TU Delft

Preserving Video Utility via Consistent Subject- and Key-derived Pseudonyms combined with Face Swapping

By Timo van Hoorn

June 2026

MSC THESIS REPORT

Preserving Video Utility via Consistent Subject- and Key-derived Pseudonyms combined with Face Swapping

By Timo van Hoorn

Computer Vision Lab
Delft University of Technology

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday, 16 June 2026 at 10:45 AM.

Student number: 5075408
Project Duration: 10 November 2025 – 16 June 2026

Thesis Committee: Dr. Jan van Gemert, TU Delft, Supervisor
P. Benschop, TU Delft, Daily supervisor
Dr. J.H.G. Dauwels, TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

Contents

PART 1

General introduction

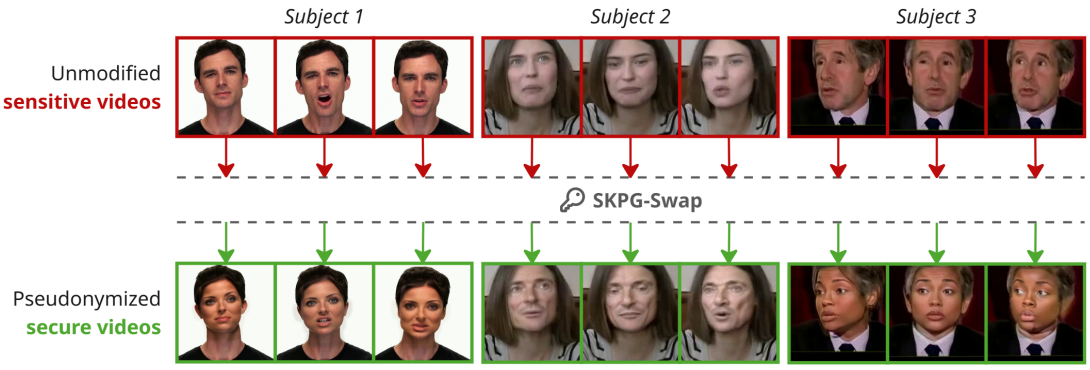

Figure 1. Pseudonymization across diverse videos using the proposed SKPG-Swap framework. The top row displays unmodified sensitive video frames for three distinct subjects. The arrows indicate the processing of these frames through SKPG-Swap, which uses the subject and a cryptographic key to generate the pseudonym. The bottom row shows the resulting pseudonymized secure video frames. This demonstrates that SKPG-Swap replaces a subject's identity with a consistent pseudonym, while also preserving context like head pose, facial expressions, and the surrounding environment.

A clinician studying how a patient's emotions evolve over months of therapy and an analyst running action recognition on surveillance footage both rely on automated computer vision tools that process faces in their video data [1, 2, 3, 4]. The human face is biometric data, and storing it without strong protective measures is forbidden under regulations such as the General Data Protection Regulation (GDPR) [5] in Europe and the Health Insurance Portability and Accountability Act (HIPAA) [6] in the United States. However, the automated tools that enable analysis require the face as input. Removing or obscuring the face for compliance therefore strips the data of the signal on which automated analysis depends [7]. A more useful approach is to replace the face with a generated one that conceals the real identity while preserving the surrounding signal. The generated face acts as a pseudonym, a stand-in identifier used in place of the real one. Replacing the face in this way is called pseudonymization. Unlike anonymization, which removes identifying information altogether, pseudonymization keeps a link back to the real identity [8]. That link is accessible only through a separately held secret, such as a cryptographic key. The individual whose face is being protected will be referred to as the subject.

For pseudonymization to keep data both analytically useful and secure, three requirements must hold simultaneously. The first is that the context must be preserved. Head pose, facial expression, and the surrounding environment all carry information that downstream tools rely on [9, 10]. The second is that the same subject must be mapped to the same pseudonym every time they appear, both within a single video and across separate videos. Without this property,

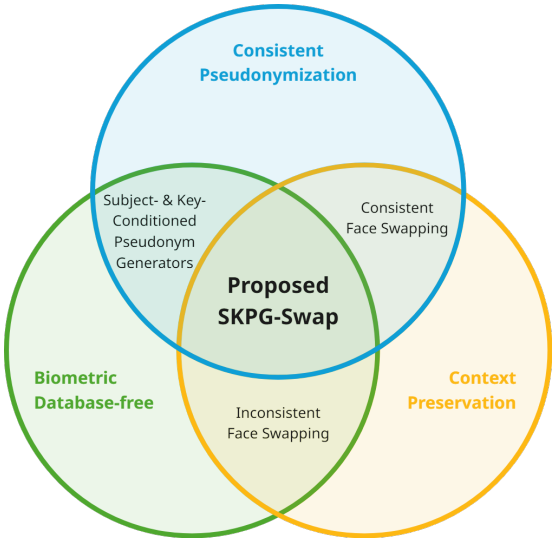

Figure 2. Venn diagram illustrating the overlap of the three requirements for secure video analysis: consistent pseudonymization, context preservation, and biometric database-free. The intersecting regions demonstrate that prior approaches force a trade-off by satisfying only two out of the three requirements. Subject- and key-conditioned pseudonym generators fail to preserve context. Consistent face swapping requires a vulnerable biometric database. Inconsistent face swapping removes the database and preserves context but fails to map subjects to consistent pseudonyms. Therefore, the proposed SKPG-Swap framework at the center intersection is the only method that satisfies all three requirements.

a patient recorded in January and again in March would be assigned two different pseudonyms, and an emotion-recognition tool would respond partly to the change in pseudonym rather than to a genuine change in the patient's mood. The third is that there may not exist a database linking real identities to their assigned pseudonyms. Such a database would be a single point of failure, and unlike a leaked password, a face cannot be reissued once exposed [11].

