Balancing Privacy and Action Performance: A Penalty-Driven Approach to Image Anonymization
-:Supplementary Material:-

The supplementary material is organized into the following sections:

1. Section A: Dataset details
2. Section B: Self supervised contrastive loss for privacy removal branch f_B
3. Section C: Additional Results.

A. Dataset

UCF101 [8] dataset is a large action annotated dataset with 101 different day-to-day human actions with 13,320 videos. All the experiments in this paper are conducted on the split-1, which contains 9,537 training videos and 3,783 testing videos.

HMDB51 [4] dataset is comparatively small dataset compared to UCF101 and comprise of 6,849 total videos collected from 51 different human actions. All the results in this papar are reported on split-1, which consists of 3,570 training videos and 1,530 testing videos.

VISPR [6] is a multi-class classification dataset designed for private attribute recognition, comprising 22,167 images annotated with 68 different private attributes, including face, gender, skin color, race, and nudity. Following prior works [2, 9], we utilize two distinct subsets of the VISPR dataset, referred as VISPR1 and VISPR2, for our experiments. Each subset contains seven different private attributes, detailed in Table 1.

VISPR1[6]	**VISPR2** [6]
$a17\_color$	$a6\_hair\_color$
$a4\_gender$	$a16\_race$
$a9\_face\_complete$	$a59\_sports$
$a10\_face\_partial$	$a1\_age\_approx$
$a12\_semi\_nudity$	$a2\_weight\_approx$
$a64\_rel\_personal$	$a73\_landmark$
$a65\_rel\_soci$	$a11\_tattoo$

Table 1. Private attribute subsets of VISPR[6] dataset used in experiments.

VPUCF [5] and VPHMDB [5] are large-scale datasets annotated with private attributes for action recognition tasks. The VPUCF dataset is built from the UCF101 dataset, consisting of 101 human action classes with a total of 13,320 videos, while the VPHMDB dataset is derived from the HMDB51 dataset, containing 51 action classes and 6,849 videos. Each video in these datasets is labeled with five private attributes: face, skin color, gender, nudity, and familial relationship. These attributes are represented as binary labels, where 1 indicates the presence of an attribute and 0 denotes its absence. The results reported in this paper are based on experiments conducted on the full dataset.

B. Self supervised contrastive loss for privacy removal branch f_B

A schematic diagram of the self-supervised contrastive loss for the privacy removal branch is depicted in Figure 1. An input video X_p is passed through the anonymizer f_A to generate the anonymized video $f_A(X_P)$. This anonymized video is then processed by a temporal frame sampler S_{fp}, which selects two frames based on sampling strategies. The sampled frame pair $S_{\text{fp}}(f_A(X_p))$ is then passed through a 2D-CNN backbone, f_B, followed by a non-linear projection head, mapping them into the representation space. This results in two projected representations, Z_i and Z'_i. The goal of the contrastive loss is to enforce high similarity between projections from the same video (Z_i, Z'_i) while pushing apart projections from different videos (Z_i, Z_j) where $j \neq i$. The NT-Xent contrastive loss [1] for a batch of N videos is formulated as:

$$L_i^B = -\log \frac{h(Z_i, Z'_i)}{\sum_{j=1}^{N} [1[j \neq i] h(Z_i, Z_j) + h(Z_i, Z'_j)]}, \quad (1)$$

where $h(u, v) = \exp(\cdot)$ is the similarity function used to compute pairwise relationships in the representation space.

For our anonymization purpose, the contrastive loss function works in the opposite manner compared to [1]. Instead of maximizing the agreement between positive pairs and minimizing the agreement between negative pairs, our objective is to increase the disagreement between positive pairs while reducing the agreement between negative

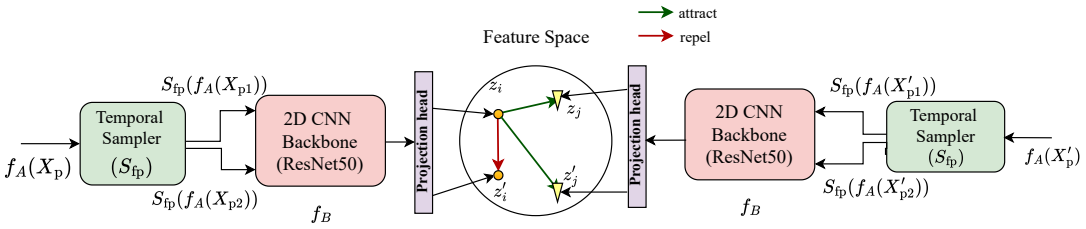


Figure 1. A contrastive learning approach to train the privacy budget task f_B. To anonymize the private attribute of the input data, the distance between the same samples of input data has been maximized, while the distance between the different samples has been minimized.

