RAPrivacy: a Readable Anonymizer for Privacy Preserving Action Recognition

Zi-Zhen Wang
eliawang.cs11@nycu.edu.tw
Yen-Lung Chu
egbert.cs08@nycu.edu.tw
Pei-Chun Tsai
peichun23.cs12@nycu.edu.tw
Kuan-Wen Chen
kuanwen@cs.nycu.edu.tw

Department of Computer Science
National Yang Ming Chiao Tung University
Hsinchu, Taiwan

Abstract

The growing awareness and demand for privacy protection have become increasingly prominent in the field of action recognition, which involves processing image sequences that capture human motion. Consequently, ensuring both accuracy and protecting sensitive information presents a significant challenge. Existing methods often face difficulties in balancing privacy preservation with effective action recognition, often resulting in images that are not perceptible to the human eye. To address these, we introduce RAPrivacy, a novel model that leverages adversarial learning in conjunction with style transfer to generate images that maintain privacy while remaining interpretable for human observers in the context of action recognition. For example, such images enable medical personnel to remotely monitor a patient's physical condition without revealing the patient's entire identity. Empirical evaluations on the VP-UCF101 and VP-HMDB51 datasets demonstrate the effectiveness of RAPrivacy, achieving competitive performance in action recognition while significantly enhancing privacy protection. Notably, to the best of our knowledge, this is the first study in the domain of privacy-preserving action recognition that ensures human-eye readability.

1 Introduction

Action recognition methods [14] have significantly advanced with the progress in deep learning. However, achieving high accuracy while ensuring privacy preservation in action recognition tasks remains challenging. Privacy-preserving action recognition [5, 8, 16] refers to the protection of sensitive attributes [32], such as skin color, face, gender, nudity, and relationships, among others.

Existing research in privacy-preserving action recognition [8, 21, 35] primarily addresses privacy protection at the frame level. This approach typically involves three stages: (1) extracting all frames from a video, (2) independently removing private information from each frame, and (3) reconstructing a new video by stitching the transformed frames. While this

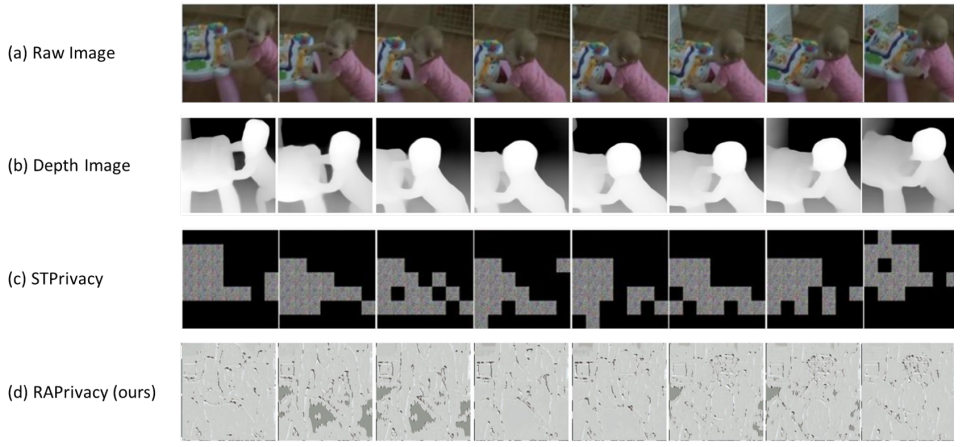


Figure 1: A visual comparison of (a) the raw image, (b) the depth image, (c) the current state-of-the-art method STPrivacy [17], and (d) our proposed RAPrivacy. For a comprehensive evaluation of RAPrivacy's ability to preserve motion clarity while ensuring privacy, we strongly encourage the reader to view the video at https://youtu.be/OSoADlsRtLw.

pipeline effectively mitigates frame-level privacy threats, it has certain limitations. First, processing each frame independently disregards the temporal dynamics between frames, resulting in significant motion discontinuities that degrade action recognition performance. Second, while this method prevents privacy leaks within individual frames, private information can still be inferred by aggregating residual information across all frames [21, 30]. Some approaches, such as STPrivacy [17], address these challenges by incorporating temporal dynamics, thus achieving a better trade-off between privacy protection and action recognition accuracy.

