SoK: Multi-Perspective-Video-Anonymization

Islam Amar
Karlsruhe Institute of Technology, Germany
islam.amar@kit.edu

Omar Moured
Karlsruhe Institute of Technology, Germany
omar.moured@kit.edu

Simon Hanisch
Technical University Dresden, Germany
simon.hanisch@tu-dresden.de

Thorsten Strufe
Karlsruhe Institute of Technology, Germany
thorsten.strufe@kit.edu

Abstract

Video data has become central in many modern systems, from public surveillance to autonomous vehicles, but its widespread use raises serious concerns about exposing people's identities and behaviors. This survey takes a structured look at how recent research has tried to address these concerns through video anonymization. We conduct a systematic review of the literature and organize existing work into a taxonomy that separates different anonymization strategies, the visual regions they target, and the assumptions they make about the video setting. By examining these categories, we highlight how current methods approach privacy protection and where they tend to fall short.

A consistent pattern across the field is that most studies are designed for single-view videos, even though many real environments involve multiple synchronized cameras. Only a handful of works consider this multi-perspective setting, revealing a clear disconnect between research and real-world needs. We also review the datasets and evaluation metrics commonly used in anonymization studies and show that they rarely capture multi-view complexity, underscoring the need for more representative benchmarks. Finally, we discuss the utility goals that anonymization methods aim to preserve and outline key gaps that future work must address to support practical, multi-camera video applications.

Keywords

Systematization, Privacy, Anonymization, Video

1 Introduction

Video data collection has become integral to our daily lives, ranging from surveillance systems to smartphone recordings. These recordings are frequently employed in a variety of applications, including face detection [164], face recognition [40, 154], human pose estimation [11, 19], and human segmentation [17, 55]. As the number of cameras continues to increase, there is an increasing number of applications for multi-perspective videos. Possible applications include multi-view action recognition in operation rooms for surgery guidance [133], tracking people through supermarkets to understand customer behavior [119, 165], or action recognition in public spaces to detect violent behavior [82]. Besides static setups, multi-view videos are also used by autonomous systems like robots [98] and autonomous cars [45] to allow for a better understanding of their surroundings. All these applications rely on fusing the input of multiple cameras, which record the same scene from different perspectives. The fusing allows for a richer scene understanding, mitigation of occlusion effects, and better tracking of subjects.

While these represent impressive improvements for the respective applications, they also further increase the already existing privacy problems of video data. Video recordings contain sensitive personal information such as the biometric traits of people like facial features [21], expressions [134], eye gaze [14], gait [167], body shape [134], or actions [122]. With multi-view videos, the capture quality and quantity are further increased, and as such, privacy is again increasingly threatened.

Recent incidents have shown that video recordings can lead to serious privacy breaches. During the COVID-19 pandemic, misconfigured cloud storage left Zoom meeting recordings publicly accessible, exposing faces, voices, and private surroundings [79]. In 2021, hackers breached Verkada Inc.'s cloud-based security systems, gaining access to live and archived feeds from over 150,000 surveillance cameras in hospitals, factories, and prisons[1]. Similarly, in 2023, former Tesla employees were reported to have shared private in-vehicle camera footage captured from customers' cars, including recordings inside garages, driveways, and within the car interior, exposing personal living spaces and sensitive situations[2].

Given the growing privacy risks associated with large-scale video collection, there is an increasing demand for anonymization techniques that aim to help protect individuals' privacy in published or shared video data. The intention is to protect individuals appearing in the video against both identity and attribute disclosure, ensuring that no observer–human or algorithmic–can recognize them or infer sensitive characteristics, such as gender, age, health status, or emotional state. While such anonymized data is commonly not truly anonymous, anonymization aims to reduce identification risks by removing or suppressing biometric cues that reveal identity, including facial appearance, body shape, and behavioral signals.

However, removing or altering these identity-revealing elements inevitably reduces the amount of information contained in the data, which may also diminish its usefulness for downstream tasks. Many applications, such as action recognition, behavior analysis, surgical training, or sports analytics, rely on fine-grained visual and temporal cues to extract meaningful insights. Therefore, anonymization must balance privacy protection with the preservation of task-relevant information.

When individuals are recorded from multiple perspectives, anonymization is more challenging. Due to richer scene capturing, biometric traits of individuals are available in higher quality, allowing for easier identification. Inconsistencies in anonymized appearance,

Proceedings on Privacy Enhancing Technologies 2026(3), 317–340

https://doi.org/10.56553/popets-2026-0084

[1] https://www.msspalert.com/news/verkada-video-surveillance-camera-breach
[2] https://www.reuters.com/technology/tesla-workers-shared-sensitive-images-recorded-by-customer-cars-2023-04-06/

motion cues, or behavioral signals across viewpoints can expose identity through cross-perspective aggregation. To prevent such identity leaks while retaining utility for multi-perspective analysis, anonymization must ensure consistent, privacy-preserving, and utility-maintaining representations across all views.

The research community has developed a wide spectrum of anonymization techniques aimed at mitigating the described threats. Early efforts relied on masking- and coarsening-based obfuscation methods such as blurring, pixelation, and masking [85, 126, 127], which concealed identity but also degraded the overall utility of data for tasks that rely on fine-grained visual or behavioral cues. While these methods can be applied consistently over time, they suppress fine-grained facial and body cues that are essential for many downstream tasks, such as emotion recognition, gaze estimation, focus of attention, social interaction understanding, or behavioral analysis, thereby limiting utility of the protected video.

These limitations have motivated a shift towards identity-replacing, utility-preserving visual anonymization, with the goal not to obscure identifiable information, but to substitute identity-bearing signals with privacy-safe alternatives while retaining the original scene's structure, semantics, and utility. Unlike classical obfuscation methods such as pixelation, blurring, or masking, which reduce information through distortion, these approaches leverage learning-based generative models in their aim to produce anonymous yet realistic visual content aligned with the original appearance and context. This emerging class of methods includes generative replacement techniques, adversarial approaches that modify visual representations to confound identification models, and diffusion-based generative frameworks capable of delivering semantically consistent, context-aware, and temporally coherent streams, that supposedly are anonymous, thereby achieving stronger privacy without sacrificing downstream utility.

In this paper, we review existing methods related to video anonymization in general, with particular emphasis on multi-view video anonymization. This focus is motivated by the growing number of real-world applications that rely on multi-camera setups, including autonomous driving, smart cities, and operating rooms. Given that only few approaches for simultaneous multi-view video anonymization exist, so far, we include those single-view approaches that potentially provide foundations for consistent multi-view anonymization. The main contributions of this survey can be summarized as follows:

- We performed the first survey that systematically analyzed video anonymization under a multi-perspective evaluation framework, explicitly addressing the challenges that arise in the synchronized multi-camera environments.
- We propose a taxonomy that classifies anonymization methods by transformation type, targeted visual content, and intended utility goals.
- We identify and analyze privacy risks specific to multi-camera setups, including multi-perspective vulnerabilities.
- We highlight the current lack of multi-view anonymization benchmarks for evaluating cross-camera consistency.

2 Related Surveys

Privacy in visual data has been studied through facial de-identification, visual content protection, and privacy-enhancing biometric systems. Corresponding surveys provide useful foundations, but largely focus on image-based or face-specific methods, or situate anonymization within broader biometric contexts. None offers a comprehensive treatment of video anonymization for emerging challenges in multi-view and multi-perspective settings, where temporal and cross-camera consistency are critical.

Several surveys examine privacy-enhancing techniques for facial biometrics, exploring how identity can be concealed while retaining utility. Meden *et al.* [110] offer a comprehensive review of Biometric Privacy-Enhancing Techniques (B-PETs) for facial data, covering de-identification, adversarial perturbations, and template protection. Likewise, recent work surveys facial privacy protection in the digital era, outlining modern threats, legal considerations, and generative AI–based de-identification methods [148].

While these efforts offer systematic coverage of facial privacy, they remain face-centric and primarily image-focused. When video is considered, it is treated as a temporal extension of images, without addressing how anonymization must preserve motion continuity or prevent identity leakage across frames or camera viewpoints. As a result, these surveys do not account for full-body identity cues, nor do they analyze anonymization beyond facial regions.

A second survey line focuses on privacy protection in visual content, typically focusing on images. Wen *et al.* [151] provide a systematic review of image privacy techniques, introducing a multi-level framework spanning data-, content-, and feature-level protection. Similarly, recent work surveys visual content privacy protection, examining threats, protection goals, and privacy-preserving transformations for images shared on digital platforms [162].

These surveys offer valuable foundations on privacy objectives, threat models, and content-level protection, but their scope remains largely image-centric. Video appears only marginally or as an extension of image use cases, with no treatment of video-specific requirements such as temporal smoothness, identity persistence, or multi-view consistency—crucial factors for anonymizing video streams in real-world settings.

Another relevant stream covers behavioral and soft-biometric anonymization. Hanisch *et al.* [58] survey anonymization techniques for behavioral data, including voice, gait, gestures, and eye movements, comparing methods by privacy goals and anonymization concepts. As behavioral biometric data is inherently a time series, it especially prevalent in videos. Hence, behavioral biometric anonymizations are also often targeted at anonymizing videos to remove specific behavioral biometric traits (gait, eye movements, gestures, etc.). Earlier work on soft biometrics analyzes how attributes such as age, gender, and body shape can be inferred from images and videos, highlighting that identity leakage extends beyond facial appearance [36]. These surveys highlight the need for full-body and behavioral anonymization, as motion and body dynamics can reveal identity beyond the face. Yet they treat behavioral anonymization only abstractly or per modality and overlook the increased identity leakage that arises in multi-camera environments.

Across these survey lines, three gaps become evident:

(1) No existing survey systematically addresses anonymization for multi-view video, where identities can be reconstructed from cross-camera aggregation. This remains unexamined despite the growing deployment of multi-camera systems in smart-city surveillance, transportation, sports, and clinical environments.
(2) Existing reviews are either face-only or image-centric, overlooking the need for full-body, behavioral, and scene-level anonymization in video.
(3) There is no unified taxonomy for video anonymization that links anonymization paradigms (obfuscation vs. generative replacement) to privacy, utility, and crucially, temporal and cross-view consistency—core requirements for effective video anonymization.

To address these shortcomings, we present the first survey dedicated to video anonymization, with a particular focus on multi-view video settings. We review anonymization methods beyond the facial region, covering full-body, behavioral, and contextual cues, and analyze how approaches balance privacy, utility for downstream tasks, visual realism, and temporal/cross-view consistency. Furthermore, we consolidate relevant datasets and evaluation methodologies, offering a structured foundation for advancing research on reliable anonymization in multi-camera video environments.

Table 2 provides a comparative summary of these related surveys, outlining their main scope, data modality, target applications, datasets, and existing limitations.

3 Background

In this section, we review the relevant terminology used in this work and the existing surveys on video anonymization.

3.1 Terminology

In the context of visual data, particularly image and video anonymization, it is important to first clarify the key concepts and terminology that form the foundation of this survey.

Privacy concerns naturally arise because identifiable biometric traits are captured the moment individuals are recorded. These traits can be broadly categorized into physical and behavioral types. Physical traits include facial features, body shape, or full-body appearance, while behavioral traits involve motion dynamics such as gait [127, 144]. Both categories can be exploited by human observers as well as automated recognition systems, making them significant sources of privacy risk.

A central part of this risk relates to information disclosure. It is helpful to distinguish between identity disclosure, which occurs when a person can be directly recognized from visual data, and attribute disclosure, where sensitive personal attributes, such as gender, emotional state, or health conditions can still be inferred, independent of direct identification [20, 58].

Anonymization in visual data refers to altering or removing identifiable cues so that individuals cannot be linked to their biometric signatures, including those that may still be retained in the processed data, with the primary goal of preventing identity disclosure [48, 115]. In image-based anonymization, this typically involves protecting specific regions of interest, most commonly the face, the full body, or contextual elements that may reveal identity depending on the intended level of privacy protection.

In the case of videos, anonymization must account for additional identity cues available beyond static appearance, as identity leakage can occur not only within individual frames but also through motion patterns, gestures, gait, and temporal correlations across time and viewpoints. Moreover, even if information is obfuscated in individual frames, identity can potentially still be recovered over time through temporal accumulation or reconstruction. Therefore, effective video anonymization must perturb or remove these multi-frame identity signals by ensuring temporal coherence across consecutive frames and maintaining consistent anonymized appearances across different perspectives to prevent re-identification through temporal or cross-perspective aggregation.

Another essential property of anonymization is utility, which denotes the degree to which data that underwent anonymization remains suitable for their intended applications. In the context of image and video anonymization, utility refers to the preservation of task-relevant visual cues such as facial expressions, gaze direction, motion dynamics, or scene structure

that are necessary for tasks like action recognition, behavioral analysis, or medical training, while minimizing the risk that these cues reveal personal identity.

3.2 Assumptions and Models

In this survey, we assume a scenario where visual data follow a structured anonymization pipeline designed to protect identity information prior to any release or external use. The data flow consists of three main stages: acquisition, anonymization, and publication or processing. During the acquisition stage, multi-view or single-view image and video streams are captured, often containing identifiable biometric cues such as facial features, body shape, or gait. In the anonymization stage, identity-revealing information is modified or removed to prevent re-identification while maintaining task-relevant utility.

In the publication or processing stage, the protected data are shared, stored, or employed for downstream analytical tasks such as activity recognition, behavioral understanding, or training.

This is the operational setting considered in this survey, where anonymization is performed after data capture but before dissemination or model training. We use this pipeline as a guiding perspective and discuss existing approaches in terms of how well they align with and support such a workflow.

Note that we explicitly exclude cryptographic approaches that target input privacy. While cryptographic methods such as secure multiparty computation and homomorphic encryption provide strong input privacy guarantees, they cannot provide convincing output privacy guarantees, which is the necessary focus to ensure privacy of the captured individuals in videos for downstream tasks.

To improve clarity, we introduce a simplified notation of video anonymization. Note, that the complete formalization is in the appendix B. For this purpose we define a video V as a sequence of images/frames $I_t \in \mathcal{I}$ as: $V = (I_t)_{t=1}^{t_{max}}$, where t_{max} denotes the last frame of the stream. Note, that for multi-perspective video there are sets of (commonly synchronized streams) $\mathcal{V} = \{V^{(c)}\}_{c=1}^{c_{max}}$ from cameras $c \in C$. De-identifying a video, or an image, requires a

segmentation function S that outputs the region containing an individual, or their identifying parts: $S : V \rightarrow r_{t,i}$ with $r_{t,i} \in \mathcal{R}$ the contiguous pixel region i in frame t and $S(V) = (r_{t,i})_{t=1}^{t_{max}}$. Where S may correspond to a face detector, full body detector, or human segmentation, depending whether the mechanism targets facial regions, full bodies, or behavioral sequences. Note, that modifying facial or body-level identity cues does not necessarily guarantee strict anonymity, as other visual or contextual attributes (e.g., clothing, background, gait, or scene context) may still facilitate re-identification. Some de-identification attempts will extract a feature representation of identifying traits, first, to then modify and render the altered versions back into the video stream. This requires a feature extraction function that takes as an input a video stream V and regions $r_{t,i}$ containing information that is, acoording to the model of the respective mechanism, considered identifying and outputs those features: $F : V \times \mathcal{R} \rightarrow \mathcal{Z}, (V, \mathcal{R}) \rightarrow z$, where $z \in \mathcal{Z}$ are latent space representations of semantic or biometric features of the person visible in region $r_{t,i}$. A de-identification, or anonymization mechanism $\mathcal{M}$ then takes as an input a raw video stream V, identifying regions r_i, and potentially feature representations z_i and outputs a privacy protected stream $\widetilde{V}$, $\mathcal{M} : V \times \mathcal{R} \times \mathcal{Z} \rightarrow \widetilde{V}$.

