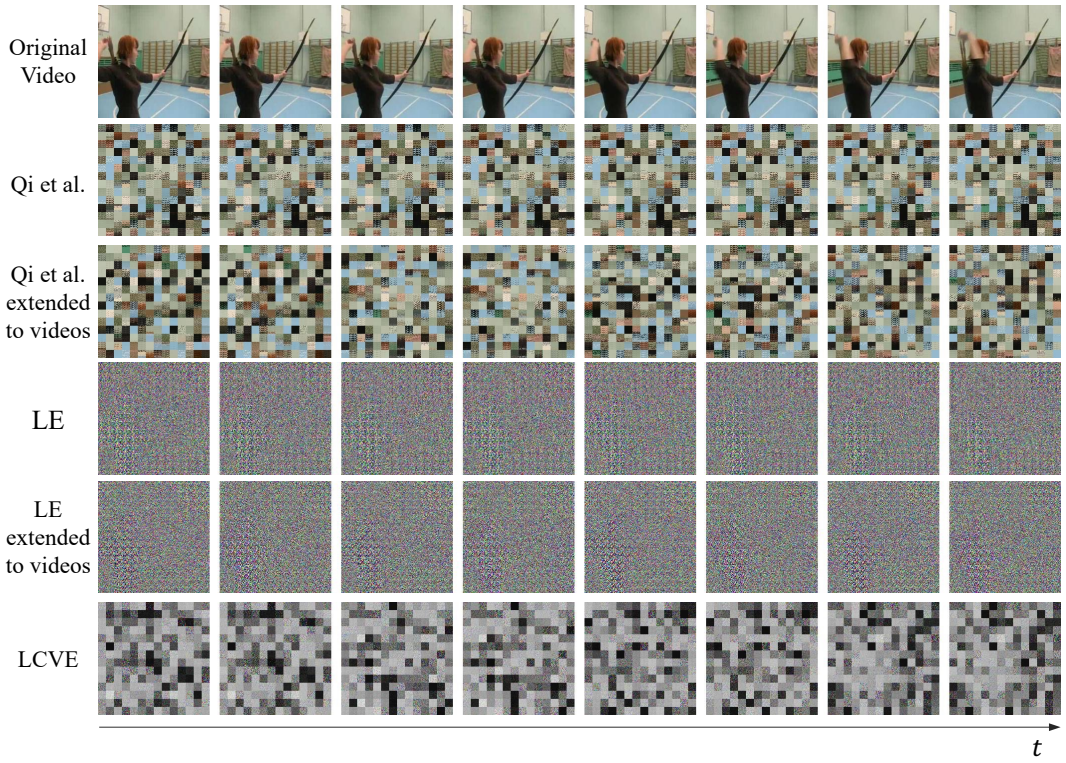

Figure 1. **Comparison of visualization samples with other encryption methods.**

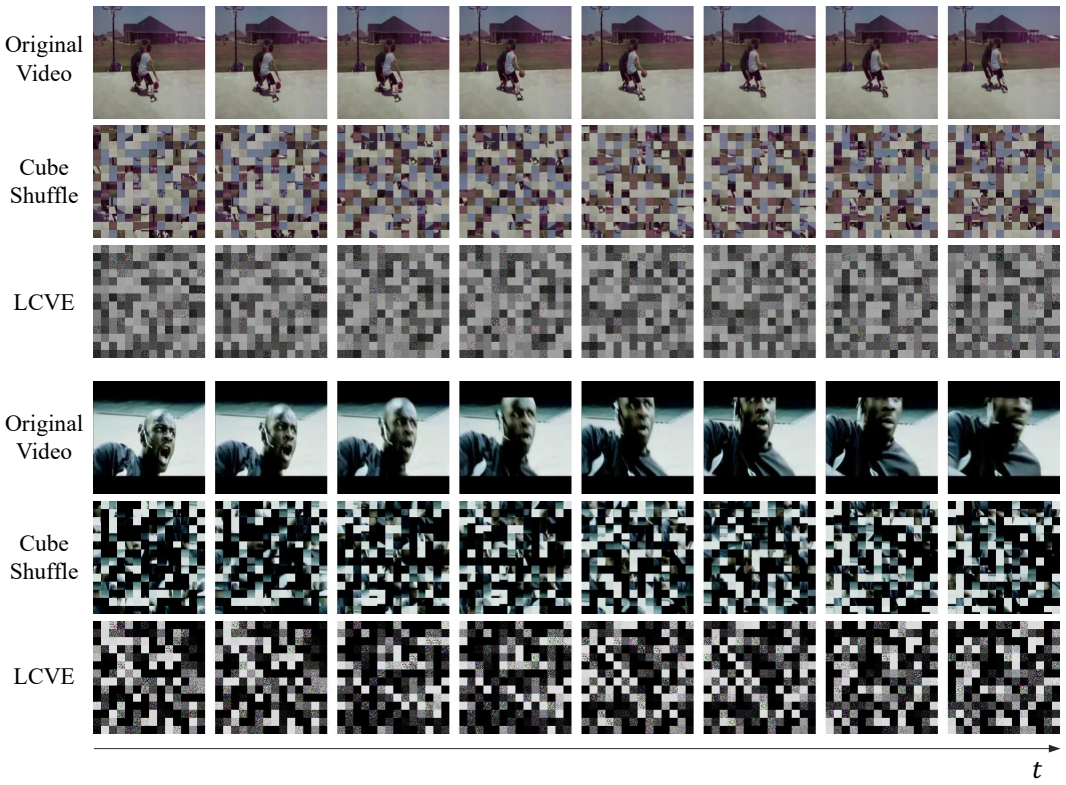

