Universiteit Leiden

Through the PRISM: Privacy-Region Investigation for Surveillance Media in Video Action Recognition

Kacper Nizielski (s4068858)

Supervisors:
Dr. Erwin Bakker & Prof. Dr. Michael Lew

Bachelor Thesis – Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

August 6, 2026

Abstract

The increasing number of surveillance systems with action recognition architectures raises privacy concerns, as videos captured in public spaces may reveal sensitive personal traits even beyond facial identity. Although prior work has studied the privacy-utility trade-off problem in video action recognition, most approaches focused on applying anonymization globally or to the full person, instead of exploring how various regions individually contribute to privacy preservation and action recognition. For example, current state-of-the-art models (STPrivacy [23] or RAPrivacy [52]) achieve strong results through learned, full-video anonymization. This thesis proposes a new, structured evaluation pipeline that isolates individual video regions under different transformation conditions. The pipeline uses a Self-Correction for Human Parsing model [24] to extract five regions per frame: head, arms, legs, full body, and background. Three classical anonymization transformation types (Gaussian blur, pixelization, and solid fill masking) are applied to each region at three strength levels, which produces 45 distinct anonymization combinations. Utility is measured by top-1 accuracy of VideoMAEv2 on the HMDB51 dataset, while privacy leakage is evaluated using three attacker architectures (ResNet-50, ResNet-101, and MobileNetV2). They are trained on the VISPR dataset and evaluated on the PA-HMDB51 dataset across five sensitive attributes: face, gender, nudity, relationship, and skin color.

The results show that region choice is the dominant factor in the privacy-utility trade-off. Full body anonymization results in the most consistent drop in privacy leakage across transform types but also one of the largest losses in utility. Background anonymization produces the worst trade-off of all, as it reduces utility while leaving privacy leakage largely unchanged or, higher than baseline in case of pixelization. Head pixelization at medium strength achieves the best balance between both tasks, and arms/legs anonymization showed an insignificant effect on either metric. Within a given region, varying transformation type or strength shows a minor effect on privacy leakage, whereas utility is more sensitive to strength. A comparison against STPrivacy and RAPrivacy supports this claim, showing that full body anonymization using classical transformation, such as pixelization, can achieve the same privacy leakage levels as state-of-the-art models.

Contents

1 Introduction **1**

2 Related Work **3**
- 2.1 Privacy-preserving action recognition . . . 3
- 2.2 Face anonymization and de-identification . . . 5
- 2.3 Person/full-body/video de-identification . . . 5
- 2.4 Obfuscation methods and privacy evaluation . . . 6
- 2.5 Region-Anonymization and Privacy . . . 8

3 Fundamentals **8**
- 3.1 De-identification and anonymization . . . 8
- 3.2 Video and Frame Representation . . . 9
- 3.3 Region Masks . . . 9
- 3.4 Privacy Transformation Functions . . . 9
- 3.5 Utility and Privacy Leakage . . . 10

4 Dataset and Preprocessing **12**
- 4.1 Datasets . . . 12
  - 4.1.1 PA-HMDB51 dataset . . . 12
  - 4.1.2 VISPR Dataset . . . 14
  - 4.1.3 VP-HMDB51 Dataset . . . 14
- 4.2 Region Extraction and Mask Generation . . . 14
- 4.3 Privacy Transformations . . . 15
- 4.4 Video Reconstruction . . . 16
- 4.5 PRISM Preprocessing . . . 16

5 Methodology **20**
- 5.1 Overview of PRISM . . . 20
- 5.2 Utility evaluation: Action recognition . . . 21
- 5.3 Privacy evaluation: Attacker models . . . 21
- 5.4 Privacy-Utility Trade-off Analysis . . . 23
- 5.5 Comparison Protocol with STPrivacy and RAPrivacy . . . 24

6 Experimental Setup **25**
- 6.1 Action Recognition Training . . . 25
- 6.2 Privacy Attacker Training . . . 26
- 6.3 Action Recognizer Training (Comparison Protocol) . . . 26
- 6.4 Privacy Attacker Training (Comparison Protocol) . . . 27

7 Results **28**

8 Discussion **43**

9 Conclusions and future work **45**

References **51**

1 Introduction

For a long time, the number of surveillance cameras in public spaces has been increasing, monitoring public activity at a wide scale [32, 43]. These devices are commonly integrated with an action-recognition feature, which is a supportive tool for various parts of daily life, such as surveillance, smart environments, public transport, and safety monitoring. Moreover, compared to still images, video provides temporal information across consecutive frames, giving more valuable information, e.g., for recognizing human actions. This temporal context helps recognize activities that may differ in motion, sequence, or duration in subsequent frames. The action-recognition feature is mainly applied to detect suspicious and dangerous activities, e.g., robberies or brawls. Hence, modern systems like these can be helpful in preventing such situations before their execution [44]. Here, computer vision models, applied to recognize such actions, are crucial, with the goal of improving public safety [5, 28].

However, people want to keep their personal data private, and many are concerned about their privacy, especially with the collection of personal data by the government and monitoring systems [2, 34, 10]. During surveillance in public spaces, sensitive attributes, such as face, may be captured, which allows to identify the person. Above that, privacy is not limited to facial identity only, as many other factors can also reveal information about an individual. From cues such as motion, clothing, pose, or even background context can be used to identify someone. Consequently, concealing one part of the body or clothing might not be sufficient to protect a person's identity. Thus, it raises questions from an ethical perspective, as capturing such personal information should be minimized and protected from third parties.

The probable solution would be to use privacy filtering in such a way that recognizing the identity of a person is significantly reduced, while maintaining the accuracy of the action recognition as much as possible. Such a challenge can be described as a privacy-utility trade-off, where both the performance of the utility task, action recognition, and the protection of the identity must be maximized. This is a formidable task for action recognition in videos, since a temporal dimension is added, and both context and motion need to be considered within it. The context provides crucial information about the actions occurring in individual frames, while motion offers additional cues that help the model recognize the action more accurately. Therefore, it is expected that stronger anonymization would not affect all tasks equally, as every task uses different information from the video and might even result in decreasing the accuracy and incorrectly recognizing an action. In addition, privacy might still be leaked even if anonymization was applied to each frame. Due to the temporal context, the transition between frames might provide some personal identity attributes, leading to their leakage [23]. Thus, there are various challenges to consider.

In recent years, privacy-utility trade-offs have been widely studied to protect identity while keeping utility performance. Utility is mostly evaluated by the performance on downstream tasks such as action recognition or object detection. This allows researchers to measure how much useful information is maintained after applying anonymization transformations and to compare privacy-utility trade-offs using different methods [9]. While developing such models, it is important that privacy should be evaluated against attacker models that try to retrieve information [53]. However, privacy evaluation is not a simple task, as every attacker model could differ from each other, and hence the system should be robust against multiple attacker models. Thus, the effectiveness of privacy protection should not be tested with only one recognition model, as the model could deceive the anonymization method, which allows other models to still be able to recognize the person

[54]. Furthermore, in the privacy-utility task, applying filters in every frame might be insufficient. Even if the identity might look safe due to being not visible in every frame, privacy leakage is still possible due to aggregation across frames [9]. Recent approaches such as STPrivacy and RAPrivacy, representing the current state-of-the-art, achieve substantial results in the researched trade-off [23, 52] (STPrivacy reports Top-1 = 50.73%, cMAP = 72.48%; RAPrivacy reports Top-1 = 50.12%, cMAP = 70.12%; both models evaluated on VP-HMDB51 dataset [23] which is one of the most used and most recent benchmark datasets for this task, where higher Top-1 indicates better utility, and higher cMAP indicates more leakage). Moreover, these state-of-the-art methods apply anonymization globally and provide little insight into which visual regions are responsible for privacy leakage or utility loss. This results in outcomes being either so heavily obfuscated that they are unintelligible to the human eye, resembling television static, or interpretable only at the video level, where actions are apparent from temporal motion but cannot be inferred from individual frames. Previous work has studied privacy-preserving action recognition, low-resolution recognition, face anonymization, and learned de-identification methods. However, these works have paid less attention to systematically comparing how different identity-related regions contribute to anonymization.

Thus, although much research has been conducted on privacy-utility visual tasks, it still remains unclear how the trade-off between privacy and action recognition utility is affected by applying anonymization transformations of different strengths to different regions of the video. Previous work shows that anonymizing only a face or training on only one attacker may be insufficient to protect privacy. However, there is still a limited understanding of how action recognition performance and privacy leakage are influenced by applying transformations to different parts of the body, the full body, and the background regions across varying transformation strengths. The thesis addresses this gap with PRISM, a novel evaluation pipeline that provides the first controlled benchmark comparing fixed regions under matched transformation types and strengths in the action recognition task.

The goal of this research is to evaluate a video processing pipeline that preserves privacy, where the identity of the person is anonymized in the video by performing privacy transformations on areas such as different parts of the body, the full body, and background, and recognizing the action being performed. Therefore, the main research question is:

- How does the performance of the action recognition model change when various strengths of privacy filters are applied to different regions of the video, such as different parts of the body, the full body, and background?

With succeeding subquestions:

- What are the effects of action recognition accuracy performance by applying different transformation strengths for each region?
- How do different regions change privacy leakage while using the same transformation strength?

Formally, each input sample is a video containing human action. The utility objective is to preserve enough information in the video for the action-recognition model to correctly classify the action, while the privacy objective is to reduce enough information that could be used by an attacker model to correctly identify a person's sensitive information.

This thesis examines how the trade-off between action recognition utility and privacy protection is influenced when privacy-preserving transformations are applied to different video regions, specifically, individual body parts, full body, and background. Particularly, it compares and evaluates the impact of transformations of varying strength on both recognition accuracy and privacy leakage. Based on this analysis, the thesis aims to provide an understanding of which regions in the video are most crucial for preserving action information and which contribute most strongly to person identity privacy leakage.

The rest of the thesis is organized as follows. The next section, Section 2, provides background and prior research done in the field of privacy-preserving visual action recognition. In Section 3, the fundamentals are explained. Section 4 provides information about datasets and the preprocessing steps, including region extraction and the privacy transformation functions. Then, the following section, Section 5 describes the methodology used for the experiment, including the evaluation pipeline, utility and privacy evaluation protocols, and trade-off analysis. Then, the experimental setup is presented in Section 6. Section 7 reports the results and analyzes the privacy-utility trade-offs across different regions and transformation strengths. Finally, Sections 8 and 9 conclude the thesis and discuss limitations and directions for future research.

2 Related Work

This section provides information about prior work relevant to privacy-preserving video action recognition. Section 2.1 describes work specifically addressing the privacy-utility trade-off in action recognition. Sections 2.2 and 2.3 provide an overview of visual de-identification, by firstly focusing on face anonymization, and then extending to full body and person-level de-identification methods. Section 2.4 presents common obfuscation techniques and how privacy protection is typically evaluated across this literature. Finally, Section 2.5 defines the research gap this thesis addresses and highlights the proposed approach.

2.1 Privacy-preserving action recognition

The privacy-preserving action recognition task aims to achieve the highest performance on the utility task while removing sensitive information that can be used to identify a person from the video. Models might use cues from the context, such as motion, pose, appearance, or background scene, to categorize actions [42, 12, 13]. Therefore, removing or altering crucial information in specific action scenarios can affect performance [40]. For example, by concealing hands, the model might not recognize that a person is holding a cup, or by obscuring the background, such as a trampoline, the model might not recognize jumping action, resulting in a worse utility score. Besides that, previous research [7, 53, 54] mainly studied transformations applied to full frames or full videos and evaluated the overall privacy-utility trade-off, with only a few exceptions explicitly decomposing privacy leakage across different visual regions [18].

An early approach to privacy-preserving action recognition attempted to represent the original visual signal more confidentially. For example, researchers tried to preserve part of the motion information required for activity recognition while reducing visible personal details using extremely low-resolution video [40], though this method resulted in a limited privacy-utility trade-off. A more recent study took an alternative representation approach based on skeleton and demonstrated

that even using abstract pose alone can leak sensitive information such as identity or gender[30]. Together, these studies present important insights about privacy preservation in action recognition, that alternative representations can still contain sensitive information.

A learned optimization approach is another direction researchers explored in attempting to solve the privacy-preserving action recognition problem. This method jointly trains an anonymization function with an action model and a privacy attacker in order to preserve information about action while suppressing privacy cues. Early works focused on privacy-preserving action detection only on faces, altering facial identity while preserving utility performance [37]. This technique was later improved and extended to generalize against more adversarial frameworks, in which models were optimized against one or more attacker models. This shows that privacy depends on the attacker models, and various models might perform worse or better on privacy leakage depending on the implementation and method they use. Hence, the privacy-preserving model must be robust against attacks from varying systems, and not just one model. Given that, anonymization and privacy evaluation should be tested on various attackers. Especially, Wu et al. formalized the privacy-preserving action recognition problem as an adversarial learning framework, in which an anonymization model trains against attacker models to be capable of inferring sensitive identity attributes, while preserving action recognition [53]. For this research, they introduced the now widely used PA-HMDB51 benchmark dataset with action and privacy annotations.

More recent methods, such as STPrivacy, being one of the current state-of-the-art models, model privacy in the spatial and temporal domain rather than applying privacy to each frame [23]. Addressing the fact that, despite applying anonymization techniques in every frame and receiving decent results, privacy leakage is still possible due to the temporal sequence. By aggregating information across frames, attacker models could identify the person in that video. Consequently, the model showed that treating anonymization as a spatio-temporal problem allows for preserving motion cues relevant for action recognition, while preventing the cross-frame aggregation attacks described above. Hence, the anonymization remains temporally consistent rather than leaving frame-independent gaps that an attacker could exploit, achieving Top-1 = 50.73% on VP-HMDB51 (see Table 10). However, STPrivacy's anonymization operates by discarding action-irrelevant tokens and transforming the remaining ones within an embedding space via adversarial training. The authors report that this produces videos featuring unidentifiable silhouettes. Due to that, the people in the video become unrecognizable to a human observer [23]. In contrast to classical pixel methods(e.g., blur, pixelization), which remain visually interpretable even after anonymization.

Other research has focused on making models more robust to unseen actions and privacy attributes while reducing annotation cost. SPAct addresses both objectives simultaneously. Such an approach is particularly important as supervised anonymization methods typically learn a function tied to a predefined set of privacy attributes, and may then fail to anonymize previously unseen attributes (For instance, a model trained to conceal skin color may still leave background information unprotected). This method's self-supervised privacy-removal branch avoids this by not depending on privacy labels at all [7]. This shows that expensive privacy annotations can be omitted while still learning an effective privacy-preserving action recognition model.

RAPrivacy, being the second current state-of-the-art model, extends the research by taking a different direction [52]. It emphasizes that the anonymized videos should remain clearly readable to a human observer, reporting Top-1 = 50.12% on VP-HMDB51 dataset (see Table 10). Therefore, the researchers behind the RAPrivacy model emphasized creating privacy-preserving, human-readable videos that are also useful to human observers. Finally, Ilic et al. argue that protecting privacy

should not be globally applied to the whole frame but rather selective on different parts of the frame, which crucially supports the main goal of this thesis [18]. However, this method learns which areas to hide using a saliency map built from privacy examples chosen by people, adjusted for each image, and kept consistent across video frames using optical flow. PRISM systematically evaluates every combination of region, transformation method, and strength, enabling controlled comparisons of how obfuscating different regions affects privacy and utility.

2.2 Face anonymization and de-identification

Face anonymization and de-identification are the most common areas of research in visual privacy. It is mostly driven by the fact that the face is the most common and easily exploited identity cue in videos. Consequently, the aim is to remove or reduce sensitive facial information and preserve anonymous information such as pose or gaze. In contrast to privacy-preserving action recognition, where the main utility task is classification of the action, the face de-identification system tries to suppress information on the face, used to identify a person.

Early work in facial anonymization indicates that hiding or disrupting the face might be insufficient. It is because various methods provide different levels of protection, and most of these methods degrade the quality of the image or downstream utility. Hence, recent approaches increasingly view privacy protection as a transformation problem rather than simple face obfuscation.

More recent literature focuses on generating realistic anonymized faces, rather than simply reducing facial information. These methods try to synthesize a new face, hiding the true identity while preserving selected parts of the original face. For example, AnonymousNet creates faces by modifying or generating them, using specific features while aiming for measurable privacy goals [25]. Similarly, Cai et al. tried to replace identity and preserve insensitive facial attributes for tasks such as expression analysis, landmarks, or head pose [6]. These methods offer a privacy-utility trade-off in which generated faces remain anonymous and photorealistic, though this approach introduces a new risk. Since anonymized faces are generated by modifying features of the original face, a resulting synthetic face could coincidentally resemble another real person. This concern was acknowledged in the face-generation literature [21]. Moreover, it moves from the idea of simple face transformation and can be used for further analysis.

Other researchers developed models in which the results of face anonymization are intentionally reversible. Different models, such as UU-Net, were designed to recover the original identity under authorized conditions [14, 35]. This functionality is relevant for surveillance or security settings because the face is not identifiable and recognizable at first, but re-identification might be required later by official access. This reversibility raises another problem of its own. An attacker might attempt to deduce patterns in the encoding-decoding function to decode others if several known key-identity pairs were exposed. This concern is recognized in the literature where anonymization is password-conditioned. The methods are explicitly engineered so different keys map to sufficiently diverse output identities [14]. Li et al. took a different direction where anonymization was designed to change visual appearance but preserve biometric discriminability [22].

2.3 Person/full-body/video de-identification

Although much research has focused on face anonymization, more recent studies showed that limiting the focus to the face alone is insufficient, and the model might still leak person identity.

Covering or altering the face might not protect identity, as a model can still recognize a person by other attributes. Information, exposed on the cameras, which might seem irrelevant to people (e.g., nationality, religion, or wealth), is still private data and might be used to classify a person's characteristics. Notably, this same capability is dual-use: while it shows a privacy risk to ordinary individuals whose face happens to be obscured, it can also support security goals, such as identifying a malicious person who has deliberately concealed their face during a crime, using e.g., clothing or gait. Besides that, the video captures a large amount of information that is not necessarily connected to physiological biometrics (e.g., face). They can also record non-biometric (e.g., clothing), behavioral biometrics (e.g., gait), soft biometrics (e.g., height), or even background, and could reveal a person's identity [8, 46]. Moreover, since the video includes an additional temporal dimension, a person can also be identified by temporal motion cues. This suggests that video may be an even greater privacy risk due to temporal continuity providing additional information that can support re-identification across frames. This is the exact vulnerability that spatio-temporal anonymization methods such as STPrivacy are designed to address, by explicitly accounting for the temporal dimension during anonymization instead of treating frames independently. Such a problem resulted in more extensive and broader research beyond face anonymization, emphasizing preserving privacy through transformation application on the whole visual scene.