pairs. This ensures that the anonymizer struggles to encode private attribute features effectively, thereby enhancing anonymization performance. In the experimental setting, this is achieve by taking the negative gradient. We have selected the positive pairs after every four frames from each video. The rationale is that selecting positive frame pairs from large temporal distances reduces the effectiveness of anonymization. This occurs because incorporating highly dissimilar positive samples in contrastive loss leads to suboptimal representation learning. A similar phenomenon has been reported in prior studies [3, 7], where using temporally distant positive pairs resulted in degraded performance.

C. Additional Results

C.1. Training of f_A, f_B and f_T

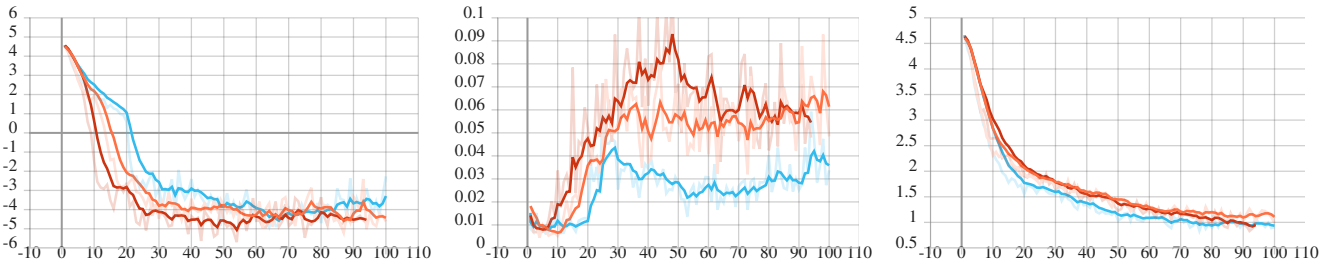
The training loss curves of the anonymizer f_A, budget task f_B, and utility task f_T are shown in Figure 2. The anonymizer is expected to converge by minimizing $\mathcal{L}_\mathcal{A}$ (refer eq. 3 of main paper), which is reflected in Figure 2(a), where the loss of f_A decreases over several training epochs. Meanwhile, the budget task loss increases, as the anonymizer aims to prevent the encoding of private attribute features of the input data. This trend is observed in Figure 2(b), where the loss of f_B increases with the increase in epochs. In contrast, the utility task loss, which is based on cross-entropy, should decrease as training progresses and eventually converge, as shown in Figure 2(c). Additionally, we observe that incorporating the penalty term B with different values allows f_A to reach convergence while preserving the critical features of the utility task and effectively obstructing the decoding of private attributes in the budget task.

C.2. Evaluate f_A^* on different action classifier f'_T

A learned anonymization function, f_A^*, should be able to train any action recognition target model, f'_T, on anonymized data without a significant drop in performance. To validate this, we conducted experiments using the learned anonymizer with different utility target models and analyzed the results, as shown in Table 2. Specifically, we evaluated R3D-18, R2plus1D, MViTv2, and I3D as utility target models. This utility target model is either trained from scratch or initialized with the pretrained weights from the Kinetics 400 dataset. From Table 2, we observe that with different penalty settings of $B = 0.3, 0.5, 0.7, 0.9$, the performance at $B = 0.3$ is closest to that of raw data. However, as B increases, meaning the level of anonymization is higher, the performance declines. This suggests that the anonymizer effectively anonymizes the incoming data, including action-related features. Notably, when R2plus1D is initialized with pretrained weights from the Kinetics-400 dataset, the action recognition performance improves significantly. This improvement occurs because the model has prior knowledge of action features before training. This experiment also suggested that the proposed anonymization training approach makes the model agnostic, and the learned anonymizer can be used with the different utility target models.

