Supplementary Material: TeD-SPAD: Temporal Distinctiveness for Self-supervised Privacy-preservation for video Anomaly Detection

Joseph Fioresi, Ishan Rajendrakumar Dave, Mubarak Shah
Center for Research in Computer Vision, University of Central Florida, Orlando, USA
{joseph.fioresi, ishanrajendrakumar.dave}@ucf.edu, shah@crcv.ucf.edu
Project Page: https://joefioresi718.github.io/TeD-SPAD_webpage/

A. Supplementary Overview

Section B: Dataset details
Section C: Implementation details
Section D: Additional results

B. Dataset Details

UCF-Crime [15] contains 1,900 (950 normal, 950 anomalous) videos with 13 different crime-based anomalies, for a total of 128 hours. The labels are included at the video level, indicating whether or not the video contains at least one anomalous event. The footage comes from real-life CCTV surveillance cameras in a variety of scenes. The average video contains 7,247 frames, which is ≈3 minutes at 30fps. The training set has a total of 800 normal videos and 810 anomalous videos, and the testing set has 150 normal and 140 anomalous videos. Both sets contain examples of all anomaly categories, with some videos having multiple anomalies.

XD-Violence [16] contains 4,754 (2405 normal, 2349 anomalous) videos with 6 different anomaly categories, total 217 hours of untrimmed footage, making it the largest weakly supervised video anomaly detection dataset. The labels are also at the video level, except they allow for each video to have more than one anomaly label. The videos also contain audio signals to allow for multi-modal anomaly detection. The videos are gathered from various types of cameras, movies, and games, resulting in a unique blend of scenes for increased difficulty. The training set contains 3,954 videos while the test set has 800 videos total, 500 anomalous and 300 normal.

ShanghaiTech [10] contains 437 videos in 13 different scenes with a total of 130 anomalous events. The training set includes 330 videos while the test set includes 107. Out of a total of ≈317,400 frames in the dataset, 17,900 are anomalous. Each anomaly also contains a pixel-level location for anomaly localization. It was published as an unsupervised anomaly detection dataset, but Zhong *et al.* [18] proposed a weakly supervised rearrangement, which is used in this work.

VISPR1 [17, 4]	
Label	**Description**
a17_color	skin color
a4_gender	gender
a9_face_complete	full face visible
a10_face_partial	part of face visible
a12_semi_nudity	partial nudity
a64_rel_personal	shows personal relationship
a65_rel_soci	shows social relationship

Table 1: Privacy attributes from subset of VISPR [11] labels as used in previous works.

VISPR [11] is a visual privacy image dataset containing 22k public Flickr images labelled with 68 different private attributes. Private attributes are determined by personally identifiable information as considered in the US Privacy Act of 1974 and the EU Data Protection Directive 95/46/EC [1]. The training and testings sets contain 10,000 and 8,000 images, respectively. For ease of comparison, we use the same VISPR attribute split used in [17, 4], seen in Table 1.

UCF101 [14] contains 13,320 videos in 101 different human action categories. In the default setting, split-1 is used. Each video shows the action directly with no filler, so the average video length is 7.21s.

Kinetics400 [7] is used as the standard video dataset for action classifier pretraining. The dataset contains a total of 306,245 videos, with over 400 examples of each of the 400 human action classes.

Method	VISPR Privacy cMAP(%)(↓)	UCF-Crime Anomaly AUC(%)(↑)	XD-Violence Anomaly AP(%)(↑)	ShanghaiTech Anomaly AUC(%)(↑)
Raw data	62.30	77.68	73.72	90.63
Downsample-2x	55.64 ↓10.69%	76.09 ↓2.05%	62.11 ↓15.75%	84.65 ↓6.60%
Downsample-4x	52.84 ↓15.18%	68.12 ↓12.31%	59.36 ↓19.48%	82.96 ↓8.46%
Obf-Blurring	58.68 ↓5.81%	75.69 ↓2.56%	59.36 ↓23.81%	89.63 ↓1.10%
Obf-Blackening	56.36 ↓9.53%	73.91 ↓4.85%	56.17 ↓26.74%	88.72 ↓2.11%
SPAct [4]	52.71 ↓15.39%	73.93 ↓4.83%	53.36 ↓27.62%	87.72 ↓3.21%
Ours	**42.21 ↓32.25%**	**74.81 ↓3.69%**	**60.32 ↓18.18%**	**90.59 ↓0.04%**

Table 2: Comparison with different privacy-preservation methods on UCF-Crime, XD-Violence and ShanghaiTech anomaly detection. Bold indicates the best trade-off results. Trade-off plots are shown in main paper Fig. 3. Downward arrows ↓ and ↓ show the relative percent change compared to the raw data.