In this study, we also address a commonly overlooked issue in privacy-preserving action recognition methods: the poor human-eye readability of generated images. For instance, when these methods are applied in hardware implementations (such as camera lenses), the resulting images (e.g., Fig. 1(c)) lack sufficient readability, making it difficult for human observers to extract meaningful content or context from them. Although these methods effectively safeguard individuals' privacy, they render the content of the captured images inaccessible, leaving the scene captured by the camera largely unknown. More importantly, if the protected images can retain a degree of readability for human eyes, they could facilitate subsequent tasks, such as assigning new classification labels to the images, allowing smoother transitions to the next processing stage.

Several studies [1, 12] related to depth imaging suggest that depth images can inherently obscure certain privacy-sensitive information. Motivated by these studies, we hypothesize that depth images could offer some level of privacy protection while maintaining good human-eye readability, and we conduct experiments to verify this. However, our experimental results (Section 4.4) show that while depth images exhibit high readability, their privacy protection performance is comparable to raw data. Specifically, the F1 scores for depth images and raw data were 0.672 and 0.673, respectively, where a lower F1 score indicates better privacy protection.

To overcome the aforementioned limitations and mitigate the elevated risk of privacy

breaches in continuous video sequences, we propose the RAPrivacy model. This model integrates a style transfer module that processes depth images, followed by the application of adversarial learning to train an action recognition module. Additionally, the model employs contrastive loss between different video sequences to enhance privacy protection within the video content.

As demonstrated in Fig. 1, images generated by STPrivacy [17] are nearly impossible to interpret visually. In contrast, depth images clearly reveal movement contours and other key details. While the static images generated by our proposed RAPrivacy (Fig. 1(d)) are not as detailed as depth images (Fig. 1(b)), they offer better readability than the images produced by STPrivacy [17]. The videos containing static images from RAPrivacy closely match the readability of videos containing depth images, demonstrating the effectiveness of our proposed model. We highly recommend watching the video at https://youtu.be/OSoADlsRtLw to appreciate the enhanced clarity achieved by our model fully. This framework shows substantial potential for facilitating secondary development tasks in practical applications.

Our RAPrivacy model demonstrates competitive action recognition accuracy compared to state-of-the-art methods [17], while also excelling in privacy protection. Specifically, our model achieves a privacy F1-score of 0.605, compared to 0.613 for the best existing methods. Further details can be found in Table 1 and supplementary material.

In this work, we define privacy in action recognition as the suppression of visual cues that may reveal personal identity or sensitive characteristics. We focus on five attributes: face, skin color, gender, nudity, and relationships, as annotated in VP-UCF101 and VP-HMDB51. Unlike identity- or noise-based methods, our attribute-centric approach minimizes these cues while preserving motion patterns essential for both machine and human interpretation. The primary contributions of this paper are as follows:

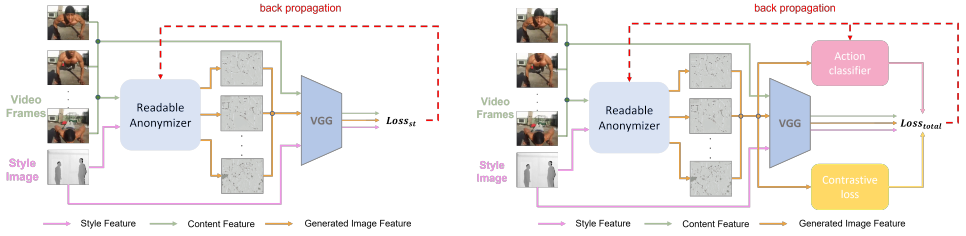
1. We propose the RAPrivacy model, which combines adversarial learning and style transfer to generate privacy-preserving, human-readable images for action recognition. To the best of our knowledge, this is the first study to produce privacy-preserving action recognition images that are readable by the human eye, offering significant practical value in action recognition.

