Selective, Interpretable and Motion Consistent Privacy Attribute Obfuscation for Action Recognition

Filip Ilic
TU Graz
filip.ilic@tugraz.at

He Zhao
York University
zhufl@eecs.yorku.ca

Thomas Pock
TU Graz
pock@tugraz.at

Richard P. Wildes
York University
wildes@cse.york.ca

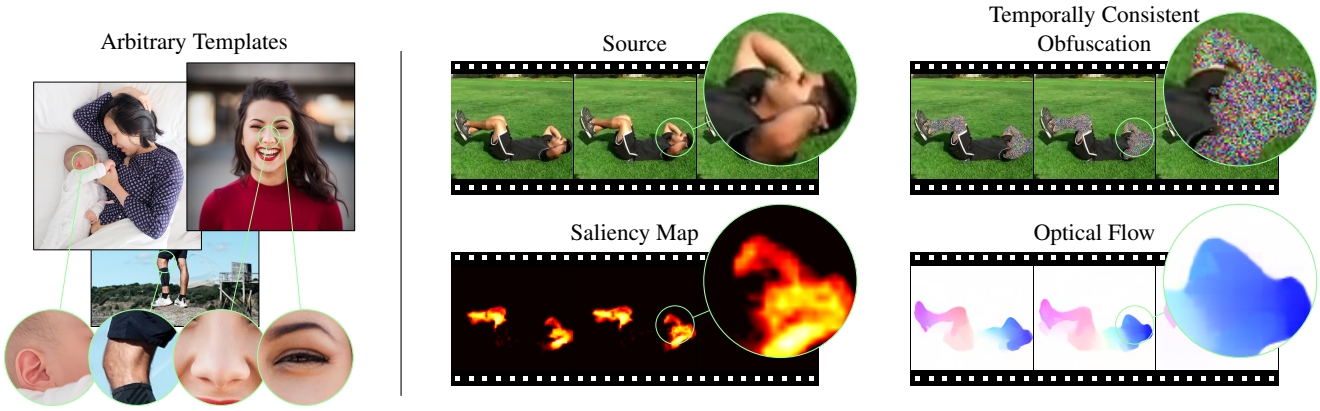


Figure 1. Our goal is to hide privacy attributes without action recognition performance dropping. Left: Arbitrary images can be used to specify an interpretable template library defined by privacy attributes. Middle: A salience map is generated from privacy templates; example illustrates use of templates for personal identification. Right: The source video is masked with noise as guided by salience and animated by source video optical flow. Salience makes masking selective to privacy sensitive regions, while preserving *scene context*; optical flow preserves *motion* – both of which are critical for action recognition. The obfuscated video can be input directly to arbitrary privacy and action recognition systems without retraining. Zoomed circles highlight details only for illustration.

Abstract

Concerns for the privacy of individuals captured in public imagery have led to privacy-preserving action recognition. Existing approaches often suffer from issues arising through obfuscation being applied globally and a lack of interpretability. Global obfuscation hides privacy sensitive regions, but also contextual regions important for action recognition. Lack of interpretability erodes trust in these new technologies. We highlight the limitations of current paradigms and propose a solution: Human selected privacy templates that yield interpretability by design, an obfuscation scheme that selectively hides attributes and also induces temporal consistency, which is important in action recognition. Our approach is architecture agnostic and directly modifies input imagery, while existing approaches generally require architecture training. Our approach offers more flexibility, as no training is required, and outperforms alternatives on three widely used datasets.

Code available f-ilic.github.io/SelectivePrivacyPreservation

1. Introduction

Advances in state-of-the-art computer vision and machine learning enable deployment of such systems in the public sphere. Accompanying these initiatives, concerns arise for the privacy of individuals that are captured in acquired imagery [11, 19, 35, 46]. In particular, considerations arise regarding attributes that individuals want to keep confidential, yet that are revealed through visual information, even though they are not critical for the functioning of the deployed system, *e.g.* identity, age, gender and race. Video-based action recognition is an area of consideration as it has potential for widespread applications in surveillance and monitoring. These concerns have sparked interest in privacy preserving action recognition [13, 24, 34, 58]. These approaches process input imagery to obscure privacy attributes while maintaining action recognition performance. Contemporary approaches typically apply their obfuscation across entire input video frames and improvements have been made within this paradigm. Notably, however, there are downsides to this paradigm, as follows.

Collateral damage. Global masking strategies indiscriminately obscure the entire image, impacting regions within the scene that may exhibit high correlations with actions, albeit lack relevance to privacy. For example, it is known that masking objects and scene context can impair action recognition performance [61], yet these are lost in global masking. Furthermore, global masking strategies do not allow for selective attribute obfuscation and generally hide all attributes at once, even when not all are of concern.

Loss of dynamic information. The large change in input modalities from global masking necessitates the retraining of the action recognition module or the design of custom modules (adversarial training), which adds to the challenge of practical deployment. Indeed, even if applied more locally, the obfuscation can compromise the motion of the actors, which also can be important in action recognition [50].

Lack of interpretability. Finally, given that state-of-the-art approaches are end-to-end trained with limited concern for interpretability, the exact nature of what is being masked and how it is achieved can remain unclear. Lack of interpretability is an important concern in privacy preservation, because its lack can compromise user trust [29, 38].

1.1. Contributions

We present an approach to privacy preserving action recognition that responds directly to current limitations in four ways, as illustrated in Fig. 1. (i) The approach is based on local detection and selective obfuscation of privacy sensitive regions. This selectivity maintains global context information that is crucial to action recognition, yet unimportant for privacy. (ii) The local processing avoids large modality shifts in the imagery and is independent of the action recognition module itself; therefore, it does not require algorithm retraining, which is sometimes infeasible. (iii) The masking preserves interframe motion; so, that information is available for action recognition. (iv) The privacy sensitive masking is interpretable by design, and allows inspection through the explicitly generated saliency maps.