C.3. Evaluate f_A^* on the pretrained f'_b on raw data

In a real-world scenario, the trained anonymization model f_A^* is not accessible to anyone. However, there is a potential risk of adversarial attacks targeting the privacy classifier pretrained on raw data, which could lead to the extraction of sensitive privacy-related information. To address this concern, we implemented an additional evaluation protocol. Specifically, we pretrained a new privacy model f'_B (ResNet50) on raw data and subsequently evaluated its performance on anonymized data processed by the learned anonymizer f_A^*. The results of this evaluation are presented in Table 3. Notably, across different penalty setting, $B = 0.3, 0.5, 0.7, 0.9$, the privacy leakage on the dataset remains largely unchanged, with only minor variations. This indicates that incorporating the penalty term in the anonymizer from the utility target model primarily impacts action recognition performance while having minimal influence on privacy leakage. As a result, the anonymizer can effectively anonymize private attributes in the input data to the maximum extent, ensuring minimal privacy leakage through the learned anonymizer. Furthermore, our model demonstrates comparable privacy-preserving performance



(a) Training loss curve of anonymizer model f_A (b) Training loss curve of budget task model f_B (c) Training loss curve of utility task model f_T

Figure 2. Training loss curves for different functions: (a) Anonymizer f_A, (b) Budget task f_B, and (c) Utility task f_T for $B = 0.5$, $B = 0.7$ and $B = 0.9$

Method		R3D-18	R2plus1D	R2plus1D (pretrained on K400)	MViTv2	I3D	C3D
Raw data		62.30	64.33	88.76	76.81	59.12	58.51
SPACT[2]		62.03 (↓ 0.27)	62.71 (↓ 1.62)	85.14 (↓ 3.62)	–	–	56.10 (↓ 2.41)
Ours	B = 0.3	**62.11** (↓ 0.19)	**63.18** (↓ 1.15)	**86.72** (↓ 2.04)	**73.21** (↓ 3.6)	**58.90** (↓ 0.22)	**57.21** (↓ 1.3)
	B = 0.5	59.01 (↓ 3.29)	60.81 (↓ 3.52)	80.97 (↓ 7.79)	71.04 (↓ 5.77)	56.32 (↓ 2.8)	56.14 (↓ 2.37)
	B = 0.7	57.98 (↓ 4.32)	58.21 (↓ 6.12)	77.12 (↓ 11.64)	69.10 (↓ 7.71)	54.28 (↓ 4.84)	53.81 (↓ 4.7)
	B = 0.9	55.28 (↓ 7.02)	57.92 (↓ 6.41)	76.98 (↓ 11.78)	67.18 (↓ 9.63)	51.11 (↓ 8.01)	50.91 (↓ 7.6)

Table 2. Comparison of different privacy-preserving methods with different f'_T architectures trained on UCF101. – indicates that experiment is not performed on the model. ↓ indicates drop from the raw data and high value of accuracy considered as better for the action recognition.

to [2], while significantly enhancing action recognition performance, as shown in Table 2.

C.4. Effect of different private attribute classifier f'_B

A learned anonymization function f^*_A is designed to protect against privacy leakage from any privacy target model f'_B. During the training of the anonymization function, we use ResNet50 as the auxiliary privacy model f_B and evaluate the effectiveness of the learned anonymizer f^*_A on various target privacy classifiers, including R3D-18, R3D-34, R3D-50, R3D-101, and R3D-152, both with and without ImageNet pretraining. As shown in Table 4, our method effectively prevents privacy leakage, regardless of the chosen target privacy model. Furthermore, across different penalty settings of B, the privacy leakage across various target privacy classifiers remains almost constant or only slightly varies. This suggests that introducing the penalty term during the training of the anonymizer does not impact the budget task model f_B. Additionally, when ImageNet pretraining is applied, shown in Table 5, privacy leakage increases across all methods. However, the relative reduction in leakage compared to the raw data baseline improves, demonstrating the robustness of our approach in mitigating privacy risks.