2. RAPrivacy uniquely employs style transfer with the style of depth images to preserve human-eye readability while maintaining privacy protection. Furthermore, adversarial learning balances privacy preservation and action recognition accuracy.

3. Experimental results on the VP-UCF101 and VP-HMDB51 datasets show that RAPrivacy outperforms state-of-the-art methods in privacy protection, while maintaining competitive action recognition accuracy and ensuring the readability of the generated privacy-preserving images.

2 Related work

Ensuring privacy is critical in action recognition due to the sensitive nature of the data involved. Privacy-preserving methods can be broadly categorized into four main approaches: blurring and downsampling techniques [22, 26], cryptographic methods [13], obfuscation-based methods [35], and machine learning-based approaches [8, 10, 17]. Blurring [3, 26] and



(a) Phase1: Style Adaptation

(b) Phase2: Action Recognition and Privacy Preserving Enhancement

Figure 2: Structure of RAPrivacy. The workflow of RAPrivacy is divided into two phases (a) and (b). In each iteration, the model processes sequences of continuous images through both phases. For a complete description of the RAPrivacy architecture, refer to Phase 2, which consists of three key components: a readable anonymizer, a video-based action classification branch, and a video-based privacy protection branch.

downsampling [6, 22] methods focus primarily on reducing the exposure of sensitive information by decreasing data resolution or applying blurring techniques. Cryptographic methods, such as homomorphic encryption [13], allow for secure computations on encrypted data, thus safeguarding privacy during data processing. Within the differential privacy framework [7, 27], individual identities are protected by introducing random noise into the data.

Obfuscation-based approaches [21, 35] utilize object detection techniques to obscure or replace sensitive regions, reducing the visibility of private features. Recent advancements have seen a shift towards machine learning-based approaches [8, 17, 31], which achieve privacy protection by learning image anonymization functions or directly modifying the images. Among the most common machine learning-based techniques for privacy preservation are adversarial learning [16] and self-supervised learning. Adversarial learning [8] employs an adversarial training framework to anonymize input videos, optimizing the trade-off between recognition accuracy and privacy protection. Some studies [8, 10] explore self-supervised privacy protection methods, which do not rely on privacy labels but instead aim to minimize the cost of action classification while maximizing the cost of privacy classification. STPrivacy [17] introduces a novel Transformer-based framework for privacy-preserving action recognition, addressing privacy concerns in both the temporal and spatial dimensions, thus effectively concealing private information and safeguarding personal privacy.

Although these methods are effective, they fail to achieve an optimal trade-off between privacy protection and accuracy, and in some cases, may not fully safeguard personal privacy or maintain efficient action recognition. In this paper, we propose RAPrivacy, which not only strikes a balance between privacy protection and action recognition, but also enhances content comprehensibility. Specifically, it enables action recognition while ensuring that personal features are not identifiable, all the while maintaining human-eye readability.

3 Methodology

The primary objective of this work is to train an anonymizer (or generator) that can protect privacy while maintaining high action recognition accuracy and ensuring readability. The proposed RAPrivacy model operates in two distinct phases. The first phase, referred

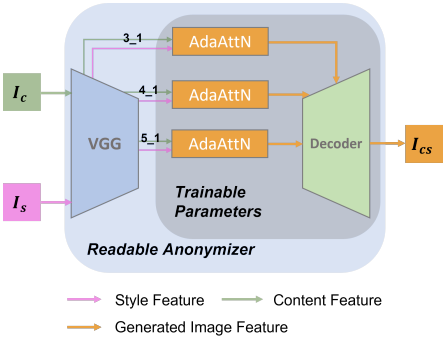


Figure 3: Architecture of the readable anonymizer (style transfer) module. In this module, I_c denotes the content image, I_s represents the style image, and I_{cs} refers to the generated image. The components within the gray area correspond to the learning components of the model, while the component within the light blue area represents the final readable anonymizer.