One of the earliest works presenting the foundations of the problem was formulated as preserving enough action information while covering the person's identity in the video [1]. This research also provided essential outcomes that privacy in the video is, by definition, more difficult than in still images, and that privacy should be protected on each frame and temporally. Moreover, the paper gives a solid understanding that privacy should be distributed through different characteristics rather than a single visible trait.

Subsequent works have strengthened this view by checking various identity information available besides the face. Especially, Dietlmeier et al. presented proof that a person may be recognized based on other visual cues, such as clothing or body appearance, within various settings and viewpoints [8]. This provides an insight that concealing other body parts, or even the surrounding context, is just as important as obfuscating the face. Dietlmeier et al.'s work investigates person re-identification by matching visual appearance across different camera viewpoints, showing that privacy leakage is distributed across different parts of the context, and that several regions of the video are used to identify a person.

Thapar et al. provided another important perspective on identity leakage, showing that the identity may be revealed even in first-person videos where the person recording never appears in the video [46]. The paper shows that identity can still leak through gain and ego-motion cues encoded in the camera trajectory itself. This provides a valuable insight into broadening the notion of video de-identification beyond visible body regions and demonstrating that privacy leakage may come only from motion signatures.

2.4 Obfuscation methods and privacy evaluation

Other researchers placed emphasis on investigating how different obfuscation methods affect the trade-off between privacy and downstream utility, instead of focusing on learning a single anonymization model. These literatures provide important understanding of the privacy-utility trade-off by not only checking whether video anonymization is possible but also showing various ways to measure privacy while preserving the utility task and whether simple transformations, such

as blur or low resolution, are adequate to prevent privacy leakage. This thesis follows the same evaluation tradition, since it tests exactly this class of classical transformations (blur, pixelization, and masking) and adopts a similarly mixed approach to measuring privacy, by combining multiple attacker architectures and metrics (binary accuracy, F1, cMAP) instead of relying on a single measure of leakage.

The most common anonymization and de-identification methods can be categorized as traditional obfuscation techniques (blur, pixelization, masking), advanced AI-based techniques (face swapping, full-body anonymization, context-aware inpainting), or procedural and data-level techniques (video cropping, metadata removal, voice alteration) [36, 4, 47, 48]. This thesis adopts methods from the first category (blur, pixelization, and masking) for the reasons detailed in Section 3.5. These methods are designed to remove identity cues recognizable by machines as simple, anonymous visual characteristics [39]. The goal of removing identity cues for a human observer refers to de-identifying a person not recognizable by this observer, whereas the main task of the de-identifying model is to transform the video in a way that the algorithm cannot recover identity information of a person. This distinction is important since a person might still be recognized by the machine, even though an identity may not be visible to a human. This thesis studies the machine recognition task, as privacy leakage is evaluated through attacker models and not by a human observer.

Most research in this area focuses on evaluating common and computationally cheap obfuscation methods such as blurring, blocking, pixelization, silhouette generation, edge filtering, or superpixels. This thesis specifically evaluates blur and pixelization, together with a solid fill masking method conceptually related to blocking. However, since the methods are straightforward, the literature indicates that they have limited effectiveness and are constructed based on evaluating functions. Prior work shows that the blocking transformation method is much better in preserving identity than blur, as it removes more visual information, but it makes the image less useful in utility tasks and for the observer [26]. This thesis's masking transformation, which fully covers a region at its highest strength, is conceptually related to blocking. Similarly, another work provides a comparison of classical transformations created primarily for surveillance and insights into the complexity of the privacy-utility trade-off [4]. Additionally, Ruchaud et al. indicate that common practical face filters, such as face blackening, pixelization, and others, may fall short on privacy protection and remain vulnerable to restoration and recognition attacks [39]. These studies provide an understanding that privacy protection should be evaluated considering the attacker and the evaluation protocol, and be based on the specific task, such as action recognition instead of face recognition.

More recent literature provides more uncommon transformations that work by abstraction or contextual suppression instead of direct identity replacement. For example, researchers invented a method of abstraction-based privacy filtering that reduces photorealistic detail while replacing selected sensitive regions with simplified clip-art representations [16]. Another research presented a technique that treats privacy as a contextual recognition problem in which, instead of anonymizing people, it learns visual models of sensitive spaces defined previously by the user and flags images taken in those environments before they are exposed to external systems [45]. A different line of research focused on privacy preservation through resolution reduction instead of anonymization [51]. In this method, the image is downsampled to multiple resolutions, and the activity recognition performance and identity leakage for cues such as face, nudity, and others. This shows that privacy is not based on a single feature.

2.5 Region-Anonymization and Privacy

In video, action recognition relies on different regions depending on the action [18]. The model might retrieve information about hands while classifying clapping or about legs and the whole body while detecting jumping. Moreover, if the model is expected to give more specific classification predictions, it might even use environment or different tools to recognize actions, such as a trampoline in classifying jumping on a trampoline or cutting vegetables with a knife. Furthermore, privacy leakage might also come from different regions of the video [27, 8, 7]. As said before, a person can be recognized by different traits, and the model might use attributes that, for the human eye, might seem irrelevant, but actually give a lot of information about the identity. Besides that, in a video, the person's identity might be spread through different frames, which can also support a machine in identifying a person [15, 31]. Even the surroundings in which a person resides might inform the machine. A model trained on more data may be able to identify a larger number of individuals. Therefore, region anonymization research is meaningful. The model should be as accurate as possible in recognizing action while using a strong but minimalist anonymization method.

The existing literature strongly indicates that privacy-utility trade-off analysis is a widely known and broadly researched problem. Despite its popularity, most prior work centers on global anonymization, full-person de-identification, or transformations learned by a neural network from training data. Only a few of them try to explicitly decompose the contribution of different visual regions [18]. There is no controlled benchmark for comparing separate regions under matched privacy strengths. Consequently, prior literature does not provide a clear answer to the research questions proposed in this thesis.

The present thesis is motivated by this gap. It proposes PRISM, a structured evaluation framework that combines five regions (head, arms, legs, full body, background) with three classical transformation types at three strength levels. This controlled design enables utility and privacy changes to be attributed to a specific named region, rather than to an opaque learned mechanism such as Ilic et al.'s saliency-based approach [18]. PRISM extends currently existing region-based anonymization research by providing a systematic analysis of how privacy and utility vary across different regions of the video.

3 Fundamentals

This section introduces the core concepts and formal notation used throughout the thesis. Section 3.1 distinguishes between de-identification and anonymization, the two related concepts about privacy protection. Sections 3.2 and 3.3 present formal representations of videos, frames, and region masks. Section 3.4 shows the general privacy transformation function. Finally, Section 3.5 defines utility and privacy leakage as two metrics used to evaluate transformations throughout the experiments.

3.1 De-identification and anonymization

One of the two processes that are crucial for the privacy-utility trade-off problem is the de-identification technique, which is explained by researchers as a process that "removes all identification information of the person from an image or video, while maintaining as much information on the

action and its context" [1]. Consequently, transforming only the face or other body part might be insufficient for preventing privacy leakage, as a person could be identified by, e.g., clothing [8, 38]. Anonymization is the other most used method in privacy protection in video. It is a process of removing or altering personally identifiable information of individuals in order for the person to be unidentifiable. The anonymization process is generally considered irreversible because, once applied, the original identity cannot be recovered from the transformed data. De-identification is often used more broadly, and while some de-identification techniques preserve a reversible link for re-identification (e.g., pseudonymization), others do not. Therefore, de-identification is not necessarily reversible or irreversible.

3.2 Video and Frame Representation

The video can be represented as a sequence of frames:

$$V = \{I_1, I_2, I_3, ..., I_T\} \tag{1}$$

where V is the video, I_t is the frame at time t, and T is the total number of frames in the video. Spatial information can be found in each frame, where the temporal information is contained in the sequence of frames.

Frames can include various spatially confined information such as appearance, body posture, skin color, clothing, and background, whereas the sequence of frames might contain movement information, such as walking, jumping, or clapping. Prior work has already shown that both of these types of information are relevant for action recognition and privacy preservation tasks.

In the experiment, the transformations are applied at the frame level. Nevertheless, the videos are reconstructed from the transformed frames so that the temporal structure is preserved.

3.3 Region Masks

The region mask indicates which pixels from the frame belong to a selected region. Hence, for a frame I_t, a binary mask for region r can be written as

$$M_t^r(x, y) \in \{0, 1\} \tag{2}$$

where (x, y) is a position of the pixel. If $M_t^r(x, y) = 1$, the pixel belongs to region r. Otherwise, it does not.

Masks allow image transformations to be applied only to selected pixels inside the mask instead of transforming the entire frame. The regions chosen in this experiment are: head, arms/hands, legs, full body, and background.

3.4 Privacy Transformation Functions

A privacy transformation modifies selected visual information in a frame with chosen filters. A transformed frame with transformation strength s can be represented as:

$$I'_t = F(I_t, M_t^r, s) \tag{3}$$

where I'_t is the transformed frame, and F is the chosen filter, I_t is the original frame at time t, M^r_t is the binary mask indicating region r at time t, and s is the strength parameter controlling the intensity of the transformation, with concrete values for each transformation type.

The transformation changes pixels inside selected regions while others remain unchanged. Such application allows for comparing the utility performance and privacy leakage of different regions under the same transformation function and strength.

The privacy transformations used for PRISM are classical visual filters, such as blur, pixelization, and a solid fill mask. These transformations were chosen because they are computationally cheap, interpretable, and can be applied consistently across different regions. Although prior studies show that these methods are insufficient in protecting sensitive information, they can still provide insight into which video regions contribute most to identity leakage.

3.5 Utility and Privacy Leakage

The effect of privacy transformations is evaluated using two metrics: utility and privacy leakage.

The utility task assesses how useful the transformed video remains for the intended task. Since action recognition is the intended task, utility is evaluated by the performance of an action-recognition model on clean and transformed videos, using various metrics.

Let $A(\cdot)$ denote the action recognition model, mapping a video to a predicted action class, and let y be the ground truth action label. For a test set of N videos:

$$U(V') = \frac{1}{N}\sum_{i=1}^{N} 1[A(V'_i) = y_i] \tag{4}$$

where $U(V')$ is the utility of the transformed videos, measured as the classification accuracy of the action recognizer, and $1[\cdot]$ is a function, equal to 1 when the prediction matches the ground truth and 0 otherwise.

Privacy is leaked when there is still remaining sensitive information after the transformation of the video. Hence, privacy leakage is evaluated using attacker models. These attacker models try to infer sensitive information from the transformed video. Depending on which information was obfuscated, the model might perform differently. If the attacker performs well, the information related to identity is still present in the video. However, if the performance decreases, the sensitive information is reduced, making these parts essential in preserving privacy. Privacy attributes in VP-HMDB51/PA-HMDB51 (see Section 4.1.1) work as binary indicators of whether sensitive information is present or inferable (e.g., predicting whether gender is inferable, not which gender), following established convention in privacy-preserving action recognition benchmarks [33, 53].

Let $P_j(\cdot)$ denote a privacy attacker j, mapping a video to a predicted sensitive attribute vector, and let a be the ground truth attribute vector. For a test set of N videos:

$$L_{j,k}(V') = \frac{1}{N}\sum_{i=1}^{N} 1[P_{j,k}(V'_i) = a_{i,k}] \tag{5}$$

where $L_{j,k}(V')$ is the privacy leakage of attacker j on attribute k measured as its prediction accuracy against ground truth attribute label

The privacy-utility trade-off presents the connection between these two outcomes. In this problem, the action-recognition performance should be preserved as much as possible while reducing

privacy leakage. Consequently, this situation indicates that the applied privacy transformation was useful.

To make these general definitions concrete, this section formalizes the following metrics used to instantiate $U(V')$ and $L_{j,k}(V')$.

Top-1 accuracy. Let $\hat{y}_i = \arg\max_c f(V_i')_c$ be the action recognizer's predicted class for video i, where $f(V_i')_c$ is the model's predicted probability for class c. Consequently, the formula is as follows:

$$\text{Top-1} = \frac{1}{N}\sum_{i=1}^{N} 1[\hat{y}_i = y_i] \tag{6}$$

Top-5 accuracy. Let $\text{Top}_5(V_i') \subseteq \{1, \ldots, C\}$ denote the set of the five classes with the highest predicted probabilities $f(V_i')_c$, i.e., the five largest values of $f(V_i')_c$ over all action classes c. The formula is as follows:

$$\text{Top-5} = \frac{1}{N}\sum_{i=1}^{N} 1[y_i \in \text{Top}_5(V_i')] \tag{7}$$

Binary accuracy. For an attacker predicting an attribute k with sigmoid score $s_{i,k} \in [0, 1]$, a binary prediction is obtained by thresholding at $\tau : \hat{a}_{i,k} = 1[s_{i,k} \geq \tau]$. The formula is as follows:

$$\text{BinAcc}_k = \frac{1}{N}\sum_{i=1}^{N} 1[\hat{a}_{i,k} = a_{i,k}] \tag{8}$$

where the specific value of τ and how $s_{i,k}$ is obtained from a video depend on the pipeline instantiation (given in Section 5.3).

F1 score. For attribute k, precision and recall are computed from the threshold predictions

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \qquad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k}, \qquad F1_k = 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}.$$

where the reported score is the mean across all K attributes:

$$\overline{F1} = \frac{1}{K}\sum_{k=1}^{K} F1_k \tag{9}$$

cMAP (class-mean Average Precision). cMAP evaluates the ranking quality of the raw scores $s_{i,k}$ across all possible thresholds. For attribute k, Average Precision is the area under the precision-recall curve traced as the threshold sweeps from 1 to 0:

$$AP_k = \int_0^1 P_k(R)\, dR, \qquad \text{cMAP} = \frac{1}{K}\sum_{k=1}^{K} AP_k \tag{10}$$

Binary accuracy and F1 measure prediction quality at a single fixed threshold, whereas cMAP measures how well the model ranks correct attributes above incorrect ones regardless of threshold.

4 Dataset and Preprocessing

This chapter describes the datasets used in this thesis and the preprocessing pipeline applied to prepare them for the experiments. Section 4.1 introduces three datasets (PA-HMDB51, VISPR, and VP-HMDB51) and explains their roles in the pipeline. Section 4.2 explains the process of extracting individual videos and converting them into masks. Section 4.3 describes the privacy transformation functions and their strength levels applied to the regions. Section 4.4 covers how transformed frames are reassembled into videos, and Section 4.5 details the additional preprocessing required for the state-of-the-art comparison experiment.

4.1 Datasets

Three datasets are used throughout this thesis. PA-HMDB51, a privacy-annotated extension of HMDB51, provides privacy attribute labels for evaluating region-based privacy transformation, while HMDB51 provides video clips and action labels used to train action recognizer. VISPR is used to pretrain the privacy attacker models on a larger set of privacy attributes before fine-tuning on PA-HMDB51. VP-HMDB51 is used for the comparison protocol against STPrivacy and RAPrivacy only, as it extends the amount of annotations to the whole HMDB51 dataset.

Dataset	**Split**	**Videos / Images**	**Purpose**
HMDB51 split 1	train	3,570	Action recognition fine-tuning
HMDB51 split 1	test	1,530	Utility evaluation (clean + transformed)
PA-HMDB51	-	515	Privacy leakage evaluation
VISPR	train	10,000	Attacker training
VISPR	val	4,167	Attacker validation
HMDB51 (STPrivacy)	train	∼4,060	Comparison: action recognizer training
PA-HMDB51 (STPrivacy)	train/test	309 / 206	Comparison: action recognizer evaluation + attacker training/evaluation
VP-HMDB51	train/test	3,570 / 1,530	Comparison: attacker training / evaluation

Table 1: Datasets used for action recognition, privacy leakage evaluation, attacker training, and STPrivacy/RAPrivacy comparison protocol.

4.1.1 PA-HMDB51 dataset

For the experiment, the PA-HMDB51 dataset, a privacy-annotated extension of HMDB51, was chosen as the main dataset for evaluating region-based privacy transformations in video action recognition. PA-HMDB51 is a widely used benchmark in privacy-preserving action recognition research and has been used by recent studies [23, 3].

The HMDB51 contains around 6,766-6,849 clips of short human-action videos, depending on the source, together with 51 human action categories, and is commonly used for action recognition [20]. HMDB51 provides three official predefined train/test splits for reproducible evaluation, whereas only split 1 is used in PRISM (see Table 1). The clips include a wide range of body movements, human interactions with objects, and human interactions with humans. The clips are stored as FMP4 compressed AVI files at 320x240 resolution and 30 fps, with moderate bitrates (∼470-770

kbps) that produce mild compression artifacts typical of MPEG-4 encoding (see Table 2). Since the official dataset is unavailable for download, the Serre Lab at Brown University provides various mirror editions gathered by the community. The dataset used in this thesis contains 6,766 clips extracted from Hugging Face, divided into 51 categories. Consequently, the PA-HMDB51 dataset was introduced for privacy-preserving video action recognition and expanded the prior dataset with privacy annotations of attributes such as skin color, face, gender, nudity, and relationship [53]. These privacy labels follow the same binary presence/absence convention. Even for attributes with several possible real-world values (e.g., gender), the dataset collapses these into a binary "can be inferred" / "cannot be inferred" label, to keep the privacy task consistent with VISPR's binary attribute annotations and to reflect whether the attacker can extract information at all. However, PA-HMDB51 is limited in size since it contains annotations only for 515 videos. Therefore, privacy attackers are only tested on this dataset, while training is performed on a separate dataset, following the same cross-dataset protocol used for PA-HMDB51 [53].

Property	**PA-HMDB51**	**VISPR**	**VP-HMDB51**
Modality	Video clips	Static images	Video clips
Codec / format	FMP4 / AVI	JPEG	FMP4 / AVI
Resolution	240p height, width 176-592 px, most common 320×240	Range 500×337 to 5760×3840	240p height, width 176-592 px, most common 320×240
Frame rate	30.0 fps	n/a	30.0 fps
Color space / bit depth	RGB, 8-bit/channel	RGB, 8-bit/channel	RGB, 8-bit/channel

Table 2: Technical statistics for the three datasets. PA-HMDB51 and VP-HMDB51 share identical underlying video files, with different annotation layers, which is why their technical rows are identical.