C.5. Visualization of anonymized images under different penalty settings of B:

To visualize the transformation produced by the learned function f^*_A, we present images under different penalty settings of B, as shown in the Figure 3, 4, 5. The visualizations indicate that the anonymized images remain primarily consistent across varying penalty values. This is because the penalty is applied explicitly to the action features, while the anonymizer retains complete flexibility to anonymize the images to the maximum extent. Analyzing the Figures 3, 4, 5, we observe that the anonymized images are not identifiable, demonstrating the effective removal of personally identifiable information by our proposed approach.

Method		VISPR1		VISPR2		PAHMDB	
		cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)
Raw		64.41	0.555	57.63	0.434	70.2	0.396
VITA [9]		22.81 (↓ 41.6)	0.243 (↓ 0.312)	26.61 (↓ 31.02)	0.184 (↓ 0.250)	57.01 (↓ 13.19)	0.231 (↓ 0.165)
SPACT [2]		27.44 (↓ 36.97)	0.076 (↓ 0.479)	20.02 (↓ 37.61)	0.046 (↓ 0.388)	58.90 (↓ 11.3)	0.094 (↓ 0.165)
Ours	B = 0.3	26.91 (↓ 37.5)	0.081 (↓ 0.474)	20.19 (↓ 37.44)	0.051 (↓ 0.383)	57.19 (↓ 13.01)	0.114 (↓ 0.165)
	B = 0.5	26.24 (↓ 38.17)	0.075 (↓ 0.480)	20.20 (↓ 37.43)	0.052 (↓ 0.382)	57.10 (↓ 13.10)	0.114 (↓ 0.165)
	B = 0.7	26.84 (↓ 37.57)	0.081 (↓ 0.474)	20.15 (↓ 37.48)	0.051 (↓ 0.383)	57.14 (↓ 13.06)	0.112 (↓ 0.165)
	B = 0.9	26.98 (↓ 37.43)	0.079 (↓ 0.476)	20.17 (↓ 37.46)	0.058 (↓ 0.376)	57.13 (↓ 13.07)	0.112 (↓ 0.165)

Table 3. Performance comparison of different methods on privacy leakage evaluation using **pretrained** f'_B settings. ↓ indicates the drop in the performance from the raw data. Lower cMAP and F1 scores indicate better privacy protection. Our method shows almost constant privacy leakage through the different penalty settings and performs better than [2].

Method		R3D-18		ResNet34		ResNet50		ResNet101		ResNet152	
		cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)
Raw data		64.38	0.538	65.30	0.555	64.41	0.555	60.70	0.526	58.83	0.485
SPACT [2]		54.83 (↓9.55)	0.457 (↓ 0.081)	54.09 (↓11.21)	0.422 (↓ 0.133)	57.43 (↓ 6.98)	0.473 (↓0.082)	52.94 (↓ 7.76)	0.409 (↓0.117)	53.27 (↓ 5.56)	0.432 (↓)
Ours	B = 0.3	52.81 (↓ 11.57)	0.431 (↓ 0.107)	52.95 (↓ 12.35)	0.412 (↓ 0.143)	57.41 (↓ 7.0)	0.451 (↓ 0.104)	51.21 (↓ 9.45)	0.391 (↓0.135)	51.25 (↓7.58)	0.422 (↓ 0.053)
	B = 0.5	52.10 (↓ 12.28)	0.423 (↓ 0.115)	52.81 (↓ 12.49)	0.401 (↓ 0.154)	57.32 (↓ 7.09)	0.457 (↓0.098)	51.45 (↓9.25)	0.392 (↓0.134)	51.44 (↓ 7.39)	0.422 (↓0.063)
	B = 0.7	52.18 (↓ 12.2)	0.429 (↓ 0.109)	52.92 (↓ 12.38)	0.410 (↓ 0.145)	57.21 (↓ 7.2)	0.452 (↓ 0.103)	51.12 (↓9.22)	0.391 (↓ 0.135)	51.48 (↓ 7.35)	0.421 (↓ 0.064)
	B = 0.9	51.98 (↓ 12.4)	0.435 (↓ 0.103)	52.91 (↓ 12.39)	0.405 (↓ 0.150)	57.22 (↓ 7.19)	0.452 (↓ 0.103)	51.22 (↓ 9.48)	0.389 (↓ 0.137)	51.51 (↓ 7.32)	0.422 (↓ 0.063)

Table 4. Comparison of different privacy-preserving methods with different f'_B architectures trained on VISPR1. ↓ indicates the drop in the performance from the raw data. Lower cMAP and F1 scores indicate better privacy protection. Our method shows almost constant privacy leakage through the different penalty settings and performs better than [2].