to as the style adaptation phase, is primarily focused on training the style transfer module. The second phase, called the action recognition and privacy-preserving enhancement phase, aims to achieve an optimal balance between action recognition accuracy and privacy protection. As illustrated in Fig. 2(b), the system comprises three main components: a readable anonymizer, a video-based action classification branch, and a video-based privacy protection branch.

The anonymizer in our model is derived from the style transfer module, which generates a new image by extracting and combining features from both style and content images. The resulting image is then fed into both the video-based action classification branch and the video-based privacy protection branch, forming an adversarial learning framework. This framework supervises the training of the RAPrivacy model as a whole, with the objective of minimizing the loss in action recognition while maximizing the loss in privacy protection. Additionally, our architecture does not rely on privacy labels, allowing it to be trained and fine-tuned on any action recognition dataset without requiring privacy-specific labels. Detailed descriptions of each component of the model are provided in the following sections.

3.1 Style transfer module

The style transfer module is based on the AdaAttN framework [18] and consists of a symmetrical encoder-decoder structure with three AdaAttN modules. The encoder is implemented using VGG-19 [24], while the decoder mirrors the structure of VGG-19 symmetrically. An illustration of the style transfer module is provided in Fig. 3.

In alignment with the strategy proposed in the AdaAttN paper [18], the style transfer module in our model adopts a multi-level approach to leverage both shallow and deep features. During each training iteration, multiple frames from the video sequences are used as content images, while a depth image, randomly selected from a large collection of depth images, serves as the style image. These images are input into VGG-19 to extract multi-level content and style features. Three AdaAttN modules are integrated at the ReLU-3_1, ReLU-4_1, and ReLU-5_1 layers of VGG. We denote the features extracted from the ReLU-x layer in VGG as $F_*^x \in R^{C\times H_*\times W_*}$, where * represents either c (content) or s (style). After extracting features at multiple layers, the features of the current layer are concatenated with the downsampled features from its preceding layers. The concatenated downsampled features at the layer l are represented as follows:

$$F_*^{1:x} = D_x(F_*^1) \oplus D_x(F_*^2) \oplus \cdots \oplus (F_*^x). \tag{1}$$

where D_x denotes the downsampling of the input features to match the shape of F_*^x, and $\oplus$ represents concatenation along the channel dimension.

After passing through the AdaAttN module, the features from different layers are used to generate attention maps by utilizing both shallow and deep content and style features. These attention maps are then employed to compute the weighted mean and variance of the style features, while the content features are normalized after computing their weighted mean and variance. The attention map is subsequently used to merge the normalized content features with the style features, ensuring consistency both locally and globally. The embedded feature representation for the layers is obtained as follows:

$$F_{cs}^x = AdaAttN(F_c^x, F_s^x, F_c^{1:x}, F_s^{1:x}). \tag{2}$$

where F_c, F_s, and F_{cs} represent the content, style, and embedded features, respectively. Finally, the multi-layer embedded features are processed by the decoder to generate stylized images I_{cs}, expressed as follows:

$$I_{cs} = Dec(F_{cs}^3, F_{cs}^4, F_{cs}^5). \tag{3}$$

, resulting in a set of new style-transformed images corresponding to the input video frames. These generated images are then fed into the action recognition branch for further processing.

3.2 Video-based action recognition branch

We utilize ResNet3D [4] as the core backbone network to facilitate action classification. The learning process is optimized using cross-entropy loss. The primary objective is to improve the classifier's accuracy, which necessitates a strategic reduction in the loss function. This focus ensures that the model achieves high precision in recognizing and categorizing diverse actions, aligning closely with the performance metrics that are critical to the goals of this research.