Existing methods can satisfy two of these three requirements. One family of methods, known as **face swapping**, replaces only the identity in a frame while keeping the context intact [12, 13]. However, because a swap needs a reference image of the pseudonym, mapping a subject to the same pseudonym consistently forces these methods to store a record linking the two. Storing these records is exactly what the database-free requirement forbids. The other family, referred to here as **subject- and key-conditioned pseudonym generators**, avoids such a database by deriving the pseudonym from the subject's face and a secret cryptographic key through a neural network [14, 15]. The network is trained to produce the same pseudonym for the same subject-key pairing, so consistency is achieved by this training objective rather than from a stored record. However, these methods do not place the generated pseudonym back onto the original frame, and the context of the video is lost. Existing methods therefore force a trade-off. Preserving the context requires accepting a vulnerable database, while methods eliminating this database lose the original video's context.

The proposed framework, named **SKPG-Swap**, closes this gap by combining the strengths of both families. It first adopts a neural network that derives the pseudonym from the subject's face and a key, so no mapping between subjects and pseudonyms ever needs to be stored. Then a face-swap renderer blends the generated pseudonym onto the original frame, preserving its

context. The result is a single pipeline that pseudonymizes diverse videos as shown in Figure 1, satisfying all three requirements as summarized by the Venn diagram in Figure 2.

Three evaluations test whether the framework meets these requirements in practice. The first measures the identity behavior of the generated pseudonyms and whether the context is preserved after pseudonymization. It checks whether the pseudonyms hide the original identities, whether the same subject reliably maps to the same pseudonym, whether different keys or different subjects produce visibly distinct pseudonyms, and whether the pseudonymized frame stays visually close to the original. The second evaluation uses an action-recognition task in which a model is asked to recognize activities such as brushing teeth or playing a flute. Because these activities depend on visual cues surrounding the face, this evaluation checks whether the model can still achieve high accuracy scores after the video is altered by the pseudonymization methods. The third evaluation uses an emotion-recognition task on separate videos of the same subject. Similar to the second task, it verifies whether the model's scores remain high on the altered videos, while also testing whether assigning the exact same pseudonym across multiple videos keeps predictions on the same subject stable. Together, the three experiments check if the framework is useful for the scenarios that motivated it.

Structure of the Report

The report is organized in three parts. The current chapter (Part 1) provides the high-level overview above and is written for a reader from outside the technical area. The core contribution is presented in Part 3 as a scientific article, containing the formal problem statement, the proposed method, the experimental setup, and the evaluation results. The article assumes that readers are familiar with the regulatory and technical domains related to pseudonymization using computer vision.

Part 2 provides the background that bridges Part 1 and Part 3, in the order in which concepts appear in the article. Section 1 covers the privacy and data-protection principles that motivate the work, including the distinction between anonymization and pseudonymization and the role of cryptographic keys. Section 2 introduces the deep-learning building blocks used in the proposed pipeline, covering neural networks, latent spaces, loss functions, and transfer learning, together with the specific architectures used in the article: Multi-Layer Perceptrons, Convolutional Neural Networks, Generative Adversarial Networks, and Diffusion models. Section 3 applies these concepts to face analysis and generation, covering face recognition, the StyleGAN architecture, face swapping, face reenactment, and adapter-based diffusion control. Section 4 introduces the identity, visual-quality, and utility metrics used in the evaluation of Part 3.

Readers already familiar with the technical area may proceed directly to Part 3. Readers from outside the field are encouraged to read Part 2 first, or to use it as a reference while reading the article.

PART 2

Preliminary Materials

This chapter introduces the core concepts and technologies required to understand the scientific article and the proposed pseudonymization framework presented in Part 3. The chapter progresses from the fundamental principles of data privacy to the underlying computer vision concepts and evaluation metrics used to validate the system.

1 Privacy and Data Protection Fundamentals

The storage and processing of sensitive information within machine and deep learning systems must strictly follow privacy laws, such as the General Data Protection Regulation (GDPR) [5] in Europe and the Health Insurance Portability and Accountability Act (HIPAA) [6] in the United States. To handle this data properly, it is essential to understand the differences between the types of data being processed and the specific methods used to keep them safe. Within the context of these regulatory frameworks, the person to whom the sensitive data belongs is formally specified as the 'data subject'. The text refers to this person as the subject.

1.1 Biometric Identifiers

A biometric identifier is a unique physical or behavioral characteristic used to distinguish one subject from another. In this thesis, the relevant biometric identifiers are in the face, although the broader category also includes traits such as gait, iris patterns, and voice [16]. Under current regulations, biometric identifiers are classified as highly sensitive personal data [5, Art. 4(1)] because they are permanent and cannot be revoked or reissued. The strict legal status of identifiable biometric data is therefore the reason that the storage and use of facial information must be minimized and protected from unauthorized access [5, Art. 5(1)(c)]. This strictness forms the central motivation for the pseudonymization framework developed in Part 3.