Method		R3D-18		ResNet34		ResNet50		ResNet101		ResNet152	
		cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)	cMAP (↓ %)	F1 (↓ %)
Raw data		69.82	0.6041	69.55	0.6447	70.76	0.6591	71.09	0.6330	69.50	0.6130
SPACT [2]		59.10 (↓ 10.72)	0.5302 (↓ 0.0739)	59.71 (↓ 9.84)	0.5227 (↓ 0.122)	60.73 (↓ 10.03)	0.5689 (↓ 0.0902)	59.24 (↓ 11.85)	0.5601 (↓ 0.0729)	60.51 (↓ 8.88)	0.5352 (↓ 0.0778)
Ours	B = 0.3	58.14 (↓ 11.68)	0.5214 (↓ 0.0827)	57.84 (↓ 11.71)	0.5122 (↓ 0.1325)	58.65 (↓ 12.11)	0.5512 (↓ 0.1079)	58.91 (↓ 12.18)	0.5502 (↓ 0.0828)	59.72 (↓ 9.78)	0.5298 (↓ 0.0832)
	B = 0.5	58.12 (↓ 11.7)	0.5211 (↓ 0.083)	57.86 (↓ 11.69)	0.5111 (↓ 0.1336)	58.54 (↓ 12.22)	0.5521 (↓ 0.107)	58.95 (↓ 12.14)	0.5509 (↓ 0.0821)	59.85 (↓ 9.65)	0.5296 (↓ 0.0834)
	B = 0.7	58.21 (↓ 11.61)	0.5212 (↓ 0.0829)	57.91 (↓ 11.64)	0.5214 (↓ 0.1233)	58.98 (↓ 11.78)	0.5525 (↓ 0.1066)	58.72 (↓ 12.37)	0.5519 (↓ 0.0811)	59.96 (↓ 9.54)	0.5284 (↓ 0.0846)
	B = 0.9	58.35 (↓ 11.47)	0.5228 (↓ 0.0813)	57.90 (↓ 11.65)	0.5224 (↓ 0.1233)	58.91 (↓ 11.85)	0.5569 (↓ 0.1022)	58.99 (↓ 12.10)	0.5558 (↓ 0.0772)	59.98 (↓ 9.52)	0.5293 (↓ 0.0837)

Table 5. Comparison of different privacy-preserving methods with different f'_B architectures trained on VISPR1. The privacy target model is **pretrained with the ImageNet** weights. ↓ indicates the drop in the performance from the raw data. Lower cMAP and F1 scores indicate better privacy protection. Our method shows almost constant privacy leakage through the different penalty settings and performs better than [2].