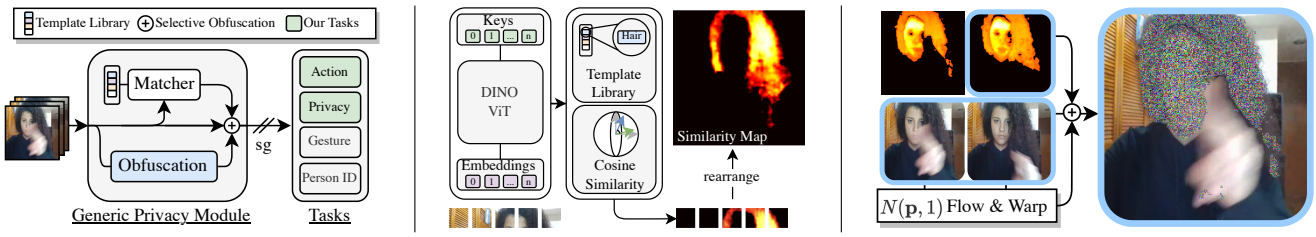
2. Related work

Privacy in machine learning. The need to protect privacy has garnered increased attention in the vision research community. Current models commonly consume large amounts of web data to learn generalizable representations [20, 33, 53], which inevitably invade personal information, such as identity and location. Moreover, in model deployment it also may be desirable to preserve privacy information. The concern for privacy is not limited to vision research, but extends across artificial intelligence, including natural language processing [5] and more general machine learning [37]. As these technologies find their way into broader society, privacy concerns must be considered.

Privacy preserving action recognition. We focus on video-based action recognition. Recent developments in this area have yielded systems capable of strong performance on challenging datasets; for review see, *e.g*. [48]. Similarly, applications, including those in privacy-sensitive scenarios (*e.g*. surveillance [10] and monitoring [51]), are being developed. To provide useful spatiotemporal signals for action recognition, video clips are mostly collected to capture actors with a great level of clarity throughout the actions, thus increasing the chance of privacy leakage. Moreover, such datasets are expanding rapidly, *e.g*. [4, 30].

In response to these concerns, research on recognizing actions while preserving private information has emerged in recent years. Early work concentrated on devising models that can work on low-resolution videos [8, 43, 44]; these methods often operate at the cost of sacrificing action recognition performance. Other work along these lines developed face-anonymization techniques to prevent models from yielding high accuracy on face recognition [28, 42]. Still, those approaches cannot easily extend to other attributes and are correspondingly limited. More recent work has focused on implicitly learning transformation functions to anonymize videos in a data-driven fashion [13, 24, 36, 59], and also extending such approaches to anomaly detection [18]. To reduce model complexity, a competing approach follows a simple, yet effective procedure [34]: Frame subtraction, followed by broadband-filtering to yield motion descriptors for action recognition, while suppressing privacy attributes. A notable limitation of that approach is the relative weakness of frame differences as motion descriptors compared to common alternatives, *e.g*. optical flow. Another common limitation to all existing work is the lack of flexibility to select arbitrary private information to hide, *i.e*. preserving privacy of only critical attributes, while maintaining the visibility of others, to avoid obfuscation of visual cues that are essential to action classification.

Template matching. Building correspondences between templates and target images via matching is a foundation for modern vision research, *e.g*. SfM [21], image correspondences [45], detection [12] and tracking [49]. Features used in matching have advanced rapidly from primarily handcrafted (*e.g*. [12, 39]) to learning-based convolutional [23] or transformers [14]. Our solution is inspired by techniques seen in feature-based template matching. We apply their insights on matching templates to localize and selectively obfuscate privacy-sensitive regions. More specifically, we adopt DINO-ViT features [6] and compute similarities using the *keys* of the last attention layer from that architecture. Our use of DINO-ViT keys (rather than queries or values) is based on previous work finding them to perform best when applied to visual correspondence [1, 40].



(a) Overview (b) Matcher (c) Temporally Consistent Selective Obfuscation

Figure 2. Overview of Method. 2a We present a privacy module that builds atop three components: (i) a semantic *template library* that contains attributes to be hidden, (ii) a descriptor *matcher* to localize template features in videos to be obscured and, (iii) an *obfuscation* method that is sensitive with respect to motion present in the scene. 2b A semantic descriptor matcher based on DINO [6]-ViT [14] keys is used to determine privacy salient regions in a video based on the template library. In our case, regions of interest correspond to those that can identify a person; however, this component can be adapted for other privacy attributes through specification of different templates. The result is a saliency map. 2c The saliency map is used as a weight to apply noise to the regions. The noise, however, is not static, but is warped with optical flow with an initial noise pattern image, $N(\mathbf{p}, 1)$, for the purposes of preserving motion information in the source video. The similarity maps of all aggregated relevant privacy attributes are used to weigh the noise and apply it to the input image, obfuscating privacy sensitive information while not destroying the underlying temporal signal.

3. Technical approach

Our approach to privacy preserving action recognition is a stand alone module that operates by preprocessing video that subsequently is input to arbitrary action and privacy recognition algorithms, *i.e*. it is independent of the recognition algorithms and does not entail any retraining of those algorithms. Our method consists of three key components; see Fig. 2a: (i) A template library that covers privacy attributes to be preserved, (ii) a matcher that produces saliency maps between selected templates and images where privacy is to be preserved and (iii) an obfuscator that uses the saliency maps to hide privacy attributes of concern in a temporally consistent fashion to preserve motion in the video. The remainder of this section details each of these components.