3.3 Video-based privacy protection branch

In the video-based privacy protection branch, pairs of distinct image sequences are utilized as positive and negative examples to facilitate the learning process. A ResNet architecture is employed to extract feature representations from these sequences. The core of our approach involves calculating the contrastive loss [11] between the feature sets, which serves as the key metric for the loss function. Given the sensitive nature of the privacy attributes inherent in the images, the objective is to maximize the ambiguity between identifiable features, thereby enhancing privacy protection. This requires maximizing the contrastive loss to effectively obscure distinguishable attributes between the compared sequences.

3.4 Loss function and adversarial training procedure

The style transfer loss is computed by incorporating both global style loss and local feature loss, in alignment with the AdaAttN [18] framework. The global style loss quantifies the discrepancy between the generated stylized image and the style image across multiple network layers by measuring the variance and mean of the VGG feature representations. This loss is

determined by summing the squared differences of these means and variances, as expressed below:

$$L_{gs} = \sum_{x=2}^{5} (\|\mu(E^x(I_{cs})) - \mu(F_s^x)\|_2 + \|\sigma(E^x(I_{cs})) - \sigma(F_s^x)\|_2). \tag{4}$$

where $E()$ represents the features extracted by the VGG encoder, and ^{x} denotes the layer index.

The local feature loss measures the consistency between the local features of the generated stylized image and the model predictions. Increased consistency leads to enhanced local visual quality. This loss is computed as the sum of the squared differences between the encoder features of the stylized image and the transformation features generated by the AdaAttN module. The local feature loss is expressed as follows:

$$L_{lf} = \sum_{x=3}^{5} \left\| E^x(I_{cs}) - AdaAttN^*(F_c^x, F_s^x, F_c^{1:x}, F_s^{1:x}) \right\|_2. \tag{5}$$

Cross-entropy [23] is utilized as the loss function for action classification, where y represents the true labels and y' denotes the predicted outcomes:

$$L_{act} = CrossEntropy(y, y'). \tag{6}$$

For the privacy protection branch, contrastive loss is employed as the loss function. This loss encourages the model to produce similar representations for videos that share the same action label but differ in privacy-sensitive attributes. Specifically, positive pairs consist of videos depicting the same action performed by different subjects, while negative pairs are drawn from videos with different actions. We follow the standard InfoNCE formulation, computing the loss over each positive and negative pair and averaging the results to obtain the final contrastive objective. The contrastive loss is expressed as:

$$L_{pri} = \frac{1}{N} \sum_{i=j,i=1}^{N} (D(x_i, x_j)^2 + max(0, m - D(x_i, x_j))^2). \tag{7}$$

where N is the number of frames sampled from a single video, m is a margin parameter that defines the minimum distance between dissimilar samples, and D represents the Euclidean distance.

As illustrated in Fig. 2, each iteration of the proposed RAPrivacy model involves two phases. In the first phase (Fig. 2(a)), $Loss_{st}$ considers only the loss associated with the image style transformation, with the action classifier frozen and not updated during this phase. In the second phase (Fig. 2(b)), $Loss_{total}$ incorporates the style transfer loss, cross-entropy loss for action classification, and contrastive loss for privacy protection. Additionally, the action classifier is updated during this phase using L_{act}. The terms $Loss_{st}$ and $Loss_{total}$ are defined as follows:

$$Loss_{st} = \lambda_g L_{gs} + \lambda_l L_{lf}. \tag{8}$$

$$Loss_{total} = \lambda_g L_{gs} + \lambda_l L_{lf} + \lambda_a L_{act} - \lambda_p L_{pri}. \tag{9}$$

This procedure is grounded in the principle of human-eye readability, which we emphasized at the outset. As such, it ensures that the generated images are stabilized to guarantee that they remain easily recognizable by the human eye.