C. Implementation Details

All code is implemented using the PyTorch [12] library.

C.1. Feature-level Privacy Leakage Tester

To test privacy leakage at the feature-level (main paper Sec. 4.6), we create a simple fully connected model f_P consisting of 5 layers: Linear(2048, 2048) → Linear(2048, 1028) → Linear(1028, 1028) → Linear(1028, 512) → Linear(512, 7). This model is trained for 50 epochs with a cross-entropy with logits loss and Adam [8] optimizer at a learning rate of 1e-4. Images are augmented similar to the test set images, then stacked 16 times to resemble a video for feature extraction input. The set of 2048 dimensional features $\mathbb{F}_{anomaly}$ from the I3D f_T model is directly input to this privacy leakage training model.

C.2. Anonymization Process

C.2.1 Input Augmentations

We utilize standard augmentations following [4]. During training, we utilize random cropping, scaling, color jittering, erasing, and horizontal flipping. During inference, we utilize center crop with a scale of 0.8.

C.3. MGFN

We use the official MGFN [3] implementation[1] for anomaly detection evaluation. Besides using only single crop features instead of ten-crop, we use their exact hyperparameters. The residual feature norm for each segment is appended with a weight of 0.1. To help mitigate potential noise, the top-k clips are considered in the loss instead of top-1, with $k = 3$. The feature dropout rate in training is 0.7. The optimizer employed is Adam [8], starting with a learning rate of 0.001 with a weight decay of 0.0005, trained for up to 1000 epochs with a batch size of 16.

[1] https://github.com/carolchenyx/MGFN

For reference, the compound MGFN loss function is:

$$L_{AD} = L_{sce} + \lambda_1 L_{ts} + \lambda_2 L_{sp} + \lambda_3 L_{mc}, \quad (1)$$

where $\lambda_1 = \lambda_2 = 1$, and $\lambda_3 = 0.001$.

The base loss starts with standard sigmoid cross entropy loss:

$$L_{sce} = -ylog(s^{i,j}) - (1-y)log(1-s^{i,j}), \quad (2)$$

where y is video-level label ($y = 1$ is anomaly, $y = 0$ is normal), $s^{i,j}$ is the computed anomaly score for frames i in segment j.

Sultani et al. [15] proposed the use of a temporal smoothness $L_{ts} = \sum_i^{(n-1)} (f(V_a^i) - f(V_a^{i+1}))^2$ and a sparsity term $L_{sp} = \sum_i^n f(V_a^i)$, where $f(V_a^i)$ is the extracted features for segment i of anomalous video V_a. These encourage infrequent anomaly detections and smoothness between representations of sequential video segments.

MGFN also includes a feature amplification mechanism paired with a magnitude contrastive (MC) loss (Eq. 3) to better enhance feature separability both within videos and between videos. The MC loss is formulated as follows:

$$L_{mc} = \sum_{p,q=0}^{B/2} (1-l)(D(M_n^p, M_n^q)) + \sum_{u,v=B/2}^{B} (1-l)(D(M_a^u, M_a^v)) + \sum_{p=0}^{B/2} \sum_{u=B/2}^{B} l(Margin - D(M_n^p, M_a^u)), \quad (3)$$

where B is the batch size, M is the feature magnitude of the corresponding segment, $D(\cdot,\cdot)$ is a distance function, and l is an indicator function. For more details about this loss, refer to [3].