3.1. Template library

To obfuscate privacy sensitive regions in images in an interpretable fashion, we need an explicit set of "attributes" to be hidden from the target video. Such a concept template library, T, can be built manually choosing image patches corresponding to features one wishes to obfuscate. In this paper, we concentrate on privacy attributes related to personal identification. Therfore, we use landmark anatomical features, corresponding to detailed facial landmarks and the hand (Fig. 3 left: *forehead, hair, eye, cheek, lips, hand*) as well as larger body parts (Fig. 3 right: *arm, torso, legs*). The choice to focus on preservation of human identity is motivated by the fact that all three of our evaluation datasets define their privacy attributes on attributes relating to person identity. Notably, however, such a template library easily can be extended depending on the particular task at hand. For example, attributes pertaining to location (e.g. street signs, distinctive scene objects) could guide development of a complimentary set of templates. In any case, given a template library, a user can selectively combine templates to obscure attributes of concern in a particular application.

Formally, we define each template τ_i as an element in the set $\mathsf{T} = \{\tau_1, \tau_2, ..., \tau_n\}$. Further, let $\tilde{\mathsf{T}} \subseteq \mathsf{T}$ be the subset of templates that the user selects from the library, T, for a particular subset of privacy attributes to be preserved for a given dataset or application. Our manual approach to template selection leads to concepts that are *interpretable by design*. In particular, we use two source images, shown in Fig. 3, one for small-scale features on the face, and one for large regions, taken from the IPN [3] and SBU [60] datasets. We choose these two datasets because of their complimentarity in template selection: IPN focuses on small scale features (*e.g*. facial and hand), while SBU focuses on larger scale features (*e.g*. larger body parts), as detailed later in Sec. 4.1. The choice of the particular template images is not critical to the functioning of our approach, because we extract semantic features from the templates that are known to generalize well across images instances [6], as described next.

3.2. Matching: Local patch descriptor templates

Semantic features. Our requirements for good features to match between privacy templates in our library and frames in an action recognition video are straightforward: We require (i) semantically rich features that generalize across (in our case) different individuals and (ii) high spatial resolution to perform privacy obfuscation in a localized manner without destroying regions that are required for action recognition. We find that DINO-ViT features fit our needs very well [6, 56]: (i) They have been trained to yield high similarity for semantically related concepts (*e.g*. objects and their parts), while suppressing the similarity of unrelated

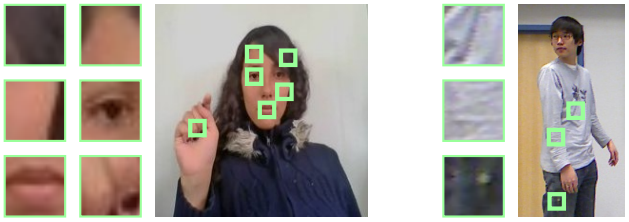


Figure 3. Template Library consisting of Patches Chosen from Anatomical Landmark Regions. These specific images are passed through a DINO-ViT feature extractor. The keys, corresponding to spatial locations of the highlighted patches, are chosen as the templates for matching to input images to obtain semantically similar regions for obfuscation.

matters (*e.g.* background). (ii) They support local patch feature extraction over high resolution images without losing global context. Elsewhere, DINO-ViT keys have proven especially useful in matching between templates and target images [1, 40]. Following these advances, we use last attention layer DINO-ViT keys computed from privacy templates (*e.g.* Fig. 3) to match against input images to compute privacy saliency maps.

Privacy saliency matching. To find privacy salient regions in an input video, a similarity map, S, is computed between each frame, I, in the video and the selected set of privacy templates, $\tilde{\mathsf{T}}$. The image is tiled with m patches, $I_j, j \in \{1, ..., m\}$, and for each patch DINO-ViT keys, $\mathsf{K}(I_j)$, are extracted. These features are matched with DINO-ViT keys, $\mathsf{K}(\tau_i)$, for all selected templates, $\tau_i \in \tilde{\mathsf{T}}$. Clipped cosine similarity is used to establish the saliency of each image patch according to

$$s_j = \frac{1}{|\tilde{\mathsf{T}}|} \sum_{i=1}^{|\tilde{\mathsf{T}}|} \max\left(0, \frac{\langle \mathsf{K}(\tau_i), \mathsf{K}(I_j) \rangle}{\|\mathsf{K}(\tau_i)\| \|\mathsf{K}(I_j)\|}\right), \forall j \in [\![1, m]\!], \quad (1)$$

where $\langle \cdot, \cdot \rangle$ is inner product and $|\tilde{\mathsf{T}}|$ is the cardinality of $\tilde{\mathsf{T}}$, the set of selected privacy templates. Clipping is used because only positive values imply salience. This calculation is performed for every patch in every image of the video. Subsequently, the m saliency patches are reassembled in the shape of the original image to produce the final saliency map,

$$S = \mathcal{R}(s_1, ..., s_m; h, w), \quad (2)$$

with $\mathcal{R}$ a function that accepts image tiles, s_j, and reshapes them into their original image format of height h and width w. Note that a separate salience map, S, is calculated for each frame in a video of interest, I. The entire process of feature extraction, matching and the resulting saliency maps are summarized in Fig. 2b. Example saliency maps in Fig. 4 illustrate the ability of our approach to capture all of our templates in a variety of scenarios.

3.3. Temporally consistent obfuscation

Our goal is to mask privacy sensitive regions in images to obfuscate them. If we were to apply masks to every frame independently, then temporal information important to action recognition, *e.g.* motion of actors, would be destroyed as well. We empirically document the issue in Sec. 4.2. In response to this challenge, we follow previous work that presented a method for producing temporally consistent spatial noise across videos that supported action recognition, while obscuring appearance information in single frames [25]. While the previous work applied noise patterns uniformly across entire frames, we instead weight them by our privacy salience maps, S, to preserve as much context information as possible. The remainder of this subsection details our approach, with an outline shown in Fig. 2c.