3.2.1 Attacker Model. We assume an attacker whose goal is to identify persons in protected video data. Formally, the attacker is given access to a protected video stream $\widetilde{V}$, or in the multi-perspective case to a set of synchronized anonymized streams $\widetilde{\mathcal{V}} = \{\widetilde{V}^{(c)}\}_{c=1}^{c_{max}}$, where each $\widetilde{V}^{(c)}$ denotes the privacy-protected output of camera c. In addition to the anonymized stream(s), the attacker has access to auxiliary information $a_{\text{aux}} \in \mathcal{A}$, where $\mathcal{A}$ denotes the space of auxiliary data. Such auxiliary information may include publicly available video data, background datasets for training biometric models, or direct reference samples of specific target individuals. The attacker further has knowledge of the anonymization mechanism $\mathcal{M}$ and its parameters. Given this knowledge, the attacker may attempt to adapt her recognition system to the transformation applied by $\mathcal{M}$. For identification, the attacker employs a biometric recognition function $\mathcal{B} : \widetilde{V} \times \mathcal{A} \rightarrow \mathcal{Y}$, where $\mathcal{Y}$ denotes the identity space. The recognition system $\mathcal{B}$ may be trained using the auxiliary dataset a_{aux}, and can exploit all observable biometric cues present in $\widetilde{V}$, including but not limited to facial features, body shape, or motion dynamics. Due to knowledge of $\mathcal{M}$, the attacker may either: (i) retrain $\mathcal{B}$ directly on anonymized data, or (ii) attempt to construct a reversal operator $\mathcal{R}$ that approximates the inverse of $\mathcal{M}$, i.e., $\hat{V} = \mathcal{R}(\widetilde{V})$. As we consider multi-view videos in this paper, we extend the attacker model to the case where access is given to multiple anonymized camera streams $\widetilde{\mathcal{V}} = \{\widetilde{V}^{(c)}\}_{c=1}^{c_{max}}$, Given this capability, the attacker can aggregate information across views to construct a joint embedding for identification. Formally, this corresponds to applying a multi-view aggregation function $\Psi : \{\widetilde{V}^{(c)}\}_{c=1}^{c_{max}} \times \mathcal{A} \rightarrow \mathcal{Z}$, which produces a combined representation that may contain richer biometric information than any single-view embedding. The biometric recognition functions $\mathcal{B}$ can then operate on the aggregated representation, enabling stronger identification attacks compared to single-view scenarios. A more detailed formalization of these assumptions, adversarial capabilities, and temporal aggregation is provided in the Appendix B.

3.3 Fundamental Concepts: Deep Learning in Visual Anonymization

Utility-preserving generative anonymization methods demand capabilities that go beyond the limitations of traditional obfuscation techniques such as blurring, pixelation, or masking. These approaches are expected to maintain syntactic and semantic consistency [43], produce realistic visual content [129], preserve task-relevant information [108], and ensure temporal and cross-perspective coherence [105]. This survey focuses on anonymization approaches that are capable of achieving these properties in visual data.

To achieve these capabilities, recent anonymization techniques are built on three major families of deep generative models. Generative Adversarial Networks (GANs) [53] are widely used for identity replacement and inpainting [18, 62], synthesizing realistic surrogate appearances while preserving scene structure. They operate through an adversarial training process between a generator and a discriminator, enabling the model to learn how to produce visually plausible identity-altered outputs that remain consistent with the original scene.

Variational Autoencoders (VAEs) [84] enable identity disentanglement by encoding images into structured latent spaces where identity-related and identity-neutral attributes can be separated [28, 95]. The fundamental idea is to train an encoder that compresses the input to a distribution, preferably with disentangled latent representations that correspond to interpretable attributes, on which the parameters for decoding can be varied, or randomly sampled. This allows anonymization systems to selectively modify or replace identity components while keeping non-identifying visual characteristics intact.

Diffusion models [68] have recently become state-of-the-art for photorealistic anonymization [90, 104] by gradually transforming random noise into an identity-free image through iterative denoising steps. These models operate by learning two complementary processes: a forward diffusion that incrementally adds noise to destroy identifiable content, and a reverse generative process that learns to invert this corruption, reconstructing clean samples while selectively omitting identity-specific details. In anonymization pipelines, this process is typically guided by conditioning signals (e.g., pose, segmentation maps, or textual prompts) so that the model removes identity-specific features while reconstructing a realistic surrogate that matches the original scene context. Their stepwise refinement process allows precise control over which visual attributes are removed, altered, or preserved, enabling targeted suppression of facial or body identity while retaining non-identifying appearance, motion, and scene cues. This makes diffusion models well-suited for high-fidelity anonymization. In addition to generative models, discriminative architectures such as CNNs [88] and transformers [39] are often employed for tasks like detection, segmentation, or pose estimation [61], providing structural or semantic guidance to the anonymization pipeline.

3.4 Evaluation Metrics

Evaluating visual anonymization requires measuring both how well a method conceals sensitive information (privacy) and how much useful content it preserves for downstream tasks (utility). Effective

anonymization must balance these often competing goals—the privacy–utility trade-off [60, 104]. Excessive modification destroys the data's usefulness for downstream tasks, but prioritizing utility risks privacy leakage. Consequently, research relies on a combination of privacy and utility metrics, supplemented by human-centered evaluations and robustness tests [1].

Privacy metrics quantify how effectively identifiable or sensitive information is suppressed in data that has been modified towards anonymity. A common measure is re-identification risk, typically assessed by testing whether recognition models can still identify individuals. However, as Hanisch *et al.* [60] note, many studies report overly optimistic results as they rely on weak adversaries-models, trained only on unprotected data, thus creating a false sense of privacy. They advocate using stronger, adaptation-aware adversaries ("parrot recognition"), which better approximate worst-case conditions. Additional metrics include attribute leakage (whether demographic or biometric traits remain inferable), adversarial robustness (resistance to inversion or linkage attacks), and human-perception studies that ask participants to match anonymized identities to originals.

Utility metrics assess whether anonymized images or videos remain suitable for downstream tasks such as classification, detection, segmentation, and pose estimation. Common measures include standard task performance indicators—accuracy, precision, recall [121], and Mean Per Joint Position Error (MPJPE) [76], computed on anonymized data. Perceptual quality metrics such as Structural Similarity Index (SSIM) [149], Peak Signal-to-Noise Ratio (PSNR) [70], Learned Perceptual Image Patch Similarity (LPIPS) [161], and Fréchet Inception Distance (FID) [64] further evaluate visual realism and structural fidelity, which are especially important for generative replacement methods. Together, these metrics capture complementary aspects of utility: task-oriented performance reflects how well anonymized data supports machine perception, while perceptual metrics quantify the preservation of structure and visual plausibility in the generated output. Human-centered utility evaluations are also used, with participants rating the realism, expressiveness, or interpretability of anonymized outputs, particularly in video settings [22, 65, 153]. However, current utility metrics often fail to capture task-specific behavioral cues essential for real-world usability. Tasks such as emotion recognition, gaze estimation, attention analysis, social interaction understanding, and activity recognition depend on subtle signals—micro-expressions, eye dynamics, and body language that anonymization may distort [27, 62, 65]. This highlights a gap in current practice: dedicated utility metrics that explicitly capture the preservation of such subtle yet task-critical cues are still missing and should be developed and integrated into video anonymization studies.

Detailed descriptions of these metrics, including mathematical definitions and implementation examples, are provided in Appendix A.

4 Scope and Surveying Strategy

We conducted a systematic review of anonymization techniques for video that aim to achieve syntactic and semantic consistency, photo-realistic appearance, and preservation of task-relevant information. Our review aims to provide a rigorous categorization and comparative analysis of existing literature in this field. To ensure methodological transparency and replicability, we incorporated key elements of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines by Moher *et al.* [111]. The overall workflow of our study collection strategy is illustrated in Figure 2 in the Appendix C, where each step of the search and refinement pipeline is explicitly shown.

A structured literature review was performed across multiple databases, including IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar. The initial seed set of studies was identified using the following search string within 2011–2025: "video anonymization" OR "video de-identification"

This formulation ensured alignment with our focus on video anonymization methods and restricted the search to studies explicitly addressing anonymization or de-identification in video data.

We removed duplicates, including pre-prints followed by their peer-reviewed versions and workshop or conference papers followed by journal articles for the corresponding, more comprehensive versions. An initial manual screening of titles and abstracts was conducted to ensure topical relevance. Subsequently, the full text was recovered through an in-depth eligibility assessment using the following inclusion criteria: 1) studies published in English. 2) studies published in peer-reviewed venues. 3) studies explicitly describing anonymization methods for video or image data. 4) studies that report evaluation metrics for privacy and/or utility. To maintain focus and ensure compatibility, we applied the following exclusion criteria: 1) studies without an available English full text. 2) Studies that provide only input privacy (e.g., encryption, secure multiparty computation, or homomorphic cryptography) but do not output protected images or video streams suitable for downstream use.

3) studies that describe anonymization at a purely conceptual level without implementations or experimental validation. 4) works addressing unrelated domains, e.g., textual anonymization or medical record redaction.

This search produced our initial seed set of studies, as illustrated in Figure 2. In addition to adhering to PRISMA guidelines, we employed three supplementary iterative steps to broaden coverage and ensure completeness. These steps are depicted as steps 2, 3, and 4 in Figure 2. We applied the *Connected Papers* tool to each included work, and we systematically screened all suggested studies against our predefined criteria. Step 3 consisted of manually inspecting citation networks using the "Cited by" function in Google Scholar, both backward (references) and forward (subsequent citing works).

These steps were repeated iteratively until no new relevant studies were identified.

Through this combined strategy, we obtained an initial corpus of 414 studies, from which 41 publications were retained after applying inclusion and exclusion criteria.

5 Taxonomy and Classification

To establish a clear conceptual foundation for analyzing visual anonymization methods, we propose a taxonomy, as depicted in Figure 1, which categorizes existing approaches according to how they manipulate the original biometric data encoding personally identifiable information. From this perspective, two overarching

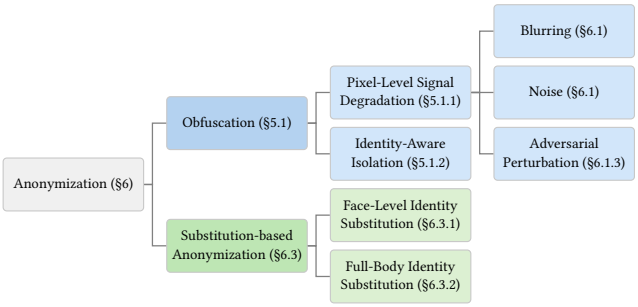


Figure 1: Taxonomy of approaches for video de-identification and anonymization.

paradigms emerge: Obfuscation-based approaches that modify existing identifying traits, and Substitution-based approaches, that mask and subsequently replace areas containing such traits. A quick comparison of the concepts is shown in Table 4. To further illustrate the practical differences between these paradigms, Figure 3 in Appendix C provides a visual comparison of representative anonymization mechanisms.

This division offers a unified perspective that connects classical pixel-level manipulations with contemporary generative anonymization methods, providing a conceptual framework grounded in the degree of dependency on the original biometric signal.

5.1 Obfuscation

Obfuscation refers to a class of anonymization techniques that alter the original biometric data to make it non-identifiable. Unlike replacement or generation-based approaches, obfuscation methods operate directly on raw video V, given the segmented regions $r_{t,i}$ containing identifying information, to ensure that all modifications remain derived from the input data itself. Formally, obfuscation corresponds to a de-identification mechanism $\mathcal{M}$ that transforms regions of identifying information within a set of video streams and outputs anonymized video streams $\tilde{V}^{(c)} = \mathcal{M}_{\text{obf}}(V^{(c)}, r_{t,i}^{(c)})$ with the aim of being anonymous.

The goal is to suppress or distort identifiable features while retaining, to varying degrees, the overall structure or semantic information of the scene. Depending on the method and transformation intensity, obfuscation may preserve global context or, in stronger forms, heavily degrade local details to maximize privacy. Obfuscation techniques can be grouped into two main subcategories: (1) pixel-level signal degradation, and (2) identity-aware isolation and manipulation.

5.1.1 Pixel-Level Signal Degradation. Pixel-level signal degradation methods anonymize visual data by directly modifying the image or video at the pixel level, without analyzing or isolating identity-specific features. In the formal framework introduced earlier, identity-bearing regions are obtained via $S : V \to r_{t,i}$ with $r_{t,i} \in \mathcal{R}$. Pixel-level methods apply transformations directly to these regions without extracting a latent representation via $F : V \times \mathcal{R} \to \mathcal{Z}$. More formally, this class corresponds to de-identification mechanisms $\mathcal{M}_{PSD} : V \times \mathcal{R} \to \widetilde{V}$ such that, for each camera stream $V^{(c)}$ and segmented region $r_{t,i}^{(c)}$, the anonymized region satisfies $\tilde{r}_{t,i}^{(c)} = \mathcal{T}_{\text{pixel}}(r_{t,i}^{(c)})$ where $\mathcal{T}_{\text{pixel}}$ operates directly in pixel space and does not rely on latent identity decomposition.

These approaches suppress or corrupt high-frequency information, the fine-grained appearance cues that make individuals recognizable, while preserving the broader scene structure. Typical transformations satisfy $\mathcal{T}_{\text{pixel}} \in \{\text{Blur, Pixelate, Noise, AdvPerturb}\}$. Common techniques in this category include resolution-reducing transformations such as pixelation and blurring, statistical perturbations such as random noise injection, and model-targeted alterations such as adversarial perturbations. As a result, pixel-level methods offer simple and computationally efficient anonymization but often at the cost of degrading utility for tasks that depend on detailed visual information.

Pixelation and *blurring* represent resolution-degrading obfuscation techniques that conceal identity by reducing the spatial detail of visual data. Both approaches operate on the principle of suppressing high-frequency information that encodes individual traits, thereby generalizing the underlying content while preserving the overall scene structure. Pixelation achieves this effect by dividing the image into coarse square blocks and replacing each block with its mean color value, producing a characteristic mosaic pattern that hides fine textures and facial features. Blurring, in turn, performs continuous smoothing through spatial filtering—typically using a Gaussian kernel—that spreads intensity values across neighboring pixels, reducing local contrast and eliminating sharp boundaries. These methods obscure identifiable cues for both human and machine recognition, but at the cost of losing fine-grained information important for downstream analysis

Noise-based anonymization (or noising) conceals identity by injecting random or structured perturbations into the original data. These perturbations are typically sampled from predefined probability distributions such as Gaussian, Laplacian, or Poisson noise. Unlike identity-aware approaches, noise-based methods do not isolate identity features but instead globally corrupt the visual signal, often degrading both privacy and utility.

Adversarial perturbation anonymizes by introducing precisely optimized, datapoint-level changes that deceive recognition or re-identification models. Unlike random noise, adversarial perturbations are computed via gradient-based optimization to shift the model's feature representation away from the true identity while retaining the utility.