C.4. Privacy Evaluation

To evaluate the privacy leakage of each anonymizer f_A, we train a ResNet50 [6] model f_B in a supervised man-

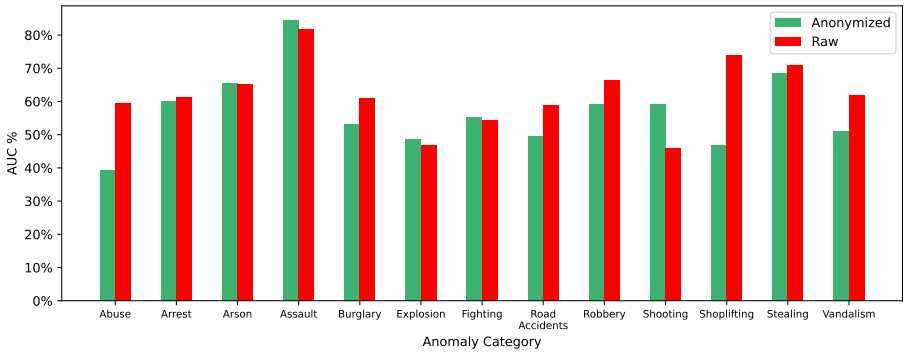

Figure 1: UCF-Crime classwise AUC performance comparison between raw and anonymized videos.

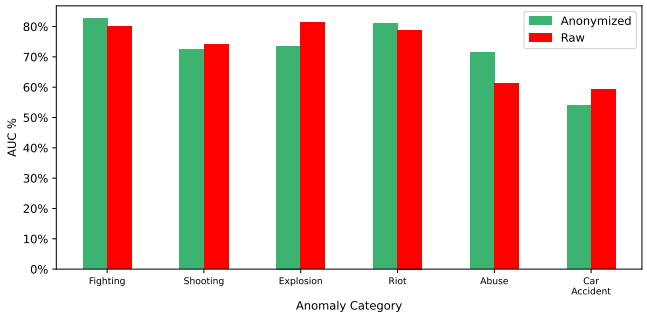

Figure 2: XD-Violence classwise AUC performance comparison between raw and anonymized videos.

ner to predict whether every input VISPR image contains each of the 7 private attributes from the split shown in Table 1. Training lasts for up to 100 epochs, stopping early if the learning rate drops to 1e-12. Learning rate starts at 1e-3, dropping to 1/5 of its current value on a training epoch where the loss does not decrease. Given an image $\mathbf{I^i} \in \mathbb{D}_{privacy}$, our baseline evaluates on $f_B(\mathbf{I^i})$, with subsequent experiments passing each image before evaluation, $f_B(f_A(\mathbf{I^i}))$.

D. Results

D.1. Quantitative Results

Table 2 compares different privacy-preserving methods and their effect on downstream anomaly detection performance. Notably, our utility loss modification allows our anonymizer to remove more privacy and improve utility performance when compared to previous methods. Compared to prior best method [4], our method is able to remove 19.9% more privacy with a slightly better utility score (1.19%).

We present class-wise performance for the anomaly detection in Fig. 1 and 2. We also show frame-level prediction scores for the anomaly detection task in Fig. 3.

Effect of temporal invariance during anonymization training: Temporal invariance objective is conceptually opposite to temporal distinctiveness objective. With invariance, the learned representations are encouraged to be similar across the temporal dimension. Temporal invariance is implemented using the formulation from [13]. Let $\mathbf{x}_{t1}^{(i)}$ and $\mathbf{x}_{t'}^{(i)}$ be the two randomly sampled clips of a video instance $X^{(i)}$. Passing such clips through utility model f_T and a non-linear projection head , we get their representations $\mathbf{z}_{t'}^{(i)}$ and $\mathbf{z}_{t'}^{(i)}$. Now the goal of the temporal invariance is to increase the mutual agreement between these two representations while maximizing the disagreement between the representation of clips of other video instances j, where $j \neq i$. This can be expressed as following equation:

$$\mathcal{L}_I = -\sum_{i=1}^{B} \log \frac{\mathrm{h}\left(\mathbf{z}_t^{(i)}, \mathbf{z}_{t'}^{(i)}\right)}{\sum_{j=1}^{B} [\mathbb{1}_{[j\neq i]} \mathrm{h}(\mathbf{z}_t^{(i)}, \mathbf{z}_t^{(j)}) + \mathrm{h}(\mathbf{z}_t^{(i)}, \mathbf{z}_{t'}^{(j)})]}, \tag{4}$$

where $\mathrm{h}(\mathbf{u_1}, \mathbf{u_2}) = \exp\left(\mathbf{u_1}^T \mathbf{u_2} / (\|\mathbf{u_1}\| \|\mathbf{u_2}\| \tau)\right)$ is used to compute the similarity between $\mathbf{u_1}$ and $\mathbf{u_2}$ vectors with an adjustable temperature parameter $\tau = 0.1$, B is batchsize. $\mathbb{1}_{[j\neq i]} \in \{0, 1\}$ is an indicator function which equals 1 iff $j \neq i$.