Figure 2. **Visualization samples on HMDB51.**

Figure 3. **Visualization samples on UCF-101.**

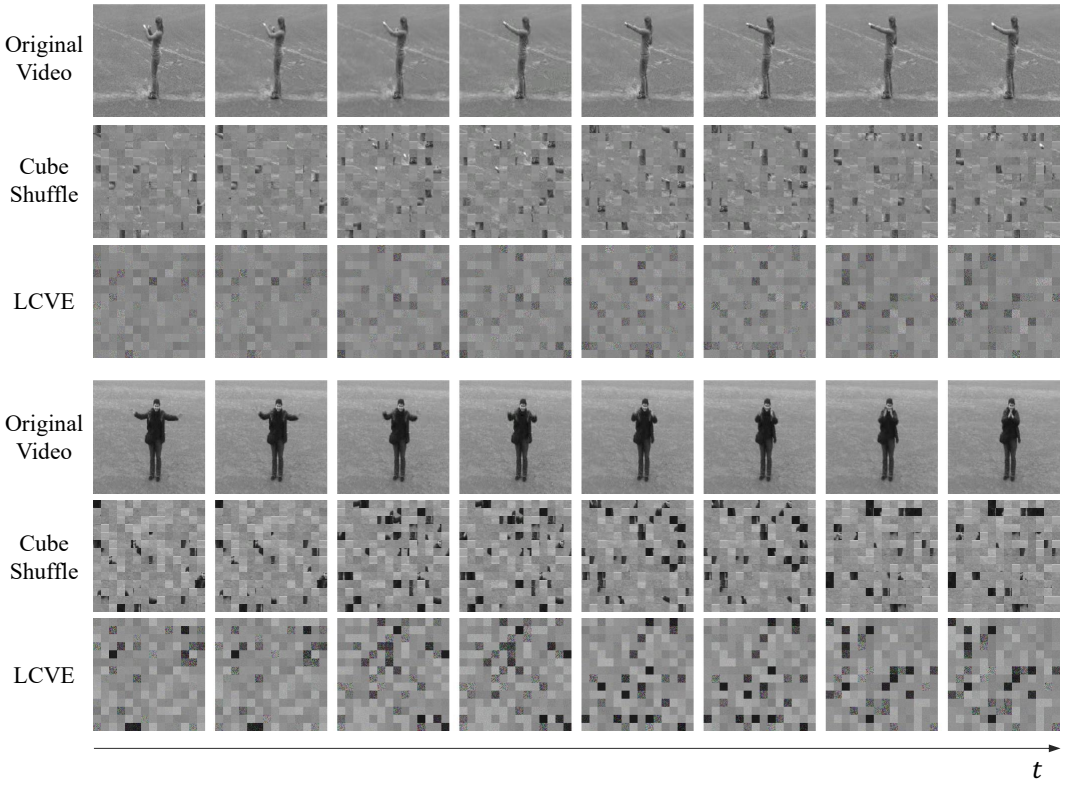