5.1.2 Identity-Aware Isolation and Manipulation. Identity-aware anonymization methods target the specific components of visual data that encode a person's identity (see Figure 4 for examples). These approaches isolate identity-bearing features, such as facial geometry, texture, or characteristic motion, and then remove, neutralize, or transform them within a learned latent space. Such methods rely on extracting a latent representation $z_{t,i}^{(c)} = E(r_{t,i}^{(c)})$, where $z_{t,i}^{(c)} \in \mathcal{Z}$ encodes semantic or biometric attributes of the individual visible in region $r_{t,i}^{(c)}$. The representation is decomposed into identity and non-identity components, $z_{t,i}^{(c)} = (z_{t,i}^{\text{id}}, z_{t,i}^{\text{non-id}})$, where $z_{t,i}^{\text{id}}$ captures identity-related features (e.g., facial structure, body shape, gait cues), and $z_{t,i}^{\text{non-id}}$ captures task-relevant attributes such as pose, motion, illumination, or scene context. Anonymization modifies the identity component, yielding $\tilde{z}_{t,i}^{(c)} = (\hat{z}_{t,i}^{\text{id}}, z_{t,i}^{\text{non-id}})$, where $\hat{z}_{t,i}^{\text{id}}$

denotes a transformed or synthetic identity representation. The anonymized region is reconstructed as $\tilde{r}_{t,i}^{(c)} = D(\tilde{z}_{t,i}^{(c)})$, and integrated into the output stream via the mechanism $\mathcal{M}$, where $D(\cdot)$ denotes a decoder or generator network that maps the modified latent representation back to pixel space.

The goal is to create output that is identity-agnostic yet semantically coherent, as it retains structural and contextual information derived from the same individual. By manipulating only identity-related factors while preserving non-identity attributes like pose, illumination, context, or temporal dynamics, identity-aware methods aim to improve utility at the cost of privacy, and to provide controls for the privacy–utility trade-off. They hence aim to produce coherent, realistic anonymized outputs.

5.2 Substitution-based Anonymization

Substitution-based anonymization refers to techniques that eliminate the original biometric information either by erasing it or by substituting it with new content not derived from the input identity. Substitution can target either the face or the full body. Face substitution replaces only the facial region to remove the appearance-based cues that contribute to identity leakage. While full-body substitution replaces the entire person, including body shape and other structural cues to suppress additional biometric signals. Unlike obfuscation methods, which merely alter the existing data, substitution-based approaches aim to break the identifiable connection between the anonymized and original representations.

Formally, substitution mechanisms $\mathcal{M}_{sub}$ aim to reduce identifiability by replacing identity-bearing regions $r_{t,i}^{(c)}$ with anonymized regions $\tilde{r}_{t,i}^{(c)}$, thereby producing modified video streams $\tilde{V}^{(c)}$. This can be written as $\tilde{V}^{(c)} = \mathcal{M}_{\text{sub}}(V^{(c)}, r_{t,i}^{(c)}, z)$, where z denotes identity features of the individual in the raw video, that are derived from the input region. For anonymization the mechanism samples new $\tilde{z} \leftarrow \mathcal{Z}$.

Masking and *swapping* share the same fundamental privacy principle of removing identifiable information from an image or video. Masking will simply delete any area within the content that may carry identifying information, whereas swapping will replace the content from those specific regions with generated, or copied content from other individuals.

Masking removes identity information by replacing the region with non-informative content, $\tilde{r}_{t,i}^{(c)} = M \odot B$, where M is a binary mask and B denotes background or constant fill. Here, $\odot$ denotes element-wise multiplication between the mask and the background. Masking therefore eliminates the original visual signal in the affected region, removing physiological and (to a large extent) behavioral biometric cues. Swapping replaces the original identity with external or synthetic identity content while preserving scene and behavioral consistency. In generative substitution methods, anonymization may be implemented using a generator model, $\tilde{r}_{t,i}^{(c)} = G(z_{\text{target}}, c_{t,i}^{(c)})$, where $G(\cdot)$ denotes a generative model (e.g., GAN-based or diffusion-based), $z_{\text{target}} \sim \mathcal{P}_{\text{id}}$ is an external identity representation sampled from a distribution of identities, and $c_{t,i}^{(c)}$ denotes conditioning information extracted from the original region. The conditioning signal is typically obtained as $c_{t,i}^{(c)} = \phi_{\text{cond}}(r_{t,i}^{(c)})$, where $\phi_{\text{cond}}(\cdot)$ extracts non-identity attributes such as pose, expression, illumination, or temporal motion cues.

Their effect on the achieved privacy is not strictly equivalent: Masking ensures the complete removal of the original visual signal, eliminating both physiological and (to a large extent) behavioral biometric cues from the affected area. In contrast, swapping may still retain certain residual behavioral or dynamic signals, such as gaze direction, facial muscle movements, or characteristic expression patterns, which can indirectly reveal identity.

The distinction becomes significant primarily from a utility perspective. Masking erases the region entirely using a binary mask or abstract shapes, which may disrupt visual consistency or downstream task performance. In contrast, swapping replaces it with realistic content that can be either obtained from a database of existing individuals or synthetically generated by a model such as a GAN or diffusion network. Swapping can be regarded as a more utility-preserving alternative to masking, but one that requires careful design to avoid leaking behavioral biometric information.

5.3 Utility Goals

While anonymization primarily aims to suppress or remove personally identifiable information, its practical deployment across real-world domains is guided by an equally important objective, the preservation of data utility. In practice, utility spans several dimensions, including visual realism, semantic and attribute preservation, machine-useful utility for downstream model performance, and the retention of behavioral and motion cues.

Utility, in our context, refers to the degree to which anonymized visual data remains informative, interpretable, and functionally valuable for subsequent machine or human analysis.

In surveillance and public safety, utility means for instance enabling object detection, multi-target tracking, and activity recognition without revealing identities. This requires preserving machine-relevant features and motion patterns; methods such as conditional GAN-based replacement or 3D mesh substitution maintain pose, dynamics, and spatial consistency while rendering individuals unrecognizable. A related safety goal is selective identification, where a secret key allows identity recovery when necessary, for example by law enforcement.

Within medical and healthcare domains, the notion of utility extends to the retention of diagnostic or clinical features. This reflects a stronger emphasis on semantic and attribute preservation, where anonymization must protect patient identity while keeping pathology-relevant information—such as retinal lesions, anatomical landmarks, or facial expressions—intact for accurate disease assessment or behavioral analysis. However, this creates an inherent tension: the same behavioral or physiological signals that are clinically informative may also serve as biometric identifiers, meaning that preserving utility can unintentionally reintroduce identity leakage.

In behavioral and interaction-based contexts, utility emphasizes preserving communicative and motion-related information. This corresponds to behavioral and motion preservation, where anonymizing sign language videos or mixed-reality motion data, systems must maintain gesture trajectories, body articulation, and

temporal smoothness to ensure that human communication or immersive interaction remains natural. However, this also reveals a deeper challenge: even when visual identity is removed by replacing the signer with an avatar, behavioral or linguistic identity cues may persist. One more privacy-preserving alternative would be to interpret the signed content into a semantic representation and then generate an avatar to perform the extracted meaning. Yet, even in such cases, individuals may remain identifiable through characteristic signing styles, phrasing, or communication patterns. This illustrates that privacy and utility exist on multiple levels of abstraction from raw visual appearance to motion dynamics and semantic content, and effective anonymization fundamentally should consider how identity may leak across these layers.

For multi-view camera anonymization, a new utility goal is cross-view appearance consistency. This refers to the anonymized representation of a person to remain visually consistent across camera views. It is important for applications such as tracking people through multiple camera views, for example, to understand consumer behavior in supermarkets [119, 165]. High perceptual realism achieved through harmonization networks or illumination-aware blending ensures that anonymized datasets remain suitable for training and evaluating downstream computer vision models. Evaluation includes embedding similarity across views, perceptual similarity metrics (LPIPS), and appearance drift scores.

In machine learning and data sharing, utility relates to the statistical and algorithmic usability of anonymized data. Anonymization should not disrupt feature distributions or degrade performance in classification, detection, or segmentation tasks. Task-aware methods, such as adaptive obfuscation or differential privacy frameworks, explicitly optimize this trade-off by maintaining model accuracy and decision invariance under anonymization. These utility dimensions are interdependent; improving one (e.g., realism or behavioral fidelity) can elevate privacy risks, while stronger privacy mechanisms may reduce model or human interpretability. Thus, while privacy mechanisms eliminate identity information, they simultaneously preserve the representational fidelity necessary for reliable, reproducible, and ethically compliant computer vision research.

6 Classification and Overview of Video Anonymization Methods

We now describe the literature on video anonymization methods according to the proposed taxonomy. Table 1 provides a comparative overview of the surveyed studies across key methodological and evaluation axes.

6.1 Pixel Level Signal Degradation Obfuscation

Pixel-level signal degradation methods aim to anonymize videos by degrading visual fidelity through blurring or noise perturbation, thereby suppressing identifiable cues. Unlike image-based obfuscation, which only affects static frames, video protection must maintain temporal consistency and cross-frame coherence, ensuring that identity is not reassembled from motion continuity or multi-view overlap. As a result, modern video approaches typically integrate object detection or segmentation networks with spatio-temporal filtering or learnable degradation modules to preserve both privacy and visual utility.

6.1.1 Blurring-based Obfuscation. The earliest video anonymization mechanisms relied on spatio-temporal blurring to suppress identifiable details while maintaining action visibility. For instance, Agrawal and Narayanan [4] applied $\mathcal{T}$: voxel-space exponential blur and line integral convolution on segmented human volumes (S), achieving full-body de-identification that preserved movement cues but introduced ghost-like distortions. Building on this, Ivašić-Kos *et al.* [77] used Gaussian filters ($\mathcal{T}_{\text{blur}}$) over multi-view silhouettes (r) to conceal gait and clothing patterns while retaining global motion, providing early insights into privacy–utility trade-offs for action recognition.

With deep learning, these concepts evolved into real-time blurring ($\mathcal{T}$) pipelines that first detect sensitive regions and then modify them to remove identifying signal. Angus *et al.* [13] implemented a YOLOv4-based detector to locate faces (S) and license plates in city surveillance footage and applied Gaussian blur ($\mathcal{T}_{\text{blur}}$) at over 60 FPS.

Similarly, Grosselfinger *et al.* [54] extended this idea to multimodal settings by improving detection across RGB, IR, and grayscale video, using YOLOv3 and OpenPose to detect faces (S) and license plates across panoramic sensors.

Recent implementations, such as Ahmed *et al.* [7], adapted YOLOv9 architectures for child face (S) anonymization in real time, focusing on social media protection and ethical data publication. However, despite its high detection accuracy, the approach still suffers from limitations in complex scenarios, such as low lighting, motion blur, and occlusions, where the model may fail to detect (S) or correctly blur faces ($\mathcal{T}_{\text{blur}}$).

FaceOff [49] presents an obfuscation-based anonymization method for surgical operating-room videos that detects faces (S) and conceals identity with blur ($\mathcal{T}_{\text{blur}}$). It employs a region-proposal two-stage detector (e.g., Faster R-CNN) to generate candidate regions ($\mathcal{R}$), classify faces, and refine bounding boxes via shared convolutional features; temporal smoothing aggregates detections across neighboring frames to recover missed faces under occlusion or rapid motion.

MedVidDeID [112] introduces an obfuscation-based anonymization pipeline for clinical encounter videos that removes text, audio, and visual Protected Health Information (PHI) within a unified workflow. It combines speech-to-text transcription, textual PHI filtering, voice transformation, and pose-guided video anonymization using selective or full-frame face (S) blurring ($\mathcal{T}_{\text{blur}}$). The system has notable limitations due to the complexity of clinical environments: rapid movements, partial occlusions, and reflections in glass surfaces or mirrors frequently disrupted detection.

6.1.2 Noise-based and Perturbation-based Obfuscation. To overcome the limitations of static blur, noise-based anonymization introduces controlled perturbations or learned degradations to obscure identity while maintaining temporal and structural realism.

For gait anonymization, Tieu *et al.* [141] introduced a Spatio-Temporal GAN (ST-GAN) that injects Gaussian noise ($\mathcal{T}_{\text{noise}}$) into gait representations (F) to produce anonymized silhouettes in which recognition-relevant gait cues are intentionally degraded while using spatial and temporal discriminators to maintain visual realism and smooth motion. The method alters contour shape but leaves

Table 1: Video Anonymization Studies

Study	TC	MC	RL	R	Model Architecture	Dataset	PM	UM	UG
Agrawal [4]	●	□	▲	B	Line Integral Convolution	CAVIAR [33], BEHAVE [24]	HCPE	HCUE	■
Ivašić-Kos [77]	●	■	▲	G	Gaussian silhouette	i3DPost [51]	RID	TP	■
Angus [13]	●	□	▲	F	YOLOv4 + Gaussian	COSMOS [13]	RID	TP	●
Grosselfinger [54]	●	■	▲	F	YOLOv3 + OpenPose	MODISSA (custom) [54]	RID	TP	●
Ahmed [7]	●	□	▲	F	YOLOv9 detection	Custom (296 images) [7]	RID	TP	●
FaceOff [49]	●	□	▲	F	Faster R-CNN	WIDER Face [158], FaceOff [49]	RID	TP	●
MedVidDeID [112]	●	□	▲	F, B	YOLOv11+Kalman-filter	clinical Audio-Video(AV)	RID	NA	★
Tieu [141]	♦	□	▲	G	GAN	CASIA-B [166]	RID	HCUE, TP	■
Tieu [139]	♦	□	▲	G	GAN	CASIA-B [166]	RID	HCUE,PQ	■
Parashar [120]	♦	□	▲	G	Random perturbation	CASIA-B [166]	HCPE	PQ	★, ■
Ahmad [5, 6]	♦	■	▲	B	Autoencoder	Event-ReId [5]	RID	TP	●
Adra [2]	♦	□	▲	F	U-Net (E2VID)	VETEX [3], FES [23]	RID	PQ,TP	●
Yalçın [157]	♦	□	♦	F	Stylization (CNN)	Custom Dataset [157]	HCPE	HCUE	◆
Tieu [140]	▼	□	♦	G	CNN encoder-decoder	CASIA-B [166]	RID	HCUE	■
Hirose [66]	▼	□	♦	G	VAE-based gait	OU-ISIR [103]	RID	HCUE	■
Hirose [67]	▼	□	♦	G	CNN + texture	OU-ISIR [103] + Custom Data [67]	RID	TP	■
Thapar [138]	▼	□	♦	G	Attention-based CNN	EPIC-KITCHENS [34], FPSI [137], IITMD [47]	RID	TP	●
SPAct [37]	▼	□	♦	B	Self-sup. adversarial	UCF101 [135], HMDB51 [89], VISPR [118]	RID	TP	●
De Coninck [38]	▼	□	♦	B	Lightweight CNN	Avenues [102], ShanghaiTech [99], WILD-TRACK [30]	RID	TP	●
RiDDLE [93]	×	□	●	F	StyleGAN2 + trans.	CelebA-HQ [80], FFHQ [81], LFW [71]	RID	TP	★
Color Transfer [142]	×	□	♦	G	Encoder-decoder	CASIA-B [166]	RID	PQ	▲, ■
CIAGAN [105]	■	□	●	F	GAN-based	CelebA [101], MOTS [146], LFW [71]	RID	PQ, TP	■, ▲
Decoupling [106]	■	□	●	F	GAN	CelebA [101], FF++ [130], LFW [71]	RID	TP	▲
SimSwap [31]	■	□	●	F	Encoder-Decoder GAN	VGGFace2 [29], FF++ [130], CelebA [101]	NA	TP	◆
E4S [100]	■	□	●	F	Encoder + GAN	CelebAMask-HQ [91], FFHQ [81]	NA	TP,PQ	◆
FaceDancer [128]	■	□	●	F	GAN-based	VGGFace2 [29], LS3D-W [25], FF++ [130]	NA	TP,PQ	◆
Wen [152]	■	□	●	F	Face synthesis net	VoxCeleb [113]	RID	TP,PQ	▲
FIVA [129]	■	□	●	F	GAN-based	VGGFace2 [29], LFW [71], FF++ [130]	RID	TP	▲
XimSwap [10]	■	□	●	F	GAN-based	VGGFace [29]	RID	TP	◆
SafeTriage [26]	■	□	●	F	GAN-based	Clinical dataset [26], FFHQ [81]	RID	HCUE, TP	●
3PFS [163]	■	□	●	F	GAN-based	VGGFace2 [29], JAAD [123]	RID	TP,PQ	◆
DisguisOR [18]	★	■	♦	B	VoxelPose + SMPL	COCO [97], Panoptic [78]	NA	HCUE, TP	●, ■
Huh [72]	★	□	♦	B	OpenPose + SMPL	SIAT [72]	NA	TP	●
Cartoonized [145]	★	□	♦	B	MediaPipe	Sign language custom Data [145]	NA	NA	■
DiffSign [87]	★	□	●	B	Diffusion + pose	Custom Sign [87], FFHQ [81]	NA	PQ, TP	■
DiffSLVA [156]	★	□	●	B	Diffusion + ControlNet	ASLLRP [114], ASL [156], VoxCeleb [113]	RID	NA	■
Zhang [160]	★	□	●	B	U-Net + LLM	How2Sign [41], ASLLRP [114]	NA	PQ	■
GTNet [57]	★	□	♦	G	GAN	CASIA-B [166], TUM-GAID [69]	RID	HCUE, TP	■

TC (Taxonomy):
- ● Blurring-based
- ♦ Noise-based
- ▼ Adv. Perturbation
- × ID-Aware Isolation
- ■ Sub. (Face)
- ★ Sub. (Body)

MC (Multi-Cam):
- □ NO
- ■ YES

RL (Realism):
- ▲ Low
- ♦ Med
- ● High

R (Regions):
- F: Face
- B: Full-body
- G: Gait

PM (Privacy Metrics):
- RID: Re-identification Risk
- HCPE: Human-Centric Privacy Evaluation
- NA: Not Available

UM (Utility Metrics):
- PQ: Perceptual Quality
- HCUE: Human-Centric Utility Evaluation
- NA: Not Available
- TP: Task-Specific Performance

UG (Utility):
- ■ Behavioral/Motion
- ● Machine-Understandable
- ★ Selective Identification
- ◆ Semantic/Attribute
- ▲ Visual Quality

deeper temporal cues such as gait speed and cycle timing largely unchanged, which may still carry identity information.