We perform experiments by modifying our utility loss to $\mathcal{L}_T = \mathcal{L}_{CE} + \omega * \mathcal{L}_I$, where ω is a loss weight.

In order to ensure that our invariance baseline is strong enough we perform several experiments varying different ω in Table 3. This demonstrates that temporal invariance

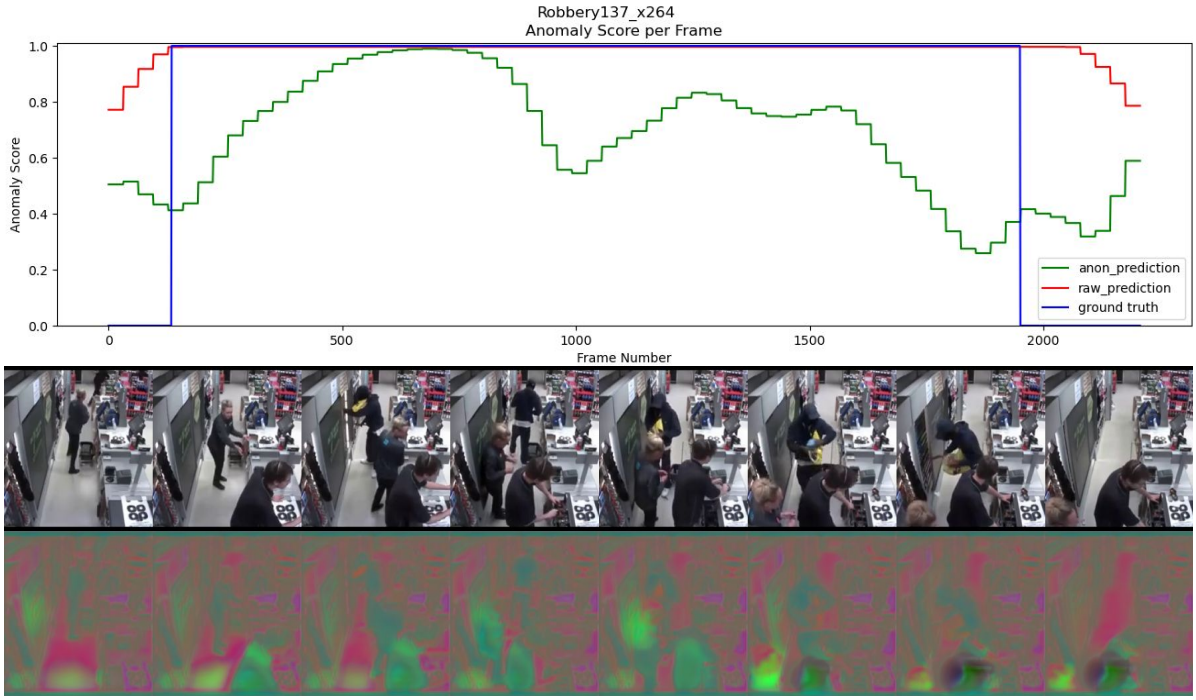

Figure 3: Frame-level anomaly score plot for video `Robbery137_x264.mp4` from UCF-Crime. Green line shows our anonymized model, red line is the raw input model, both compared to the blue ground truth line. The below visualizations shows uniformly sampled frames from the video.

is not well-aligned with the anomaly detection utility task. For insights, look to main paper Sec. 4.