Figure 4. **Visualization samples on KTH Dataset.**

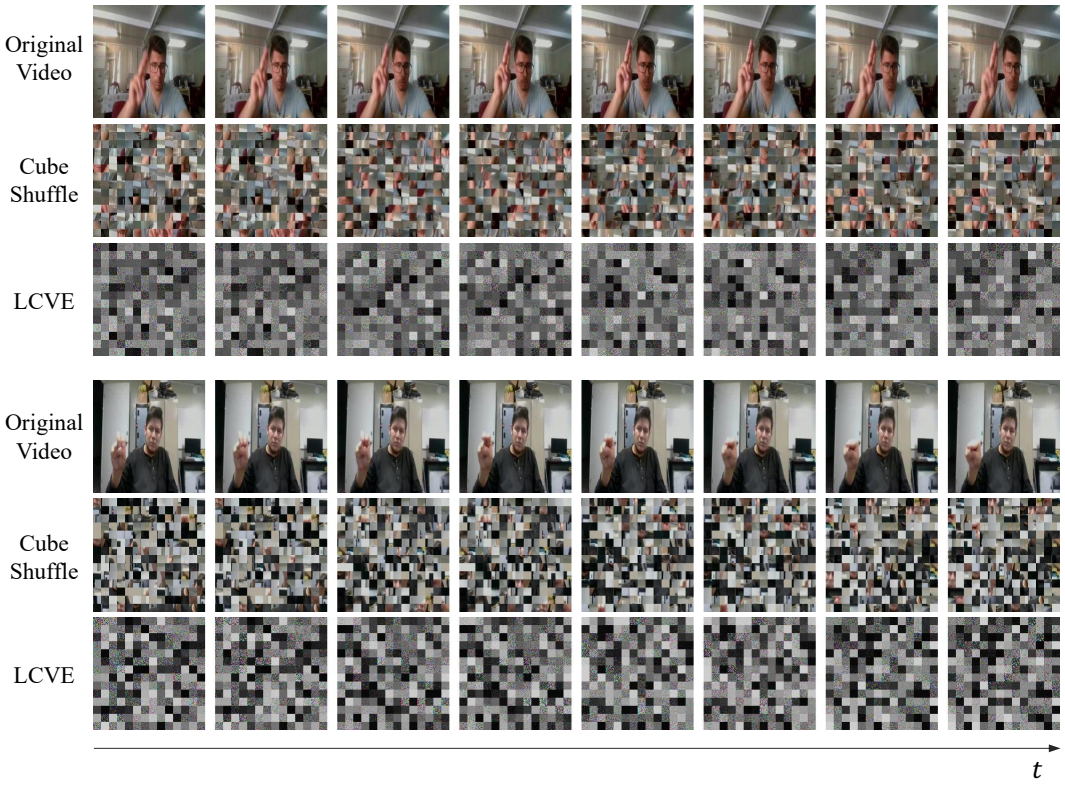

Figure 5. **Visualization samples on IPN Hand Dataset.**