For the comparison with STPrivacy, PA-HMDB51 was additionally split into 60/40 train/test sets (309 training videos, 206 test videos), which matches one of the SOTA evaluation protocols [23] (see Table 1). This differs from the main experiment, where all 515 annotated PA-HMDB51 videos are used as a test set, with no train/test split applied to the dataset (see Table 1).

The dataset was chosen to evaluate performance and privacy preservation based on applying transformations with varying strengths to different regions of the video. PA-HMDB51 is especially suitable for studying utility tasks and privacy leakage since it contains both human actions, which support utility evaluation, and privacy-related information, required for the analysis of sensitive information.

Moreover, the chosen dataset provides a realistic contrast between various body poses, clothing, camera viewpoint, motion, and background, which provides better region-based analysis. Especially recognizing an action may depend on different visual cues. For example, actions such as walking or running may rely heavily on the legs, while clapping may rely on the hands and arms. In other scenarios, the background or surroundings may provide useful contextual information. In such a way, applying the same transformation to all regions may hide unnecessary sensitive information or remove important information in action recognition. Moreover, this variation is useful for this research as privacy leakage may come from several body parts, clothing, motion pattern, or background context, not only from the face.

4.1.2 VISPR Dataset

In addition to PA-HMDB51, the VISPR dataset is used for pretraining the privacy-attribute attacker models. The VISPR dataset was first presented as part of the Visual Privacy Advisor work and contains around 22,000 images annotated with 68 privacy-related visual attributes [33]. These annotations are useful in learning general privacy cues, such as skin color, gender-related information, relationships, and other sensitive visual information. The VISPR dataset is used for training the attacker model since it provides a large amount of privacy supervision data, and the PA-HMDB51 dataset is limited in size. Therefore, the privacy attacker models are trained on VISPR and evaluated on both clean and transformed PA-HMDB51 to test privacy leakage.

4.1.3 VP-HMDB51 Dataset

In addition to PA-HMDB51 and HMDB51, the VP-HMDB51 dataset is used in this thesis for the comparison with the current state-of-the-art models. The dataset was originally introduced in the paper that proposed STPrivacy, which is currently considered the state-of-the-art model [23]. In contrast to PA-HMDB51, which annotates only a subset of 515 videos, the VP-HMDB51 dataset extends it and provides privacy annotations for all 6,766 clips of HMDB51. Hence, PA-HMDB51 is a strict subset of VP-HMDB51, so all 515 annotated PA-HMDB51 videos are also contained within VP-HMDB51. More than three annotators labeled each video across the same five sensitive attributes, which already exist in PA-HMDB51: face, skin color, gender, nudity, and relationship [23]. Both datasets cover exactly these five attributes, with none added or omitted. VP-HMDB51 differs from PA-HMDB51, however, in different label formatting, since VP-HMDB51 attributes are binary, and PA-HMDB51 allows an uncertain value of 0.5 when annotators disagreed.

4.2 Region Extraction and Mask Generation

Before applying transformations to different video parts, these regions are first found and extracted. In order to do that, the Self-Correction for Human Parsing (SCHP) model was used [24]. SCHP is a parsing approach in which every pixel is assigned to a semantic body part category and was proposed to improve the reliability of human parsing models under noisy annotations.

SCHP was selected for its wide adoption and off-the-shelf availability as a pretrained extractor, which made it relatively suited for reliable per-frame region extraction, particularly given the limited computational resources available for this experiment. More recent methods (e.g., SCHNet [29], Hulk [50]) report higher mIoU on the LIP benchmark (64-66% vs. SCHP's 59.36%). However, these improvements stem from foundation-model backbones (e.g., SAM/CLIP-based). SCHNet, in particular, has no public implementation available, and Hulk's released codebase is built around a unified multi-task framework requiring substantial setup, making neither well-suited as a lightweight drop-in extractor for downstream masking pipelines like ours. Integration of the newer parsers is left for future work.

The output of the SCHP is converted into masks for the regions. The regions are defined as follows:

- head: face, hair, and other head labels when available
- arms/hands: left arm, right arm, and hand pixels when available
- legs: left leg, right leg, and foot/shoe pixels when available

- full body: union of all detected human part labels
- background: inverse of the full body mask

The masks are generated each frame. Such a solution is more reliable for detecting body parts than applying bounding boxes. Full body masking is created by combining all other human labels, while the background masking is applied to the inverse of the full body mask.

4.3 Privacy Transformations

After region detection, privacy transformations are applied only to selected regions. The rest of the frame remains unchanged. Gaussian blur, Pixelization, and solid fill are transformation types used in the experiment. Each transformation is applied with three strength levels: low, medium, and high. All transformations are applied to frames at their native resolution (240p height, width of varying 288-576px depending on the clip; most commonly 320×240), in BGR, 8-bit per channel.

Gaussian blur is a technique that removes local texture detail, resulting in smoothing image. Pixelization is a method in which the spatial resolution inside a selected region is reduced by downsampling and then upsampling this region. Masking function fills the selected region with a solid color at a fixed opacity. The fill color is set to black (RGB(0,0,0)).

The strength settings are defined as follows:

Transformation	**Regions**	**Low strength**	**Medium strength**	**High strength**
Gaussian blur	head,arms,legs	kernel size 7	kernel size 15	kernel size 31
Gaussian blur	full body, background	kernel size 15	kernel size 31	kernel size 63
Pixelization	all regions	resize to 25%	resize to 12.5%	resize to 6.25%
Masking (RGB(0,0,0))	all regions	50% opacity	75% opacity	100% opacity

The low and high strengths were chosen to compare two extreme settings. The medium level was chosen to observe whether performance changes gradually as the transformation becomes stronger.

The values within each transformation type follow a consistent progression. For pixelization, the three levels (25%, 12.5%, 6.25%) form a geometric progression, with each halving the resolution of the previous level, so that the medium level acts as a midpoint for testing proportional degradation. For masking, opacity increases linearly (50%, 75%, 100%), evenly spaced with full occlusion as the natural upper bound at high strength.

Parameters for Gaussian blur are selected differently, since a shared kernel size does not scale consistently across regions of very different physical size. A Gaussian blur kernel of size k draws on a neighborhood of radius k/2 pixels around each modified pixel. When this radius exceeds the local thickness of the region, the kernel's neighborhood is drawn from outside the region, and then the transformation is effectively governed by the surroundings. To address this, median local region thickness was measured from SCHP masks across the dataset.

For the purpose of selecting comparable blur kernel sizes, the five regions are additionally grouped into two size tiers. This grouping is used only to parameterize blur strength. Head, arms, and legs form a small tier, whereas full body and background form a large tier. Both tiers use the same kernel progression pattern ($2^n - 1 = 7, 15, 31, 63$).

4.4 Video Reconstruction

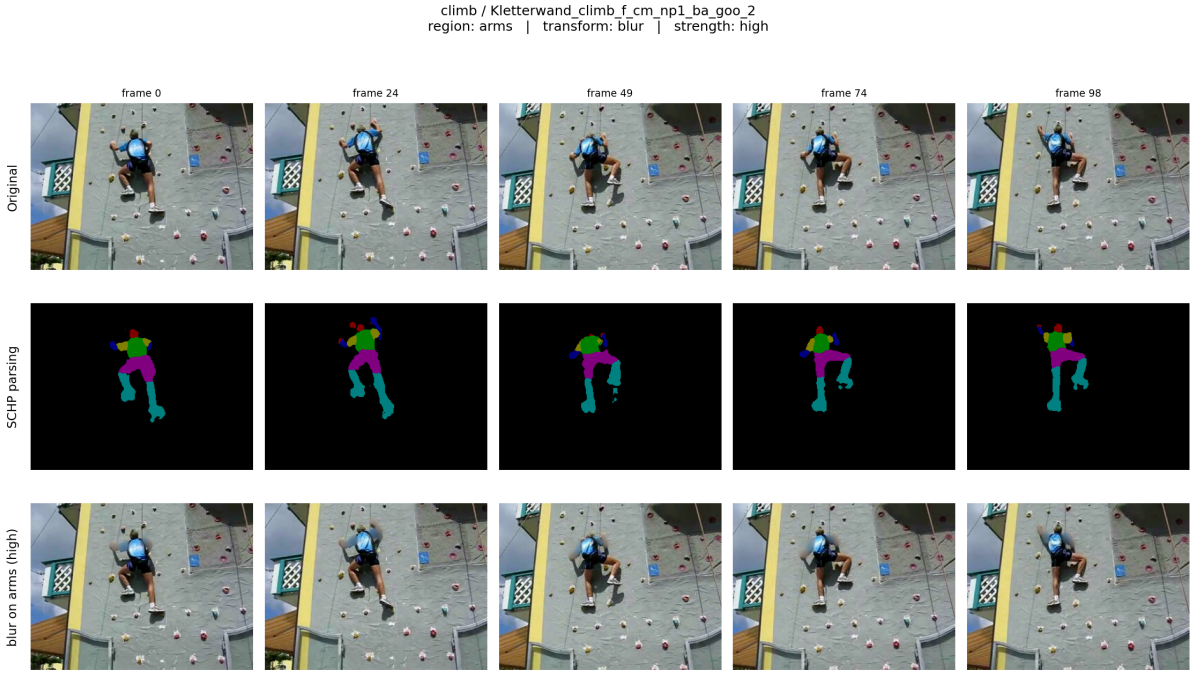
The videos, before region extraction, were decoded into individual frames. Consequently, the regions were extracted, and transformations were applied to each frame. After the transformation had been applied, the frames were reassembled into a video using the same AVI format, while preserving the original resolution and frame rate. The video was re-encoded with the MJPG codec instead of the original codec, because each frame is transformed independently and MJPG compresses each frame independently as well. This avoids artifacts that might appear if the codec relies also on neighboring frames. Similarly, the MJPG codec was applied to untransformed videos as well to ensure that measured differences in both tasks are derived from applied anonymization methods rather than unexpected side effects of the compression process.

4.5 PRISM Preprocessing

PRISM's pipeline (described fully in Section 5.1) evaluates region-based privacy transformations by training an action recognizer (VideoMAEv2) and a set of privacy attackers (ResNet-50, ResNet-101, MobileNetV2), then measuring their performance on both clean and transformed video. This section describes how the training data for these models is prepared differently for the main experiment and the comparison with STPrivacy and RAPrivacy.

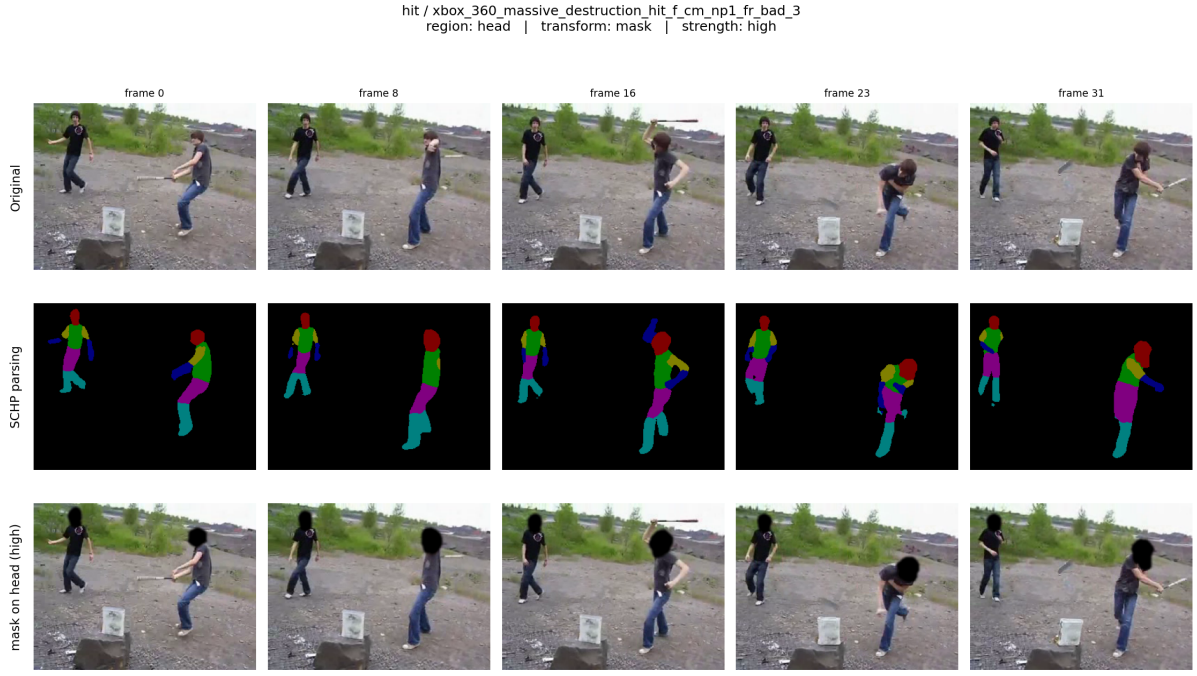
PRISM trains VideoMAEv2 and privacy attackers once on clean video and then evaluates them on transformed test videos. This process simulates a filter applied to a model that was already deployed. To ensure a reliable and valid comparison with STPrivacy and RAPrivacy, the model was trained directly on the transformed data, since both methods train their models in that way. Consequently, for the eight selected combinations with the most significant results used in this comparison, the transformed videos were also applied in the training, rather than only used in the test split.

PRISM's main pipeline (Section 5.1) trains an action recognizer and privacy attackers once on clean video, then evaluates them on transformed test videos without retraining. This simulates a filter applied to a model that was already deployed. The action recognizer is fine-tuned once on HMDB51 split 1's training set (3,570 videos) and evaluated on clean and transformed HMDB51 split 1 test videos (1,530). The attackers are trained once on clean VISPR and evaluated on clean and transformed PA-HMDB51 (515 videos). STPrivacy and RAPrivacy instead train directly on transformed data, so for the eight comparison combinations, both the action recognizer and privacy attacker are retrained per condition on transformed video rather than trained once on clean data. For the STPrivacy protocol on the transformed PA-HMDB51 60/40 split (309/206), and for the RAPrivacy protocol on the transformed HMDB51 split-1 (3,570/1,530) with VP-HMDB51 labels.

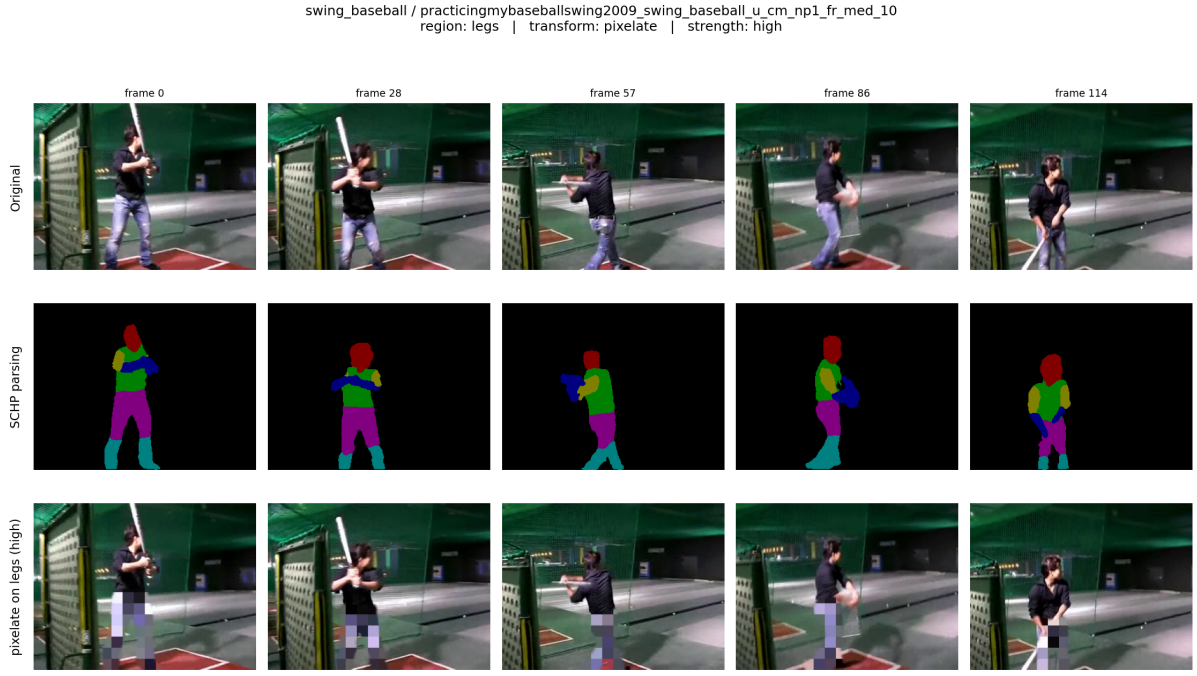


(a) Climb sequence with blur transformation on arms (high strength)

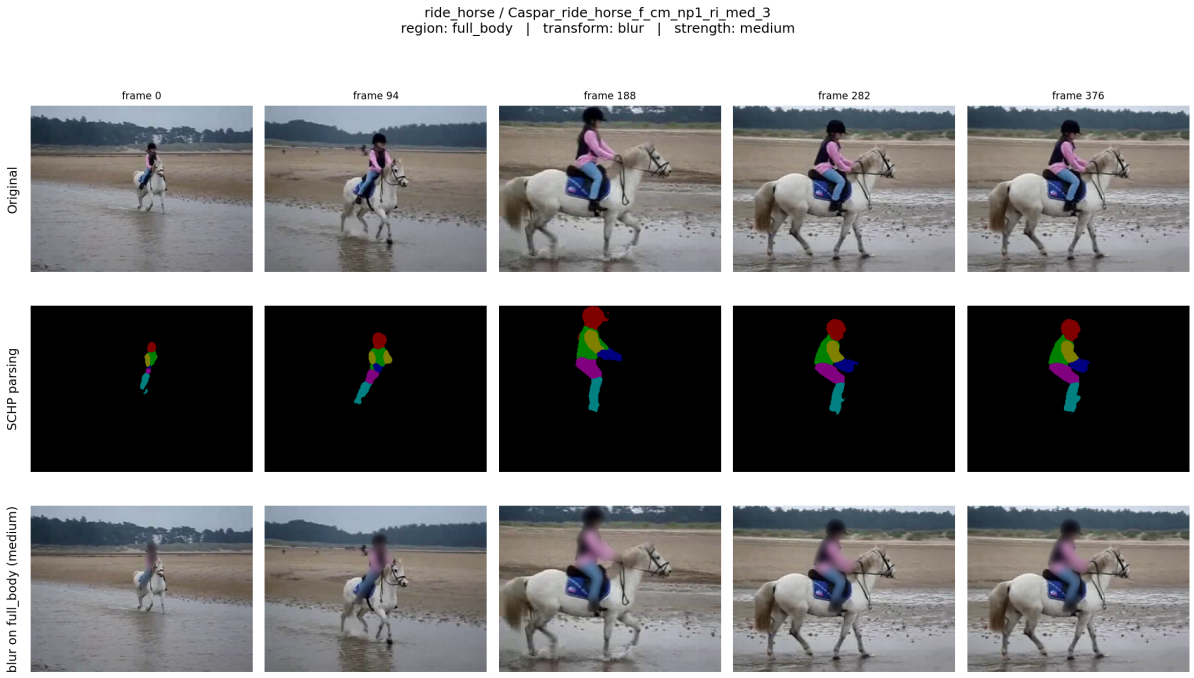
(b) Catch sequence with pixelization transformation on background (high strength)



(c) Hit sequence with mask transformation on head (high strength)



(d) Swing baseball sequence with pixelization transformation on legs (high strength)



(e) Ride horse sequence with blur transformation on full body (medium strength)

Figure 1: Visualization of image transforms applied across five frames from various videos. The subfigures show original frames, SCHP body-part segmentation, and the transformed result.