Noise pattern initialization. Let $\mathbf{p} = (x, y)$ be image coordinates and t time. We initialize a noise image, $N(\mathbf{p}, 1)$, with the same dimensions as a frame from the input video (*i.e.* $h \times w \times 3$), with h height, w width and 3 the number of colour channels. The dataset mean, μ, and standard deviation, σ, are used to define a uniform distribution from which individual pixel intensities are drawn according to

$$N(\mathbf{p}, t = 1) \sim \mathcal{U}[\mu - \sigma, \mu + \sigma]. \quad (3)$$

Motion consistent noise. To create *motion consistent noise* the initial random frame, $N(\mathbf{p}, 1)$, is warped forward with flow fields derived from the original video, $I(\mathbf{p}, t)$, as extracted by an optical flow algorithm. Let $\mathbf{v}(\mathbf{p}, t) = (u(\mathbf{p}, t), v(\mathbf{p}, t))$ be the flow field that maps points, $\mathbf{p}$, in frame t to those in frame $t - 1$, with u and v the horizontal and vertical components of the flow. Then, a motion consistent noise sequence is generated as

$$N(\mathbf{p}, t) = N\big(\mathbf{p} + \mathbf{v}(\mathbf{p}, t), 1\big) \quad (4)$$

with $t \in \{2, ..., T\}$ and T the number of frames in the original input video.

Selective privacy obfuscation. The computed video sequence, N, shows no single frame appearance related to the original video as it is random noise; however, when viewed as a video it reveals the motion present in the original, *cf*. [25]. Direct use of this synthesized video obscures privacy attributes; however, it also obscures other context information that could be of use in action recognition. So, instead we selectively apply N to every frame I in the video by using the privacy salience maps S, according to

$$O(\mathbf{p}, t) = I(\mathbf{p}, t) + \Big(S(\mathbf{p}, t) \times \big(N(\mathbf{p}, t) - I(\mathbf{p}, t)\big)\Big). \quad (5)$$

The resulting video, O, contains selectively obfuscated regions, built with interpretable templates by design and contains motion information that (in principle) does not differ from the original input video.

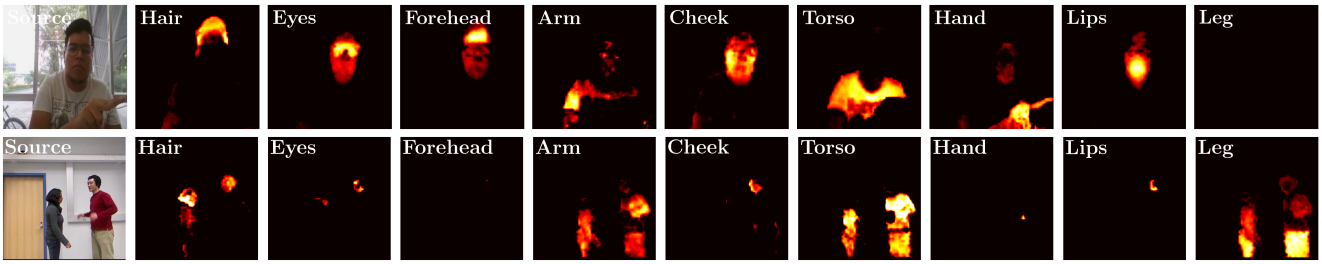


Figure 4. Saliency Maps for Descriptors in the Template Library. The manual selection of these templates allows for *interpretability* of the obfuscated parts of the image *by design*. The matched DINO-ViT features capture rich semantic information and allow for detailed spatial localization due to the nature of vision transformers. These saliency maps can then be combined for obfuscating any combination of templates depending on the task at hand.

ACTION RECOGNITION ON SOURCE DATASETS

Network	$f \times r$	IPN	KTH	SBU
C2D [57]	8×8	74.56	83.00	87.50
CSN [55]	32×2	91.73	88.33	90.91
E2S X3D L [25]	16×5	**95.15**	**95.33**	91.11
E2S X3D M [25]	16×5	89.38	92.33	86.36
E2S X3D S [25]	13×6	85.16	92.00	82.98
I3D [7]	8×8	83.15	89.67	82.95
MVIT[15]	16×4	88.00	90.00	**92.55**
R2+1D [54]	16×4	89.80	87.33	81.91
Slow [17]	8×8	85.76	89.00	88.64
SlowFast [17]	32×2	88.06	87.33	90.00
X3D L [16]	16×5	94.41	94.00	90.22
X3D M [16]	16×5	91.67	93.00	81.91
X3D S [16]	13×6	87.81	90.67	75.00
Average		88.05	90.15	86.31

PRIVACY PRESERVATION FOR SOURCE DATASETS

Network	IPN	KTH	SBU
ResNet_{18} [22]	88.46	87.33	90.43
ResNet_{50} [22]	92.31	90.00	**96.81**
ResNet_{101} [22]	**94.23**	**94.00**	84.04
$\text{ViT}_{b/16}$[14]	**94.23**	**94.00**	79.79
$\text{ViT}_{b/32}$[14]	90.38	**94.00**	78.72
Average	91.92	91.87	85.96

Table 1. Top-1 Accuracy for all Privacy and Action models on the original unmodified (source) videos, *i.e.* without privacy obfuscation. **Bold** and underline indicate first and second best, resp. Number of frames and temporal sampling rate indicated as $f \times r$.