To mitigate these issues, the authors later proposed an RGB reconstruction framework [139] combining a DCGAN-based anonymizer (A-NET) that removes identifiable gait cues (F) with a Siamese colorization network (C-NET) that restores appearance, enabling realistic anonymized RGB gaits even from low-quality silhouettes.

Parashar *et al.* [120] present a reversible gait anonymization framework that applies controlled geometric and textural perturbations to gait silhouettes, enabling privacy-preserving data sharing with identity restoration through stored transformation parameters. Geometric anonymization slightly scales and morphs silhouette shapes—modifying limb proportions. Textural anonymization replaces each silhouette's pixel texture with an averaged multi-subject template. Combined, these transformations produce visually plausible yet identity-free silhouettes that can later be reconstructed from the saved parameters. The authors assess utility in terms of preserving natural movement.

A parallel research line integrates learned noise and blurring into event-based vision. Ahmad *et al.* [5, 6] proposed an autoencoder-based anonymization framework that processes event voxels—spatio-temporal volumes of brightness changes captured by event cameras to conceal identity before video reconstruction. The anonymizer is jointly trained with an attacker network (E2VID) that attempts to reconstruct gray-scale images and with downstream task networks for person re-identification (ReID) and human pose estimation (HPE).

Adra and Dugelay [2] introduced E2PRIV, that embeds anonymization directly into the event-to-video reconstruction model rather than adding it as a post-processing step, where data from event cameras, sensors capturing only brightness changes rather than full frames, are converted into videos. Built upon the E2VID model,

which employs a U-Net (encoder–decoder) architecture for image reconstruction, the system integrates dynamic Gaussian blur ($\mathcal{T}_{\text{blur}}$) and noise injection ($\mathcal{T}_{\text{noise}}$) layers to obscure facial details (S) during reconstruction. While the approach reduces identifiable details, the authors show that face verification accuracy drops substantially for VGG-Face and DeepFace but remains high for ArcFace, indicating that some identity cues survive anonymization, and the resulting videos appear visually degraded to human observers.

Yalçın *et al.* [157] proposed a novel AI-driven stylization method for video-based anonymization that replaces traditional blurring ($\mathcal{T}_{\text{blur}}$) or pixelation ($\mathcal{T}_{\text{pixelate}}$) with artistic transformations. Their approach modifies facial geometry (S) and applies aesthetic rendering models that both anonymize appearance and preserve emotional expressiveness in interview videos. Through two user studies, they demonstrated that stylized anonymization not only conceals identity but also enhances empathic engagement and emotional understanding among viewers.

6.1.3 Obfuscation–Adversarial Perturbation-Based Video Anonymization. This class of methods follows the principle of concealing identity by injecting optimized, task-aware perturbations, in pixels, motion fields, or latent features, to suppress re-identification while preserving downstream utility.

Video-based perturbation must preserve temporal consistency, ensuring coherence across frames and robustness against sequence-level aggregation. Consequently, these methods operate in spatio-temporal or feature spaces, generating motion-consistent adversarial signals that conceal identity while maintaining realism, continuity, and analytical usability.

Early approaches focused on gait anonymization through structural perturbation. Tieu *et al.* [140] employed a convolutional encoder–decoder to merge original and "noise gait" contours ($\mathcal{T}_{\text{noise}}$), producing anonymized silhouettes that disrupt identity-specific motion while preserving natural walking patterns.

Building on this, Hirose *et al.* [66] proposed a (VAE) framework that disentangles shape and phase components of gait and applies targeted perturbations to each, achieving identity suppression without destroying temporal smoothness. Their later extension integrates a CNN-based encoder–decoder with a texture transfer module, allowing realistic RGB reconstruction while maintaining anonymization of static and dynamic gait cues[67].

Beyond silhouette representations, Thapar *et al.* [138] introduced an adversarial motion perturbation framework for egocentric videos, where wearer identity is embedded in optical flow and camera egomotion. Using a deep attention-based recognition network (EgoIdNet) that learns wearer-specific motion patterns, and an ArcFace loss function that enforces strong separation between different identities in the feature space, they backpropagate identity gradients into small homography-based motion perturbations, effectively anonymizing the wearer while keeping activity recognition intact.

Adversarial learning expanded perturbation to feature-level anonymization. SPAct [37] is a self-supervised framework to remove privacy cues from video without requiring privacy annotations. The system has three components: an anonymizer that transforms video frames (V) to hide identity, a privacy discriminator that detects remaining identity cues, and an action recognition branch to preserve task utility. They are trained adversarially and contrastively: the anonymizer removes private information while retaining motion cues for activity understanding. However, SPAct relies on basic self-supervised frameworks designed mainly for action recognition, limiting its extension to tasks like detection or anticipation.

Similarly, De Coninck *et al.* [38] developed a lightweight adversarial framework for privacy-preserving video analytics on edge devices. Their approach trains a compact obfuscator that transforms input frames (V) into privacy-protected representations, while a competing deobfuscator attempts to reconstruct the original image; this adversarial pairing forces the obfuscator to remove identity-related details.

6.2 Identity-Aware Isolation and Manipulation

This class of methods conceals identity by separating identity-related factors from other semantic or temporal representations in visual or motion data. Unlike image anonymization, which disentangles identity (z^{id}) from appearance attributes ($z^{\text{non-id}}$) (e.g., pose, lighting, texture), video-based approaches must additionally decouple temporal dynamics such as gait, motion trajectories, and behavioral patterns. Models typically employ encoder–decoder or variational architectures that learn structured latent spaces ($\mathcal{Z}$), using adversarial or contrastive objectives to isolate and suppress identity components (z^{id}) while retaining non-identifying motion and appearance information ($z^{\text{non-id}}$). Operating in spatio-temporal latent spaces, these methods obfuscate identity cues while preserving realism and continuity. Compared with pixel-level obfuscation, they offer a better privacy–utility trade-off but remain constrained by high computational cost and modality-specific design.

Early latent disentanglement frameworks such as RiDDLE [93] introduced the idea of identity encryption in latent space ($\mathcal{Z}$), where facial (S) identity embeddings are separated from expression and appearance features within a StyleGAN2 generator. Through a transformer-based encryptor conditioned on passwords, the model obfuscates the identity component (z^{id}) while preserving photo-realistic texture and motion coherence ($z^{\text{non-id}}$), enabling reversible and diverse anonymized faces. This work established the feasibility of latent-level obfuscation but remained limited to face-level anonymization and computationally expensive inversion.

Building on this foundation, Color Transfer Gait Anonymization [142] extends disentanglement from appearance to the spatio-temporal motion domain. Rather than encrypting identity embeddings, it separates appearance from motion by using anonymized silhouettes to encode gait structure and an encoder–decoder color-transfer network to reconstruct natural color and texture. This restores a realistic appearance while maintaining privacy.

6.3 Substitution-Based Video Anonymization

This class of methods conceals identity by replacing the original face or body with a newly generated surrogate, rather than degrading or masking it. While image-based swapping operates on individual frames, video anonymization requires temporal stability, consistent surrogate identities across time and viewpoints, and seamless blending into complex backgrounds. Accordingly, these approaches integrate identity-guided generative models, tracking modules, or 3D reasoning to maintain visual realism, continuity, and control over the anonymization process. In this section, we

categorize substitution-based video anonymization methods by the identity replacement region: face-level substitution (Sec 6.3.1) and full-body-level substitution approaches (Sec 6.3.2).

6.3.1 Face-Level Identity Substitution. We review face-level substitution, in which anonymization is achieved by replacing faces in the video with surrogates. Early frameworks such as CIAGAN [105] established the foundation for controlled replacement through a conditional GAN architecture. CIAGAN receives pose landmarks as input and synthesizes realistic surrogate identities (ξ) using a U-Net (encoder-decoder) generator guided by a Siamese identity discriminator, which enforces both pose preservation and identity independence. A key advancement lies in its ability to generate new, unseen identities while maintaining expression and temporal coherence through landmark smoothing. However, CIAGAN still depends heavily on accurate landmark detection and struggles under occlusions or extreme poses, limiting its robustness in unconstrained video environments.

To overcome the entanglement between visual realism and identity synthesis, Decoupling Identity and Visual Quality [106] proposed a two-stage GAN framework with an AnonymizationNet for generating diverse identities and a HarmonizationNet for blending them naturally into the scene. This separation allows training on high-quality datasets without identity labels, improving photorealism while maintaining privacy. Its recurrent extension with a temporal discriminator ensures smoother transitions across frames. Nevertheless, this dual-network design increases computational cost, and blending errors occasionally appear in cluttered or dynamic scenes. Moreover, temporal consistency is enforced only in the blending stage, not in the core anonymization module, allowing frame-wise identity drift.

A major direction in identity-replacement focuses on high-fidelity face swapping, where identity cues are fully substituted while preserving target pose, expression, and illumination. SimSwap [31] accomplishes this using a shared encoder, feature-level identity-injection module, and decoder that reconstructs a surrogate face aligned with the target's appearance factors. E4S [100] advances this through region-wise disentanglement, pairing a pose-alignment

module with a regional encoder–decoder and a mask-guided generator that injects surrogate identity features (ξ) into local facial components. FaceDancer [128] further improves robustness via adaptive feature-fusion attention and identity-feature regularization to better handle occlusion, extreme poses, and illumination changes. Across these methods, anonymization is achieved by replacing identity-bearing geometry and texture with surrogate identity vectors while preserving non-identity attributes needed for downstream tasks.

Wen *et al.* [152] propose a generative identity-replacement approach that leverages temporal redundancy in face videos to achieve high-quality anonymization with minimal flicker. The system uses a dense motion-flow generator (G) to estimate relative motion between consecutive frames. Anonymization occurs by creating a synthetic face (ξ) for the first frame and then propagating this anonymized frame through the entire sequence using motion-flow warping. Privacy is supported by large reductions in identity similarity, while utility is preserved through high face-detection rates and minimal flickering compared to per-frame generative methods.

FIVA [129] extended face replacement toward identity tracking and adversarial defense. The architecture integrates an Identity Tracking Module (ITM) that assigns stable pseudo-identities throughout a video, combined with anchor-based sampling to generate embeddings far from both the source and reconstructed identities. Moreover, FIVA introduces defensive noise layers to resist reconstruction or inversion attacks, making it one of the first models to explicitly address anonymization reversal. While it achieves temporal consistency, its reliance on adversarial perturbations and high-complexity modules makes real-time deployment challenging. XimSwap [10] presents a lightweight encoder–decoder face-swapping system for real-time video anonymization on edge devices. It employs a compact face-detection (S) and feature-extraction (F) pipeline together with an identity-injection module that reconstructs each face using a surrogate identity (ξ) while preserving the original pose and expression. Anonymization replaces all appearance-based identity cues so only anonymized frames are transmitted off-device, while utility is maintained through stable pose tracking and low landmark error that supports downstream analysis.

SafeTriage [26] introduces a generative identity-replacement method for clinical facial videos that preserves patient motion while removing appearance-based identity cues. The system integrates a conditional generator that creates a surrogate face (ξ) aligned to the patient's pose and structure with a motion-transfer network that retargets facial dynamics—such as asymmetry, lip motion, and head pose—onto this synthetic identity. Privacy is validated through reduced face-embedding similarity, and utility is supported by clinical expert evaluations and stable stroke-triage performance on synthetic data.

3PFS [163] proposes an identity-replacement anonymization method for pedestrian videos that uses real surrogate faces (ξ) rather than synthetic ones, preserving natural appearance and soft attributes such as age and gender. The pipeline combines a pedestrian detector, face detector (S), and pre-processing modules (deblurring and facial refinement) with an attribute-aware face-selection algorithm and an encoder–decoder face-swapping network guided by a pose-preservation loss. Anonymization proceeds by detecting each pedestrian's face, enhancing it, selecting a surrogate with matching soft attributes but maximal identity distance, and blending the swapped face into the frame; tracking ensures temporal consistency.

Complementing generative methods, Lee et al. [92] propose face-transformation–based anonymization for ASL with three variants: (1) a mask-overlay filter that replaces the face with a cartoon 3D mask (ξ), and (2–3) two face-swap pipelines that detect facial regions, transfer expressions, and generate a composite face via latent-space ($\mathcal{Z}$) editing while preserving the signer's motion. One variant further removes the torso and background to suppress clothing and scene cues. A user study shows that replacing facial identity while retaining non-manual markers preserves intelligibility and perceived message accuracy. Finally, BLANKET [56] adapts swapping for infant video anonymization by coupling diffusion-based inpainting (for surrogate identity (ξ) generation) with temporal face swapping (FaceFusion). This hybrid pipeline maintains expression, gaze, and head-pose consistency, achieving realistic anonymization suitable for developmental and psychological studies. Its limitation

lies in diffusion's computational cost and reduced distinctiveness among infant faces.

6.3.2 Full-Body Identity Substitution. Here we review full-body substitution techniques that replace person's visual with a surrogate ones. Beyond 2D replacement, DisguisOR [18] extends anonymization to the 3D domain for complex multi-camera environments such as operating rooms. Rather than altering faces in individual frames, the method reconstructs a 3D digital body model from synchronized color–depth recordings and aligns it to each person's position and orientation in the scene. The facial region is then substituted with a generic identity-free template, and the anonymized model is rendered back into all camera views to generate realistic, view-consistent results. However, the approach depends on depth sensors and multi-stage alignment and rendering pipelines, which introduce high computational cost and currently hinder real-time or large-scale deployment.

A recent line of work integrates anonymization with data minimization by reducing the sensitive visual information transmitted across systems. In this direction, Huh *et al.* [72] propose a scene-preserving framework that replaces each person in a video with a 3D digital avatar (ξ) whose body structure and facial expressions replicate the original. Instead of sending full video streams, the system transmits compact numerical representations of these avatars from low-power edge hardware to a central processor, limiting exposure to raw imagery while retaining the spatial and behavioral cues needed for action and interaction analysis.