5 Methodology

5.1 Overview of PRISM

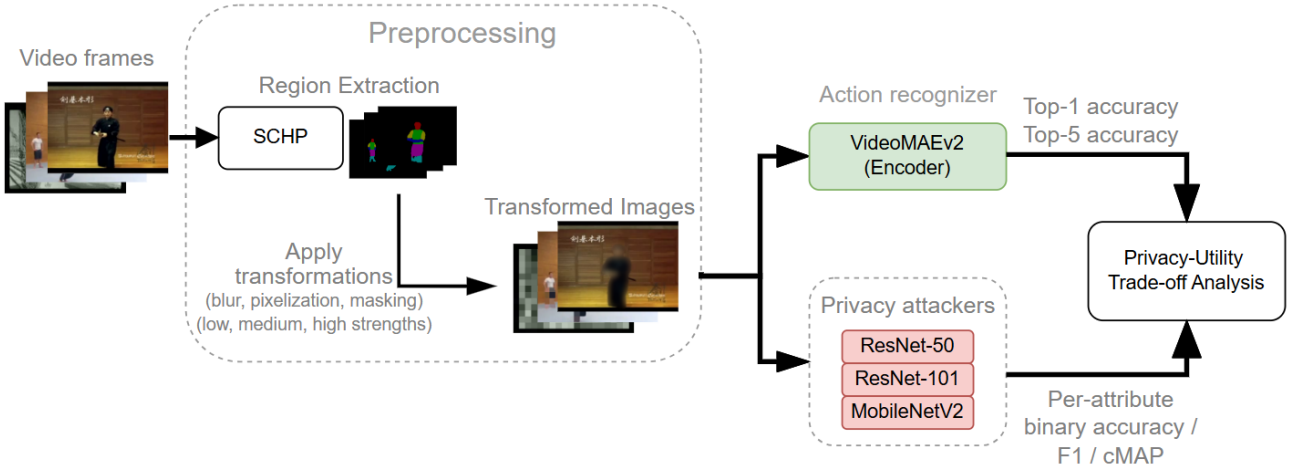


Figure 2: Overview of the privacy-preserving action recognition pipeline. Videos are changed into frames and preprocessed using the SCHP architecture to extract different regions, and then transformed using three methods: blur, pixelization, and solid fill to obfuscate chosen regions (head, arms, legs, full body, background). The transformed frames/video is evaluated by an action recognizer, VideoMAEv2, for Top-1 accuracy, while privacy attackers' leakage performance (ResNet-50, ResNet-101, MobileNetV2) is evaluated using per-attribute binary accuracy(i.e., each of the five sensitive attributes is predicted independently as present or not, formally defined in Sections 3.5 and 5.3), F1, and cMAP. The results are combined for a privacy-utility trade-off analysis.

At a high level, PRISM's evaluation framework consists of three components: an input video, a privacy transformation applied to selected regions of the video, and a final stage in which the transformed video is evaluated by both the action recognition model and the privacy attacker model. Since this thesis focuses on different regions of the video rather than transformations applied to full frames, the anonymization transformation is applied only to selected regions (head, arms/hands, legs, full body, background), and both utility and privacy metrics are evaluated accordingly to these region transformations. Specifically, this framework is instantiated as a pipeline consisting of four stages.

The evaluation pipeline consists of four stages, as shown in Figure 2:

1. Raw frames are preprocessed using SCHP [24] to extract chosen regions. Then privacy-preserving transformations are applied through all combinations of regions, transformation functions, and strengths.

2. In the utility stage, utility performance is evaluated by sending transformed videos to VideoMAEv2 [49], an action recognition model that classifies actions and reports Top-1 accuracy.

3. In the privacy stage, privacy leakage is evaluated by passing transformed videos to different attacker models: ResNet-50, ResNet-101 [17], and MobileNetV2 [41], which try to infer sensitive information. The architectures were chosen since they are well-established and reproducible baselines rather than being state-of-the-art image classifiers. Moreover, the goal is to reliably measure relative privacy leakage across transformation and region conditions rather than maximizing attacker strength.

4. Utility and privacy results are combined for privacy-utility trade-off analysis.

To the best of our knowledge, PRISM is the first framework to provide a controlled comparison and evaluation of privacy-utility trade-offs when using anonymization of fixed regions, isolating each region individually under matched transformation types and strengths, unlike prior learned or global anonymization approaches [18, 23].

5.2 Utility evaluation: Action recognition

The action recognition model chosen for this experiment is the encoder part of VideoMAEv2 with a ViT-Giant-Patch14 backbone, which reports 88.1% top-1 accuracy on HMDB51 [49]. Despite no longer representing the state-of-the-art, the model was selected because it achieves strong performance on this benchmark and because its pre-trained encoder weights are publicly available. Since our PRISM experiments focus on the evaluation of the performance degradation in accuracy after applying segment-noise anonymization transformations, the model was a practical and sufficient choice, ensuring a reasonably comparative baseline.

The model is initialized from a checkpoint, previously pretrained on the Kinetics-710 dataset, which is then fine-tuned on the HMDB51 dataset using split 1 as part of this study's setup (Section 4.1.1, see Table 1).

The fine-tuning process is performed once on clean, untransformed training data, and the weights remain frozen during evaluation. Following the mirror version of the HMDB51 evaluation protocol, the validation and test split are identical, and no separate validation split is available (see Table 1). Consequently, the same checkpoint is used to check model performance on both clean and transformed test videos without retraining. The goal of this decision is to imitate a realistic scenario where privacy-filtered data is applied to an existing model in order to measure the performance drop of the model on unseen anonymized data.

The performance of the model is based on top-1 classification accuracy over the 51 action classes on the test split. In addition to top-1 accuracy, top-5 accuracy is reported to indicate whether the correct action remains among the model's five most confident predictions.

The baseline model for this experiment is the accuracy metric on untransformed test videos. Other accuracies received from the transformation variants will be evaluated and compared against the baseline. This comparison is performed across all 45 transformation variants. It covers every combination of region, transformation type, and strength level.

5.3 Privacy evaluation: Attacker models

Privacy leakage is measured by classifier models that try to correctly predict sensitive attributes from video frames. The lower attacker accuracy implies that transformed videos are more effective in protecting privacy.

The experiment consists of 3 architectures used as privacy attackers: ResNet-50, ResNet-101, and MobileNetV2, since evaluating on one attacker model was indicated as insufficient, as the system may fool only one specific model and remain vulnerable to others. Hence, using multiple architectures tests better generalization of the privacy protection. All models are initialized from ImageNet-pretrained weights.

The choice of architectures follows a pattern seen repeatedly in privacy-preserving action recognition literature. ResNet-50 and ResNet-101 are used recurrently as standard privacy-attribute classifiers across the field, from Wu et al. [54] through to Ilic et al. [18] in 2024. Also, pretraining on VISPR for further fine-tuning is an established practice, used by Dave et al. [7] and Fioresi et al. [11].

Architecture	**Depth**	**Parameters**	**GFLOPs**	**Feature dim**	**ImageNet top-1**	**Design family**
ResNet-50	50 / 54	23.5M	4.09	2048	80.86%	Residual CNN
ResNet-101	101 / 105	42.5M	7.80	2048	81.89%	Residual CNN
MobileNetV2	53 / 53	2.2M	0.30	1280	72.15%	Lightweight CNN

Table 3: Description of the attacker models' parameters. Parameter counts reflect the attacker models used in PRISM (head is replaced with a 5-unit sigmoid output). ImageNet top-1 accuracy reflects the IMAGENET1K_V2 pretrained weights used for initialization. The table presents the models' depth, parameter count, computational cost, pooled feature dimension, ImageNet top-1 accuracy, and design family.

Each attacker is a multi-label binary classifier that predicts, for each of the five PA-HMDB51 privacy attributes (face, gender, nudity, relationship, and skin color), whether information about that attribute is inferable from the video. Each attribute is treated as an independent binary prediction, using sigmoid output to assign each attribute its own probability, and BCEWithLogitsLoss as the loss function. That allows the model to simultaneously predict any combination of attributes. The metric used in this task is binary accuracy for each attribute with a sigmoid output threshold set at $\tau = 0.5$, formally defined in Section 3.5.

The attacker models are trained on the VISPR dataset, which provides around 22,000 images annotated with 68 detailed privacy attributes, and then are fine-tuned on PA-HMDB51, containing annotations only for 515 videos (see Table 1). Since PA-HMDB51 contains only 5 attributes, the specific attributes from the VISPR dataset were mapped. The mapping is defined as follows:

VISPR attributes	**PA-HMDB51**
a9_face_complete, a10_face_partial	face
a4_gender	gender
a12_semi_nudity, a13_full_nudity	nudity
a64-a69 (personal, social, professional, competitors, spectators, views)	relationship
a16_race, a17_color	skin_color

Table 4: Mapping between VISPR attributes and PA-HMDB51 attributes.

A PA-HMDB51 attribute label is assigned to a VISPR image if at least one of the mapped VISPR labels appears in that image annotation. Each model was trained for 20 epochs using the

AdamW optimizer with a learning rate 1×10^{-4}, a batch size of 64, ImageNet normalization, and standard data augmentation consisting of random cropping and horizontal flipping.

The models are evaluated on frames extracted from untransformed and transformed PA-HMDB51 videos, with corresponding annotations assigned to each frame. For each video, 16 frames are sampled uniformly, matching VideoMAEv2's frame count for a consistent evaluation window across the action recognizer and privacy attacker models. Since the attacker architectures are 2D image classifiers with no mechanism for processing video, such a score is obtained by computing a sigmoid score independently for each frame and mean-pooling across the 16 frames, then thresholding the pooled score at $\tau = 0.5$ (Section 3.5) to obtain a binary prediction for each attribute. Averaging frame-level model outputs to obtain a video-level prediction is a well-established technique in video classification, following the same late-fusion-by-averaging approach by Simonyan and Zisserman [42] and widely adopted in subsequent work [19]. VideoMAEv2, on the other hand, processes all frames jointly. Performance is measured using binary accuracy for each attribute, mean F1 score, and class-mean Average Precision (cMAP) (see Section 3.5) against the PA-HMDB51 ground-truth labels. F1 averages the precision-recall balance for each attribute across the five attributes, providing a more robust measure under class imbalance, whereas cMAP averages the area under the precision-recall curve computed from raw sigmoid scores across attributes. In any of these scores, the accuracy drop compared to the untransformed baseline indicates how much sensitive information was removed by the transformation.

5.4 Privacy-Utility Trade-off Analysis

For each combination (region, transform, strength), two scores are evaluated: the utility measure achieved from the action recognition top-1 accuracy from VideoMAEv2, and the privacy leakage measure achieved from the per-attribute binary accuracy from the privacy attacker models.

Since three attacker models are evaluated, all three scores are noted. However, only a single privacy score per combination is derived and compared in the trade-off by taking the maximum per-attribute accuracy across the three models. This worst-case approach follows the design principle established by Wu et al. [53], who argue that the performance of a single model is not a reliable measure of privacy protection, since such defense must remain robust against the strongest available attacker rather than an average one. By averaging across attackers, it would be possible for a weak attacker score to mask a strong attacker performance. This indicates which transformation most significantly diminishes leakage against the strongest attacker, while still comparing the ability of other attackers to extract sensitive information. Moreover, the results provide insight into whether privacy protection generalizes across models. Hence, privacy attributes are kept separate throughout the analysis and are not averaged into a single score.

The trade-off is analyzed by comparing utility and worst-case privacy scores across combinations. Especially, the experiment is focused on three comparisons:

- Across regions at fixed transform and strength - which region causes the largest drop, and which preserves the most privacy.
- Across strengths at fixed region and transform - if increasing strength gradually reduces more privacy leakage, and if it reduces utility.

- Across transform types at fixed region and strength - do various transformations differ in privacy-utility trade-off.

5.5 Comparison Protocol with STPrivacy and RAPrivacy

The main pipeline (Sections 5.1-5.4) and state-of-the-art models' evaluation protocols differ drastically in design, which is why a direct numerical comparison between the main results and their published results would not be valid. The main evaluation pipeline is designed as a robustness test, where a single action recognizer and a set of privacy attackers are trained once on clean data and then tested with anonymized video that was not available in the training split. STPrivacy and RAPrivacy follow a protocol in which both the action recognizer and privacy attacker are trained and tested directly on transformed video, using the same backbone for both tasks. Because these training protocols differ considerably, directly comparing the main experiment results with SOTA's would make it difficult to distinguish true privacy-utility differences from effects caused by the evaluation protocol. Moreover, because the code for both methods' complete training is not publicly available at this moment, PRISM was adapted to approximate their reported protocol as closely as possible. Consequently, the backbone and training procedure and data split (Table 1) were altered from the main experiment in order to replicate the state-of-the-art model's setup and to make a reliable comparison possible.

Both the action recognizer and the privacy attacker are switched from the ViT-Giant action recognizer and ResNet/MobileNetV2 attackers used in the main pipeline to a ViT-S/16 backbone, matching the backbone for further comparison. The action recognizer is a VideoMAEv2 encoder, initialized from a 2D ImageNet-pretrained ViT-S/16 checkpoint.To match the method documented by STPrivacy and the 3D tublet embedding required by VideoMAEv2, the ImageNet checkpoint's patch-embedding kernel is inflated by repeating it along the temporal axis and dividing the resulting weights by the tubelet size, following the same method used by STPrivacy. The privacy attacker is a timm ViT-S/16 image classifier, initialized from ImageNet-pretrained weights. In this protocol, the privacy attacker is trained directly on privacy labels from the target dataset, rather than being pretrained on VISPR and fine-tuned through the attribute mapping in Table 4, as in the main pipeline.

Combination	**Rationale**
Clean	Baseline.
Pixelate / head / medium	Best overall privacy-utility trade-off.
Blur / head / high	Head-region performance degrades gradually with transformation strength.
Blur / full body / high	Strong privacy protection.
Pixelate / full body / medium	The best F1 among all 45 combinations.
Mask / full body / high	All three transformations produce similar results for the full-body region.
Blur / arms / high	Anonymizing the arms has little effect on either metric.
Pixelate / background / high	Represents the worst case.

Table 5: Selected combinations for the STPrivacy/RAPrivacy comparison protocol, chosen from the 45 combinations evaluated in the main experiment. Legs are excluded from this comparison, as the main experiment showed arms and legs behave almost identically across both metrics.

Because the comparison protocol requires training each combination individually, only 8 of the 45 combinations are evaluated. The selected conditions include the baseline, the combination with

the best privacy-utility trade-off, and combinations that achieved the strong individual metrics in the main experiment. Besides, the chosen combinations cover most regions (head, arms, full body, background, with legs excluded), including all possible high-strength conditions (see Table 5).

Protocol	Task	Training data	Test data	Notes
STPrivacy (PA-HMDB51)	Action recognizer	Random 60% of HMDB51 (∼4,060 videos)	Remaining 40% of PA-HMDB51 (206)	HMDB51 is used for training because PA-HMDB51 is only a small subset.
STPrivacy (PA-HMDB51)	Privacy attacker	Random 60% of PA-HMDB51 (309)	Remaining 40% of PA-HMDB51 (206)	Uncertain labels (0.5) are mapped to 0 for compatibility with binary classification.
RAPrivacy (VP-HMDB51)	Both	Official HMDB51 split-1 training set (3,570)	Official HMDB51 split-1 test set (1,530)	Uses the privacy labels provided by VP-HMDB51.

Table 6: Training and evaluation combinations used for the STPrivacy and RAPrivacy comparison protocols. The configurations were adapted to follow the procedures described in the research papers as closely as possible since the implementations of complete training are not publicly available. The 60/40 splits of HMDB51 and PA-HMDB51 were generated independently, each using the same fixed random seed for reproducibility.

Utility and privacy leakage are computed similarly to the main pipeline: top-1 accuracy for the action recognizer, and F1 and cMAP for the privacy attacker.