Because facial images, generated pseudonyms, and mathematical embeddings are biometric identifiers that can be used to link identities, any database storing them is explicitly treated as a vulnerable biometric database. These identifiers are central to the methods discussed in this thesis.

1.2 Anonymization and Pseudonymization

When processing sensitive data, protecting the identity of a subject typically involves anonymization or pseudonymization. The two approaches differ in whether the link to the subject is permanently destroyed or hidden behind additional information. Figure 3 illustrates the difference between anonymization and pseudonymization conceptually.

Anonymization is the processing of personal data so that it can no longer be linked to the subject [17]. In the context of visual data, anonymization permanently destroys or obfuscates the original biometric identifiers. The defining property is irreversibility: once data is anonymized, all logical and mathematical links to the subject are permanently removed. Because re-identifying the subject is computationally and practically infeasible, anonymized data are no longer legally

classified as personal data. Consequently, they fall outside the scope of legal frameworks like the GDPR [5, Recital 26].

Pseudonymization, in contrast, replaces original biometric identifiers with generated ones, known as pseudonyms, in a process that can be reversed. This reversal relies on a key that links the pseudonym back to the subject. Because subjects could still be re-identified by anyone with access to that key, pseudonymized data is legally considered personal data and must be handled under strict compliance rules [5, Art. 4(5)]. Pseudonymization therefore offers weaker legal protection than anonymization. However, it preserves the ability to map subjects to consistent pseudonyms, which is a core requirement of the proposed framework evaluated in Part 3.

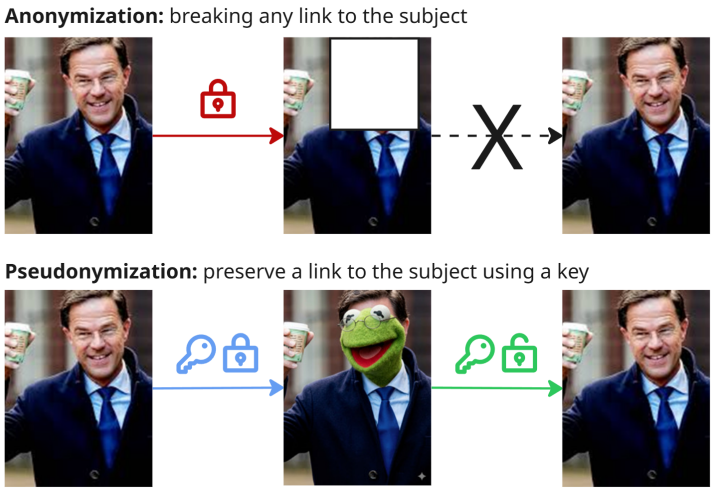

Figure 3. Conceptual comparison of anonymization and pseudonymization. (Top Row) Anonymization irreversibly destroys the original biometric identifiers, making the identity permanently inaccessible and the data unlinkable to the subject. (Bottom Row) Pseudonymization replaces the subject's identity with a generated substitute, which can be reversed and linked back to the subject using a key. While pseudonymization preserves the linkability, it requires strict handling of a key to be compliant.

1.3 Cryptographic Systems

To comply with the legal rules for handling pseudonymized data, systems often rely on cryptography to mathematically protect the link between an identity and its pseudonym. Cryptography is a method of securing data so that only authorized parties can access or link it [18]. It uses cryptographic keys to control how data is transformed. A key acts as a secret mathematical seed made of a random string of bits.

For example, in the context of pseudonymization, an algorithm can take the biometric identifiers of a subject and process them alongside a secret cryptographic key [19]. The key dictates the functional output of the algorithm, ensuring that the resulting pseudonym's face is securely bound to the subject. Consequently, the key becomes the exclusive linking mechanism between the subject and the pseudonymized output.

A key-based design enables the strict separation of information required by privacy regulations. For pseudonymized data to be legally compliant, it must be stored entirely separately from the additional information needed to link it back to the subject [5, Art. 4(5)]. By using a secret

cryptographic key, the pseudonymized data can be safely stored and analyzed in computing environments while the key remains completely isolated in a secure location.

2 Foundational Deep Learning Concepts

To implement the secure pseudonymization pipelines required by privacy frameworks, systems rely on deep learning models. This section introduces the conceptual foundations first: the neural-network learning process, latent spaces and embeddings, and loss functions. This is then followed by four specific architectures: Multi-layer Perceptrons, Convolutional Neural Networks, Generative Adversarial Networks, and Diffusion models. The section concludes with the principles of pre-trained models and transfer learning, which explain how these architectures can be combined without training from scratch. Together, these deep-learning building blocks form the foundation on which the proposed framework is built and the specific architectures it employs.

2.1 Deep Learning and Neural Networks

Deep learning is a subfield of machine learning that uses artificial neural networks [20]. A neural network is structured in multiple interconnected layers, typically including an input layer, several hidden layers, and an output layer. The hidden layers allow the network to progressively extract more complex and abstract patterns, where early layers capture simple features and deeper layers combine them into higher-level representations.