4 Experiments

4.1 Datasets

In this section, we conducted experiments on two datasets that are labeled with both action and privacy-related attributes, namely the VP-UCF101 and VP-HMDB51 datasets, in accordance with the state-of-the-art (SOTA) method [17].

The VP-UCF101 dataset is an extension of the UCF101 [25] dataset, which includes privacy attribute annotations. It consists of 13,320 videos categorized into 101 types of sports actions. Similarly, the VP-HMDB51 dataset is an extension of the HMDB51 [15] dataset, incorporating privacy attribute annotations. This dataset contains 6,849 videos, categorized into 51 types of human actions. Each video in both datasets is annotated with both action and privacy labels [17]. The action labels identify the activity being performed by individuals in the video, while the privacy labels include five distinct attributes: face, skin color, gender, nudity, and familial relationship. These privacy labels are binary, denoted as 0 or 1, indicating whether identification is permissible in the video. All experiments were conducted using the official training and testing splits provided by the datasets.

Additionally, for the training process, we required a series of depth images to serve as style images. These images were obtained from the SBU [33] dataset, which is specifically designed for action recognition and offers a comprehensive collection of depth images. The depth images (Fig. 2(c) in the supplementary material) in the SBU dataset capture human movements and postures in detail, making them suitable for effectively training our RAPrivacy model to recognize actions while ensuring privacy preservation.

4.2 Evaluation metrics

The performance of action recognition is evaluated based on the Top-1 accuracy, with the classification results determined by averaging the predictions from 5 clips and 3 crops [28]. A higher Top-1 accuracy value indicates better action recognition and superior performance. Each privacy label consists of five binary attributes, thereby framing privacy recognition as a multilabel classification problem.

The privacy protection performance was assessed using class-wise mean average precision (cMAP) and F1 score, where lower values of these metrics signify better privacy protection.

4.3 Implement details

During the training process, a 3D ResNet-50 is utilized as the video-based action classifier. For evaluation, in order to ensure fairness, we adopt the same settings and parameters as those used in the previous method. Both the action recognition classifier and the privacy recognition classifier are based on the ViT-S architecture [8, 20, 29]. Optimization is performed using the AdamW optimizer [19], with a weight decay of 0.05. The video and tablet sizes are set to $16\times112\times112\times3$ and $2\times16\times16\times3$, respectively.

The frame sampling rate for the VP-UCF101 dataset is set to 4, while for the VP-HMDB51 dataset, it is set to 2. The learning rate is linearly scaled by 0.001 times the batch size divided by 512, and reduced via cosine annealing. All experiments are conducted on 3 NVIDIA RTX A5000 GPUs.

Method	Privacy Learning	Readability	VP-HMDB51			VP-UCF101		
			Top-1 (↑)	F1 (↓)	cMAP (↓)	Top-1 (↑)	F1 (↓)	cMAP (↓)
Raw Data		✓	51.44	0.673	75.58	84.20	0.684	76.62
Depth Image		✓	51.49	0.672	75.48	83.57	0.681	76.52
Downsample-2×		✓	40.80	0.601	71.35	72.79	0.620	71.49
Downsample-4×		✓	31.32	0.594	69.79	56.07	0.615	69.85
Blackening [8]		✓	38.27	0.649	74.06	69.41	0.660	75.37
StrongBlur [8]		✓	40.91	0.655	74.33	73.94	0.672	75.58
WeakBlur [8]		✓	47.24	0.663	75.11	77.31	0.678	76.03
Collective [34]	✓		46.88	0.651	74.12	78.01	0.663	75.22
VITA [32]	✓		48.11	0.638	73.89	78.49	0.657	75.36
SPAct [8]	✓		48.56	0.642	73.78	78.40	0.651	75.29
STPrivacy [17]	✓		**50.73**	<u>0.613</u>	<u>72.48</u>	**82.55**	<u>0.634</u>	<u>73.79</u>
RAPrivacy (ours)	✓	✓	<u>50.12</u>	**0.605**	**70.12**	<u>78.7</u>	**0.626**	**70.50**

Table 1: Comparison of known actions. The best results of learning-based methods are highlighted in bold, while the second-best results are underlined.