In the communication domain, anonymization methods increasingly aim to preserve the linguistic integrity of signing while concealing the signer's(people performing sign language) identity. Cartoonized Anonymization [145] replaces real signers with stylized animated avatars (ξ) by tracking body and hand motion through pose-estimation models and retargeting these motions to a cartoon character. This maintains essential communicative cues, hand articulation, head movement, and gesture timing, while removing visual identity, and its non-photorealistic rendering helps avoid the uncanny-valley issues often observed in neural synthesis.

Building on this, DiffSign [87] introduces a diffusion-based signer-replacement pipeline that extracts 2D keypoints, retargets them to a 3D parametric avatar (ξ) for stable pose trajectories, and then conditions a generative diffusion model using edge maps and pose controls, with an appearance adapter injecting features from a reference image to define the surrogate signer. This produces photorealistic identity-free videos while retaining expressive hand–body articulation and temporal smoothness.

Further advancing diffusion-based anonymization, DiffSLVA [156] proposes a text-guided framework that avoids pose-estimation entirely by using low-level edge features (HED) through ControlNet to guide large pre-trained diffusion models. The system integrates cross-frame attention for temporal coherence and a specialized facial-expression enhancement module to preserve linguistically crucial non-manual markers such as brow movement and mouth shapes. Anonymization is achieved by generating a fully synthetic signer (ξ) that follows the original motion but appears as a computer-generated individual defined by a text prompt.

Zhang *et al.* [160] present a modular ASL video generation system that produces photorealistic signing from English by jointly modeling manual and non-manual cues. It has three stages: an LLM module predicting glosses and facial grammar, a motion synthesis stage using Motion Matching for coherent poses, and an image-to-image module rendering photorealistic frames conditioned on signer identity. While not built for anonymization, it can serve as one by replacing real signers with synthetic, identity-agnostic avatars (ξ) that preserve linguistic structure, yielding privacy gains with minimal utility loss. A study with 30 DHH participants shows high understandability, but notes facial quality and motion smoothness issues that may reduce realism and comfort.

Across these methods, privacy arises from comprehensive identity removal, while utility is preserved through accurate motion transfer and retention of linguistic cues. However, they still rely on reliable motion cues or edge extraction, and diffusion-based generation remains computationally demanding. Additionally, they lack explicit mechanisms for multi-view consistency or protection against strong re-identification attacks, meaning some behavioral signatures, such as signing style, may remain traceable. User studies also highlight that mismatched facial attributes, removal of torso context, or residual visual artifacts may reduce naturalness and comfort for viewers, particularly in face-swap.

In parallel, a smaller line of work focuses on substitution-based gait anonymization, which targets motion-based identity signals rather than appearance. GTNet [57] employs a conditional GAN with dense-pose fusion and attention-based discriminators to transform gait sequences into new identity-free versions while retaining pose and texture. It effectively reduces gait recognition accuracy while maintaining motion naturalness, yet still struggles under real-world occlusions or varied clothing.

7 Datasets

There is a clear benchmark-to-application gap in the video anonymization domain. Most evaluation datasets are built from still images or short clips from single scenes. In reality, deployed systems operate on multi-view surveillance cameras, wearable devices, or medical recordings, where identity must be continuously suppressed despite long temporal persistence and camera view shifts. This section discusses benchmark misalignment and its impact on temporal evaluation reliability. We report only the top 10 most utilized for these tasks, as identified in this survey. A comprehensive overview is provided in Table 3 in the Appendix C .

We would like to note, first, that there is a broad range of datasets with facial images (also included in the appendix), with varying perspectives and diversity. Some video anonymization approaches indeed are evaluated on such data, which raises concerns regarding the level of protection these approaches actually achieve.

Video datasets remain misaligned with anonymization goals despite containing person labels. Considering face datasets, FaceForensics++ [130] and MEAD [147] provide videos of talking identities, with the latter recording faces under 7 view shifts, yet both are captured in strictly controlled, single-identity settings. Considering gait, CASIA-B [166] and TUM-GAID [69] provide walking videos from multiple viewpoints; the latter includes a depth modality beyond RGB, as well. Still, these datasets mainly contain single-person walks or short talking segments with minimal crowds, occlusion, or interaction, which limits real temporal testing. Large-scale person

frames sourced from Pexels [104] improve scale and appearance variety but remain identity-unlabeled and temporally unlinked, with no continuous identity tracks across camera transitions. Overall, only a minimal number of datasets allow real measurement of temporal identity anonymization at scale. It would hence be a big service to the community if datasets covering long, continuous tracks, multi-view video, and identity labels where made available, preferably with labelled activity or other features for downstream tasks, for a convincing privacy—and utility—evaluation.

8 Discussion & Takeaways

Visual anonymization has progressed rapidly from classical pixel-level transformations to advanced deep learning–based generative methods. Early techniques—blurring, pixelation, masking—offered basic privacy but substantially degraded utility and realism. Modern approaches built on GANs, VAEs, and diffusion models can synthesize realistic, identity-free imagery while preserving expressions, motion, and contextual structure. This shift enables more controllable, task-aware anonymization that selectively removes identity cues while maintaining downstream usability for detection, pose estimation, and behavioral analysis.

Despite substantial progress, key structural limitations remain. Most methods assume single-view facial settings and anonymize frames independently, without modeling cross-view geometry or enforcing temporal consistency—an inherent constraint of 2D architectures lacking multi-view correspondence. As a result, few approaches address full-body, gait, or multi-view video anonymization. We argue that combining visual anonymization (e.g., face or body) with behavioral anonymization (e.g., gait or gaze) is the next step toward holistic video protection; for instance, integrating motion anonymization with inpainting could provide stronger combined safeguards. This single-view focus conflicts with real-world deployments in smart cities and medical settings, where synchronized multi-view capture requires view-consistent anonymization to preserve both privacy and downstream utility.

Across the literature, obfuscation-based methods show clear limitations. They preserve coarse layout and context but degrade the fine-grained visual features needed for downstream analysis. Blurring, pixelation, and noise distort body articulation, facial geometry, and motion cues, reducing structural and temporal fidelity. Moreover, obfuscation does not fully remove identity information: residual biometric signals such as gait, body proportions, and low-frequency facial patterns often persist and re-emerge across time or viewpoints, making these methods unsuitable for multi-view or long-term video anonymization.

Removal-based approaches, including GAN- and diffusion-based replacement, address these limits by generating new surrogate identities that break structural and semantic ties to the original, strengthening resistance to cross-view linking and temporal aggregation. However, they rely on accurate segmentation, tracking, and 3D/pose estimation, so errors in these stages cause temporal or view inconsistencies. They are also computationally intensive, limiting real-time use and large-scale evaluation.

Another challenge is the reliability of reported results. Many studies test privacy against weak adversaries, often using recognition models trained only on original data, which overestimates protection. As Hanisch *et al.* [60] note, such protocols ignore adversaries who adapt to anonymized data. Evaluations are also frequently limited to small datasets, restricting assessment of cross-view consistency and long-term temporal coherence.

From an evaluation perspective, privacy and utility do not follow a simple trade-off; the relationship varies widely across anonymization families and methods. Some approaches, such as blurring and pixel-level degradation, achieve privacy by suppressing fine-grained facial and body cues, but this significantly reduces utility for downstream tasks. Others, particularly identity-replacement and latent-space methods, maintain high realism yet may still leak residual identity cues by preserving behavioral or dynamic signals that enable re-identification across frames or viewpoints.

In summary, despite advances in realism, controllability, and methodological rigor, significant gaps remain in achieving multi-view consistency, developing lightweight architectures, and establishing standardized evaluation protocols for fair comparison and reproducibility [18]. These limitations highlight multi-view and multi-perspective anonymization as central open research challenges. Future research should explore geometry-aware or 3D-consistent generative anonymization frameworks capable of maintaining temporal and spatial coherence across multiple cameras, a necessary step towards effective and scalable anonymization solutions that can operate securely in real-world environments such as smart-city CCTV networks, autonomous systems, and medical data pipelines.

9 Conclusion

This survey aims to clarify the expanding landscape of video anonymization. We reviewed the field systematically and organized existing methods into a taxonomy that highlights how approaches relate, what identity cues they protect, how they operate, and the trade-offs they entail, rather than presenting techniques in isolation. A clear pattern in the literature is the continued focus on single-view videos, despite real environments rarely operating that way. In public spaces, hospitals, and transportation hubs, individuals are routinely captured by multiple cameras, yet only a few studies consider such settings, creating a significant gap between current research and practical deployment.

Our analysis of datasets and evaluation metrics shows that most benchmarks are limited and fail to reflect multi-camera conditions, underscoring the need for stronger datasets and clearer evaluation standards, especially for methods intended to operate across multiple viewpoints. Lastly, we emphasized that anonymization must balance privacy with the utility required for downstream tasks. Taken together, these insights point toward future research that treats multi-perspective data as a core requirement, develops techniques suited to the complexity of real-world video systems, and finally improves methodology towards more rigorous, and convincing privacy evaluations.

References

[1] Sara Abdulaziz, Giacomo D'amicantonio, and Egor Bondarev. 2025. Evaluation of Human Visual Privacy Protection: Three-Dimensional Framework and Benchmark Dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[2] Mira Adra and Jean-Luc Dugelay. 2025. E2PRIV: Privacy-Preserving Event-to-Video Reconstruction with Face Anonymization. In *2025 13th International Workshop on Biometrics and Forensics (IWBF)*.

[3] Mira Adra, Nelida Mirabet-Herranz, and Jean-Luc Dugelay. 2024. Beyond RGB: Tri-Modal Microexpression Recognition with RGB, Thermal, and Event Data. In *International Conference on Pattern Recognition*.

[4] Prachi Agrawal and PJ Narayanan. 2011. Person de-identification in videos. *IEEE Transactions on Circuits and Systems for Video Technology* (2011).

[5] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. 2023. Person re-identification without identification via event anonymization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[6] Shafiq Ahmad, Pietro Morerio, and Alessio Del Bue. 2024. Event anonymization: Privacy-preserving person re-identification and pose estimation in event-based vision. *IEEE Access* (2024).

[7] Hunar A Ahmed, Mohammed H Ahmed, Jafar Majidpour, and Saman M Omer. 2025. Enhanced child face anonymization using YOLO framework for real-time privacy protection in videos. *Neural Computing and Applications* (2025).

[8] Rouqaiah Al-Refai, Pankaja Priya Ramasamy, Ragini Ramesh, Patricia Arias-Cabarcos, and Philipp Terhörst. 2025. A Comprehensive Re-Evaluation of Biometric Modality Properties in the Modern Era. *Journal of Evidence-Based Medicine* (2025).

[9] Mohammed Talha Alam, Fahad Shamshad, Fakhri Karray, and Karthik Nandakumar. 2025. FaceAnonyMixer: Cancelable Faces via Identity Consistent Latent Space Mixing. *arXiv preprint arXiv:2508.05636* (2025).

[10] Alberto Ancilotto, Francesco Paissan, and Elisabetta Farella. 2024. Ximswap: Many-to-many face swapping for tinyml. *ACM Transactions on Embedded Computing Systems* (2024).

[11] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[12] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*.

[13] Alex Angus, Zhuoxu Duan, Gil Zussman, and Zoran Kostić. 2022. Real-time video anonymization in smart city intersections. In *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*.

[14] Sarker Monojit Asish, Arun K Kulshreshth, and Christoph W Borst. 2022. User identification utilizing minimal eye-gaze features in virtual reality applications. In *Virtual Worlds*.

[15] Atta Badii, Ahmed Al-Obaidi, Mathieu Einig, and Aurélien Ducournau. 2013. Holistic privacy impact assessment framework for video privacy filtering technologies. *Signal and Image Processing: An International Journal* (2013).

[16] Atta Badii, Touradj Ebrahimi, Christian Fedorczak, Pavel Korshunov, Tomas Piatrik, Volker Eiselein, and Ahmed A Al-Obaidi. 2014. Overview of the MediaEval 2014 Visual Privacy Task.. In *MediaEval*.

[17] Olivier Barnich and Marc Van Droogenbroeck. 2010. ViBe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing* (2010).

[18] Lennart Bastian, Tony Danjun Wang, Tobias Czempiel, Benjamin Busam, and Nassir Navab. 2023. DisguisOR: holistic face anonymization for the operating room. *International Journal of Computer Assisted Radiology and Surgery* (2023).

[19] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 2014. 3D pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[20] Michele Bezzi. 2010. An information theoretic approach for privacy metrics. *Trans. Data Priv.* (2010).

[21] Pascal Birnstill, Sebastian Bretthauer, Simon Greiner, and Erik Krempel. 2015. Privacy-preserving surveillance: an interdisciplinary approach. *Int'l Data Priv. L.* (2015).

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

[23] Ulzhan Bissarinova, Tomiris Rakhimzhanova, Daulet Kenzhebalin, and Huseyin Atakan Varol. 2024. Faces in event streams (FES): An annotated face dataset for event cameras. *Sensors* (2024).

[24] Scott Blunsden and Bob Fisher. 2010. The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA* (2010).

[25] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*.

[26] Tongan Cai, Haomiao Ni, Wenchao Ma, Yuan Xue, Qian Ma, Rachel Leicht, Kelvin Wong, John Volpi, Stephen TC Wong, James Z Wang, et al. 2025. SafeTriage: facial video de-identification for privacy-preserving stroke triage. In *International Conference on Information Processing in Medical Imaging*.

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

[28] Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. 2021. Personalized and invertible face de-identification by disentangled identity information manipulation. In *Proceedings of the IEEE/CVF international conference on computer vision*.

[29] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE international conference on automatic face & gesture recognition*.

[30] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. 2018. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[31] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*.

[32] William L Croft, Jörg-Rüdiger Sack, and Wei Shi. 2019. Differentially private obfuscation of facial images. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*.

[33] James L Crowley, Patrick Reignier, and Sebastien Pesnel. 2004. Context aware vision using image-based active recognition.

[34] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*.

[35] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2020. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[36] Antitza Dantcheva, Petros Elia, and Arun Ross. 2015. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security* (2015).

[37] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. 2022. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[38] Sander De Coninck, Wei-Cheng Wang, Sam Leroux, and Pieter Simoens. 2024. Privacy-preserving visual analysis: training video obfuscation models without sensitive labels. *Applied Intelligence* (2024).

[39] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[40] Ming Du, Aswin C Sankaranarayanan, and Rama Chellappa. 2014. Robust face recognition from multi-view videos. *IEEE transactions on image processing* (2014).

[41] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[42] Frédéric Dufaux and Touradj Ebrahimi. 2010. A framework for the validation of privacy protection solutions in video surveillance. In *IEEE International Conference on Multimedia and Expo*.

[43] Anil Egin, Andrea Tangherloni, and Antitza Dantcheva. 2025. Now You See Me, Now You Don't: A Unified Framework for Expression Consistent Anonymization in Talking Head Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[44] Adám Erdélyi, Thomas Winkler, and Bernhard Rinner. 2018. Privacy protection vs. utility in visual data: An objective evaluation framework. *Multimedia tools and applications* (2018).

[45] Sudeep Fadadu, Shreyash Pandey, Darshan Hegde, Yi Shi, Fang-Chieh Chou, Nemanja Djuric, and Carlos Vallespi-Gonzalez. 2022. Multi-view fusion of sensor data for improved perception and prediction in autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

[46] Chao Fan, Saihui Hou, Jilong Wang, Yongzhen Huang, and Shiqi Yu. 2023. Learning gait representation from massive unlabelled walking videos: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[47] Alircza Fathi, Jessica K Hodgins, and James M Rehg. 2012. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*.