4. Empirical evaluation

The privacy obfuscated video, (5), serves directly as input to action and privacy recognition. No specialized development, training or other modification of the recognition algorithms nor adaptation of the privacy preservation system is necessary. Indeed, a major difference compared to competing state-of-the-art approaches (*e.g.* [34, 43, 59]) is that *we do not retrain* networks with our obfuscated data, while *they do retrain*. We exploit the fact that privacy attributes are generally *independent* from action recognition cues, which is enabled by our unique *selective obfuscation* approach.

4.1. Protocol

Datasets. We use three datasets commonly used for investigating action recognition and privacy, IPN [3], SBU [60] and KTH [47], which pose different challenges concerning both action and privacy.

IPN is a large-scale video-based hand gesture recognition dataset that consists of 50 actors performing 13 static or dynamic gestures, against three different backgrounds [3]. The performed gestures are used as the action labels and actor genders are used as the privacy labels.

SBU is a video dataset depicting eight human interactions. Each video is a video of actor-pairs, in the same laboratory environment. Action labels are derived from the actor interactions and the privacy attributes are the unique pairings of seven different actors resulting in 13 privacy labels.

KTH has 25 actors doing one of six actions [47]. Each action is performed four times in different environments. The six action classes are used for action recognition and 25 actor identities serve as privacy labels. In the originally proposed splits, videos of any one actor are fully contained in one set, since the intent of the KTH dataset was not concerned with actor identity recognition. For privacy identity, each actor must appear in both sets; so, the data is split such that each action is performed twice by each person in the training set, and once in the validation and test sets.

Metrics. Privacy recognition results are obtained following standard practice [34], averaging outputs over multiple frames (32 for IPN and KTH, and 16 for SBU) from the same video. Action recognition results are obtained by reporting the top-1 accuracy, as per convention [24, 58, 59].

It also is interesting to gauge the trade-off between action and privacy recognition. Let $0 \leq a \leq 1$ be normalized action recognition performance derived from dividing accuracy percentages by 100, and let p be defined analogously for privacy recognition. To quantify their trade-off we use a linear combination of a and $1 - p$ and define

$$f_\lambda(a, p) = (1 - \lambda)a + \lambda(1 - p), \tag{6}$$

with $0 \leq \lambda \leq 1$ weighing the relative importance of action

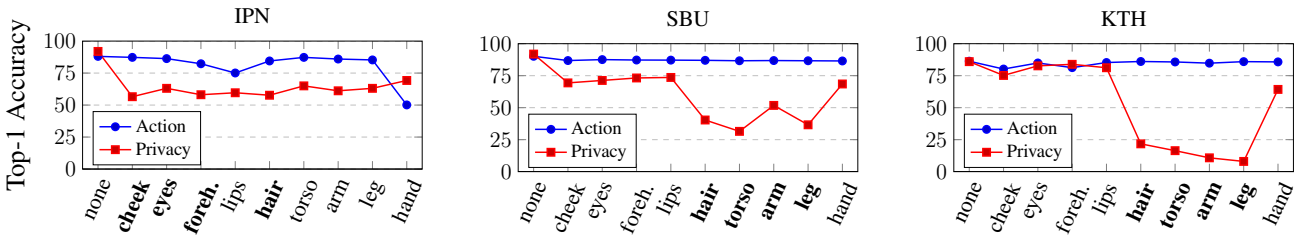


Figure 5. Obfuscation with a Single Attribute and the Impact on Performance. Attribute importance is dataset dependent. For example, notice how the 'Hand' template contributes to a large decrease in action recognition performance on IPN, as the action is determined soley by the hand, whereas on SBU it does not. Optimally **blue** is high and **red** is low. Corresponding qualitative examples of saliency maps for each individual template are shown in Fig. 4. **Bold** text along the abscissa of each plot indicates the templates used for the final results.

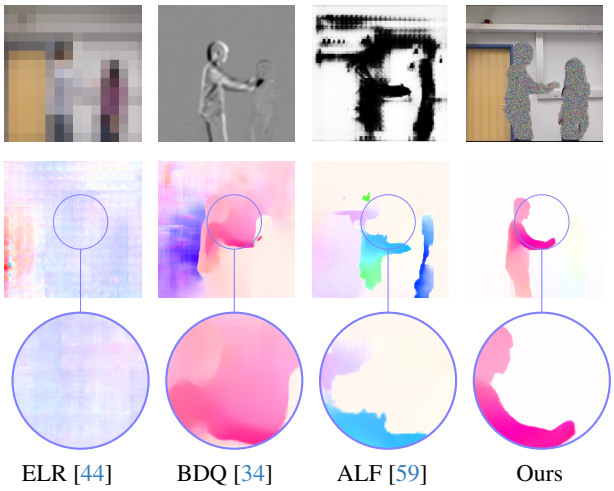


Figure 6. Optical flow from two consecutive frames from competing approaches. Only our approach retains dynamic information that supports high-quality optical flow recovery. Top row: Single frames as processed by comparison algorithms. Middle row: Full frame optical flow recovery. Bottom row: Zoomed optical flow details. Optical flow shown in Middlebury colour coding [2].

recognition vs. privacy. This metric is suitable to create a linear ranking as a, and $(1-p)$ are optimally 1 ensuring $f_\lambda \in [0,1]$. If privacy or action recognition are not equally important, then they can be weighted accordingly; *e.g.* if privacy is critical, then λ can be increased. In our main experiments we use $\lambda = 0.5$, abbreviated as $f_{0.5}$ in Tab. 2 and show an ablation over different values of λ in Fig. 7.