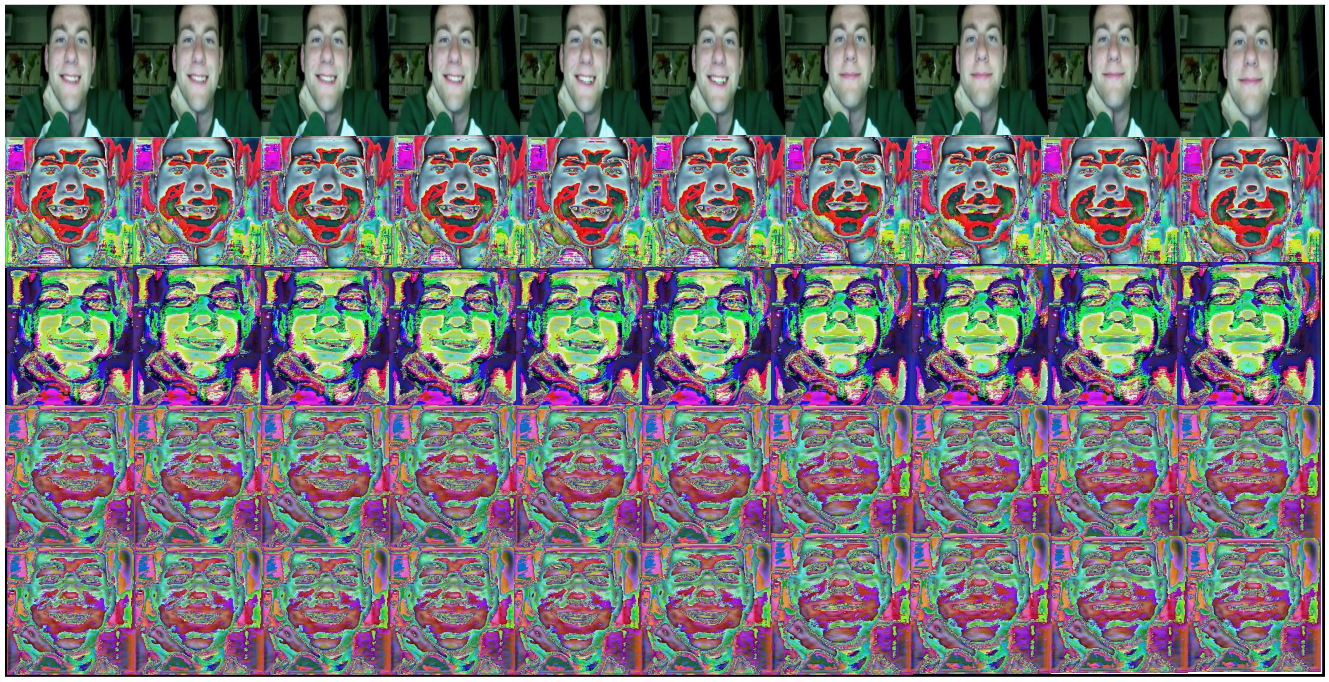


Figure 3. Anonymized frames of smiling action from the HMDB51 dataset across different penalty settings. Top to bottom: Raw image, followed by $B = 0.3$, $B = 0.5$, $B = 0.7$, and $B = 0.9$.