4.4 Comparison analysis

Table 1 presents the experimental results of our approach. As shown, while the action recognition performance of our method slightly lags behind the current SOTA [17], it still demonstrates superior accuracy compared to other learning-based methods. Regarding privacy protection, although the method employing 4× downsampling achieves the best privacy performance, it significantly sacrifices action recognition accuracy. Our approach, in contrast, outperforms all other methods in terms of privacy protection. A notable advantage of our method is its human-eye readability, a characteristic that is absent in other existing approaches.

As previously mentioned, the impact of depth images on privacy protection has not been clearly established, motivating us to conduct experiments to evaluate their efficacy in this regard. In these experiments, all original images were converted to depth images using ZoeDepth [2], and subsequently subjected to the same video-based action recognition and privacy evaluation as the original images. Table 1 shows that, in terms of video action recognition, the performance of depth images is nearly identical to that of raw images. However, the privacy protection achieved with depth images did not significantly differ from that of raw data. This outcome can be attributed to the fact that depth images retain contours, trajectories, and detailed information on aspects such as interactions and gender, which undermines the effectiveness of privacy protection.

4.5 Comprehensive evaluation and analysis

To assess the generalizability of our trained anonymizer, we conducted a transfer evaluation. Specifically, the notation VP-UCF101→VP-HMDB51 indicates that the anonymizer was trained on the VP-UCF101 dataset, followed by validation of action classification and privacy detection on the VP-HMDB51 dataset.

RAPrivacy shows strong cross-dataset generalizability, surpassing prior SOTA (STPrivacy) in both recognition accuracy and privacy protection. A user study involving 23 par-

ticipants was conducted to evaluate the human-eye readability of anonymized videos generated by RAPrivacy. Participants achieved a recognition accuracy of 98.27% without access to the original content, confirming the model's ability to preserve semantic clarity while ensuring identity protection. Compared to prior approaches, RAPrivacy demonstrated superior visual interpretability, as further evidenced by the video at https://youtu.be/OSoADlsRtLw.

On the other hand, ablation studies highlight the depth style as the optimal choice for balancing interpretability and privacy. Despite minimal accuracy trade-offs, the model delivers coherent anonymized videos well-suited for downstream tasks. Due to space constraints, detailed experimental results such as a human-eye readability user study, a privacy preservation ablation analysis, and visual comparisons with SPAct and STPrivacy, are provided in the supplementary material. We strongly encourage readers to refer to these additional results, as they directly address key concerns regarding interpretability, comparative performance, and the effectiveness of our contrastive loss design.

5 Conclusion

In this study, we propose RAPrivacy, an end-to-end generative model that innovatively integrates adversarial learning and style transfer to generate images. This model not only preserves high action recognition accuracy but also enhances privacy protection while ensuring human-eye readability. As such, it overcomes the limitations of previous techniques that compromise image clarity to achieve better privacy protection.

The experimental results demonstrate the effectiveness of RAPrivacy across various datasets. The model achieves superior privacy protection performance while maintaining competitive action recognition accuracy. Consequently, RAPrivacy represents a significant advancement in the field of privacy-preserving action recognition, establishing a new benchmark for future developments. The findings of this study offer valuable insights for future research on privacy-preserving, high-accuracy action recognition systems.

Acknowledgement

This work was supported in part by the National Science and Technology Council, Taiwan (113-2221-E-A49-164-MY3, 114-2221-E-A49 -054 -MY3 and 113-2634-F-A49-007-).

References

[1] J Andrew Onesimu and J Karthikeyan. An efficient privacy-preserving deep learning scheme for medical image analysis. *Journal of Information Technology Management*, 12(Special Issue: The Importance of Human Computer Interaction: Challenges, Methods and Applications.):50–67, 2020.