A defining characteristic of neural networks is their learnability. During training, the network processes data and produces an output, which is evaluated against the desired outcome using a mathematical loss function that quantifies the error of the model. Through an iterative optimization process known as backpropagation [21], the network adjusts its internal parameters to minimize the loss. The behavior the network ultimately learns is therefore determined entirely by the training data and the choice of loss.

2.2 Latent Spaces and Embeddings

A recurring concept across deep learning architectures is the latent space. A latent space is a learned, lower-dimensional representation in which the network encodes the most informative properties of its input data. Rather than operating on raw, high-dimensional inputs such as image pixels, networks compress these inputs into compact vectors that capture their underlying structure. Each input is represented as a single point, or embedding, within the latent space, where the geometric distance between points reflects the semantic similarity of the corresponding inputs. Two images of the same subject, for example, are mapped to nearby embeddings, while images of different subjects are mapped further apart, as illustrated in Figure 4.

Different networks learn different latent spaces, each shaped by their training objective and architecture. A face recognition network learns an identity latent space in which biometric similarity dominates the geometry, whereas a generative model learns a latent space whose directions correspond to visual attributes such as pose or lighting. Because these spaces are structured differently, embeddings cannot be exchanged directly between models. A recurring task in modern pipelines is therefore to learn a mapping that translates representations from one latent space into another, a function typically performed by a neural network.

Combining vectors from different latent spaces introduces an additional practical issue: the vectors may not lie on the same numerical scale. A common fix is L2 normalization, which

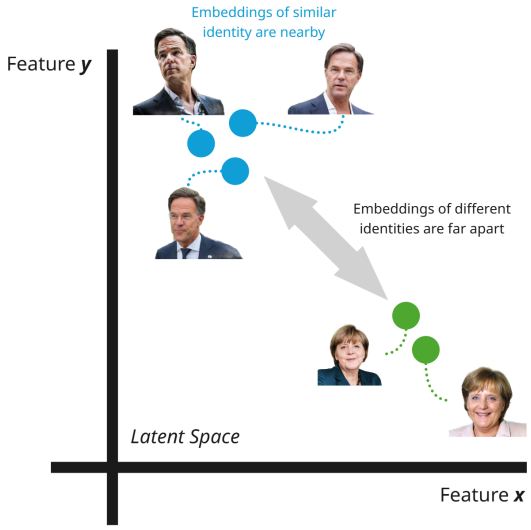

Figure 4. Conceptual illustration of facial embeddings in a learned latent space. The two axes represent two learned dimensions of the space. Real facial latent spaces are far higher-dimensional, but the principle is shown here in two dimensions for clarity. Each colored point corresponds to the embedding of a single face image, with color indicating the underlying identity. Multiple images of the same subject (blue) are mapped to nearby points, while images of a different subject (green) are mapped far apart, as indicated by the gray distance arrow. The figure illustrates the central property used throughout this thesis: identity similarity translates into geometric proximity in the latent space, which is what makes cosine-similarity comparison of embeddings a meaningful proxy for biometric matching.

rescales a vector to unit length while preserving its direction, computed as

$$\hat{v} = \frac{v}{\|v\|_2}, \tag{1}$$

where $\hat{v}$ is the normalized vector, v is the original latent facial embedding, and $\|v\|_2$ represents the Euclidean norm (magnitude) of the vector. L2 normalization is widely applied to facial embeddings because identity information is encoded in the angular relationships between vectors rather than in their magnitudes [22]. When vectors from different sources are concatenated or added, L2 normalization ensures that no single input dominates a downstream computation purely because of its larger numerical scale. The same balancing argument is later used inside the proposed Projector, where a face embedding and a cryptographic key of different magnitudes must be combined into a single input vector.

2.3 Loss Functions and Training Objectives

The behavior of a deep learning model is determined by the loss function used during training. These loss functions are specifically chosen to optimize a specific training objective. Many modern systems are not optimized against a single objective but instead balance several simultaneous objectives. This balance is achieved through a multi-task learning objective, in which multiple

individual loss terms are combined into a single weighted sum [23] given by

$$\mathcal{L}_{tot} = \sum_{i=1}^{n} \lambda_i \mathcal{L}_i, \tag{2}$$

where $\mathcal{L}_{tot}$ is the combined loss, $\mathcal{L}_i$ represents an individual loss term, and λ_i is the corresponding coefficient that controls its relative weight during training. This formulation allows a model to satisfy several competing requirements at once, as the relative weighting of these terms determines the final trade-off the network learns. In the proposed framework, for instance, the Projector utilizes this formulation to train against five simultaneous objectives.

2.4 Multi-Layer Perceptrons

The Multi-Layer Perceptron (MLP) is a fundamental class of feedforward artificial neural networks [24]. In an MLP, information propagates strictly forward from the input nodes through the hidden layers to the output nodes, where each layer consists of fully connected neurons that apply linear transformations followed by non-linear activation functions. Among the basic neural network architectures, the MLP is therefore the one most relevant to the Projector component of the proposed framework.