Temporal Invariance Loss Weight ω	**VISPR** Privacy cMAP(%)(↓)	**UCF-Crime** Anomaly AUC(%)(↑)
0	52.71	73.93
0.1	51.62	69.35
0.5	46.51	65.84
1.0	**45.64**	**69.52**
2.0	52.2	64.4

Table 3: Comparison of using different loss weights of the temporal invariance contrastive loss during the anonymization process. Bold indicates best trade-off.

Effectiveness of different f_T architectures: For all experiments in the main paper, we follow previous works and use I3D [2]. Table 4 shows experiments with different f_T architectures to ensure that our anonymization function is suitable for varying architectures. Since the downstream anomaly detection task relies on input features, it is important to note that our I3D implementation outputs features of dimensionality 2048, while MViTv2 [9] and R3D-18 [5] output 768 and 512, respectively. These experiments used the same hyperparameters as our best I3D experiment, the models may achieve a better trade-off with hyperparameter tuning.

f_T Model Architecture λ	**VISPR** Privacy cMAP(%)(↓)	**UCF-Crime** Anomaly AUC(%)(↑)
I3D	42.21	74.81
MViTv2	24.21	69.22
R3D-18	33.58	70.67

Table 4: Comparison of different f_T architectures for both the proxy utility task and feature extraction.

D.2. Qualitative Results

We present qualitative results of our anonymization function in Fig. 4 and 5. More visualization can be found in the attached videos of the supplementary material.

D.3. Training Progression

We show outputs of our anonymization framework at different epochs of anonymization training in Fig. 6 and 7. We can clearly observe that as the training progresses, our framework is able to anonymize better.

Figure 4: Visualization of anomalous clip (shooting) from XD-Violence dataset video `Fast.Furious.6.2013__#00-45-40_00-47-13_label_B2-0-0.mp4`.

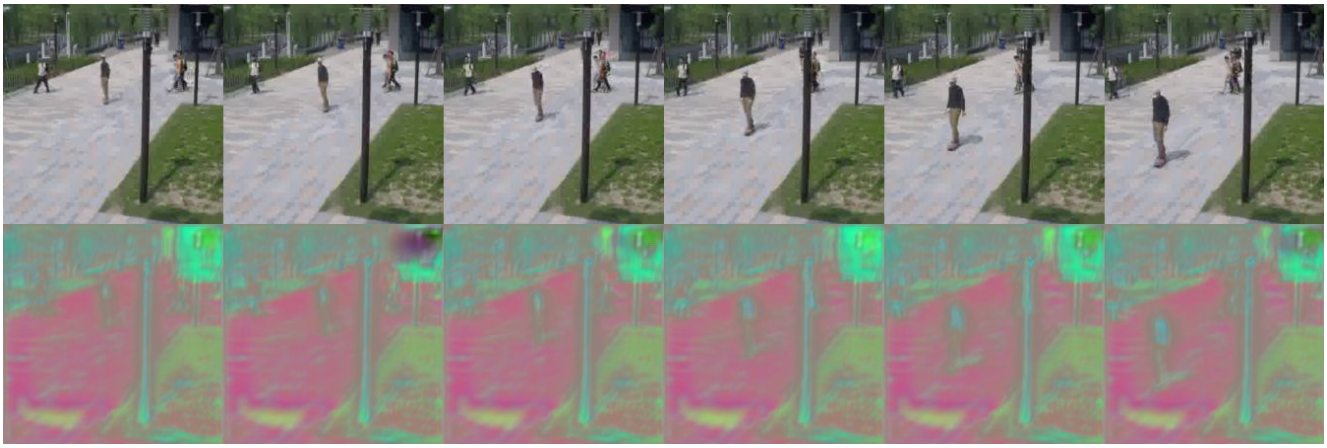

Figure 5: Visualization of anomalous clip (skateboard passing) from ShanghaiTech dataset video `08_0178.avi`.