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. URL https://arxiv.org/abs/2302.12288.

[3] Daniel J Butler, Justin Huang, Franziska Roesner, and Maya Cakmak. The privacy-utility tradeoff for remotely teleoperated robots. In *In proceedings of the IEEE International Conference on human-robot interaction*, pages 27–34, 2015.

[4] Chun-Fu Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2021.

[5] Sheng-Yang Chiu, Yu-Ting Huang, Chieh-Ting Lin, Yu-Chee Tseng, Jen-Jee Chen, Meng-Hsuan Tu, Bo-Chen Tung, and YuJou Nieh. Privacy-preserving video conferencing via thermal-generative images. In *In proceedings of the IEEE International Conference on Robotics and Automation*, pages 9478–9485. IEEE, 2023.

[6] Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, and Li Fei-Fei. Privacy-preserving action recognition for smart hospitals using low-resolution depth images. *arXiv preprint arXiv:1811.09950*, 2018.

[7] Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, et al. Advancing differential privacy: Where we are now and future directions for real-world deployment. *arXiv preprint arXiv:2304.06929*, 2023.

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

[9] 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. In *In proceedings of the IEEE Conference on International Conference on Learning Presentations*, 2020.

[10] Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In *In proceedings of the IEEE Conference on International Conference on Computer Vision*, pages 13598–13609, 2023.

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742. IEEE, 2006.

[12] Ankit Jain, Rajendra Akerkar, and Abhishek Srivastava. Privacy-preserving human activity recognition system for assisted living environments. *IEEE Transactions on Artificial Intelligence*, 2023.

[13] Miran Kim, Xiaoqian Jiang, Kristin Lauter, Elkhan Ismayilzada, and Shayan Shams. Hear: Human action recognition via neural networks on homomorphically encrypted data. *arXiv e-prints*, pages arXiv–2104, 2021.

[14] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *In proceedings of the IEEE Conference on International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[16] Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *In proceedings of the Conference on European Conference on Computer Vision*, pages 518–534. Springer, 2022.

[17] Ming Li, Xiangyu Xu, Hehe Fan, Pan Zhou, Jun Liu, Jia-Wei Liu, Jiahe Li, Jussi Keppo, Mike Zheng Shou, and Shuicheng Yan. Stprivacy: Spatio-temporal privacy-preserving action recognition. In *In proceedings of the IEEE Conference on International Conference on Computer Vision*, pages 5106–5115, 2023.

[18] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *In proceedings of the IEEE Conference on International Conference on Computer Vision*, pages 6649–6658, 2021.

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

[20] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[21] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *In proceedings of the Conference on European Conference on Computer Vision*, pages 620–636, 2018.

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

[23] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[25] 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.

[26] Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy. Human pose estimation on privacy-preserving low-resolution depth images. In *In proceedings of the International conference on medical image computing and computer-assisted intervention*, pages 583–591. Springer, 2019.

[27] Rishabh Subramanian. Have the cake and eat it too: Differential privacy enables privacy and precise analytics. *Journal of Big Data*, 10(1):117, 2023.

[28] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[29] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *In proceedings of the IEEE International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[30] 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.

[31] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *In proceedings of the conference on European Conference on Computer Vision*, pages 606–624, 2018.

[32] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2126–2139, 2020.

[33] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *In proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012.

[34] Dalin Zhang, Lina Yao, Kaixuan Chen, Guodong Long, and Sen Wang. Collective protection: Preventing sensitive inferences via integrative transformation. In *In proceedings of the IEEE International Conference on Data Mining*, pages 1498–1503. IEEE, 2019.

[35] Zhixiang Zhang, Thomas Cilloni, Charles Walter, and Charles Fleming. Multi-scale, class-generic, privacy-preserving video. *Electronics*, 10(10):1172, 2021.