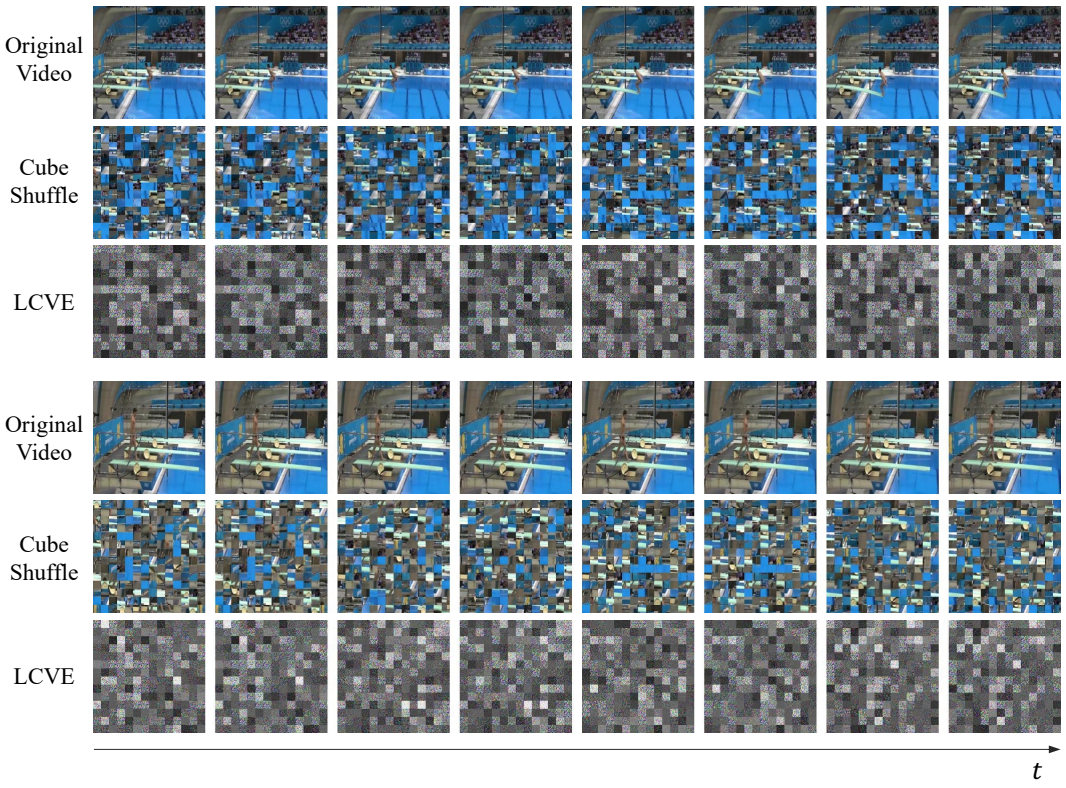

Figure 6. **Visualization samples on Diving48.**

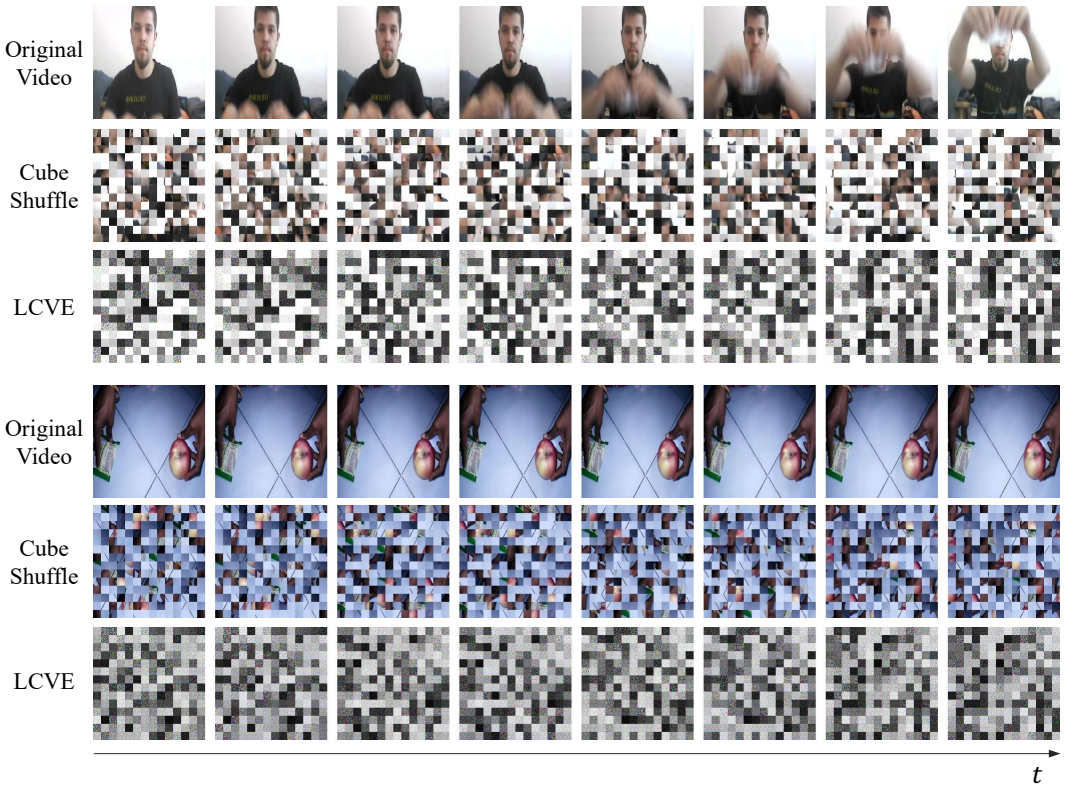