[48] Andrea Fieschi, Pascal Hirmer, Christoph Stach, and Bernhard Mitschang. 2025. Characterising and Categorising Anonymization Techniques: A Literature-Based Approach. In *Proceedings of the International Conference on Information Systems Security and Privacy (ICISSP)*.

[49] Evangello Flouty, Odysseas Zisimopoulos, and Danail Stoyanov. 2018. Faceoff: anonymizing videos in the operating rooms. In *International Workshop on Computer-Assisted and Robotic Endoscopy*.

[50] Simson Garfinkel et al. 2015. De-identification of Personal Information:. (2015).

[51] Nikolaos Gkalelis, Hansung Kim, Adrian Hilton, Nikos Nikolaidis, and Ioannis Pitas. 2009. The i3dpost multi-view and 3d human action/interaction database. In *Conference for Visual Media Production*.

[52] Marta Gomez-Barrero, Javier Galbally, Christian Rathgeb, and Christoph Busch. 2017. General framework to evaluate unlinkability in biometric template protection systems. *IEEE Transactions on Information Forensics and Security* (2017).

[53] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* (2020).

[54] Ann-Kristin Grosselfinger, David Münch, and Michael Arens. 2019. An architecture for automatic multimodal video data anonymization to ensure data protection. In *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies III*.

[55] Monica Gruosso, Nicola Capece, and Ugo Erra. 2021. Human segmentation in surveillance video with deep learning. *Multimedia Tools and Applications* (2021).

[56] Ditmar Hadera, Jan Cech, Miroslav Purkrabek, and Matej Hoffmann. 2025. BLANKET: Anonymizing Faces in Infant Video Recordings. In *2025 IEEE International Conference on Development and Learning (ICDL)*.

[57] Agrya Halder, Pratik Chattopadhyay, and Sathish Kumar. 2023. Gait transformation network for gait de-identification with pose preservation. *Signal, Image and Video Processing* (2023).

[58] Simon Hanisch, Patricia Arias-Cabarcos, Javier Parra-Arnau, and Thorsten Strufe. 2025. Anonymization techniques for behavioral biometric data: a survey. *Comput. Surveys* (2025).

[59] Simon Hanisch, Julian Todt, Jose Patino, Nicholas Evans, and Thorsten Strufe. 2023. A false sense of privacy: Towards a reliable evaluation methodology for the anonymization of biometric data. *Proceedings on Privacy Enhancing Technologies* (2023).

[60] Simon Hanisch, Julian Todt, Jose Patino, Nicholas Evans, and Thorsten Strufe. 2024. A False Sense of Privacy: Towards a Reliable Evaluation Methodology for the Anonymization of Biometric Data. *Proceedings on Privacy Enhancing Technologies* (2024).

[61] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*.

[62] Fabio Hellmann, Silvan Mertes, Mohamed Benouis, Alexander Hustinx, Tzung-Chien Hsieh, Cristina Conati, Peter Krawitz, and Elisabeth André. 2024. Ganonymization: A gan-based face anonymization framework for preserving emotional expressions. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

[63] Jane Henriksen-Bulmer and Sheridan Jeary. 2016. Re-identification attacks—A systematic literature review. *International Journal of Information Management* (2016).

[64] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* (2017).

[65] Jan Hintz, Jacob Rühe, and Ingo Siegert. 2024. AnonEmoFace: Emotion Preserving Facial Anonymization.. In *ICISSP*.

[66] Yuki Hirose, Kazuaki Nakamura, Naoko Nitta, and Noboru Babaguchi. 2019. Anonymization of gait silhouette video by perturbing its phase and shape components. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

[67] Yuki Hirose, Kazuaki Nakamura, Naoko Nitta, and Noboru Babaguchi. 2022. Anonymization of human gait in video based on silhouette deformation and texture transfer. *IEEE Transactions on Information Forensics and Security* (2022).

[68] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* (2020).

[69] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. 2014. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation* (2014).

[70] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *20th international conference on pattern recognition*.

[71] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[72] Jungwoo Huh, Jiwoo Kang, Jongwook Woo, and Sanghoon Lee. 2025. A Novel Intelligent Video Surveillance System Using Low-Traffic Scene-Preserving Video Anonymization. *ACM Transactions on Intelligent Systems and Technology* (2025).

[73] Håkon Hukkelås and Frank Lindseth. 2023. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.

[74] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*.

[75] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. 2023. Realistic full-body anonymization with surface-guided GANs. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*.

[76] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* (2013).

[77] Marina Ivasic-Kos, Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas. 2014. Person de-identification in activity videos. In *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*.

[78] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*.

[79] Dima Kagan, Galit Fuhrmann Alpert, and Michael Fire. 2020. Zooming into video conferencing privacy and security threats. *arXiv preprint arXiv:2007.01059* (2020).

[80] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

[81] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[82] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. 2024. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications* (2024).

[83] Younggun Kim, Sirnam Swetha, Fazil Kagdi, and Mubarak Shah. 2025. Safe-LLaVA: A Privacy-Preserving Vision-Language Dataset and Benchmark for Biometric Safety. *arXiv preprint arXiv:2509.00192* (2025).

[84] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[85] Sander R Klomp, Matthew Van Rijn, Rob GJ Wijnhoven, Cees GM Snoek, and Peter HN De With. 2021. Safe fakes: Evaluating face anonymizers for face detectors. In *16th IEEE International Conference on Automatic Face and Gesture Recognition*.

[86] Pavel Korshunov, Andrea Melle, Jean-Luc Dugelay, and Touradj Ebrahimi. 2013. Framework for objective evaluation of privacy filters. In *Applications of Digital Image Processing XXXVI*.

[87] Sudha Krishnamurthy, Vimal Bhat, and Abhinav Jain. 2024. DiffSign: AI-Assisted Generation of Customizable Sign Language Videos With Enhanced Realism. In *European Conference on Computer Vision*.

[88] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* (2012).

[89] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *International conference on computer vision*.

[90] Han-Wei Kung, Tuomas Varanka, Sanjay Saha, Terence Sim, and Nicu Sebe. 2025. Face anonymization made simple. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[91] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

[92] Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. American sign language video anonymization to support online participation of deaf and hard of hearing users. In *Proceedings of the 23rd international ACM SIGACCeSS conference on computers and accessibility*.

[93] Dongze Li, Wei Wang, Kang Zhao, Jing Dong, and Tieniu Tan. 2023. RiDDLE: Reversible and Diversified De-identification with Latent Encryptor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[94] Jingzhi Li, Lutong Han, Ruoyu Chen, Hua Zhang, Bing Han, Lili Wang, and Xiaochun Cao. 2021. Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *Proceedings of the 29th ACM international conference on multimedia*.

[95] Tao Li and Lei Lin. 2019. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.

[96] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[97] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*.

[98] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. 2021. Multi-view Fusion for Multi-level Robotic Scene Understanding. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[99] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition.
[100] Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. 2023. Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8578–8587.
[101] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision.*
[102] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision.*
[103] Yasushi Makihara, Hidetoshi Mannami, Akira Tsuji, Md Altab Hossain, Kazushige Sugiura, Atsushi Mori, and Yasushi Yagi. 2012. The OU-ISIR gait database comprising the treadmill dataset. *IPSJ Transactions on Computer Vision and Applications* (2012).
[104] Simon Malm, Viktor Rönnbäck, Amanda Håkansson, Minh-ha Le, Karol Wojtulewicz, and Niklas Carlsson. 2023. Rad: Realistic anonymization of images using stable diffusion. In *Proceedings of the 23rd Workshop on Privacy in the Electronic Society.*
[105] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.*
[106] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2022. Decoupling identity and visual quality for image and video anonymization. In *Proceedings of the Asian Conference on Computer Vision.*
[107] Blaz Meden, Ziga Emersic, Vitomir Struc, and Peter Peer. 2017. κ-same-net: neural-network-based face deidentification. In *International Conference and Workshop on Bioinspired Intelligence (IWOBI).*
[108] Blaž Meden, Manfred Gonzalez-Hernandez, Peter Peer, and Vitomir Štruc. 2023. Face deidentification with controllable privacy protection. *Image and Vision Computing* (2023).
[109] Blaž Meden, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc, and Peter Peer. 2017. Face deidentification with generative deep neural networks. *IET Signal Processing* (2017).
[110] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. 2021. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security* (2021).
[111] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Bmj* (2009).
[112] Sriharsha Mopidevi, Kuk Jin Jang, Basam Alasaly, Sydney Pugh, Jean Park, Ashley Batugo, Sy Hwang, Eric Eaton, Danielle Lee Mowery, and Kevin B Johnson. 2025. MedVidDeID: Protecting privacy in clinical encounter video recordings. *Journal of Biomedical Informatics* (2025).
[113] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).
[114] Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022. Asl video corpora & sign bank: Resources available through the american sign language linguistic research project (asllrp). *arXiv preprint arXiv:2201.07899* (2022).
[115] Gregory S Nelson. 2015. Practical implications of sharing data: a primer on data privacy, anonymization, and de-identification. In *SAS global forum proceedings.*
[116] Hong-Wei Ng and Stefan Winkler. 2014. A data-driven approach to cleaning large face datasets. In *IEEE international conference on image processing (ICIP).*
[117] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. 2016. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision.*
[118] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision.*
[119] Marina Paolanti, Rocco Pietrini, Adriano Mancini, Emanuele Frontoni, and Primo Zingaretti. 2020. Deep understanding of shopper behaviours and interactions using RGB-D vision. *Machine Vision and Applications* (2020).
[120] Anubha Parashar and Rajveer Singh Shekhawat. 2022. Protection of gait data set for preserving its privacy in deep learning pipeline. *IET Biometrics* (2022).
[121] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).
[122] Y Pratheepan, Joan V Condell, and Girijesh Prasad. 2010. Individual identification from video based on" behavioural biometrics". In *Behavioral biometrics for human identification: Intelligent applications.*
[123] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. 2017. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE international conference on computer vision workshops.*
[124] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. 2024. A review on visual privacy preservation techniques for active and assisted living. *Multimedia Tools and Applications* (2024).
[125] Moritz Rempe, Lukas Heine, Constantin Seibold, Fabian Hörst, and Jens Kleesiek. 2025. De-identification of medical imaging data: a comprehensive tool for ensuring patient privacy. *European radiology* (2025).
[126] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. 2018. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV).*
[127] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. 2016. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication* (2016).
[128] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. 2023. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision.*
[129] Felix Rosberg, Eren Erdal Aksoy, Cristofer Englund, and Fernando Alonso-Fernandez. 2023. FIVA: facial image and video anonymization and anonymization defense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*
[130] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision.*
[131] Peter Rot, Klemen Grm, Peter Peer, and Vitomir Štruc. 2023. PrivacyProber: Assessment and detection of soft–biometric privacy–enhancing techniques. *IEEE Transactions on Dependable and Secure Computing* (2023).
[132] Omair Sarwar, Bernhard Rinner, and Andrea Cavallaro. 2019. A privacy-preserving filter for oblique face images based on adaptive hopping Gaussian mixtures. *IEEE Access* (2019).
[133] Adam Schmidt, Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. 2021. Multi-view surgical video action detection via mixed global view attention. In *International conference on medical image computing and computer-assisted intervention.*
[134] Md Shopon, Sanjida Nasreen Tumpa, Yajurv Bhatia, KN Pavan Kumar, and Marina L Gavrilova. 2021. Biometric systems de-identification: Current advancements and future directions. *Journal of Cybersecurity and Privacy* (2021).
[135] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
[136] Philipp Terhörst, Marco Huber, Naser Damer, Peter Rot, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. 2020. Privacy evaluation protocols for the evaluation of soft-biometric privacy-enhancing technologies. In *International conference of the biometrics special interest group (BIOSIG).*
[137] Daksh Thapar, Chetan Arora, and Aditya Nigam. 2020. Is sharing of egocentric video giving away your biometric signature?. In *European Conference on Computer Vision.*
[138] Daksh Thapar, Aditya Nigam, and Chetan Arora. 2021. Anonymizing egocentric videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*
[139] Ngoc-Dung T Tieu, Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. An RGB gait anonymization model for low-quality silhouettes. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).*
[140] Ngoc-Dung T Tieu, Huy H Nguyen, Hoang-Quoc Nguyen-Son, Junichi Yamagishi, and Isao Echizen. 2017. An approach for gait anonymization using deep learning. In *2017 IEEE workshop on information forensics and security (WIFS).*
[141] Ngoc-Dung T Tieu, Huy H Nguyen, Hoang-Quoc Nguyen-Son, Junichi Yamagishi, and Isao Echizen. 2019. Spatio-temporal generative adversarial network for gait anonymization. *Journal of Information Security and Applications* (2019).
[142] Ngoc-Dung T Tieu, Junichi Yamagishi, and Isao Echizen. 2020. Color transfer to anonymized gait images while maintaining anonymization. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).*
[143] Julian Todt, Simon Hanisch, and Thorsten Strufe. 2024. SEBA: Strong Evaluation of Biometric Anonymizations. *arXiv preprint arXiv:2407.06648* (2024).
[144] Quang Nhat Tran, Benjamin P Turnbull, and Jiankun Hu. 2021. Biometrics and privacy-preservation: How do they evolve? *IEEE Open Journal of the Computer Society* (2021).
[145] Christina O Tze, Panagiotis P Filntisis, Anastasios Roussos, and Petros Maragos. 2022. Cartoonized anonymization of sign language videos. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP).*
[146] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition.*
[147] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European conference on computer vision.*

[148] Miaomiao Wang, Sheng Li, Xinpeng Zhang, and Guorui Feng. 2025. Facial privacy in the digital era: A comprehensive survey on methods, evaluation, and future directions. *Computer Science Review* (2025).
[149] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* (2004).
[150] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
[151] Wenying Wen, Ziye Yuan, Yushu Zhang, Tao Wang, Xiangli Xiao, Ruoyu Zhao, and Yuming Fang. 2024. Image Privacy Protection: A Survey. *arXiv preprint arXiv:2412.15228* (2024).
[152] Yunqian Wen, Bo Liu, Rong Xie, Jingyi Cao, and Li Song. 2021. Deep motion flow aided face video de-identification. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*.
[153] Leslie Wöhler, Satoshi Ikehata, and Kiyoharu Aizawa. 2024. Investigating the Perception of Facial Anonymization Techniques in 360° Videos. *ACM Transactions on Applied Perception* (2024).
[154] Lior Wolf, Tal Hassner, and Itay Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. In *CVPR*.
[155] Allison Woodruff, Dirk Balfanz, Will Drewry, and Mariana Raykova. 1998. A Risk Assessment Framework for Digital Identification Systems. *Applied Ergonomics* (1998).
[156] Zhaoyang Xia, Yang Zhou, Ligong Han, Carol Neidle, and Dimitris N Metaxas. 2024. Diffusion models for Sign Language video anonymization. In *Proceedings of the LREC-COLING 2024 11th workshop on the representation and processing of sign languages: Evaluation of sign language resources*.
[157] Özge Nilay Yalçın, Vanessa Utz, and Steve DiPaola. 2024. Empathy through aesthetics: Using ai stylization for visual anonymization of interview videos. In *Proceedings of the 3rd Empathy-Centric Design Workshop: Scrutinizing Empathy Beyond the Individual*.
[158] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
[159] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).
[160] Han Zhang, Rotem Shalev-Arkushin, Vasileios Baltatzis, Connor Gillis, Gierad Laput, Raja Kushalnagar, Lorna C Quandt, Leah Findlater, Abdelkareem Bedri, and Colin Lea. 2025. Towards AI-driven Sign Language Generation with Non-manual Markers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.
[161] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.
[162] Ruoyu Zhao, Yushu Zhang, Tao Wang, Wenying Wen, Yong Xiang, and Xiaochun Cao. 2025. Visual content privacy protection: A survey. *Comput. Surveys* (2025).
[163] Zixian Zhao, Xingchen Zhang, and Yiannis Demiris. 2024. 3PFS: Protecting Pedestrian Privacy Through Face Swapping. *IEEE Transactions on Intelligent Transportation Systems* (2024).
[164] Guangyong Zheng and Yuming Xu. 2021. Efficient face detection and tracking in video sequences based on deep learning. *Information Sciences* (2021).
[165] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. 2022. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
[166] Shuai Zheng, Junge Zhang, Kaiqi Huang, Ran He, and Tieniu Tan. 2011. Robust view transformation model for gait recognition. In *18th IEEE international conference on image processing*.
[167] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou. 2021. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*.