While MLPs are traditionally used for classification and regression tasks, generative pipelines frequently use them as projectors. As a projector, the MLP is trained to learn complex transformations that translate vector representations from one specific latent space into another. This transformation capability makes MLPs effective for combining distinct inputs that originate from entirely different latent spaces and projecting the merged result into a new, unified latent space. In the proposed framework, an MLP is trained to combine a facial identity embedding and a secret cryptographic key from their respective latent spaces, projecting the merged input into the latent space of a generative model.

2.5 Convolutional Neural Networks

While Multi-Layer Perceptrons are effective for processing structured vector data, they scale poorly to the high-dimensional, grid-like arrangement of image data. A standard fully connected network would require a large number of parameters to process a high-resolution image, since every pixel would have to be connected to every neuron. Convolutional Neural Networks (CNNs) solve this problem by replacing fully connected layers with a specialized mathematical operation known as convolution [25]. A convolution applies small, learnable filters that slide across the input image to detect local patterns such as edges, textures, and shapes. By sharing parameters across spatial positions, this drastically reduces the overall parameter count.

A typical CNN is composed of multiple convolutional layers followed by pooling layers that reduce the spatial dimensions of the data. This stacked structure allows the network to build a visual hierarchy of features, in which early layers identify simple lines and subsequent layers recognize complex structures such as eyes, mouths, and entire facial components, as illustrated in Figure 5. Within the context of this thesis, CNNs serve as the primary tool for extracting biometric identifiers from images, functioning as the identity encoders used both in the proposed framework itself and for the final evaluations.

2.6 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are one of the two main families of generative architectures used to generate realistic visual data [27]. A GAN consists of two separate neural

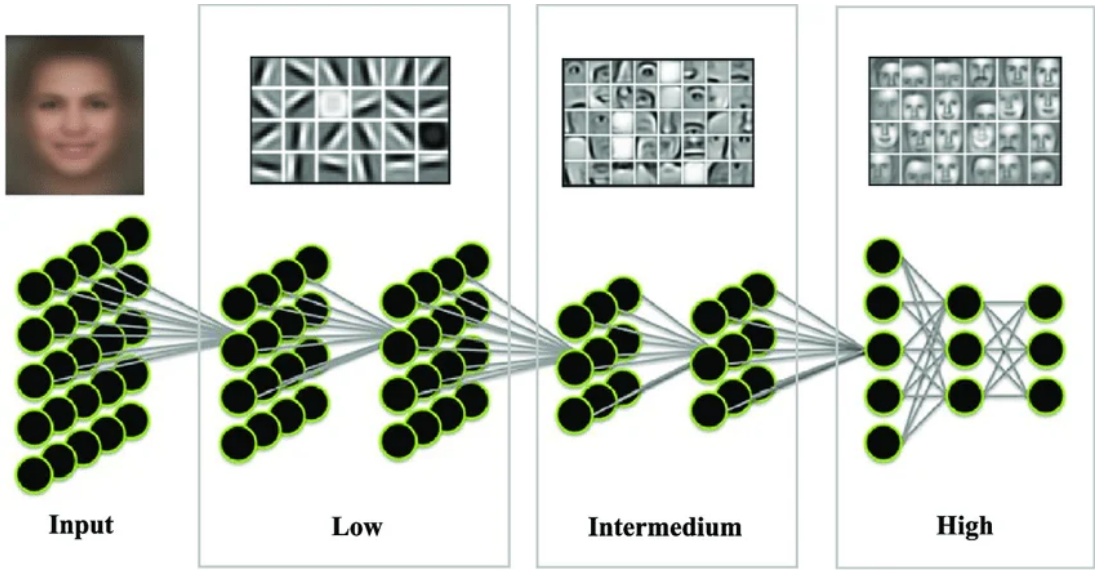

Figure 5. Hierarchical feature extraction in a Convolutional Neural Network applied to face images. The input image (left) is processed by a sequence of convolutional layer groups, schematically drawn as stacks of nodes and labeled *Low*, *Intermedium*, and *High* from left to right. The grayscale tiles above each group visualize the patterns that the filters at that depth respond most strongly to. Low-level layers respond to simple oriented edges, Intermedium layers respond to mid-level facial parts such as eyes, noses, and mouths, and high-level layers respond to whole-face configurations. The figure illustrates the hierarchical feature extraction that makes CNNs effective for face recognition: the network learns to compose simple local patterns into the structural identity cues that the encoder later compresses into a facial embedding. Figure adapted from Li et al. [26].

networks with opposing objectives. The first network, the generator, produces data from an input vector sampled from a latent space. The second network, the discriminator, evaluates data and attempts to distinguish the produced outputs of the generator from real-world data samples.

During training, the two networks participate in a competition in which the gain of one network is balanced by the loss of the other, as illustrated in Figure 6. The discriminator updates its parameters to maximize its accuracy in identifying fake data, while the generator updates its parameters to minimize that accuracy and effectively fool the discriminator. The adversarial objective forces the generator to produce increasingly realistic outputs over time, until the produced samples become difficult for the discriminator to distinguish from real ones. Once training reaches an optimal state, the generator can produce data that closely matches the original training distribution. This capability to match the original distribution makes GANs highly effective for masking subjects' identities, which is why the StyleGAN2 architecture is used as the generative component in the proposed framework.