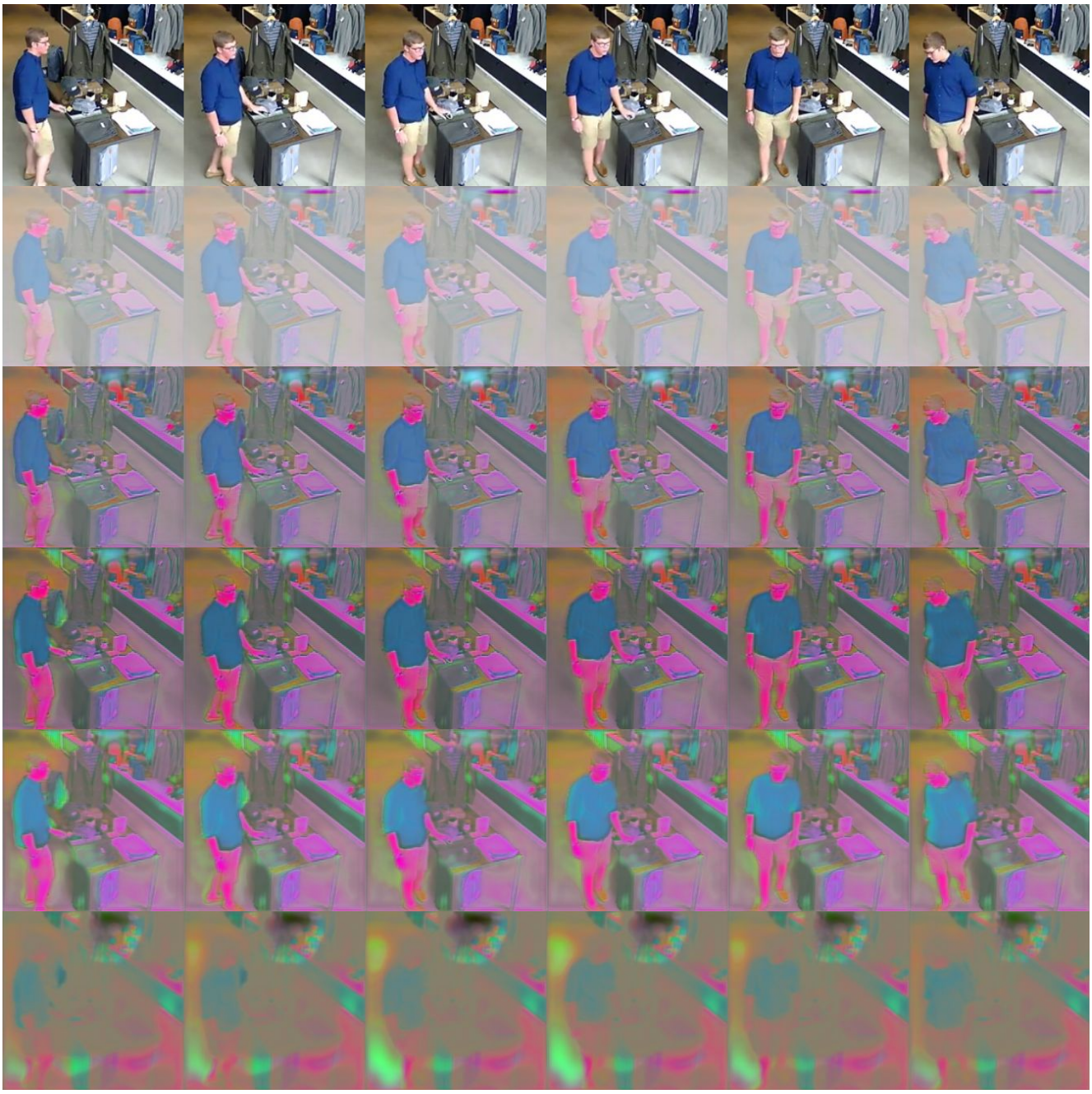

Figure 6: Training progression per epoch of the anonymization process. In order from top to bottom, visualization after f_A on epoch 1, 6, 9, 12, 15, and 20 is shown.