Acknowledgments

The authors would like to thank the anonymous reviewers for their very insightful input, which helped improve the text notably. Generative AI-based tools have been used in revising the text to improve flow and correct any typos, grammar errors, and awkward phrasing. Funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/2 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden; This work was funded by the Topic Engineering Secure Systems of the Helmholtz Association (HGF) and supported by KASTEL Security Research Labs, Karlsruhe.

A Evaluation Metrics

This section explores the evaluation metrics in detail, providing both mathematical definitions and real-world context.

A.1 Privacy Metrics

Privacy metrics quantify the extent to which identifiable or sensitive information has been suppressed in the anonymized image or video. These are typically categorized into four groups: re-identification risk, attribute leakage, adversarial robustness, and human perception-based evaluation.

A.1.1 Re-identification Risk. This metric directly assesses whether anonymized data can still be matched back to an individual using automated systems. For instance, a pre-trained face recognition model $f(\cdot)$ might be used to predict the identity of anonymized input $\tilde{x}_i$. The recognition accuracy is defined as:

$$\text{ReID}_{\text{acc}} = \frac{1}{N}\sum_{i=1}^{N} 1\left[f(\tilde{x}_i) = y_i\right] \tag{1}$$

where y_i is the ground-truth identity and $1[\cdot]$ is the indicator function. A lower ReID accuracy implies better anonymization. For example, in anonymized surveillance footage, a ReID accuracy below 10% would be considered successful if the original model scored above 95%.

In retrieval scenarios (e.g., using anonymized queries to find identities from a gallery), mean Average Precision (mAP) is used:

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q} \text{AP}(q) \tag{2}$$

Here, Q is the number of queries, and $\text{AP}(q)$ is the average precision for query q. Lower mAP indicates greater privacy. This is crucial in applications like smart city surveillance, where re-identification of individuals must be minimized even if the system uses query-based analytics.

Verification scenarios (e.g., matching faces across frames) use the verification rate, measuring the percentage of correctly identified positive pairs from anonymized inputs. A lower verification rate reflects reduced identity exposure.

A.1.2 Attribute Leakage. Even if identity is hidden, residual cues may allow classifiers to infer demographic traits like gender, age, or ethnicity. This is known as attribute leakage. Suppose $g(\cdot)$ is a classifier trained to predict attribute a_i from anonymized sample $\tilde{x}_i$. The classification accuracy becomes:

$$\text{Attr}_{\text{acc}} = \frac{1}{N}\sum_{i=1}^{N} 1\left[g(\tilde{x}_i) = a_i\right] \tag{3}$$

A lower attribute classification accuracy implies better anonymization. For example, an anonymized facial dataset where gender is classified with only 55% accuracy (versus 98% on the original) demonstrates effective leakage suppression.

To go beyond classifier performance, Mutual Information (MI) is used to assess statistical dependence:

$$I(X;\tilde{X}) = \sum_{x,\tilde{x}} p(x,\tilde{x}) \log\left(\frac{p(x,\tilde{x})}{p(x)p(\tilde{x})}\right) \tag{4}$$

A lower MI indicates that the transformation disrupts dependency between the original and anonymized data — useful for data publishing or dataset release policies.

A.1.3 Adversarial Robustness. Privacy risks often arise not from legitimate users but from adversaries trying to recover identity or attributes. Therefore, anonymization systems must be evaluated for robustness against attacks. Three key metrics are:

- **Attack Success Rate (ASR)**: The proportion of samples where an attacker can reverse the anonymization process (e.g., using a GAN to deblur faces).
- **Membership Inference Accuracy**: Measures how well an attacker can determine whether a specific sample was part of the anonymizer's training set — a common threat in machine learning privacy.
- **Model Inversion Fidelity**: If the attacker reconstructs an image $\hat{x}_i$ from an anonymized version $\tilde{x}_i$, we use:

$$\text{Fidelity} = \frac{1}{N}\sum_{i=1}^{N} \text{SSIM}(x_i, \hat{x}_i) \tag{5}$$

where SSIM is a perceptual similarity score (explained later). High fidelity = high privacy risk.

A.1.4 Human-Centric Privacy Evaluation. Humans are perceptually sensitive and can often detect identity leakage missed by automated systems. User studies are therefore essential. The Human Re-identification Rate measures the success rate of human participants in matching anonymized images to their originals. A low rate suggests strong anonymization. Additionally, subjective privacy scores, typically gathered via Likert-scale surveys (e.g., 1–5 or 1–10), reflect how well anonymization is perceived. For example, users may rate anonymized faces based on "recognizability" or "visual identity retention."

A.2 Utility Metrics

Utility metrics evaluate whether anonymized images or videos remain usable for downstream vision tasks, including detection, classification, segmentation, or pose estimation.

A.2.1 Task-Specific Performance. This class of metrics checks how well anonymized content supports practical AI tasks. For classification tasks (e.g., action recognition), we compute:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} 1\left[h(\tilde{x}_i) = t_i\right] \tag{6}$$

where $h(\cdot)$ is the downstream task model and t_i is the ground-truth label.

For object detection or segmentation, we use Precision, Recall, and F1-score:

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}, \tag{7}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

These are especially important in medical video anonymization or surveillance, where detection of actions or objects must remain accurate after anonymization.

In semantic segmentation, Mean Intersection over Union (mIoU) is widely used:

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C} \frac{|S^c \cap S_c|}{|S^c \cup S_c|} \tag{9}$$

where S^c and S_c are the predicted and ground truth regions for class c.

In human pose estimation, PCK (Percentage of Correct Keypoints) and MPJPE (Mean Per Joint Position Error) are key:

$$\text{PCK@r} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\frac{\|\hat{p}_i - p_i\|_2}{s_i} \le r\right], \quad \text{MPJPE} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{p}_i - p_i\|_2 \tag{10}$$

These determine whether body structure remains intact post-anonymization — essential in sports, clinical gait analysis, or behavior understanding.

A.2.2 Perceptual Quality. Even if task performance is preserved, the visual plausibility of anonymized content matters. Four key metrics are used:

- **SSIM (Structural Similarity Index)**: Measures image degradation using structural features:

$$\text{SSIM}(x,\tilde{x}) = \frac{(2\mu_x\mu_{\tilde{x}} + C_1)(2\sigma_{x,\tilde{x}} + C_2)}{(\mu_x^2 + \mu_{\tilde{x}}^2 + C_1)(\sigma_x^2 + \sigma_{\tilde{x}}^2 + C_2)} \tag{11}$$

Higher SSIM suggests greater visual similarity.
- **PSNR (Peak Signal-to-Noise Ratio)**: Compares pixel intensities. Higher PSNR implies lower distortion:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{M_{\text{AX}}^2}{\text{MSE}}\right) \tag{12}$$

- **LPIPS (Learned Perceptual Image Patch Similarity)**: Measures perceptual similarity using deep features. Lower LPIPS = higher quality.
- **FID (Fréchet Inception Distance)**: Measures how close the anonymized data distribution is to real data:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{13}$$

Lower FID values (e.g., <20) indicate strong realism, which is critical in synthetic face anonymization or GAN-based replacement.

A.2.3 Human-Centric Utility Evaluation. Beyond numbers, humans evaluate realism, usefulness, and emotional expression. This includes:

- **Perceived Task Relevance**: Can humans still recognize emotion, action, or scene content?
- **Realism Ratings**: Participants rate how plausible or "real" anonymized visuals appear (Likert-scale).

- **Preference Testing**: When multiple anonymization outputs are available, participants select the most effective one, providing insights into the privacy–realism trade-off.

A.3 Privacy–Utility Trade-off

Anonymization inevitably introduces tension between privacy and utility. To evaluate this balance, researchers often plot Privacy–Utility Curves, where increasing privacy (e.g., lowering ReID accuracy) may degrade task performance (e.g., classification accuracy). To make this relationship quantifiable, composite metrics like Privacy–Utility Product (PUP) are used:

$$\text{PUP} = P \cdot U, \quad \text{where} \quad P = 1 - \text{ReID}_{\text{acc}}, \quad U = \text{Task Accuracy} \tag{14}$$

Higher PUP values represent better trade-offs. For example, a method yielding 95% anonymization success (5% ReID accuracy) while preserving 85% task utility would be preferred over one with 99% anonymization but only 50% utility.

More advanced formulations like Adjusted FID aim to embed both realism and privacy leakage into a unified score. Such metrics are increasingly used in benchmark studies to rank anonymization algorithms under practical deployment conditions.

A.4 Robustness Evaluation

A crucial dimension in assessing privacy-enhancing techniques is their reliance on attempts to undo or bypass the anonymization process. Many existing evaluations rely on simplified settings in which recognition models, originally trained on non-anonymized data, are tested directly on anonymized outputs. While this baseline approach is informative for common applications such as social media tagging or advertising, it doesn't fully capture adversarial conditions. Robustness must therefore also be examined under more challenging attack models.

The literature discusses several types of adversarial strategies:

Imitation (parrot) attacks: here the adversary has full knowledge of the anynimization process and applies it to their own training data. By learning in the transformed space, the attacker builds recognition or attribute classifiers that operate effectively on anonymized inputs [107, 109, 132].

Reconstruction attacks: In this scenario, the attacker seeks to reverse the anonymization process itself. With access to the underlying mechanism (white-box), the goal, it approximate the original appearance by solving the inverse problem. A more constrained version(grey, or black box) attempts recovery with only partial or no information about the transformation, yet may still succeed in revealing sensitive cues.

linkage attacks: even without direct inversion, adversaries may combine anonymized data with auxiliary sources, such as background context, clothing, hairstyle, or repeated instances of the same subject. Such side information can be exploited to re-associate anonymized data with the original individual [32, 50, 63, 117].

While the first two attack types are relatively straightforward to simulate and commonly used in robustness testing, linkage attacks remain difficult to evaluate, as they often rely on contextual elements not deliberately anonymized. Overall, these evaluations aim to quantify the likelihood that concealed attributes, identity, or otherwise, could be reconstructed or inferred, a risk of sometimes referred to as attribute recovery.

A.5 Evaluation Frameworks and Performance Studies

The evaluation of biometric privacy-enhancing technologies has steadily progressed from early face recognition, based tests toward comprehensive, adversary-aware methodologies. initial studies by Dufaux and Ebrahimi [42] relied on standardized face recognition benchmarks (FERET, CSU FIES) to assess pixelation, blurring, and scrambling, finding scrambling more effective than naive obfuscation. Korshunov et al. [86] advanced this approach by balancing privacy (via recognition accuracy) with utility (via detection accuracy) across multiple datasets, while Badii et al. [15, 16] introduced holistic criteria, efficacy, consistency, disambiguity, intelligibility, and aesthetics, tailored to surveillance contexts. Erdélyi et al. [44] further formalized protocols for frame-by-frame and aggregated video evaluation, and Terhörst et al. [136] expanded the scope to soft-biometric PETs, proposing training-free and learning-based protocols inspired by Kerckhoffs's principle and recognition, attribute visualization plots. Recent works reflect a shift toward strong adversary models and practical deployment. Hanisch et al. [59] emphasized the weakness of "naive" adversary assumptions and introduced reliable methodologies with worst-case attacker models, while Gómez-Barrero et al. [52]focused on unlinkability evaluation in biometrics template protection. Complementary tools such as PrivacyProber [131] stress-test soft-biometric PETs through attribute recovery, and SEBA [143] consolidate state-of-the-art evaluation into an extensible, open-source toolkit for privacy-utility analysis. The trend continues in 2025 with frameworks like FaceAnonyMixer [9] enabling recoverable and replaceable face anonymization. and PRISM [83]benchmarking biometric leakage in multimodal large language models. Terhörst et al. [8] proposed a large-scale re-evaluation of biometric modalities using empirical and expert-driven assessments, while broader frameworks now consider behavioral biometrics(voice, gait, gaze) [58] and risk assessments in digital ID systems [155]. Standardization efforts through ISO/IEC JTC 1/SC 37 further complement academic initiatives by providing globally recognized protocols. Collectively, these developments show a clear evolution of evaluation frameworks that have matured from narrow recognition-based metrics into holistic, adversary-aware, multimodal, and deployment-oriented protocols, capturing both privacy guarantees and utility preservation across increasingly diverse application domains. Capturing both privacy guarantees and utility preservation across increasingly diverse application domains.

B Assumption, Anonymization & Attacker Formalization

Video Representation. We model a video stream as a sequence of frames:

$$V = \{I_t\}_{t=1}^{t_{max}} \tag{15}$$

where:

- $F_t \in \mathbb{R}^{h \times w \times 3}$ is the RGB frame captured at time t

- T is the number of frames
- h, w denote frame height and width

Multi-Perspective Video Setup. In multi-camera environments, synchronized video streams are captured from multiple viewpoints:

$$\mathcal{V} = \{V^{(c)}\}_{c=1}^{c_{max}} \tag{16}$$

where:

- $V^{(c)}$ is the video captured by camera c
- C is the number of cameras

Each stream is:

$$V^{(c)} = \{I_t^{(c)}\}_{t=1}^{t_{max}} \tag{17}$$

Video Anonymization Transformation. An anonymization method is modeled as a transformation:

$$\mathcal{M} : V \times \mathcal{R} \times \mathcal{Z} \rightarrow \widetilde{V} \tag{18}$$

where:

- V is the raw video stream,
- $\mathcal{R}$ is the identifying regions,
- $\mathcal{Z}$ is the potential feature representations
- $\tilde{\mathcal{V}}$ is the protected video stream

Obfuscation-Based Manipulation : Obfuscation methods directly modify the original biometric signal without replacing it. Formally, obfuscation transformations satisfy:

$$\tilde{V}^{(c)} = \mathcal{M}_{\text{obf}}(V^{(c)}, r_{t,i}^{(c)}) \tag{19}$$

where $\tilde{V}^{(c)}$ remains a transformed version of the original biometric signal contained in $r_{t,i}^{(c)}$, and no external identity information is introduced.

Pixel-Level Signal Degradation. Pixel-level degradation methods suppress identity by directly altering pixel statistics without isolating identity features:

$$\tilde{r}_{t,i}^{(c)} = \mathcal{T}_{\text{pixel}}(r_{t,i}^{(c)}) \tag{20}$$

where $\mathcal{T}_{\text{pixel}}$ denotes transformations such as:

$$\mathcal{T}_{\text{pixel}} \in \{\text{Blur}, \text{Pixelate}, \text{Noise}, \text{AdvPerturb}\} \tag{21}$$

Identity-Aware Isolation and Manipulation. identity-ware anonymization isolates identity and no-identity factors in a latent representation:

$$z_{t,i}^{(c)} = E(r_{t,i}^{(c)}) \tag{22}$$

where

- $r_{t,i}^{(c)}$ is the image region corresponding to person i at the time t in camera c,
- $E(\cdot)$ is an encoder network that maps image regions into a latent representation,
- $z_{t,i}^{(c)}$ is the latent representation encoding visual attributes of person i.