Figure 7. **Visualization samples on Something-Something V2.**

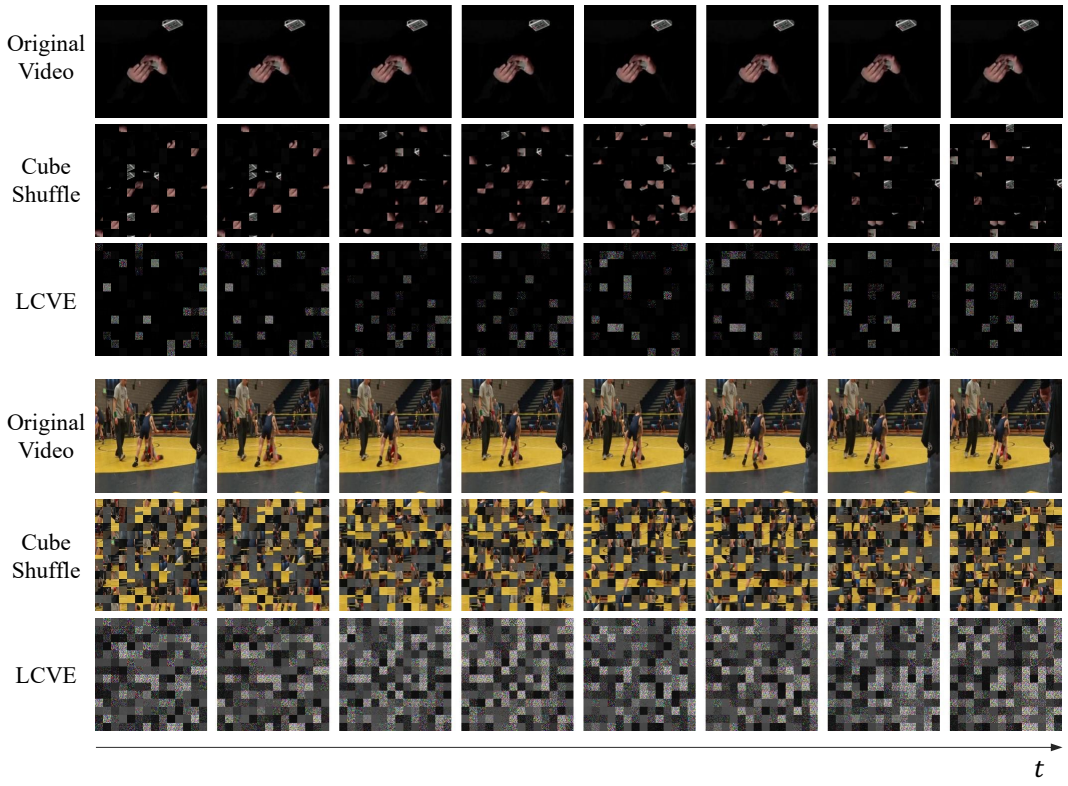

Figure 8. **Visualization samples on Kinetics400.**