2.7 Diffusion Models

Diffusion models represent another effective class of generative deep learning architectures that have recently achieved state-of-the-art results in visual data generation [29]. Unlike the

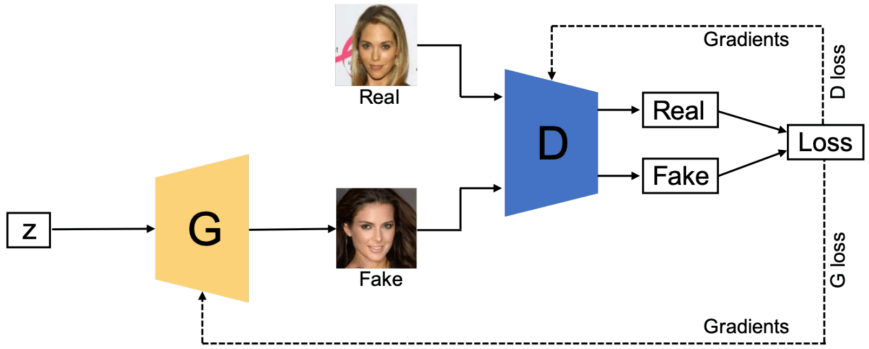

Figure 6. Architecture and training loop of a Generative Adversarial Network. A latent vector z is sampled from a simple distribution and passed through the generator network G (yellow), which transforms it into a generated "fake" face image. A real face image and the generated fake are both passed to the discriminator network D (blue), which classifies each input as Real or Fake. Both classifications feed into a single loss function from which two opposing gradient signals are derived: the discriminator loss (D loss) updates D to better distinguish real from fake, while the generator loss (G loss) updates G to better fool D. The shared loss with opposing gradients is what gives a GAN its adversarial characteristic. Convergence is reached when D can no longer reliably tell real and fake samples apart, which allows the generator to produce the realistic pseudonymized faces required by the proposed framework. Figure adapted from Wang et al. [28].

competitive training dynamic used in Generative Adversarial Networks, diffusion models are built on a mathematical framework that describes how data changes using probability. The training pipeline consists of two distinct phases. In the forward diffusion process, a mathematical algorithm systematically adds Gaussian noise to a real data sample over a series of steps, until the original image is fully transformed into a distribution of random noise.

The core of the architecture lies in the reverse diffusion process. A neural network is trained to sequentially remove the added noise, step by step, by estimating and subtracting the exact amount of noise added at each interval. By optimizing this denoising objective, the network learns to construct entirely new, high-quality images from pure random noise. Figure 7 illustrates both directions: the progressive corruption in the forward process and the iterative denoising in the reverse process. As a result, diffusion models can generate detailed and diverse visual outputs, making them a powerful tool for face generation.

Large-scale diffusion frameworks extend the basic formulation in two key ways. First, rather than operating on raw pixels, models such as Stable Diffusion [31] run the diffusion process inside a compressed latent space learned by a separate autoencoder. This substantially lowers training and inference costs while preserving image quality. Stable Diffusion is the most widely used instance of this paradigm and has been pretrained on large-scale image-text pairs. Second, the denoising network can be conditioned on additional inputs, such as text prompts or image features, that guide generation toward specific outputs. The conditional formulation makes diffusion models highly controllable. Together, these two extensions form the foundation of the adapter-based face-swap component in the proposed framework, which utilizes the publicly released weights of Stable Diffusion alongside the conditional Face-Adapter [13].

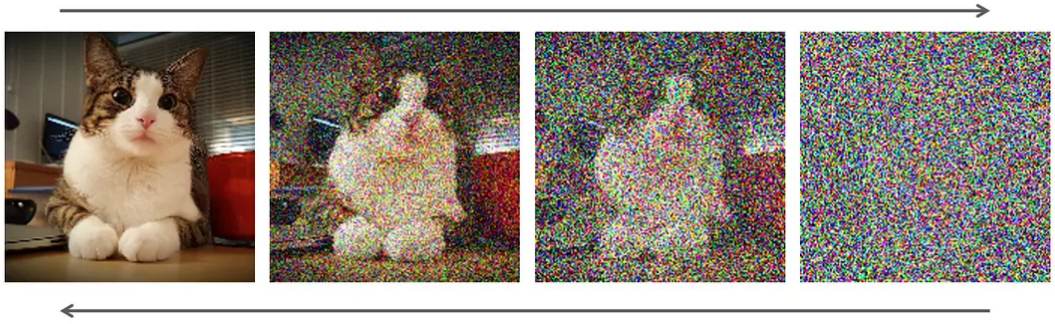

Figure 7. Forward and reverse diffusion processes. The forward process (top arrow) progressively corrupts an image by adding Gaussian noise over multiple steps until only pure noise remains. The reverse process (bottom arrow) learns to iteratively denoise, recovering high-fidelity images from pure random noise. The reverse process is what makes diffusion models useful for face generation. Figure adapted from Vahdat and Kreis [30].

2.8 Pre-trained Models

Training a deep neural network from scratch typically requires large amounts of labeled data and substantial computational resources. To avoid this cost, pipelines frequently rely on pre-trained models. These are networks whose parameters have already been optimized on large, general-purpose datasets and made publicly available [32]. These pre-trained components capture broadly applicable representations, such as general facial structure or visual realism, that can be reused across downstream tasks.