The latent representation is decomposed into identity and non-identity components :

$$z_{t,i}^{(c)} = (z_{t,i}^{\text{id}}, z_{t,i}^{\text{non-id}}) \tag{23}$$

where:

- $z_{t,i}^{\text{id}}$ encodes identity-related features(e.g., facial structure, body shape, gait cues),
- $z_{t,i}^{\text{non-id}}$ encodes task-relevant attributes (e.g., pose , motion, illumination, context).

Anonymization removes or modifies identity components:

$$\tilde{z}_{t,i}^{(c)} = (0, z_{t,i}^{\text{non-id}}) \quad \text{or} \quad (\hat{z}_{t,i}^{\text{id}}, z_{t,i}^{\text{non-id}}) \tag{24}$$

where:

- 0 denotes identity suppression,
- $\hat{z}_{t,i}^{\text{id}}$ denotes a transformed or synthetic identity representation.

The anonymized region is reconstructed as:

$$\tilde{r}_{t,i}^{(c)} = D(\tilde{z}_{t,i}^{(c)}) \tag{25}$$

where:

- $D(\cdot)$ is a decoder or generator that reconstructs the anonymized image region.

Substitution-Based Manipulation: Substitution methods remove the original biometric signal and replace it with new content not derived from the input identity:

$$\tilde{V}^{(c)} = \mathcal{M}_{\text{sub}}(V^{(c)}, r_{t,i}^{(c)}, z), \tag{26}$$

where:

- $r_{t,i}^{(c)}$ is the image region corresponding to person i at time t in camera c,
- $\mathcal{M}_{\text{sub}}(\cdot)$ is a substitution-based anonymization operator,
- $\tilde{V}^{(c)}$ is the anonymized output video stream after substitution,
- z denotes identity features of the individual in the raw video.

External identity content z may be generated or selected from external sources:

$$z \sim \mathcal{P}_{\text{ext}} \tag{27}$$

where $\mathcal{P}_{\text{ext}}$ denotes an external identity distribution (e.g., synthetic indeitties generated by a generative model or identities samples from a database).

In generative susbistution methods, anonymization meybe implemented using a geneator model:

$$\tilde{r}_{t,i}^{(c)} = G(z, c_{t,i}^{(c)}) \tag{28}$$

where:

- $G(\cdot)$ is a generative model (e.g., GAN or diffusion model),
- $c_{t,i}^{(c)}$ denotes conditioning information such as pose, illumination, or scene context.

Masking: Masking removes identity information by replacing the region with non-informative content:

$$\tilde{r}_{t,i}^{(c)} = M \odot B \tag{29}$$

where M is a binary mask and B is background or constant fill.

Swapping / Generative Substitution: Swapping replaces the original identity with external or synthetic identity content while preserving scene and behavioral consistency.

$$\tilde{r}_{t,i}^{(c)} = G(z_{\text{target}}, c_{t,i}^{(c)}) \tag{30}$$

where:

- $\tilde{r}_{t,i}^{(c)}$ is the anonymized output region after substitution,
- $G(\cdot)$ is a generative model (e.g., GAN-based or diffusion-based generator),
- $c_{t,i}^{(c)}$ is conditioning information derived from the original input , such as pose , motion, illumination , or scene context.

The target identity representation may be obtained from external identity sources:

$$z_{\text{target}} \sim \mathcal{P}_{\text{id}} \tag{31}$$

where $\mathcal{P}_{\text{id}}$ denotes a distribution of external identity representations (e.g., identities from a database or synthetic identities generated by a model).

The conditioning signal is typically extracted from the original input region :

$$c_{t,i}^{(c)} = \phi_{\text{cond}}(r_{t,i}^{(c)}) \tag{32}$$

where $\phi_{\text{cond}}(\cdot)$ extracts non-identity attributes such as pose, expression, or temporal motion cues.

This formulation generalizes face swapping, full-body replacement, and diffusion-based identity substitution methods.

Adversary Scenarios and Multi-View Evaluation Formulation.

Cross-View Re-Identification Attack: This scenario evaluates whether an adversary can recover identity by aggregating anonymized observations across cameras.

Let the adversary extract identity embeddings:

$$e_{t,i}^{(c)} = \phi_{\text{ReID}}(\tilde{r}_{t,i}^{(c)}) \tag{33}$$

where:

- $\tilde{r}_{t,i}^{(c)}$ is the anonymized region of person i at time t in camera c,
- $\phi_{\text{ReID}}(\cdot)$ is a re-identification feature extractor,
- $e_{t,i}^{(c)}$ is the identity embedding extracted from anonymized data.

Cross-view identity aggregation is modeled as:

$$e_i^{\text{agg}} = \mathcal{F}_{\text{agg}}\left(\{e_{t,i}^{(c)}\}_{c,t}\right) \tag{34}$$

where:

- $\mathcal{F}_{\text{agg}}(\cdot)$ is an aggregation function combining embeddings across cameras and time.

Identity recovery is then modeled as:

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} \text{sim}(e_i^{\text{agg}}, e_y) \tag{35}$$

where:

- $\hat{y}_i$ is the predicted identity label,
- $\mathcal{Y}$ is the set of candidate identities,
- e_y is the reference embedding for identity y,
- $\text{sim}(\cdot,\cdot)$ is a similarity function (e.g., cosine similarity).

Temporal Aggregation Attack: This scenario evaluates whether identity information accumulates across times.

$$e_i^{\text{temp}} = \mathcal{F}_{\text{temp}}\left(\{e_{t,i}\}_{t=1}^{T}\right) \tag{36}$$

where:

- $e_{t,i}$ is the identity embedding of person i at time t,
- $\mathcal{F}_{\text{temp}}(\cdot)$ is a temporal aggregation function across frames,
- e_i^{temp} is the aggregated temporal identity representation.

Temporal identity leakage can be measured as:

$$\mathcal{L}_{\text{temp-privacy}} = \text{sim}(e_i^{\text{temp}}, e_i^{\text{true}}) \tag{37}$$

where:

- e_i^{true} is the ground-truth identity embedding,
- $\mathcal{L}_{\text{temp-privacy}}$ measures temporal identity leakage.

C Additional Tables & Figures

Table 2: Comparison of related surveys on visual anonymization (image and video).

Survey	Scope and Focus	Image/Video	Applications	Datasets	Limitations
Rempe *et al.*(2022) [125]	Survey of privacy-preserving techniques in medical imaging (MRI, CT, X-ray)	Image (medical)	Medical anonymization	✓	Limited to healthcare domain
Zhao *et al.*(2025) [162]	Visual privacy protection survey with adversarial taxonomy (CV and HV threats)	Image/Video	Visual content privacy threats	✓	High-level coverage, limited technical depth
Hanisch *et al.*(2025) [58]	Comprehensive survey of anonymization for behavioral biometrics (e.g., gait, EEG); includes video	Multi-modal (incl. video)	Behavioral biometrics, privacy-utility trade-off	✓	Focuses only on behavioral biometrics (voice, gait, EEG, etc.); does not address video-based anonymization
Wen *et al.*(2024) [151]	Survey on image privacy protection with three-level taxonomy (data-, content-, and feature-level)	Image	Social media, cloud, general imaging	✓	Does not address temporal or multi-view video anonymization
Ravi *et al.*(2024) [124]	Review of visual privacy preservation techniques for Active and Assisted Living (AAL)	Image/Video	Smart homes, healthcare monitoring, AAL	✓	Domain-specific to AAL, limited coverage of general anonymization methods
wang *et al.*(2025) [148]	Comprehensive survey on facial privacy methods (appearance-, identity-, reversible-, and system-level)	Image/Video	Face privacy, recognition systems, digital platforms	✓	Focused on face; limited discussion of full-scene or multi-camera video anonymization
Meden *et al.*(2021) [110]	Comprehensive survey on privacy–enhancing techniques for face biometrics (B–PETs), including image- and video-based deidentification, adversarial, and template-level methods; introduces multi-level taxonomy and evaluation framework	Image/Video	Facial biometrics	✓	Focuses on the facial domain; includes video studies limited to face-level anonymization, not full-scene or multi-camera video anonymization
Ours	Unified multi-dimensional taxonomy of image and video anonymization; covering GAN, diffusion, task utility, and realism	Video	Visual privacy, task preservation, realism	✓	Covers full spectrum of visual anonymization techniques

Table 3: Summary of commonly used datasets for human-related visual data.

Dataset	Year	Participants/ identities	# Images / Frames
(1) Face Datasets			
CelebA [101]	2015	10,177	202,599
LFW [71]	2007	5,749	13,233
CelebA-HQ [80]	2017	6,217	30,000
FFHQ [81]	2019	70,000	70,000
VGGFace2 [29]	2018	9,131	3.31M
CASIA-WebFace [159]	2014	10,575	494,414
CelebAMask-HQ [91]	2020	6,217	30,000
FDF [74]	2019	74,054	1,080,000
FaceScrub [116]	2014	530	106,863
FaceForensics++ [130]	2019	–	1,000 Video
(2) Full Body Datasets			
Market1501 [116]	2015	1,501	23000
MPII Human Pose (MPII) [12]	2014	40, 522	24, 920
Flickr Diverse Humans (FDH) [73]	2023	-	1.5 milion
Pexel-Humans [104]	2023	-	500
COCO-Body [75]	2023	-	53830
DeepFashion-CSE [75]	2023	-	50900
MSMT17 [150]	2018	4,101	-
CUHK03 [96]	2018	1,360	13, 164
(3) Gait Datasets			
CASIA-B [166]	2005	124	13,640 video sequences
BEHAVE [24]	2010	125	4 video clips /90,000 frames in total
OU-ISIR [103]	2012	200	5000
EPIC-Kitchens [35]	2020	32	11.5M frames
IITMD-WFP [137]	2021	31	Gait
TUM-GAID [69]	2012	305	3370 video sequences
i3DPost [51]	2009	8	832
GaitLU-1M [46]	2023	1 million	1 million
GREW [167]	2021	26,345	128,671
GREW [165]	2022	4,000	25,309

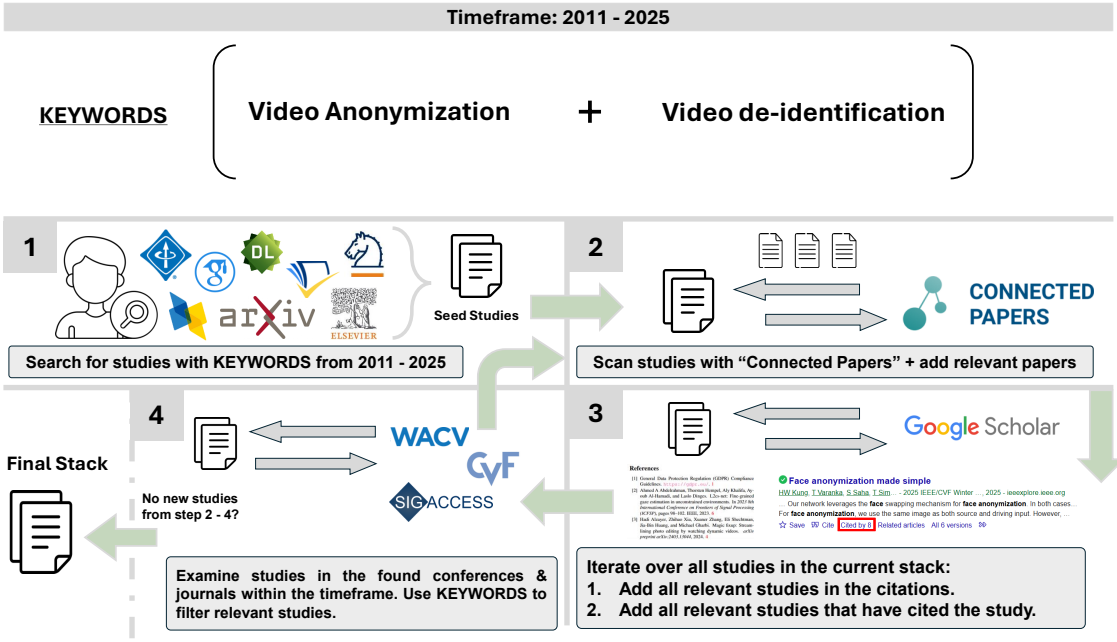
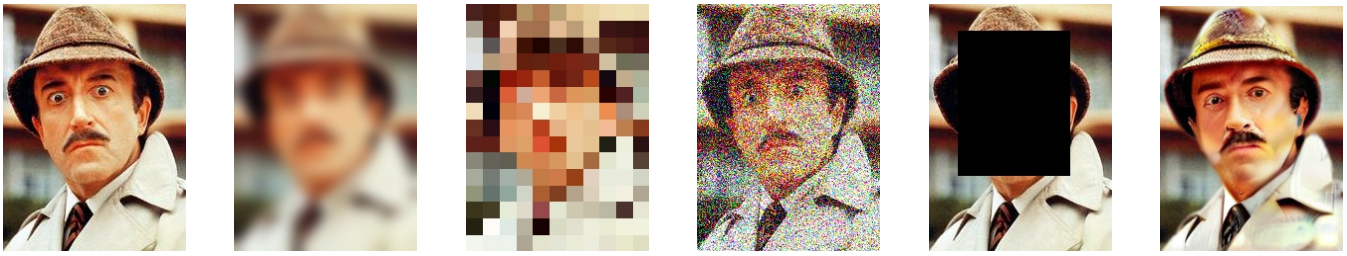


Figure 2: Search strategy adopted in our systematic review for selecting relevant studies on video, image, and face anonymization. The diagram illustrates the iterative process, starting from keyword-based search (2011–2025), expansion via Connected Papers, citation network screening, and venue-specific filtering, leading to the final stack of systematically reviewed works.



(a) Original (b) Blur (c) Pixelation (d) Noise (e) Mask (f) Swapping

Figure 3: Visual comparison between the original image and different privacy-preserving transformations, including classical obfuscation, and identity replacement via face swapping. Note, that for Swapping we show an example of RAD [104].

(a) Original image

(b) Identity-Aware Isolation and Manipulation Anonymized samples

Figure 4: Comparison between the original image and multiple anonymized face images generated using the Identity-Aware Isolation and Manipulation method [94].

Table 4: Overview of anonymization taxonomy categories, summarizing their core principles along with key advantages, limitations, and privacy protection levels.

Category	Core Principle	Advantages	Disadvantages	Privacy Level
CATEGORY: OBFUSCATION				
1. Pixel-level degradation	**Directly modifies pixels** to suppress identity cues without explicitly isolating factors.	• Simple and computationally efficient • Preserves coarse scene structure	• Degrades utility for fine details • Global corruption harms privacy/utility	Low
Blurring / Pixelation	Smoothing or block averaging to hide facial/texture details.	• Fast and widely used • Consistent across frames	• Loses fine-grained info • Requires reliable region detection • Identity remains recognizable	Low
Noise-based	Injecting random or structured perturbations (Gaussian/Poisson).	• Straightforward implementation • Tunable by noise level	• Globally corrupts the signal • Sensitive to parameter choice • Recognizers remain effective	Low
Adversarial	Add optimized gradient-based perturbations to push identity away.	• More targeted than random noise • Better utility preservation	• Model and attack dependent • Brittle under some recognizers	Moderate
2. Identity-aware isolation	Isolate identity factors (geometry/texture) and reconstruct via latent space.	• Better privacy–utility control • Coherent, realistic outputs	• Residual identity may remain • Higher complexity and dependency on good factorization	Moderate–High
CATEGORY: SUBSTITUTION-BASED ANONYMIZATION				
Masking	Delete identity regions using a binary or abstract mask to remove original signal.	• Complete removal of original signal • Strong privacy intuition	• Disrupts visual continuity • Visually conspicuous • Original identity removed	High
Swapping / Replacement	Replace identity regions with surrogate content (GAN/diffusion).	• Utility-preserving structure • Visually coherent content	• Leakage via behavioral cues • Needs temporal stability • Behavioral leakage possible	Moderate