6 Experimental Setup

6.1 Action Recognition Training

VideoMAEv2 fine-tuning followed the official fine-tuning configuration released by the authors for ViT-Giant on HMDB51, initialized from the same K710-pretrained checkpoint used in this thesis (`vit_g_hybrid_pt_1200e_k710_ft.pth`). Training was performed using one NVIDIA A100 80GB GPU per task, provided by the ALICE supercomputer at LIACS. The hyperparameters used are presented as follows:

Model: ViT-Giant-Patch14, with a modified classification head adjusted for 51 classes for the HMDB51 dataset

Checkpoint: K710-pretrained (`vit_g_hybrid_pt_1200e_k710_ft.pth`)

Optimizer: AdamW, $\beta = (0.9, 0.999)$

Learning rate: 5×10^{-4}, with cosine decay

Layer-wise learning-rate decay: 0.90

Weight decay: 0.05

Epochs: 15, with a 5-epoch warm-up period

Batch size: 3, with gradient accumulation over 3 steps

Stochastic depth rate: 0.35

Classification head dropout rate: 0.5

Input: 16 frames sampled at a stride of 2, resized to 224×224

Evaluation: single segment, single crop

These values match VideoMAEv2's official hyperparameter values. Two adaptations were made to adjust the training to available hardware. First, the official script distributes training across 8 GPUs with a per-GPU batch size of 3, whereas this thesis uses a single GPU. This allows gradients to be accumulated over three training steps before updating the model, which increases the effective batch size from 3 to 9 while remaining within the memory limit of the GPU. Second, the official evaluation uses 5 temporal segments and 3 spatial crops, producing multiple predictions for each video that are subsequently averaged. To reduce computational cost, this thesis evaluates each video using only one segment and one crop.

6.2 Privacy Attacker Training

The privacy attacker models (ResNet-50, ResNet-101, MobileNetV2) were trained using standard transfer learning hyperparameters for fine-tuning ImageNet-pretrained CNN classifiers with the AdamW optimizer. All attacker architectures were initialized from torchvision's IMAGENET1K_V2 pretrained weights. The hyperparameter values below reflect commonly used defaults for this class of problem. Training was performed using an NVIDIA RTX 2080Ti 11GB GPU per task, provided by the ALICE supercomputer at LIACS.

Architectures: ResNet-50, ResNet-101, MobileNetV2, each initialized from ImageNet-pretrained weights

Optimizer: AdamW

Learning rate: 1×10^{-4}

Weight decay: 1×10^{-4}

Batch size: 64

Epochs: 20

Loss function: `BCEWithLogitsLoss` (multi-label binary classification, one output per attribute)

Input: resized to 224×224, normalized using ImageNet statistics

Training augmentation: `RandomResizedCrop`, `RandomHorizontalFlip`

Evaluation preprocessing: resize to 256×256, center crop to 224×224

Training dataset: VISPR training split

6.3 Action Recognizer Training (Comparison Protocol)

For the comparison protocol, the action recognizer was trained with the ViT-S/16 architecture to match the STPrivacy and RAPrivacy protocol. The hyperparameters were adapted to the VideMAEv2 configuration, as the official authors do not release an official fine-tuning script for

ViT-S. Training was performed on an NVIDIA A100 (80GB) GPU, provided by the ALICE supercomputer at LIACS.

Architecture: ViT-S/16

Optimizer: AdamW

Learning rate: 5×10^{-4}

Layer-wise learning-rate decay: 0.75

Weight decay: 0.05

Epochs: 65, with an 8-epoch warm-up period

Data augmentation: fully disabled (color jitter, RandAugment, label smoothing, random erasing, mixup)

Stochastic depth (drop path) rate: 0.1

Classification head dropout rate: 0.0

Input: 16 frames sampled at a stride of 2, matching the main pipeline

Batch size: 16

The layer-wise learning-rate decay, stochastic depth rate, and classification head dropout rate match the VideoMAEv2 official fine-tuning configuration for ViT-Base. A decay of 0.75 is also used in the VideoMAEv2 ViT-Large recipe. This allows consistency across all VideoMAEv2 backbones smaller than ViT-Giant. The weight decay matches the default value built into the VideoMAEv2 fine-tuning script and is used consistently across all VideoMAEv2 backbone sizes. The learning rate was kept identical to the ViT-Giant learning rate to keep the two action recognizer setups consistent. The number of epochs, warm-up period, and the decision to fully disable data augmentation were chosen empirically. Previous training runs using shorter schedules and standard augmentation failed to converge.

6.4 Privacy Attacker Training (Comparison Protocol)

For the comparison protocol, the privacy attacker was trained using the ViT-S/16 architecture to match the STPrivacy and RAPrivacy protocol. As with the action recognition model adjusted for the comparison protocol, the values below were chosen either for consistency with the main experiment or due to hardware constraints. Training was performed on an NVIDIA RTX 2080Ti (11GB) GPU, provided by the ALICE supercomputer at LIACS.

Architecture: ViT-S/16

Optimizer: AdamW

Learning rate: 1×10^{-4}

Weight decay: 0.05

Epochs: 20

Loss function: `BCEWithLogitsLoss`

Batch size: 8

Evaluation: predictions mean-pooled over 16 uniformly sampled frames per video to obtain an overall video-level score

Training data: privacy labels from the target dataset directly (PA-HMDB51 or VP-HMDB51, depending on the protocol), rather than VISPR pretraining with attribute mapping as in the main pipeline.

The learning rate and loss function match the main experiment's privacy attackers. The weight decay, by contrast, deviates from the main experiment and matches the value used for the ViT-S/16 backbone, making the weight decay value more appropriate for ViT. The number of epochs matches the training in the main experiment for consistency. The batch size was constrained by the available hardware

7 Results

Utility and Privacy Leakage by Transformation Type

- Across all three transform types, arms and legs stay close to baseline regardless of strength, with differences rarely exceeding 1-2 percentage points.
- *Blur*: head degrades minimally (WC 88.9% to 87.7%); full body drops substantially in both utility (86.7% to 79.1% at high strength) and privacy leakage (approximately 19–28pp) but still leaks on some attributes; background degrades utility (86.7% to 78.0% at high strength) with minimal privacy benefit.
- *Pixelization*: medium-strength head achieves the best trade-off in the entire table (81.7% worst-case); full body gives the strongest attribute-level protection (34pp drop); background is the most abnormal result, with a moderate utility drop (86.7% to 74.8% at high strength) while worst-case leakage rises to 90%, worse than the clean baseline.
- *Masking*: head stays close to baseline, with low-strength head masking (WC 89.2%) actually 0.6pp above the clean baseline; in contrast, full body gives the strongest privacy reduction across all strengths (WC 84.6% at low strength to 83.3% at high strength), but shows a drastic drop in utility at high strength (85.7% to 70.7%); background at high strength has the lowest utility in the whole masking section (68.0%) with comparatively weak privacy gains (the worst trade-off of this transform type).

Table 7 provides the accuracy for baseline and all other combinations within every privacy attacker and action recognition model, and worst-case privacy leakage. Worst-case presented in the results refers to the configuration that results in the highest privacy leakage, measured by the best accuracy achieved through all attacker models and attribute combinations. The outcomes in bold refer to the best accuracy for each column, whereas the underlined value shows the second best.

			Utility		Privacy leakage					
Transform	Region	Strength	Top-1↑	Top-5↑	RN50↓	RN101↓	MNV2↓	WC Acc↓	WC F1↓	cMAP↓
clean	-	-	86.7	98.3	71.8	74.3	71.1	88.6	61.3	77.1
blur	head	low	86.5	98.1	70.9	74.6	70.4	88.9	61.1	76.6
blur	head	medium	86.1	98.1	69.3	74.4	69.8	88.3	60.8	75.7
blur	head	high	85.2	97.8	68.4	73.6	68.9	87.7	59.3	74.3
blur	arms	low	86.8	98.2	71.6	74.6	71.2	88.8	61.6	76.8
blur	arms	medium	86.8	98.2	71.5	74.7	70.8	88.9	60.5	76.2
blur	arms	high	86.6	98.3	71.4	74.7	71.0	88.9	60.1	75.6
blur	legs	low	87.0	98.4	71.8	74.6	71.1	88.1	61.6	76.9
blur	legs	medium	87.0	98.4	71.6	74.5	70.8	87.6	61.5	76.9
blur	legs	high	86.9	98.3	71.1	74.1	70.6	86.8	60.9	77.3
blur	full_body	low	85.8	98.2	62.7	71.5	67.0	83.7	56.1	73.6
blur	full_body	medium	84.4	97.5	52.2	65.8	58.8	83.1	47.5	70.9
blur	full_body	high	79.1	96.1	44.3	55.0	48.7	82.9	32.6	67.2
blur	background	low	86.3	97.8	68.4	73.3	69.8	87.9	60.7	76.0
blur	background	medium	84.2	97.1	66.3	71.7	64.3	85.4	59.5	76.3
blur	background	high	78.0	94.5	66.2	67.8	61.2	86.2	54.5	76.3
pixelate	head	low	86.3	98.1	67.1	71.9	66.9	84.3	55.4	73.6
pixelate	head	medium	85.5	97.7	64.9	69.3	66.8	**81.7**	52.9	72.4
pixelate	head	high	82.3	96.7	65.7	69.7	66.3	83.1	53.4	72.8
pixelate	arms	low	86.3	98.0	69.5	73.9	70.0	87.3	59.2	75.4
pixelate	arms	medium	85.9	98.2	69.5	73.3	69.8	85.7	58.3	74.8
pixelate	arms	high	85.7	98.2	70.1	73.5	70.1	85.9	58.8	74.6
pixelate	legs	low	**87.2**	98.4	69.7	74.8	70.2	86.8	61.6	76.7
pixelate	legs	medium	86.9	98.2	69.8	73.3	70.2	85.8	60.2	76.3
pixelate	legs	high	86.6	98.4	70.5	73.5	70.2	85.4	60.7	76.3
pixelate	full_body	low	86.0	98.0	43.6	59.6	41.3	83.3	39.9	68.9
pixelate	full_body	medium	81.8	96.2	**37.8**	**38.7**	**39.5**	82.9	**8.3**	63.1
pixelate	full_body	high	73.8	92.9	38.1	38.9	40.4	83.1	11.3	**62.8**
pixelate	background	low	86.6	98.1	71.0	75.2	71.9	88.1	61.9	77.2
pixelate	background	medium	83.3	97.2	72.7	74.5	72.1	88.1	62.2	78.8
pixelate	background	high	74.8	93.1	73.9	74.8	72.7	90.0	62.7	78.7
mask	head	low	86.5	98.4	70.5	74.6	70.1	89.2	60.9	76.1
mask	head	medium	85.9	98.2	69.4	73.9	69.9	88.9	60.2	75.2
mask	head	high	81.6	95.8	65.7	73.4	68.4	88.1	58.9	73.5
mask	arms	low	86.2	98.2	71.1	74.1	70.7	88.1	61.4	76.4
mask	arms	medium	85.8	98.0	70.5	73.8	70.9	87.6	61.1	76.2
mask	arms	high	85.2	98.4	69.4	73.7	70.7	86.8	60.7	75.8
mask	legs	low	86.9	98.4	71.4	74.7	71.1	88.6	61.6	76.7
mask	legs	medium	86.3	98.4	71.3	74.2	70.8	88.4	61.7	76.1
mask	legs	high	85.8	98.1	70.7	74.2	70.6	87.8	62.0	76.0
mask	full_body	low	86.7	97.9	67.2	72.9	69.3	84.6	58.9	75.3
mask	full_body	medium	85.7	97.9	58.5	70.2	65.4	83.9	54.7	73.3
mask	full_body	high	70.7	89.8	40.1	57.1	56.3	83.3	42.1	66.2

			Utility		Privacy leakage					
Transform	Region	Strength	Top-1↑	Top-5↑	RN50↓	RN101↓	MNV2↓	WC Acc↓	WC F1↓	cMAP↓
mask	background	low	85.4	98.2	71.5	74.2	71.5	88.4	60.6	76.6
mask	background	medium	82.9	97.4	70.7	74.3	71.0	87.3	61.3	76.9
mask	background	high	68.0	87.2	62.3	70.1	64.9	86.6	57.7	76.7

Table 7: Utility and privacy leakage for all combinations. Utility (VideoMAEv2) reports top-1 accuracy after transformation application, while privacy attackers (ResNet-50, ResNet-101, and MobileNetV2) report attacker accuracy. All values are accuracy percentages(%). Arrows indicate whether higher (↑) or lower (↓) is the better outcome for each metric. Bold figures show the best outcome for the dedicated column, while underlined shows the second best. Worst-case is the maximum attacker accuracy achieved across all evaluated attributes and attacker models, representing the strongest observed privacy leakage. RN50/RN101 = ResNet-50/101, MNV2 = MobileNetV2, WC = Worst-case.

Cross-Metric and Cross-Transformation Patterns

Region	**Top-1 range**	**WC Acc range**	**F1 range**	**cMAP range**
Head	81.6-86.5	81.7-89.2	52.9-61.1	72.4-76.6
Arms	85.2-86.8	85.7-88.9	58.3-61.6	74.6-76.8
Legs	85.8-87.2	85.4-88.6	60.2-62.0	76.0-77.3
Full body	70.7-86.7	82.9-84.6	8.3-58.9	62.8-75.3
Background	68.0-86.6	85.4-90.0	54.5-62.7	76.0-78.8

Table 8: Range (minimum–maximum) of each metric within each region, summarized from Table 7. Values are not directly comparable across metrics, as each is measured on its own scale.

- Applying pixelization on full body region at medium and high strength resulted in the lowest F1 scores (8.3 and 11.3, respectively) and lowest cMAP (63.1 and 62.8, respectively) across all combinations. Head pixelization at medium strength is the strongest reduction for that region (F1 = 52.9, cMAP = 72.4). Arms and legs remain closest to baseline across both metrics in every condition (see Table 8).
- Background anonymization affects F1 and cMAP inconsistently across transformation types. Blur reduces both metrics below baseline, while pixelization raises above it (F1 to 62.7, cMAP to 78.7), which makes it the only condition where these metrics exceed the baseline.
- At high strength, all three transform types converge to similar privacy scores on full body (blur 82.9%, pixelate 83.1%, mask 83.3%), showing region matters more than transform type.
- Background is the most damaging region at high strength for blur (78.0%) and masking (68.0%) for utility task, whereas full body is more damaging once pixelization is applied (73.8% vs. background's 74.8%)
- ResNet-101 is the strongest attacker overall, though ResNet-50 drops most sharply under full-body transformations.

Figure 3 shows the accuracy scores for all the models evaluated on untransformed HMDB51/PA-HMDB51 test sets. VideoMAEv2 received 86.7% top-1 and 98.3% top-5 accuracy on the HMDB51 dataset, being a foundation of the utility metric to which all configurations will be compared. The three privacy attacker models perform consistently, achieving mean per-attribute accuracy scores ranging from 71.1% to 74.3% on the PA-HMDB51 dataset. Among the attacker models, ResNet-101 showed the strongest privacy leakage at 74.3%, whereas ResNet-50 and MobileNetV2 are slightly weaker, achieving 71.8% and 71.1%. Table 7 indicates that the worst-case privacy leakage across all attackers and attributes is 88.6%, performing much higher than the attribute averages of every attacker. Consequently, the baseline provides a good starting point, since the results show that models are non-trivial but also imperfect.

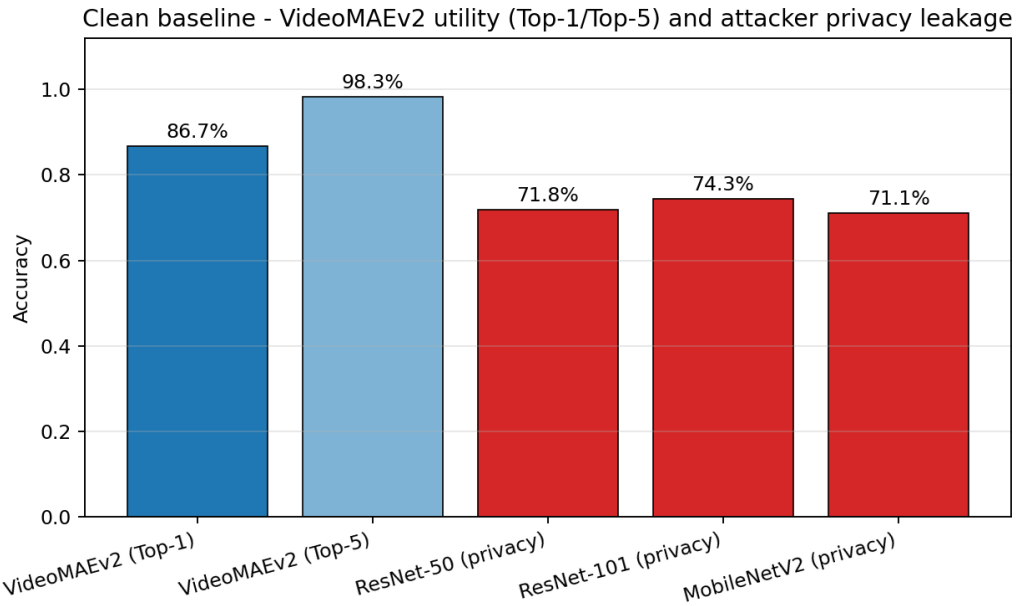


Figure 3: Baseline metrics for both action recognizer and all privacy attackers on untransformed videos. The action recognizer (blue bar), VideoMAEv2, achieves 86.7% in top-1 accuracy on the HMDB51 test set. The privacy attackers (red bars), ResNet-50, ResNet-101, and MobileNetV2 achieve around 71-74% accuracies, averaged over the five privacy attributes, on clean PA-HMDB51 videos.

Privacy-Utility Trade-off

- Points cluster by region rather than by transform type, confirming region choice as the dominant factor in the trade-off. Most combinations are clustered near the baseline in the middle-right of the plot, highlighting limited change from the baseline for most of the combinations.
- Full body is the region that most consistently shifts toward stronger privacy protection, forming a trail across all three transforms at a clear utility cost, though pixelization alone shifts head even further; arms and legs cluster tightly at baseline regardless of transform. At high strength, the three transform types on full body converge to similar privacy scores (WC

Acc 82.9–83.3%) while differing considerably in utility: mask (70.7%), pixelate (73.8%), blur (79.1%).

- Head splits visually by transform type: pixelate moves toward better privacy, while blur and masking stay near baseline regardless of strength. Head pixelization across all three strengths (Top-1 86.3/85.5/82.3%, WC Acc 84.3/81.7/83.1%) is the only cluster that consistently falls in the lower-right zone, confirming the best trade-off.
- Background pixelization is the single worst point in the plot, with worst-case leakage rising to 90.0% while utility drops to 74.8%. Blur and masking on the background also fail to reduce leakage at high strength (WC Acc 86.2% and 86.6%) while utility drops severely (78.0% and 68.0%).