When a pre-trained model is incorporated into a larger pipeline, its parameters can either be further updated during training or kept fixed. A component whose parameters remain unchanged is referred to as frozen, while a component whose parameters are still optimized is referred to as trainable. Freezing a model preserves the representations it has already learned, avoids unnecessary computational overhead, and prevents large pre-trained networks from overfitting on smaller downstream datasets [33]. As a result, many architectures are constructed by combining several frozen pre-trained components with a small, trainable component, such as an MLP, that learns to adapt or connect them. This pattern is central to the proposed framework, where only the lightweight Projector is trained while the Encoder, Generator, and face-swap components remain frozen.

3 Deep Learning in Face Analysis and Generation

In this thesis, foundational deep learning models are specifically adapted to analyze identities and generate faces. This section covers two complementary tasks: recognizing the identity in a face and generating new ones. Face recognition outlines the transition from image pixels to encoded identity representations, known as facial embeddings. Face generation then describes how generative models produce realistic faces that can mask a subject's identity. This progression begins with the StyleGAN architecture, moves through face swapping and face reenactment, and concludes with adapter-based diffusion control. Together, these deep learning concepts drive the core mechanisms of the proposed framework, utilizing facial embeddings for both the pipeline and identity evaluation, StyleGAN for creating pseudonyms, and diffusion control to

render a pseudonym into the original frame.

3.1 Face Recognition and Encoding

Modern face recognition systems use Convolutional Neural Network architectures to extract biometric identifiers from images. When an image of a face is provided to the network, the convolutional layers identify a hierarchy of patterns, ranging from simple textures to the complex structural relationships between facial components. Two prominent examples of such architectures are FaceNet [34] and ArcFace [35], both of which are integrated into the proposed framework.

Once the convolutional layers have extracted these features, the model maps the output into a standardized latent space. The final encoded vector, called a facial embedding, contains the essential information required to uniquely identify a subject. State-of-the-art recognition models are trained so that the cosine similarity between two embeddings in this space reflects how similar the underlying faces are: embeddings of the same subject align in direction, while embeddings of different subjects separate. This angular property is used twice in the proposed framework: once as the signal that trains the Projector, and once as the basis for the re-identification (Re-ID) metrics used in the evaluation of Part 3.

3.2 Face Generation

While recognition networks compress faces into embeddings, generative models like GANs and diffusion models learn the distribution of human faces and can render an identity embedding as a realistic image of a face. The remainder of this section covers the specific architectures and applications the framework uses, beginning with StyleGAN.

3.2.1 StyleGAN Architecture and Advanced Latent Spaces

Standard GANs map a single latent vector directly to an image, which gives little control over the visual properties of the result. The StyleGAN architecture [36, 37] restructured the generation process around a sequence of intermediate latent spaces. A learned mapping network first transforms an initial latent vector from a basic Gaussian space Z, or an extended initial space Z^+, into the intermediate W space, which disentangles visual attributes such as head pose and hair color. To gain even finer control, StyleGAN2 is commonly extended to the W^+ space, in which a distinct latent vector from W can be assigned to each individual layer of the generation network. Before entering the synthesis network, these vectors pass through learned affine transformations (denoted A). These affine transformations map the vectors into the Style space S, which contains the specific parameters that are used to generate the final output. Figure 8 summarizes this progression. While the full W^+ space increases the expressive range of the generator, learning independent vectors for every layer makes the optimization process highly complex. Because minor identity tweaks can be learned effectively without this overhead, the proposed framework does not use independent style layers, but instead repeats a single vector in W across all layers to balance generation quality with a stable learning objective.

A second useful property of the W space is that it has a well-defined notion of the average of all samples it was trained on, namely the average latent vector w_{avg}. Latent vectors located close to w_{avg} correspond to highly realistic outputs, because w_{avg} sits in the densest, most realistic region of the latent space. SKPG-Swap anchors its projected vectors near w_{avg} to keep generated pseudonyms as realistic as possible.

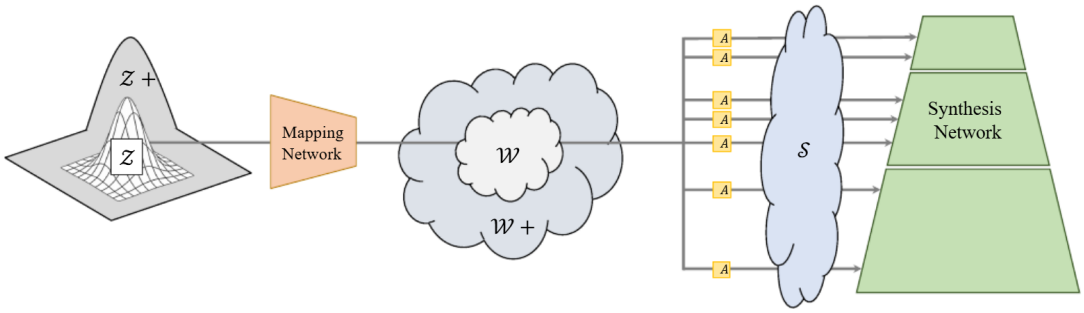

Figure 8. StyleGAN2 architecture and latent space progression. A latent code is sampled from the normally distributed latent space Z (or the extended Z^+) and transformed via a learned mapping network to the intermediate latent space W. The W space can be extended to W^+, where distinct latent vectors are applied to each layer of the generation network. These vectors pass through learned affine transformations (denoted A) to form the Style space S, which directly modulates the synthesis network. These sequential transformations enable fine-grained control over individual visual attributes, allowing StyleGAN2 to disentangle and manipulate face properties with greater precision than standard GANs. Figure adapted from Bermano et al. [38].