Competing obfuscation approaches. We briefly describe NAIVE BASELINE approaches that solely rely on full-frame pixelation and bluring as well as actor masking that have been studied across the literature [13, 58]. We also highlight other State-of-The-Art approaches (SOTA) [34, 43, 59] that we compare against. Visual examples of the SOTA approaches are shown in Fig. 6; also shown are optical flows of consecutive frames for qualitative comparison.

Mask Obfuscation. All our datasets involve people performing actions and the privacy attributes are related to people. A naive way to preserve privacy in such videos is to completely mask out the actors. To do so, we use the YOLOv8 implementation [27] based on the original YOLO [41]. We report results based on the masked region filled with the mean intensity of the image covered by the mask.

Pixelation. Pixelation applies average pooling over regions of dimensions $x \times x$ on the input image sequence. We choose two scales of $x \in \{4, 16\}$ for the patch-size to pool. In the result tables these methods are abbreviated as $\text{Pix}_{x \times x}$.

Blur. Blur applies a Gaussian blur to the images that have been rescaled to 224×224 pixels. The parameters of the Gaussian are the kernel size, κ, and the standard deviation of the kernel, σ. We choose values for weak and strong blurs, $\frac{\kappa=13}{\sigma=10}$ and $\frac{\kappa=21}{\sigma=10}$, respectively, as consistent with other work on obfuscation methods [59].

ELR initially reduces frames to Extreme Low Resolution and subsequently applies a set of learned inverse super-resolution transforms to support action recognition [43].

ALF is an Adversarial Learning Framework for action recognition that takes into account a privacy budget [59].

BDQ is a privacy-preserving encoder that sequentially Blurs, Differences and Quantizes frames. The blur and quantization parameters are learned to maximize action recognition while minimizing privacy recognition [34].

Notably, all the compared SOTA methods operate across entire frames (*i.e.* without selectivity) and have limited interpretability due to their learning-based obfuscation.

Implementation details.

Training. No large scale dataset exists that allows for joint training of action and privacy classification on the same videos. Therefore, we pretrain our action and privacy networks on Kinetics400 [7] and Imagenet1k [33], respectively, and then finetune for specific datasets. We use the AdamW [31] optimizer with a learning rate of $3e^{-4}$. The networks are trained with a patience scheme of 100 epochs that monitors the loss on the validation set. For action recognition, videos are temporally uniformly subsampled according to the architecture (Tab. 1, '$f \times r$' column).

Method	IPN ↑Action	IPN ↓Privacy	IPN ↑$f_{0.5}$	IPN Δ	KTH ↑Action	KTH ↓Privacy	KTH ↑$f_{0.5}$	KTH Δ	SBU ↑Action	SBU ↓Privacy	SBU ↑$f_{0.5}$	SBU Δ
RESULTS ACROSS NAIVE BASELINE OBFUSCATION METHODS												
Original	88.05	91.92	0.48		90.15	91.87	0.49		86.31	85.96	0.50	
Masking	36.45	64.87	0.36		37.54	35.22	0.51		52.62	30.1	0.61	
Pixelate$_{4\times4}$	85.59	73.65	0.56		83.64	59.39	0.62		73.24	46.27	0.63	
Pixelate$_{16\times16}$	49.81	65.76	0.42		38.33	38.43	0.50		25.53	29.84	0.48	
Blur $^{\kappa=13}_{\sigma=10}$ (weak)	76.24	67.40	0.54		44.59	37.73	0.53		72.75	35.69	0.69	
Blur $^{\kappa=21}_{\sigma=10}$ (strong)	58.80	65.92	0.46		30.38	31.84	0.49		57.77	34.21	0.62	
Ours	87.11	51.93	**0.68**	+0.11	88.67	5.46	**0.92**	+0.29	86.74	13.19	**0.87**	+0.18
RESULTS ACROSS SOTA OBFUSCATION METHODS												
BDQ [34]	81.00	59.00	0.61		91.11	7.15	0.92		84.04	34.18	0.75	
ALF [59]	76.00	65.00	0.56		85.89	19.27	0.83		82.00	48.00	0.67	
ELR[44] s=16	70.82	64.32	0.53		91.22	88.86	0.51		96.27	82.97	0.57	
ELR[44] s=32	52.96	63.29	0.45		85.57	82.56	0.52		92.42	64.89	0.64	
ELR[44] s=64	31.63	62.70	0.34		56.21	58.35	0.49		80.05	43.61	0.68	
Ours †	85.25	51.67	0.67	+0.06	89.44	4.31	**0.93**	+0.01	84.04	11.70	0.86	+0.11
Ours	87.11	51.93	**0.68**	+0.07	88.67	5.46	0.92	±0.00	86.74	13.19	**0.87**	+0.12

Table 2. Comparison between Ours vs. NAIVE BASELINE (top) and SOTA (bottom) Approaches as Top-1 Accuracy for both Action and Privacy labels, as well as $f_{0.5}$ Introduced in Sec. 4.1. We show Ours and NAIVE BASELINE results averaged across all recognition algorithms presented in Tab. 1. Competing SOTA approaches only present results on one specific algorithm for action recognition (I3D [7]) and privacy attribute detection (ResNet50 [22]) as those approaches are network specific. To showcase a fair comparison we also present results with the same single recognition algorithms indicated as "Ours †". First and second best results indicated by **bold** and underline, respectively. Relative performance delta (Δ) is indicated in **green** (improvement) and **orange** (tie) between best and second best method.

Note that we never perform any training on videos obfuscated by our approach. To implement warping, (4), we use pretrained RAFT to extract optical flow [52].