Figure 4 shows the privacy-utility trade-off scatter plot for all 45 anonymization combinations, where the x-axis shows utility and the y-axis shows the worst-case privacy leakage. The most wanted configurations appear in the lower-right corner of the figure, where the combination has high utility but a low privacy leakage score. The black star in the plot presents the untransformed baseline. The color encodes region, and the shape encodes transformation type.

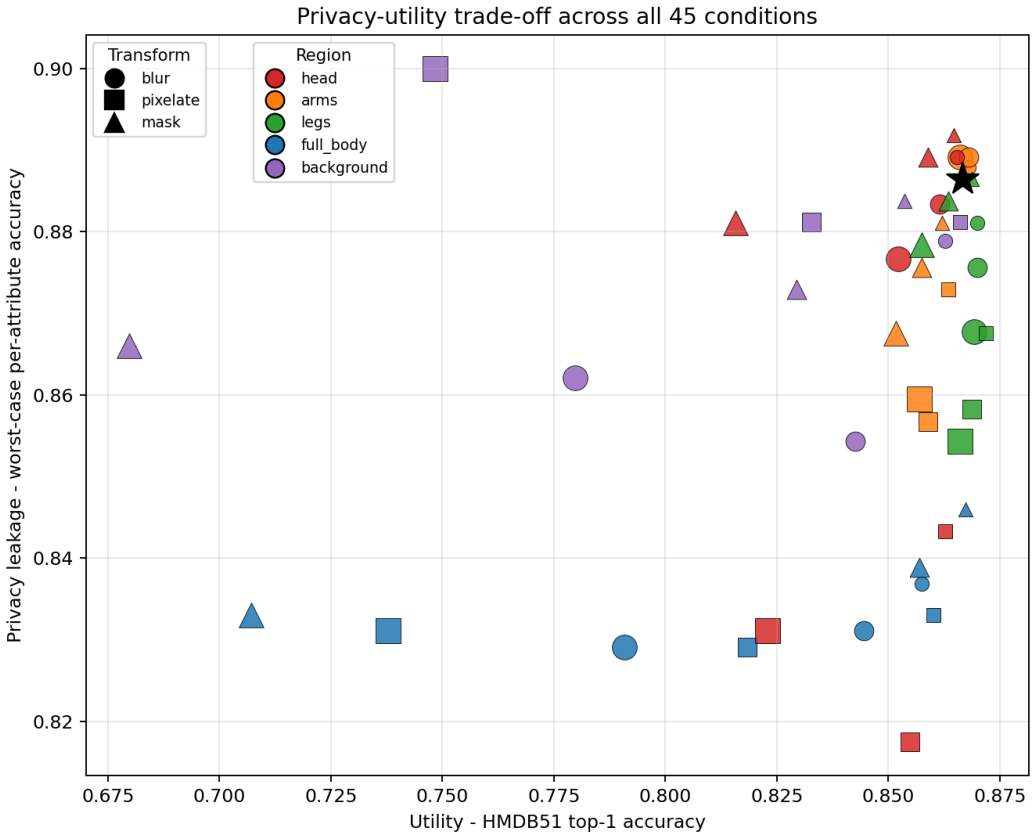
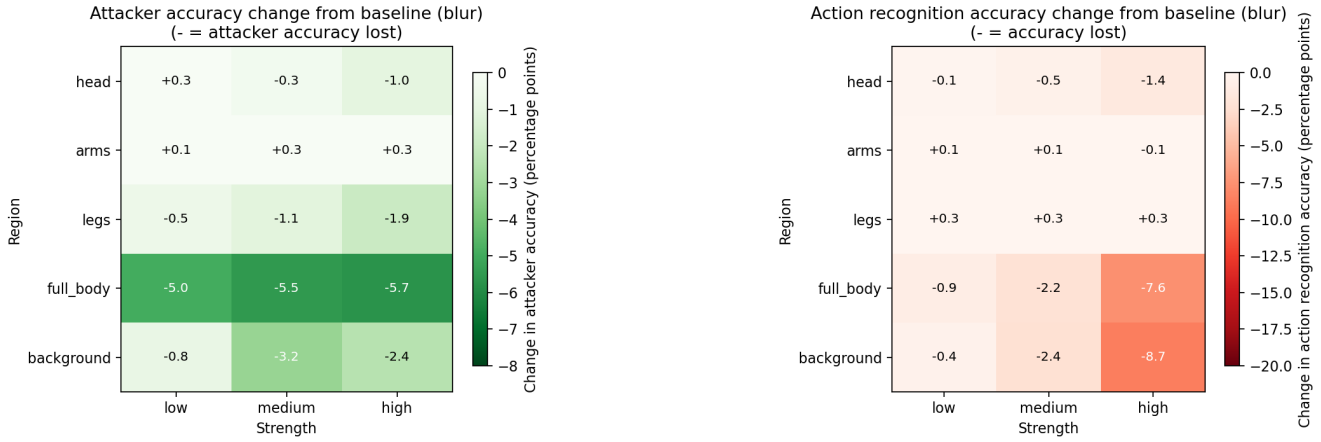


Figure 4: Privacy-utility trade-off across 45 anonymization combinations. The star indicates the untransformed baseline model. Points in the lower right corner show the most wanted configurations. Full body transformation most successfully lowers privacy leakage, however, at a cost of reducing utility accuracy. Most of the configurations fall within the range of the untransformed baseline, showing that transformations generally preserve the same privacy-utility trade-off as the baseline. The head pixelization combination indicates the best privacy-utility trade-off.

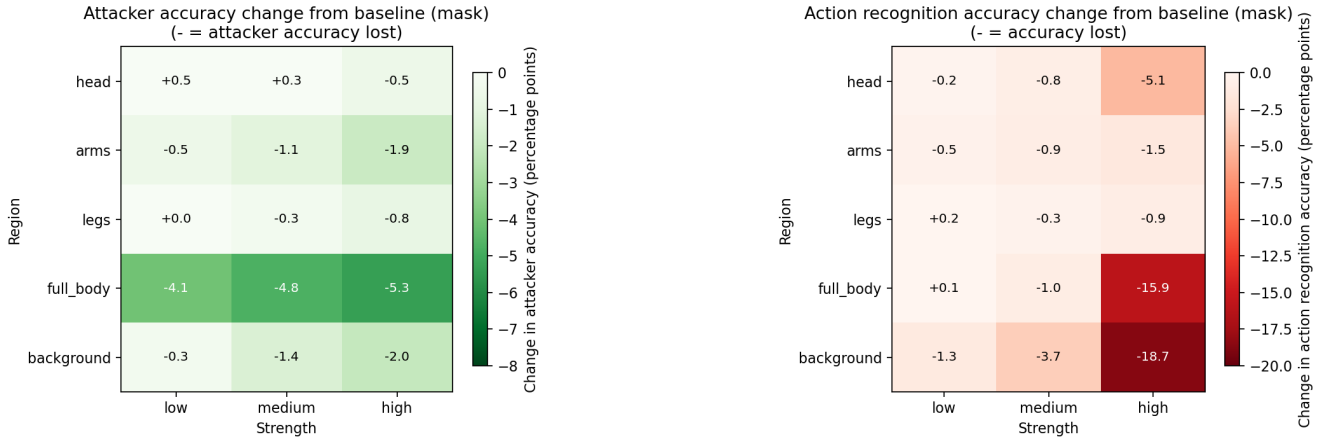
Accuracy Change from Baseline

- In the privacy heatmap, full body causes the largest shifts for blur and masking (-5.7pp and -5.3pp at high strength, respectively), while head produces the largest shift under pixelization (-6.9pp at medium strength). In the utility heatmap, background produces the single largest drop of any cell in the entire figure (-18.7pp, masking/high), exceeding full body's largest utility drop (-15.9pp, masking/high).
- Arms and legs show minimal variation in both heatmaps across all transform types (±3.2pp of baseline in the most extreme case, pixelation/legs/high), indicating that these regions have limited influence on either metric.
- Masking shows the head and background regions decreasing gradually as strength increases (head privacy: +0.5 to -0.5pp; background privacy: -0.3 to -2.0pp), similar to blur's head region (privacy: +0.3 to -1.0pp), whereas blur's background changes inconsistently (-0.8, -3.2, -2.4pp). Head masking at high strength presents an unprofitable trade-off: utility drops moderately (-5.1pp) while privacy leakage barely changes (-0.5pp).
- Pixelization's head and full body results are inconsistent, with medium strength reducing privacy leakage more than high strength (head: -4.3, -6.9, -5.5pp; full body: -5.3, -5.7, -5.5pp).
- High-strength background pixelization produces the single largest increase in privacy leakage relative to baseline of any cell (+1.4pp, the only positive value in the entire privacy heatmap), while also causing a large utility drop (-11.8pp).

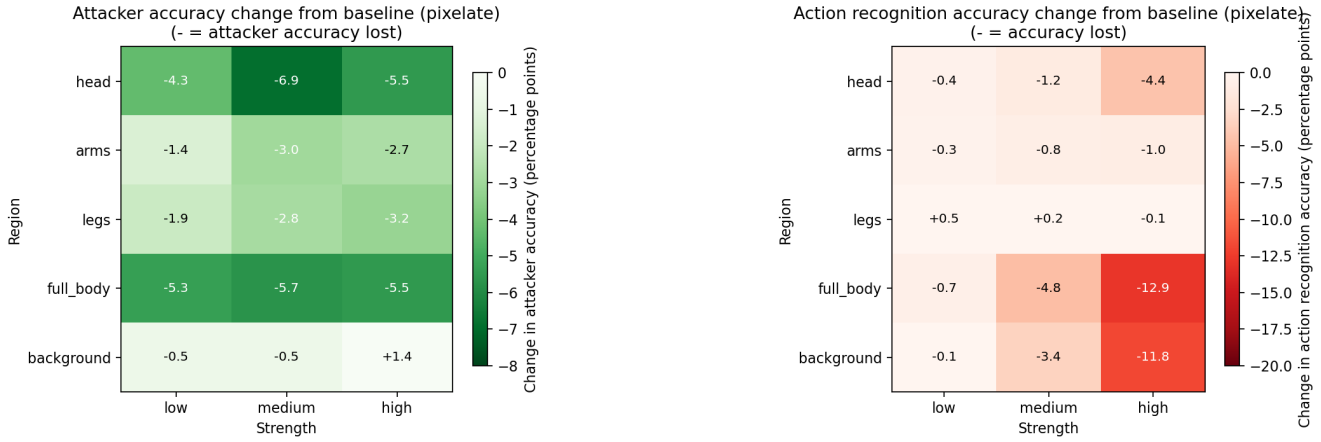
Figure 5 displays heatmaps of accuracy differences from baseline across all 45 combinations. The top row shows the changes in privacy leakage, whereas the bottom row shows changes in utility. Negative values indicate accuracy lost relative to baseline, where darker cells indicate a larger drop and lighter cells show a lighter drop.



(a) Blur



(b) Solid fill



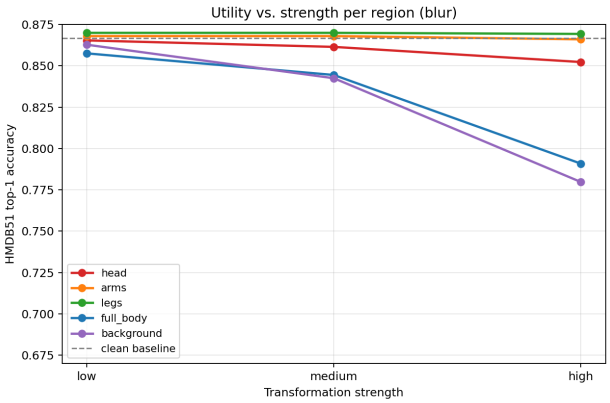
(c) Pixelization

Figure 5: Accuracy differences in action-recognition utility and privacy leakage relative to the clean baseline for all evaluated transformation configurations. Negative values indicate a reduction in accuracy relative to the baseline. Each heatmap reports the transformed region on the vertical axis and the transformation strength on the horizontal axis. Full-body transformations generally produce the largest accuracy reductions. Background and head transformations also produce substantial reductions in some cases, whereas arm and leg transformations tend to have less influence.

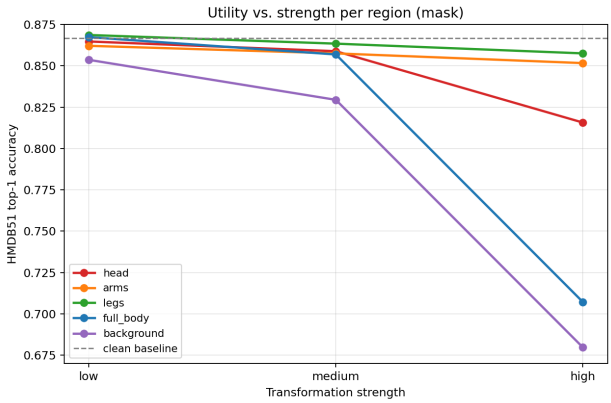
Utility Degradation by Region and Strength

- Arms and legs remain flat and near baseline across all strengths and transform types, with the largest deviation being -1.5pp.
- Head degrades gradually with strength, with a steeper decline at high strength, particularly under masking (86.5% $\rightarrow$ 81.6%) and pixelization (86.3% $\rightarrow$ 82.3%), while blur/head declines more mildly (86.5% $\rightarrow$ 85.2%).
- Full body and background decline sharply at high strength across all transform types. Full body is worst for pixelization (73.8% vs. background's 74.8%), while background is worst for blur (78.0% vs. full body's 79.1%) and masking (68.0% vs. full body's 70.7%).
- Blur and pixelization degrade gradually with strength, while masking shows a much sharper, non-linear drop at high strength.
- Region determines whether utility is strength-sensitive at all: arms and legs remain flat regardless of strength, while full body and background decline sharply as strength increases. Strength shapes the magnitude of degradation within a strength-sensitive region, but region determines whether that sensitivity exists in the first place.

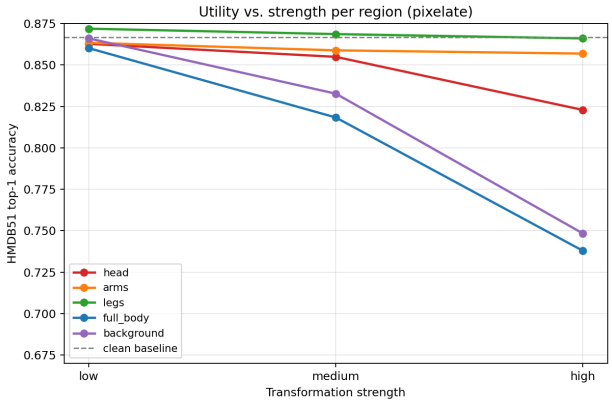
Figure 6 presents top-1 action recognition accuracy across all transformation strengths for each region. The dashed line represents the baseline, and each colored line represents each region.



(a) Blur transformation



(b) Solid-fill transformation



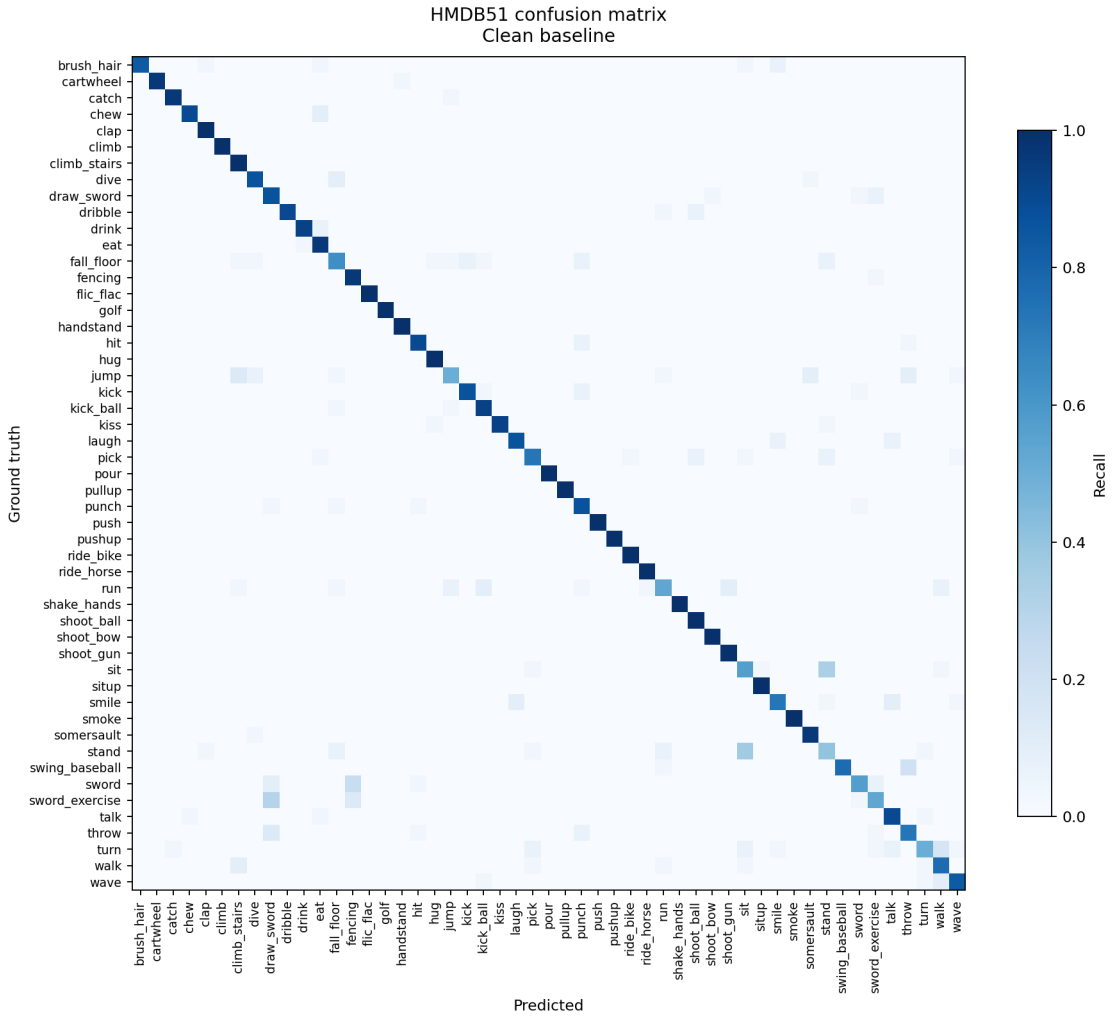
(c) Pixelization transformation

Figure 6: Top-1 action-recognition accuracy across transformation strengths and transformed regions. The dashed line denotes the untransformed baseline. Utility generally decreases as transformation strength increases, although the magnitude of the reduction depends on the transformed region and transformation type. Blur causes a gradual utility reduction, with full-body and background transformations producing the largest decreases. Solid fill produces the strongest degradation, particularly for high-strength full-body and background transformations. Pixelization also causes a gradual decline, with full-body and background transformations producing the largest reductions. Arm and leg transformations generally remain closest to the baseline.