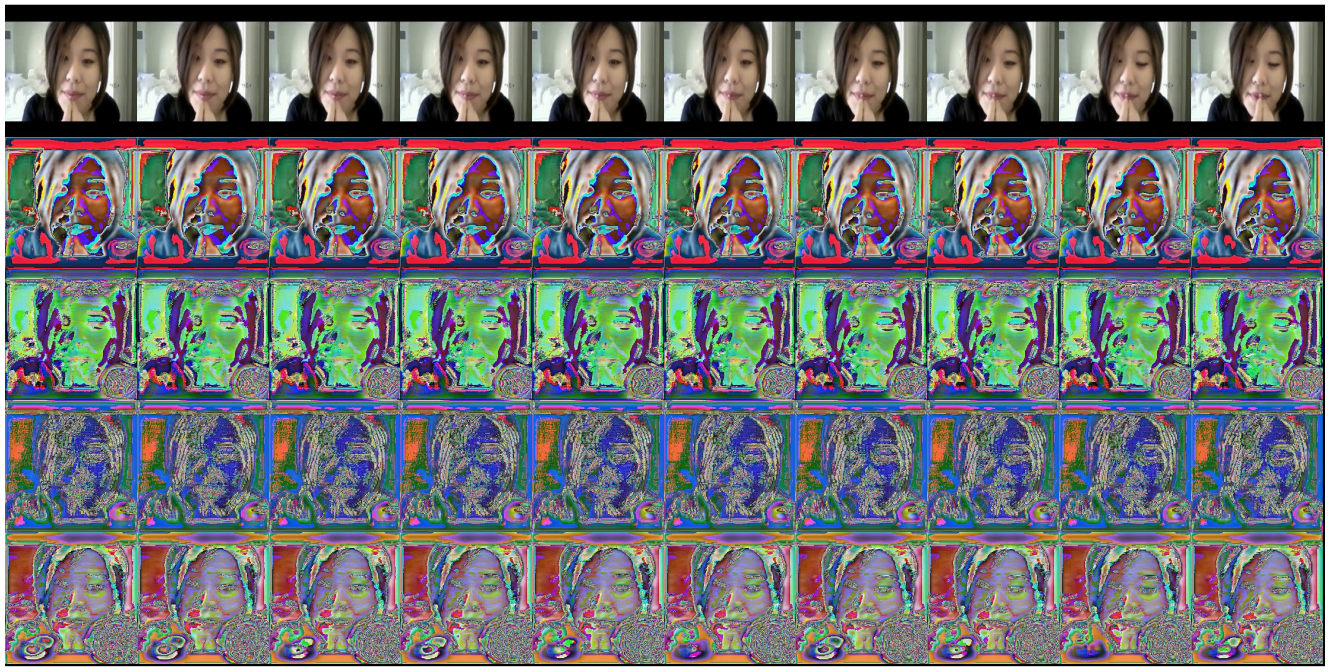


Figure 4. Anonymized frames of apply lipstick action from the UCF101 dataset across different penalty settings. Top to bottom: Raw image, followed by $B = 0.3$, $B = 0.5$, $B = 0.7$, and $B = 0.9$.

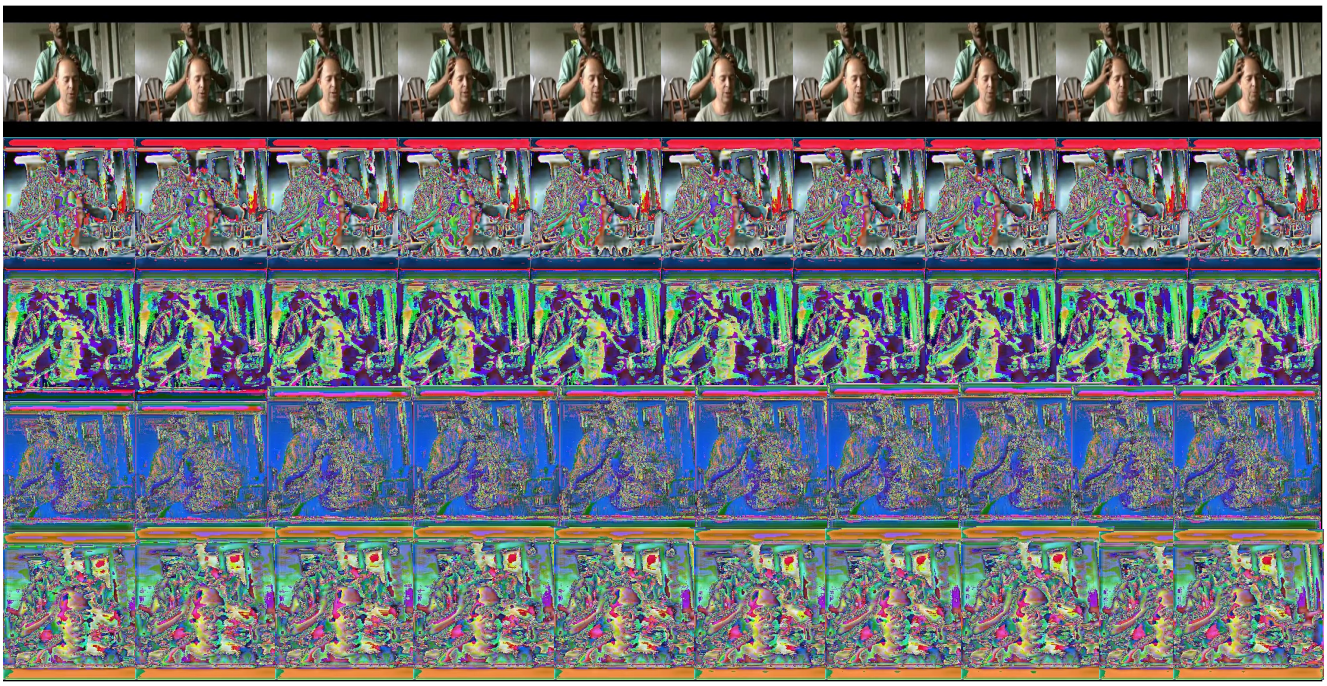


Figure 5. Anonymized frames of head massage action from the UCF101 dataset across different penalty settings. Top to bottom: Raw image, followed by $B = 0.3$, $B = 0.5$, $B = 0.7$, and $B = 0.9$

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[2] 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, 4

[3] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3299–3309, 2021. 2

[4] 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. 1

[5] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5106–5115, 2023. 1

[6] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*, pages 3686–3695, 2017. 1

[7] 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. 2

[8] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[9] 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. 1, 4