Privacy template selection. Our approach affords selective combination of a predefined set of privacy templates for a given dataset or application. For the experiments, we choose a subset of templates, $\tilde{\mathsf{T}}$, from our complete identity preserving template library, T, shown in Fig. 3, to optimize performance on a given dataset. Figure 5 shows template-wise performance based on both action recognition and privacy for each dataset. For SBU and KTH, action recognition is stable with respect to templates; however, the same four templates notably reduce privacy recognition. For consistency on IPN, we also select the four templates that most reduce privacy recognition, even while preserving action recognition; although, the distinction is less striking. This selection process results in templates $\tilde{\mathsf{T}} = \{torso, arm, leg, hair\}$ for SBU and KTH vs. $\tilde{\mathsf{T}} = \{cheek, eyes, forehead, hair\}$ for IPN. Our code yields more results on the overlap of individual saliency maps.

Runtime. The constant overhead to produce our obfuscated videos is $\approx 100ms$/frame on a NVIDIA RTX4080 GPU.

4.2. Results

We evaluate our approach to privacy preserving action recognition on 13 different action recognition models and five different privacy attribute recognition models. Competing approaches evaluate on at most five action models [34, 43, 59]. Table 1 shows results on the original videos from IPN [3], SBU [60] and KTH [47], *i.e.* without any privacy preserving processing. Table 2 shows results comparing our approach vs. NAIVE BASELINES and SOTA. Comparison is given as the average across recognition algorithms presented in Table 1. Notably, while our approach compromises at most 1.48% action recognition accuracy compared to performance on the original video (88.67% vs. 90.15% on KTH), it *always* greatly improves privacy.

Naive Baselines. Compared to the baselines, our approach scores highest with respect to the $f_{0.5}$ metric across all three datasets: IPN (**+0.11**), KTH (**+0.29**), and SBU (**+0.18**). NAIVE BASELINES show that action recognition correlates highly with privacy performance, as $f_{0.5}$ hovers around 0.5. The notable exception is Pix$_{4\times4}$ that performs well on IPN and KTH, and is the second best NAIVE BASELINE on SBU. A notable benefit of all these approaches is that they do not require retraining of the recognition algorithms. Still, none have a selective ability to obfuscate only certain parts.

State of the art. A noteworthy difference between our vs. the alternative SOTA approaches is that they only work with the recognition algorithms on which they were trained. While this fact disadvantages our approach, as it lacks such retraining, it still outperforms the competing methods on all datasets: IPN (**+0.06**), KTH (**+0.01**), and SBU (**+0.11**).

It is valuable to consider obfuscation approaches in terms of the relative importance of action recognition vs. privacy preservation. Figure 7 compares results for the SOTA ap-

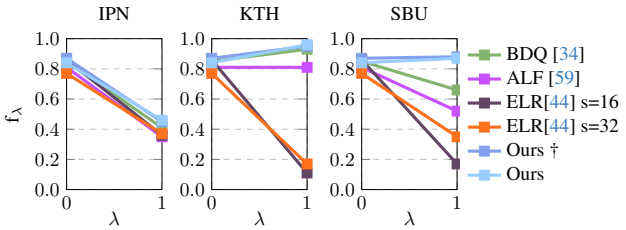


Figure 7. Performance across all datasets with varying λ; see (6). Values of λ closer to 0 weigh action recognition higher, whereas values closer to 1 increase the importance of privacy preservation.

proaches as that trade-off is varied in terms of f_λ, (6). The sweep of λ shows that our obfuscation approach performs better than any competing approach across the entire range. Especially with increasing λ, the gap to other approaches widens as more emphasis is put on privacy preservation.

Importance of selective obfuscation. Our approach is unique in its ability to selectively obfuscate particular regions within a video based on specific privacy attributes. This ability yields two benefits:

(i) Selectivity aids in interpretability. Each selected template results in a saliency map, (2), which allows for visual inspection of what information is being obscured. Figure 4 highlights this benefit as the heat maps reveal the degree of obfuscation to be applied on a template-by-template basis.

(ii) Individual privacy templates can be chosen and combined for best performance; see Fig. 5. Depending on the dataset, different privacy templates differently impact the performance of privacy and action recognition. Action recognition on SBU and KTH is relatively robust to privacy template selection; however, privacy preservation is best when a subset of templates (*torso*, *leg*, *arm*, *hair*) is selected and the rest remain unobfuscated. This fact derives directly from the saliency calculation, (1): If multiple individual saliency maps have small values, then other strong responses are scaled down in the combined final saliency map. Subsequently, this effect leads to less noise being applied in our selective obfuscation, (5), which can deteriorate privacy preservation. In contrast, action recognition on IPN is compromised if the *hand* template is selected (as expected with a gesture centric dataset), which documents that simply obfuscating the entire person leads to inferior action performance compared to selectively applied obfuscation.

Importance of motion consistent noise. Action recognition is complicated as different datasets, and even individual actions within a dataset might require different recognition capabilities and can rely to different degrees on the modeling of motion [9]. This uncertainty on the role of motion in action recognition is exacerbated by the fact that different deep learning-based architectures are successful in capturing motion to varying degrees [32].

To see the importance of our proposed motion consistent noise for privacy preservation while maintaining good

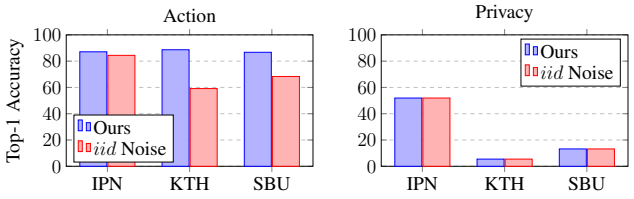


Figure 8. Comparing Temporally Consistent vs. iid Obfuscation for Action Recognition and Privacy Preservation. Action recognition performance decreases if noise is not temporally consistent.

action recognition, we compare to another version of our pipeline that obfuscates video frames using independent, identically distributed (iid) noise; see Fig. 8 and project page for video results. For all datasets, action recognition performance decreases if the noise is iid. This result is sensible, because the iid noise does not capture the temporal dynamics of the source video. In contrast, privacy preservation is robust to noise in either case. There is less impact on IPN action recognition, as the critical hand motion is never obfuscated, which underlines the importance of selective masking. Further insight on why motion consistent noise supports better action can be had through consideration of Fig. 6, where the ability of such noise to support optical flow estimation is illustrated.