Classification Behavior

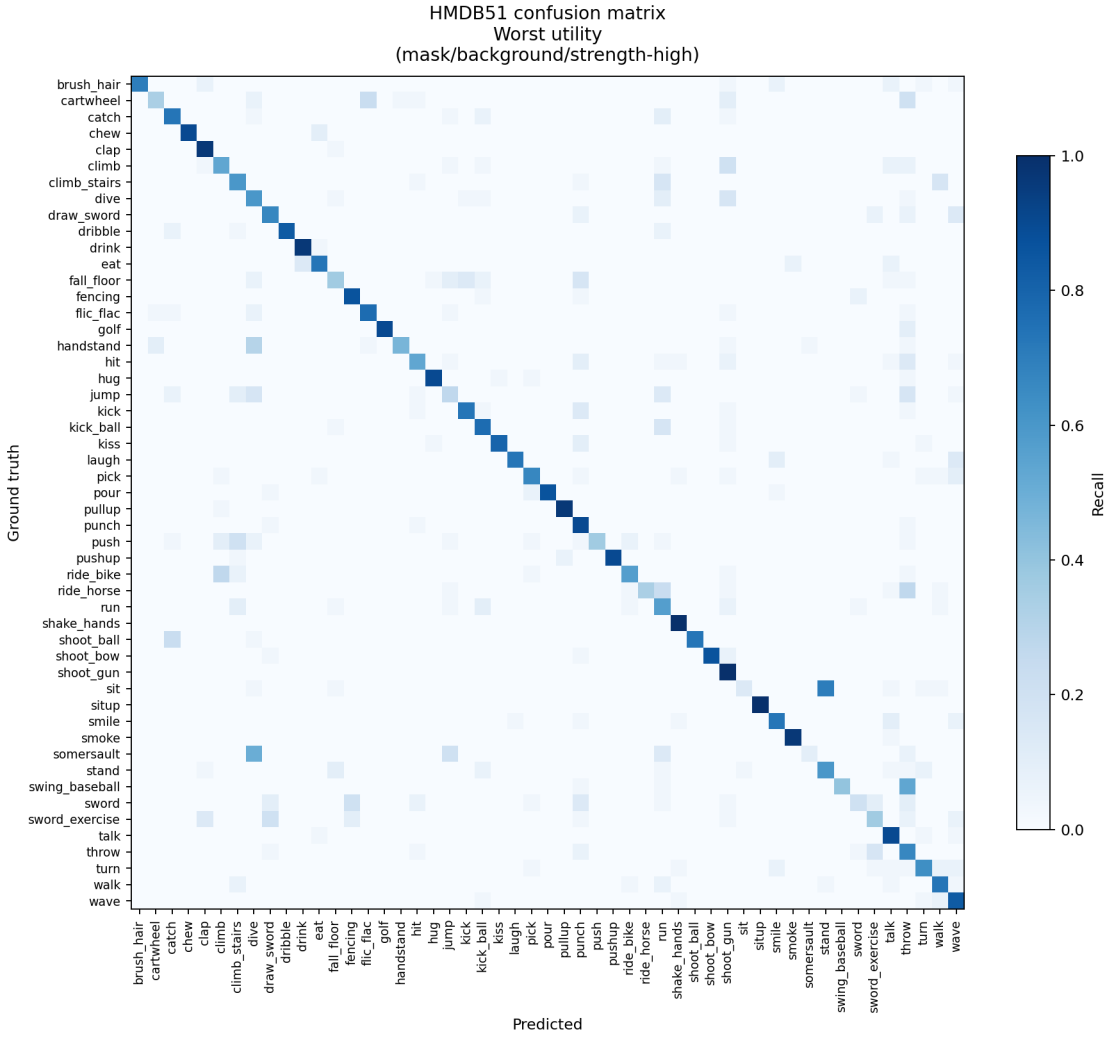
- All three confusion matrices retain a dominant diagonal (no transformation causes a huge amount of misclassification errors).
- Medium-strength head pixelization (Figure 7c) is nearly indistinguishable from the clean baseline matrix (Figure 7a).
- The worst-case condition (high-strength background masking, Figure 7b) shows more scattered errors despite an 18.7pp utility drop, but the diagonal remains clearly visible and there is no systematic confusion between specific action pairs.

Figure 7 presents confusion matrices for combinations with the worst utility (7b), the best privacy-utility trade-off (7c), as well as the baseline (7a).



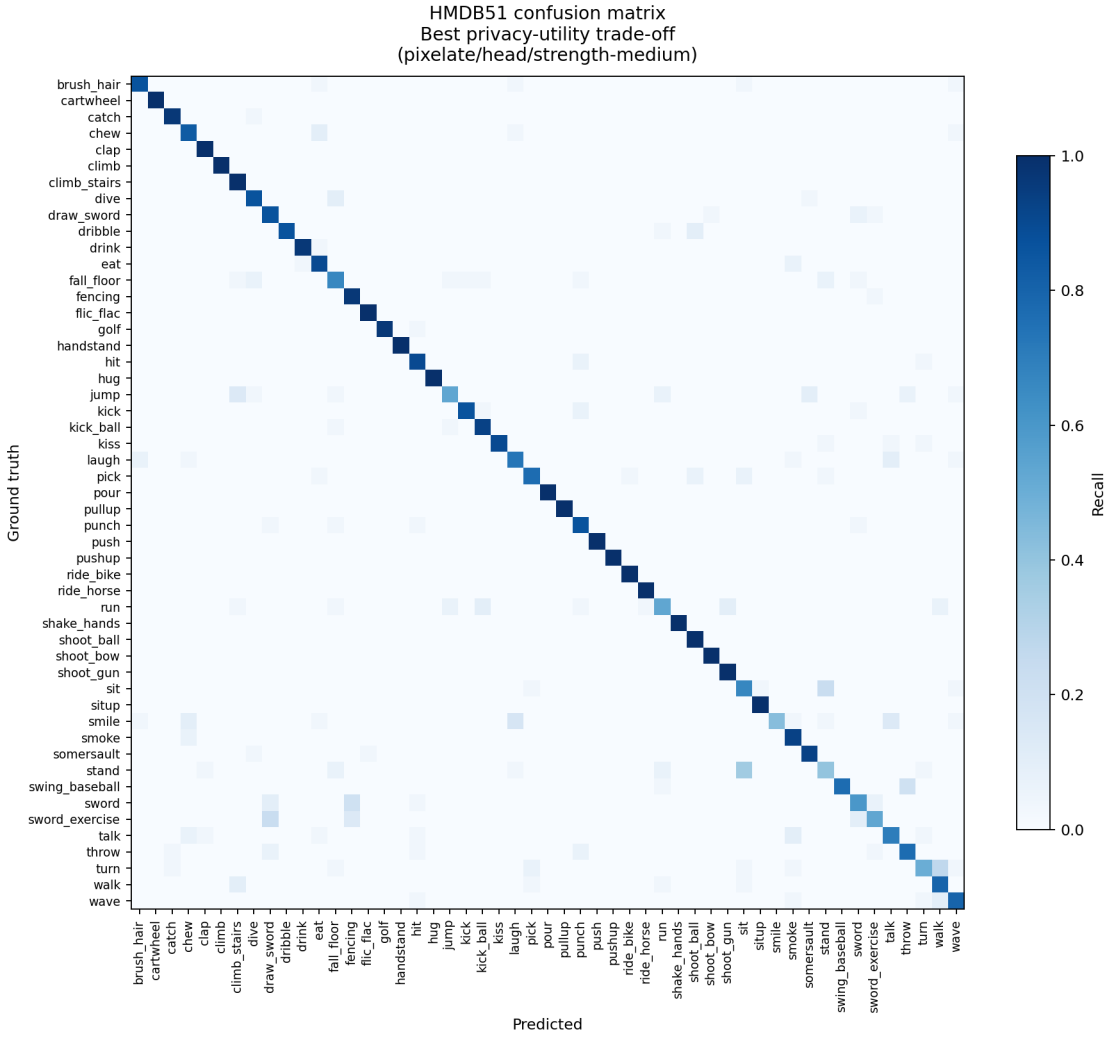
(a) Clean baseline

Figure 7: Confusion matrices for the clean baseline and selected PRISM combinations on the HMDB51 dataset. The high-strength background masking combination produces the lowest utility and slightly more misclassifications, whereas head pixelization with medium-strength provides the best privacy-utility trade-off and remains broadly similar to the clean baseline.



(b) High-strength background solid fill

Figure 7: Confusion matrices for the clean baseline and selected PRISM combinations on the HMDB51 dataset. The high-strength background masking combination produces the lowest utility and slightly more misclassifications, whereas head pixelization with medium-strength provides the best privacy-utility trade-off and remains broadly similar to the clean baseline.



(c) Medium-strength head pixelization

Figure 7: Confusion matrices for the clean baseline and selected PRISM combinations on the HMDB51 dataset. The high-strength background masking combination produces the lowest utility and slightly more misclassifications, whereas head pixelization with medium-strength provides the best privacy-utility trade-off and remains broadly similar to the clean baseline.

Privacy Leakage by Attribute

- Full-body masking reduces mean privacy leakage the most (72.4% → 51.2%), and background results in moderate reduction (72.4% → 65.8%). Head shows a modest reduction (72.4% → 69.2%), while arms (71.3%) and legs (71.8%) remain near baseline across all three attacker architectures, indicating these regions do not independently carry much sensitive privacy information.
- Full-body masking is most effective specifically for gender and skin color, the two most visually characteristic attributes. Head masking reduces face-attribute leakage only slightly by comparison, whereas full-body masking reduces it moderately.
- Nudity remains stable without any change by any region's transformation.
- Relationship attribute shows strong resistance to all transformations across every region, suggesting this attribute may depend on motion instead of static appearance cues that classical frame-level transformations do not target.

Figure 8 shows a bar chart of the mean privacy leakage per attribute across regions with a high strength level for all attacker architectures, whereas Figure 9 presents a bar chart with the worst-case privacy leakage per attribute using solid fill with a high strength level across five sensitive attributes.

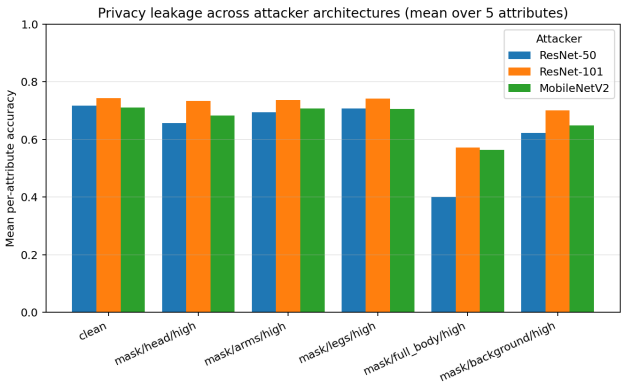


Figure 8: Mean privacy leakage per attribute across regions with high strength level for all attacker architectures. Applying the solid fill method on the head, arms, or legs gives similar results to the baseline. Only full body (having the best results) and background show a significant decrease in leakage. The results suggest consistency between different regions.

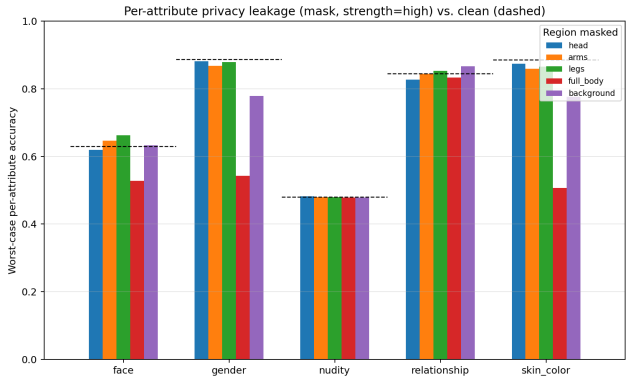


Figure 9: Worst-case privacy leakage per attribute using solid fill with high strength level across five sensitive attributes. Baseline is indicated by dashed lines. Solid fill applied on the full body most effectively reduces leakage for gender and skin colour. Nudity remains unchanged for all regions. The relationship shows high resistance to all transformation methods, indicating its leakage does not depend on appearance cues.

Comparison with STPrivacy and RAPrivacy

As a supplement to the main experiment, Table 9 and Table 10 report comparisons with state-of-the-art models. To match the STPrivacy and RAPrivacy evaluation setup, both the

action recognizer and the privacy attacker use a ViT-S/16 backbone and are trained directly on transformed videos. Table 9 reports results against STPrivacy's reported result on the PA-HMDB51 dataset, whereas Table 10 reports results against both STPrivacy's and RAPrivacy's reported results on the VP-HMDB51 dataset.

Method	Top-1 ↑ (%)	F1 ↓	cMAP ↓ (%)
PRISM(ViT-S/16) combinations on PA-HMDB51			
Clean baseline	**62.62**	0.595	67.76
Pixelate / Head / Medium	60.68	0.587	64.30
Blur / Head / High	55.83	0.580	62.18
Blur / Full body / High	50.97	0.534	58.03
Pixelate / Full body / Medium	57.28	**0.437**	**56.68**
Mask / Full body / High	58.25	0.547	58.05
Blur / Arms / High	57.28	0.567	66.82
Pixelate / Background / High	53.88	0.646	69.63
State-of-the-art method on PA-HMDB51			
STPrivacy [23]	50.61	0.523	68.76

Table 9: STPrivacy comparison on PA-HMDB51 dataset. Action recognition utility task is measured by top-1 accuracy, while F1 and cMAP measure privacy leakage by the attacker. Lower F1 and cMAP indicate stronger privacy protection, whereas higher top-1 accuracy highlights better action recognition. STPrivacy is included as the state-of-the-art reference [23].

- The clean baseline (62.62) exceeds STPrivacy's result (50.61) in the accuracy task. This gap may be partly explained by train/test overlap: since PA-HMDB51's 515 videos are a subset of HMDB51, and STPrivacy's paper describes training the action recognizer on the full HMDB51 dataset, it is unclear whether the 206-video PA-HMDB51 test set used in this protocol was excluded from the action recognizer's training data. If such overlap exists, STPrivacy's results would be affected in the same way, since this thesis follows the training protocol as faithfully as possible. In that case, this inflates Top-1 accuracy relative to genuine generalization performance.
- Pixelate/full body/medium achieves lower F1 and cMAP than STPrivacy (0.437/56.68 vs. 0.523/68.76) while keeping higher Top-1 (57.28 vs. 50.61); blur/full body/high and mask/full body/high also beat STPrivacy on cMAP (58.03, 58.05 vs. 68.76) but not F1 (0.534, 0.547 vs. 0.523), showing the ranking is inconsistent across metrics.
- Blur/arms/high stays close to baseline, and pixelate/background/high is the only condition worse than baseline on both F1 and cMAP (0.646, 69.63), reproducing patterns from the main experiment.

Method	Top-1 ↑ (%)	F1 ↓	cMAP ↓ (%)
PRISM(ViT-S/16) combinations on VP-HMDB51			
Clean baseline	33.20	0.706	77.39
Pixelate / Head / Medium	31.90	0.703	76.80
Blur / Head / High	30.39	0.639	73.16
Blur / Full body / High	26.99	0.591	72.40
Pixelate / Full body / Medium	29.48	0.680	72.98
Mask / Full body / High	29.35	**0.589**	72.45
Blur / Arms / High	30.91	0.645	76.91
Pixelate / Background / High	26.86	0.728	77.29
State-of-the-art methods on VP-HMDB51			
STPrivacy [23]	**50.73**	0.613	72.48
RAPrivacy [52]	50.12	0.605	**70.12**

Table 10: State-of-the-art comparison on the VP-HMDB51 dataset. PRISM results include the clean baseline and selected transformation combinations. Action-recognition utility is measured using top-1 accuracy, whereas F1 and cMAP measure privacy leakage through the privacy attacker. Higher top-1 accuracy indicates better utility, while lower F1 and cMAP indicate stronger privacy protection. STPrivacy and RAPrivacy are included as state-of-the-art reference methods [23, 52].

- None of the eight evaluated conditions reach STPrivacy's (50.73) or RAPrivacy's (50.12) Top-1, and even the clean baseline (33.20) trails both by roughly 17 points.
- Mask/full body/high achieves the strongest privacy reduction of any condition (F1 = 0.589, cMAP = 72.45), having better F1 than both RAPrivacy (0.605) and STPrivacy (0.613), and comparable cMAP to both (70.12, 72.48).
- Pixelate/background/high again produces the worst overall behaviour: the largest utility drop of any condition (26.86), with F1 (0.728) rising above the clean baseline (0.706) (the only VP-HMDB51 condition where privacy leakage does not improve under transformation).
- Blur/arms/high stays close to baseline (30.91 vs. 33.20 clean).

Looking at the results across both tables together, a few insights are worth noting:

- Blur/full body/high is the worst-utility condition in Table 9 (50.97%), while pixelate/background/high is the worst in Table 10 (26.86% vs. blur/full body/high's 26.99%). This replicates the main experiment's finding that full body and background anonymization are the most damaging to utility.
- The baseline nearly halves between protocols (62.62% on PA-HMDB51's 206-video test set vs. 33.20% on split-1's 1,530-video test set) despite similar training set sizes (4,060 vs. 3,570 videos) and identical architecture. This might indicate the full, balanced 51-class split-1 test set is substantially harder for this backbone than the smaller PA-HMDB51 subset. The STPrivacy's and RAPrivacy's unpublished exact training methods may also contribute to differences in the results.
- Blur/full body/high behaves consistently in relative terms across both tables: an 18.6% relative utility drop from baseline in Table 9 (50.97 vs. 62.62) and an 18.7% relative drop in Table 10 (26.99 vs. 33.20).