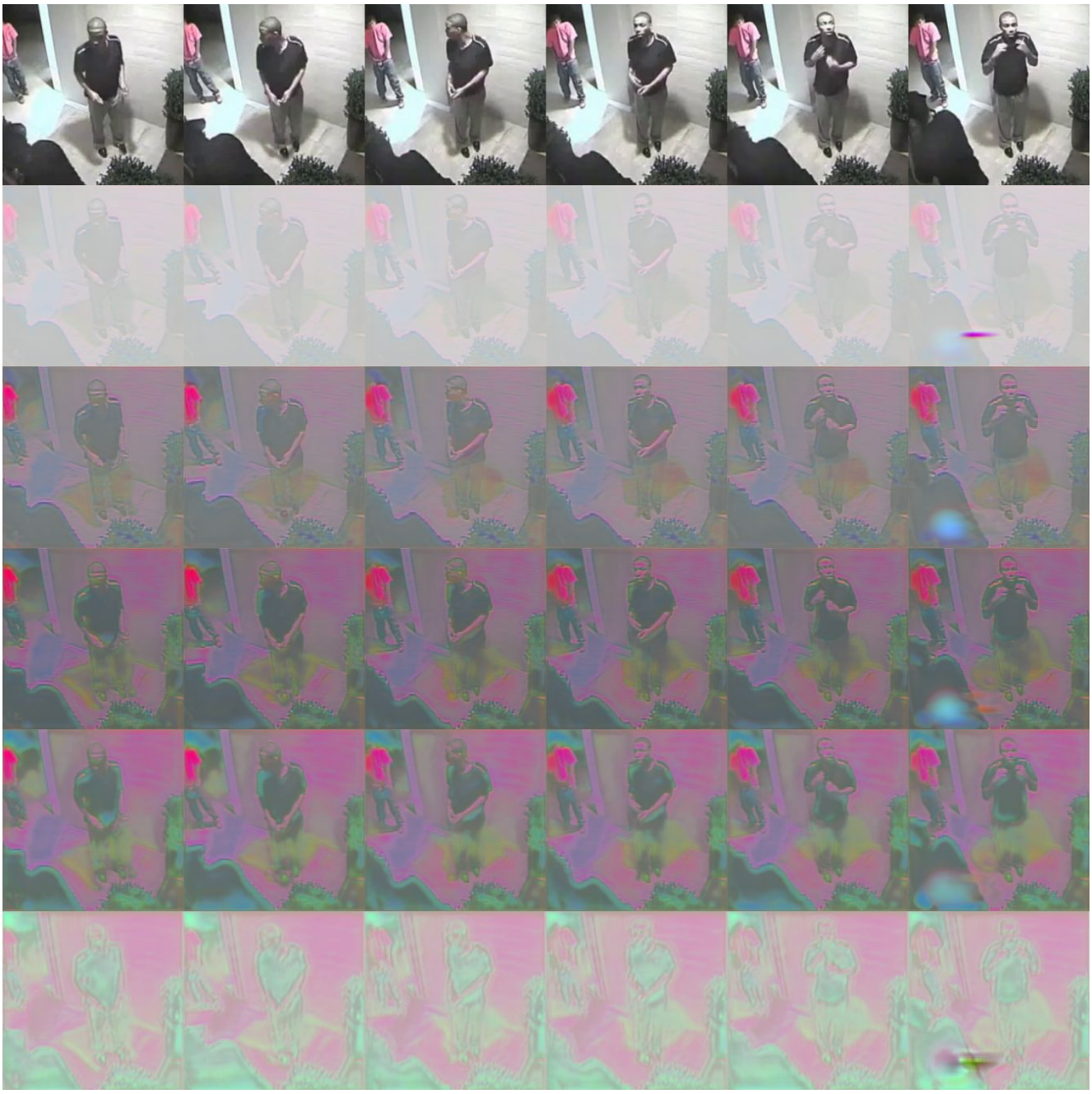

Figure 7: Training progression per epoch of the anonymization process. In order from top to bottom, visualization after f_A on epoch 1, 6, 9, 12, 15, and 20 is shown.

References

[1] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Oct. 1995. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4

[3] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton Fok, Xiaojuan Qi, and Yik-Chung Wu. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 387–395, 2023. 2

[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*, 2022. 1, 2, 3

[5] 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 (CVPR)*, pages 6546–6555, 2018. 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset, May 2017. arXiv:1705.06950 [cs]. 1

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[9] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 4

[10] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[11] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[12] 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. 2

[13] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 3

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, Dec. 2012. arXiv:1212.0402 [cs]. 1

[15] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, Salt Lake City, UT, June 2018. IEEE. 1, 2

[16] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[17] 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 (TPAMI)*, 2020. 1

[18] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1237–1246, 2019. 1