Limitations. Inevitably, there will be a trade-off between action recognition and privacy, as relevant information may be shared. In our approach, that trade-off could happen because our temporally consistent noise maintains dynamic information important for action recognition; however, it also might support motion based identification, *e.g.* gait recognition. Current protocols in privacy preserving action recognition do not consider motion-based identification. However, gait and related motion-based measurements are weak biometrics [26]; so, favouring action recognition may be apt. Moreover, if it becomes a concern, then it may be possible to apply motion perturbations that impede personal identification while maintaining action recognition.

5. Conclusion

Our work highlights that it is not necessary to train action recognition and privacy networks in an adversarial fashion for effective obfuscation of privacy attributes while maintaining strong action recognition performance. We show that a system based on local privacy templates, deep features that capture template semantics and selective noise obfuscation that is animated with source video motion can uphold privacy without hindering action recognition. Our approach is unique compared to alternative recent approaches in terms of interpretability and independence from particular action and privacy recognition algorithms. Our nine manually chosen templates, in combination with our proposed obfuscation technique outperforms other state-of-the art approaches across three different datasets.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *ECCVW*, 2022. 2, 4

[2] Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *IJCV*, 92:1–31, 2011. 6

[3] 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 *ICPR*, 2021. 3, 5, 7

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium*, 2021. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 5, 6, 7

[8] Jiawei Chen, Jonathan Wu, Janusz Konrad, and Prakash Ishwar. Semi-coupled two-stream fusion ConvNets for action recognition at extremely low resolutions. In *WACV*, 2017. 2

[9] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't I dance in the mall? Learning to mitigate scene bias in action recognition. *NeurIPS*, 32, 2019. 8

[10] Robert T. Collins, Alan J Lipton, and Takeo Kanade. Introduction to the special section on video surveillance. *IEEE TPAMI*, 22(8):745–746, 2000. 2

[11] Damien L Crone, Stefan Bode, Carsten Murawski, and Simon M Laham. The socio-moral image database (SMID): A novel stimulus set for the study of social, moral and affective processes. *PloS one*, 13(1):e0190954, 2018. 1

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[13] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. SPAct: Self-supervised privacy preservation for action recognition. In *CVPR*, 2022. 1, 2, 6

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3, 5

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, 2021. 5

[16] Christoph Feichtenhofer. X3D: Expanding architectures for efficient video recognition. In *CVPR*, 2020. 5

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *ICCV*, 2019. 5

[18] Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Ted-SPAD: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In *ICCV*, 2023. 2

[19] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, and Omer Ovenc. Scalable detection of offensive and non-compliant content/logo in product images. In *WACV*, 2020. 1

[20] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 2

[21] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5, 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[24] Carlos Hinojosa, Miguel Marquez, Henry Arguello, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. PrivHAR: Recognizing human actions from privacy-preserving lens. In *ECCV*, 2022. 1, 2, 5

[25] Filip Ilic, Thomas Pock, and Richard P Wildes. Is appearance free action recognition possible? In *ECCV*, 2022. 4, 5

[26] Anil K Jain, Patrick Flynn, and Arun A Ross. *Handbook of Biometrics*. Springer Science & Business Media, 2007. 8

[27] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 6

[28] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *International Conference on Biometrics*, 2015. 2

[29] Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre-Majorique Léger. Designing for confidence: The impact of visualizing artificial intelligence decisions. *Frontiers in Neuroscience*, 16: 883385, 2022. 2

[30] 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. 2

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[32] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal net-

works encode: Quantifying static vs. dynamic information. In *CVPR*, 2022. 8

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 2, 6

[34] Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *ECCV*, 2022. 1, 2, 5, 6, 7, 8

[35] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020. 1

[36] 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 *ICCV*, 2023. 2

[37] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys*, 54 (2):1–36, 2021. 2

[38] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. Trustworthy AI: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59, 2022. 2

[39] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 2

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 6

[42] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *ECCV*, 2018. 2

[43] Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *AAAI*, 2017. 2, 5, 6, 7

[44] Michael Ryoo, Kiyoon Kim, and Hyun Yang. Extreme low resolution activity recognition with multi-siamese embedding learning. In *AAAI*, 2018. 2, 6, 7, 8

[45] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2

[46] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, 2023. 1

[47] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004. 5, 7

[48] Vijeta Sharma, Manjari Gupta, Anil Pandey, Deepti Mishra, and Anil Kumar. A review of deep-learning-based human activity recognition on benchmark video datasets. *Applied Artificial Intelligence*, 36(1):2093705, 2022. 2

[49] Jianbo Shi and Carlo Tomasi. Good features to track. In *CVPR*, 1994. 2

[50] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 27, 2014. 2

[51] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, 2018. 2

[52] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 7

[53] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[54] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 5

[55] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, 2019. 5

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3

[57] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *CVPR*, 2018. 5

[58] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *ECCV*, 2018. 1, 5, 6

[59] 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 TPAMI*, 44(4):2126–2139, 2020. 2, 5, 6, 7, 8

[60] 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 *CVPR*, 2012. 3, 5, 7

[61] Yingying Zhu, Nandita M Nayak, and Amit K Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *CVPR*, 2013. 2