- Table 10's results should be weighted more heavily than Table 9's: RAPrivacy follows HMDB51's official, reproducible split, whereas PA-HMDB51 Top-1 accuracy may be inflated by the train/test overlap discussed above.

8 Discussion

Main Findings

To the best of our knowledge, PRISM is the first evaluation pipeline to provide a controlled comparison of fixed regions within matched transformation types and strengths. Previous research has either evaluated global transformations [23, 52] or used saliency maps [18] to learn where to obfuscate. However, no methodology offers a systematic, controlled comparison of how effective fixed, isolated regions actually are.

The results indicate that region selection is the main factor that shapes the privacy-utility trade-off, outweighing the effects of both transformation type and strength. Full-body combinations present this pattern, where the privacy scores are almost identical across transformation types, even though each transformation works differently. Consequently, this pattern suggests that it is the region that contains the sensitive identity information, rather than a specific pattern removed by a specific transformation. Hence, the choice of which region to anonymize is more important than which method to use to anonymize the sensitive information. These results give useful guidance about future work, suggesting that investing in more sophisticated transformation methods may lead to poor outcomes if the region is selected carelessly.

Medium-strength head pixelization performed best overall in the privacy-utility trade-off task. Unlike blur and masking, it disrupts spatial resolution enough to substantially reduce attacker accuracy on face and gender attributes, while only slightly affecting action recognition performance. Background anonymization, on the other hand, is the least consistent transformation method across all transformation types. Obfuscating the background significantly worsens the utility performance, as action recognition depends significantly on scene context. Consequently, in most cases, the gained privacy is too little to compensate for such loss, producing ineffective results.

Per-Attribute Insights

Per-attribute results show that the relationship stays resistant against every transformation, in every region. This suggests it is encoded in motion context rather than static appearance. This highlights that classical frame-level transformation types used in this experiment cannot suppress the relationship trait. Nudity also remains unaffected by anonymization techniques, which might be due to the lack of nudity-labelled samples in PA-HMDB51, instead of the robustness of the attribute to visual transformations. This indicates that those accuracy scores for these specific nudity attributes are unreliable and should be interpreted with caution.

Comparison with State-of-the-Art

Besides the main experiment, the results were compared with the state-of-the-art models: STPrivacy and RAPrivacy. To achieve this, both the action recognizer and the privacy attacker

were configured to use the same backbone, training procedure, and dataset splits. Such a comparison protocol supported the claim that region is the main factor of privacy reduction. On the PA-HMDB51 dataset, the pixelate/full_body/medium combination achieves lower F1 and lower cMAP than the STPrivacy model. On VP-HMDB51, the mask/full_body/high combination achieves a lower F1 than both state-of-the-art models and a cMAP comparable with STPrivacy's. This finding shows that a classical transformation applied to the full body can provide privacy protection comparable to a spatio-temporal anonymizer, even without temporal modeling. Hence, the region is indeed a more important aspect in privacy leakage than the transformation itself.

Caveats and Open Questions

There are still some caveats needed to be remembered while comparing the results. First, the two attacker paradigms differ in kind, since STPrivacy's attacker models spatio-temporal dependencies jointly with its token-sparsification anonymizer, whereas the attacker used throughout this thesis is a per-frame classifier with mean-pooling. The comparisons in Tables 9 and 10 reflect differences in both modeling approaches and transformation methods. They should therefore not be interpreted as evidence that classical transformations are always better than learned anonymization. Second, STPrivacy's random PA-HMDB51 split seed is not published, so the test split in Table 9 is most likely different from splits in the original STPrivacy evaluation. Hence, this might introduce an unavoidable source of noise. Third, the action recognizer used in the comparison protocol was adapted to match training procedures of both RAPrivacy and STPrivacy as closely as possible. Therefore, the observed utility difference between both dataset cannot be attributed to a pretraining advantage on either side. The most plausible explanation is that STPrivacy's and RAPrivacy's exact training hyperparameters were not published, so there might be a difference between setups, and therefore an exact replication of their setup is not possible.

In comparison, evaluated on the VP-HMDB51 dataset, mask/full_body/high achieves a better F1 than RAPrivacy (0.589 vs. 0.605) while its cMAP remains slightly higher (72.45 vs. 70.12). Hence, a classical transformation targeted on one region can therefore outperform RAPrivacy anonymization on F1, though RAPrivacy still remains the best in cMAP. This probably reflects RAPrivacy's explicit design priority of human readability.

It is also worth noting that the arms region does not fully replicate findings under this comparison's backbone. The main experiment, using a Kinetics-710-pretrained ViT-Giant backbone, found that arms and legs have minimal impact on utility. Under this comparison's weaker model, this pattern only partially holds. On VP-HMDB51 (1,530 test videos), blur/arms/highcauses a modest drop in utility (30.91% vs 33.20% clean). Similarly, on PA-HMDB51 (206 test videos), blur/arms/high drops to 57.28%, within the range of full body combinations in Table 9 (50.97-58.25%), showing a comparably modest deviation from the idea that arms do not matter in this trade-off, without arms clearly outweighing full body transformations. This presents a consistent pattern in which one plausible cause might be that a weaker model has fewer visual features to rely on rather than a heavily pretrained model, making it more sensitive to any regional obfuscation such as arms.

Limitations

These findings should be interpreted in light of several limitations. Firstly, all regions were transformed independently, which means that there could be a potential positive effect once interconnecting various regions remains unexplored, such as anonymizing both head and arms simultaneously. Second, transformations were applied only at a frame level, while omitting the temporal dimension and its leakage. Previous research has shown that aggregation across frames can still recover sensitive information from video, even if individual frames are completely anonymized. Lastly, the size of privacy labels in the PA-HMDB51 dataset is relatively small, as it contains only 515 annotated videos. That suggests that per-attribute privacy scores may be sensitive to noise and that conclusions achieved at the attribute level should be treated with caution.

9 Conclusions and future work

Conclusions

This thesis explored the privacy-utility trade-off in video action recognition, examining the influence of anonymization of different regions (body parts, full body, and background) on both action recognition and privacy leakage. To support this analysis, an evaluation pipeline was developed. First, a region extraction tool was used to identify the video regions targeted for anonymization. The transformed videos were then evaluated using an action recognizer and three privacy attackers. In the end the metrics are compared.

The main research question was to check how action recognition performance changes when privacy filters of varying strength are applied to different video regions. Furthermore, the subquestions delved deeper to explore the effect of transformation strength per region on utility, and the effect of region choice on privacy leakage under certain conditions. PRISM addresses these questions by providing region-specific comparisons with different strengths and anonymization methods.

The experimental results show that region choice is the most important factor in the privacy-utility trade-off task, with transformation type and strength having a more limited impact:

- Anonymization applied to full body demonstrates the most consistent privacy protection across the whole experiment, at a cost of large action recognition accuracy loss.
- Arms and legs regions showed that there is a small or almost no effect for both metrics once transformations are applied to these regions.
- The medium-strength pixelization applied to the head resulted in the best trade-off, considerably reducing attacker accuracy while still preserving most of the utility performance.
- Background anonymization considerably reduced utility accuracy without increasing privacy and consequently produced the worst results.
- Full body anonymization was the most effective at reducing privacy leakage related to gender and skin color, while relationship and nudity remained unchanged or slightly changed across all combinations.
- The worst-case privacy metric confirmed that anonymization using classical transformations is not robust and should not be applied alone, as it did not fully eliminate privacy leakage.

The comparison protocol further supports the region-based pattern. Transformations applied to full body reach levels comparable to or lower than both state-of-the-art methods in privacy leakage tasks. This evidence is stronger on VP-HMDB51 than on PA-HMDB51, where an unpublished split, seed, and train/test overlap in the action recognizer's training data limit findings. Transformations applied to arms leave privacy leakage close to baseline across both attacker architectures and training protocols. This supports the finding that this region carries little identity information regardless of backbone. However, its impact on utility is dependent on the backbone: it remains minimal on VP-HMDB51 but shows a larger drop on PA-HMDB51. Finally, the comparison shows that full body and background anonymization are both highly damaging to utility, showing the same results as the main experiment, which strengthens the importance of region choice.

Contributions of this thesis include:

- PRISM, the first controlled, region-based benchmark for the privacy-utility trade-off in action recognition, systematically isolating fixed regions under matched anonymization methods.
- Identification of medium-strength head pixelization as the combinations that provided the best privacy–utility trade-off among all evaluated combinations.
- Demonstration that region choice has greater impact on the privacy-utility trade-off than either transformation strength or anonymization type.
- A comparison protocol showing that classical, full-body transformations can reach privacy leakage levels comparable to current state-of-the-art methods (STPrivacy, RAPrivacy).

Future work

Future research may consider more advanced anonymization methods, such as learned anonymization (e.g., adversarial training), to be applied to different parts of the regions, rather than classical transformations. Furthermore, the functions might be extended to include anonymization in the temporal dimension, since previous work indicated that identity can leak through temporal aggregation even though individual frames are transformed. Future comparisons with methods like STPrivacy should ensure the action recognizer's training data excludes any videos present in the privacy evaluation test set, to avoid the train/test overlap identified in this thesis's PA-HMDB51 comparison protocol. Future work could also explore finer-grained privacy attribute labels (e.g., specific skin color) rather than the binary presence/absence framing adopted here and in prior benchmarks, which may better capture the full extent of privacy leakage. Also, it could be relevant to investigate why the relationship attribute is resistant to all visual transformations, since it may rely on non-appearance approaches, such as motion context. In addition, future work could focus on combinations of different regions simultaneously rather than one region at a time. It could help understand interaction effects and investigate if such combinations show unexpected results, such as the legs+arms combination achieving low privacy leakage. Moreover, the comparison was limited to eight representative combinations selected from the full set of 45 combinations due to the cost of retraining per condition. In the future, this could be extended to the full set of 45 combinations, which could give a more complete picture of how classical transformations applied on various regions compare to learned anonymization across the entire space. Besides that, the exploration of adaptive region selection rather than fixed regions is also important. The system could try to

identify which region carries the most identity information per video and target. Moreover, future work may require the use of a more effective attribute extractor, since although current models perform well in extracting regions, they still make mistakes, such as assigning hand pixels to the legs during a somersault. Lastly, future research might connect human perceptual studies to the field of privacy-preserving action recognition to complement the machine-based evaluation. As noted before, some research's goal is to emphasize human readability, as indicated in RAPrivacy research.

References

[1] Prachi Agrawal and P. J. Narayanan. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(3):299–310, 2011.

[2] AK Research & Consulting. Research on privacy concerns and data sharing. Technical report, Office of the Privacy Commissioner, New Zealand, Wellington, New Zealand, April 2024. Prepared for the Privacy Commissioner.

[3] Pascal Benschop, Jan van Gemert, Jelte P. Mense, and Justin Dauwels. Motion representations for privacy-aware cross-domain action recognition. *Frontiers in Imaging*, 5:1846329, 2026.

[4] Pascal Birnstill, Daoyuan Ren, and Jürgen Beyerer. A user study on anonymization techniques for smart video surveillance. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2015.

[5] Michael Cabanillas-Carbonell, Jhordan Rivera, and Jhoel Munoz. Artificial intelligence in video surveillance systems for suspicious activity detection and incident response: A systematic review. *Advances in Science and Technology Research Journal*, 19:389–405, 2025.

[6] Zikui Cai, Zhongpai Gao, Benjamin Planche, Meng Zheng, Terrence Chen, M. Salman Asif, and Ziyan Wu. Disguise without disruption: Utility-preserving face de-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 918–926, 2024.

[7] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. SPAct: Self-supervised privacy preservation for action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20132–20141, 2022.

[8] Julia Dietlmeier, Joseph Antony, Kevin McGuinness, and Noel E. O'Connor. How important are faces for person re-identification? In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6912–6919, 2021.

[9] Ádám Erdélyi, Thomas Winkler, and Bernhard Rinner. Privacy protection vs. utility in visual data. *Multimedia Tools and Applications*, 77(2):2285–2312, 2018.

[10] European Commission. Attitudes on data protection and electronic identity in the European Union. Special Eurobarometer 359, European Commission, Brussels, June 2011. Wave 74.3.

[11] Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. TeD-SPAD: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13552–13563, 2023.

[12] Ruohan Gao, Bo Xiong, and Kristen Grauman. Im2Flow: Motion hallucination from static images for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5937–5947, 2018.

[13] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with R*CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1080–1088, 2015.

[14] Xiuye Gu, Weixin Luo, Michael S. Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *Computer Vision – ECCV 2020*, volume 12368 of *Lecture Notes in Computer Science*, pages 727–743. Springer, 2020.

[15] Simon Hanisch, Evelyn Muschter, Admantini Hatzipanayioti, Shu-Chen Li, and Thorsten Strufe. Understanding person identification through gait. *Proceedings on Privacy Enhancing Technologies*, 2023(1):177–189, 2023.

[16] Eman T. Hassan, Rakibul Hasan, Patrick Shaffer, David Crandall, and Apu Kapadia. Cartooning for enhanced privacy in lifelogging and streaming videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1333–1342, 2017.

[17] 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 (CVPR)*, pages 770–778, 2016.

[18] Filip Ilic, He Zhao, Thomas Pock, and Richard P. Wildes. Selective, interpretable, and motion consistent privacy attribute obfuscation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18730–18739, 2024.

[19] Will Kay, João 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, 2017. arXiv preprint arXiv:1705.06950.

[20] Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011.

[21] Minh-Ha Le and Niklas Carlsson. StyleID: Identity disentanglement for anonymizing faces. *Proceedings on Privacy Enhancing Technologies*, 2023(1):264–278, 2023.

[22] Jingzhi Li, Lutong Han, Ruoyu Chen, Hua Zhang, Bing Han, Lili Wang, and Xiaochun Cao. Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3891–3899, 2021.

[23] 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 (ICCV)*, pages 5106–5115, 2023.

[24] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2022.

[25] Tao Li and Lei Lin. AnonymousNet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 56–65, 2019.

[26] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1343–1351, 2017.

[27] Yu-Jhe Li, Xinshuo Weng, and Kris M. Kitani. Learning shape representations for person re-identification under clothing change. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2432–2441, 2021.

[28] Pinghan Liang, Yadi Liu, Yuchen Guo, and Fanqi Zeng. The heterogeneous impact of public security cameras on safety perceptions in cities: Evidence from China. *PNAS Nexus*, 4(10):pgaf331, 2025.

[29] Kunliang Liu, Jianming Wang, Rize Jin, Wonjun Hwang, and Tae-Sun Chung. SCHNet: SAM marries CLIP for human parsing, 2025. arXiv preprint arXiv:2503.22237.

[30] Saemi Moon, Myeonghyeon Kim, Zhenyue Qin, Yang Liu, and Dongwoo Kim. Anonymization for skeleton action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15028–15036, 2023.

[31] Athira Nambiar, Alexandre Bernardino, and Jacinto C. Nascimento. Gait-based person re-identification: A survey. *ACM Comput. Surv.*, 52(2), 2019. Article 33, 34 pages.

[32] Clive Norris, Mike McCahill, and David Wood. The growth of CCTV: A global perspective on the international diffusion of video surveillance in publicly accessible space. *Surveillance & Society*, 2(2/3):110–135, 2004.

[33] 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)*, pages 3686–3695, 2017.

[34] Phoenix Strategic Perspectives Inc. 2024-2025 public opinion research on privacy issues. Public Opinion Research Report POR No. 093-24, Office of the Privacy Commissioner of Canada, Gatineau, Quebec, March 2025. Government of Canada catalogue no. IP54-118/2025E-PDF.

[35] Hugo Proença. The UU-Net: Reversible face de-identification for visual surveillance video footage. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):496–509, 2022.

[36] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. A review on visual privacy preservation techniques for active and assisted living. *Multimedia Tools and Applications*, 83(5):14715–14755, 2024.

[37] Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. Learning to anonymize faces for privacy-preserving action detection. In *Computer Vision – ECCV 2018*, volume 11205 of *Lecture Notes in Computer Science*, pages 639–655. Springer, 2018.

[38] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131–151, 2016.

[39] Natacha Ruchaud and Jean-Luc Dugelay. Automatic face anonymization in visual data: Are we really well protected? In *Proceedings of the IS&T International Symposium on Electronic Imaging: Image Processing: Algorithms and Systems XIV*, pages 1–7, 2016.

[40] Michael S. Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4255–4262, 2017.

[41] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018.

[42] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576, 2014.

[43] Robert Socha and Bogusław Kogut. Urban video surveillance as a tool to improve security in public spaces. *Sustainability*, 12(15):6210, 2020.

[44] G Sreenu and M A Saleem Durai. Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *Journal of Big Data*, 6(1):48, 2019.

[45] Robert Templeman, Mohammed Korayem, David Crandall, and Apu Kapadia. PlaceAvoider: Steering first-person cameras away from sensitive spaces. In *Proceedings of the 21st Annual Network and Distributed System Security Symposium (NDSS)*, 2014.

[46] Daksh Thapar, Aditya Nigam, and Chetan Arora. Anonymizing egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2320–2329, 2021.

[47] Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M. Gavrila, and Peter H. N. de With. Privacy protection in street-view panoramas using depth and multi-view imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10581–10590, 2019.

[48] Julien Voisin, Christophe Guyeux, and Jacques M. Bahi. The metadata anonymization toolkit, 2012. arXiv preprint arXiv:1212.3648.

[49] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023.

[50] Yizhou Wang, Yixuan Wu, Weizhen He, Xun Guo, Feng Zhu, Lei Bai, Rui Zhao, Jian Wu, Tong He, Wanli Ouyang, and Shixiang Tang. Hulk: A universal knowledge translator for human-centric tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5672–5689, 2025.

[51] Yuntao Wang, Zirui Cheng, Xin Yi, Yan Kong, Xueyang Wang, Xuhai Xu, Yukang Yan, Chun Yu, Shwetak Patel, and Yuanchun Shi. Modeling the trade-off of privacy preservation and activity recognition on low-resolution images, 2023. arXiv preprint arXiv:2303.10435.

[52] Zi-Zhen Wang, Yen-Lung Chu, Pei-Chun Tsai, and Kuan-Wen Chen. RAPrivacy: A readable anonymizer for privacy preserving action recognition. In *Proceedings of the 36th British Machine Vision Conference (BMVC)*, 2025. Paper 117.

[53] 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, 2022.

[54] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Computer Vision – ECCV 2018*, volume 11220 of *Lecture Notes in Computer Science*, pages 627–645. Springer, 2018.