3.2.2 Face Swapping

Face swapping is a specific application of face generation that replaces the biometric identifiers of a subject in a target image with those of a subject in a source image, while preserving the original non-identity attributes and environmental conditions of the target [39]. This process requires disentangling identity-specific features from other attributes. Preserving this original context is essential for maintaining the analytical utility of the data, ensuring that downstream visual tasks remain unaffected after pseudonymization.

3.2.3 Face Reenactment

Face reenactment is a related but distinct application of face generation. In contrast to face swapping, face reenactment keeps the identity of the subject in the target image fixed and instead transfers motion-related information from the subject in the source image. This driving information typically includes the head pose and facial expressions of another subject. The resulting output therefore shows the fixed identity performing the motion of the subject in the source image.

The distinction between face swapping and face reenactment is illustrated in Figure 9, which shows the Face Reenactment and Face Swap results obtained for a specific pseudonym and subject. As the figure demonstrates, only face swapping preserves the original background context required for downstream analytical utility. For this reason, while the Face-Adapter component [13] supports both *face-swap* and *face-reenactment* modes, face swapping is the proposed rendering strategy in Part 3.

3.2.4 Controllable Diffusion and Adapters

Standard diffusion models perform well in generating high-quality images from random noise, typically guided by generic textual prompts. However, applying these models to targeted tasks like face swapping requires precise structural control and identity injection that basic text

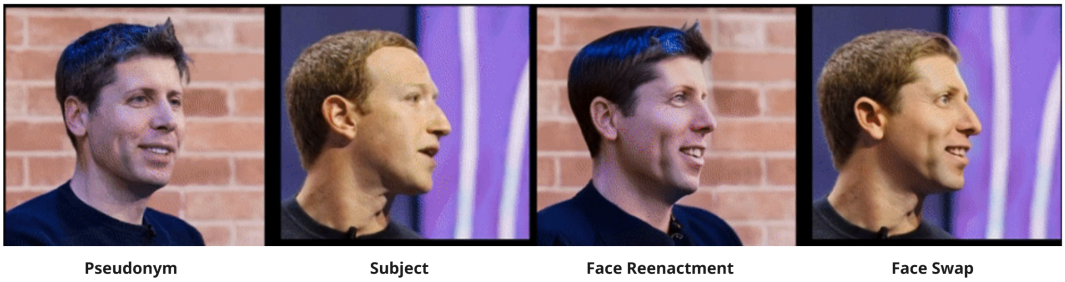

Figure 9. Comparison of face reenactment and face swapping in the pseudonymization setting. Pseudonym is the generated identity that replaces the subject, and Subject is the original person to be protected, captured in a different environment. Face Reenactment keeps the pseudonym identity but transfers the subject's head pose and expression, leaving the pseudonym re-posed on its own background rather than placed in the subject's frame. Face Swap instead renders the pseudonym identity into the subject's frame while keeping the subject's pose, expressions, and environment intact. Only face swapping returns the pseudonym to the subject's original context, which is why it is the rendering step used to preserve utility in the proposed framework. Figure adapted from Han et al. [13], with textual labels added for clarity.

prompts cannot provide. Retraining the entire base diffusion model for these specific conditions is computationally expensive.

Recent diffusion frameworks therefore use external adapter components that guide a frozen base diffusion model without modifying its weights [40]. These adapters inject specific structural context, such as environmental surroundings, head pose, and biometric identifiers, into the generation process to achieve precise face swapping. By utilizing the Face-Adapter component [13], the proposed framework can swap a pseudonym into a video frame at high quality without requiring any computationally expensive retraining of the underlying diffusion model.

4 Identity Evaluation Concepts

Having introduced the components that generate pseudonyms, the remaining question is how to evaluate them. Although generative architectures are used to produce the visual appearance of pseudonyms, a successful pseudonymization framework also requires a mathematical method to represent and evaluate the underlying identities. The central concern is the risk of biometric re-identification: whether a generated pseudonym can still be linked back to its subject. Quantifying this risk requires measuring the distance between the facial embeddings of subjects and pseudonyms, and summarizing those measurements into re-identification (Re-ID) metrics that describe the overall identity behavior of the framework. This section first introduces the cosine-similarity comparison and the Re-ID metrics built on it. Next, it presents the visual-quality metrics that measure how the surrounding context is preserved. Finally, it covers the utility metrics that test whether the pseudonymized data remains useful for downstream tasks.

4.1 Identity Evaluation and Re-identification

The Re-ID metrics quantify the identity behavior of a pseudonymization framework by comparing the facial embeddings of subjects and pseudonyms. They target four specific properties of the

generated identities. Three of these properties require the compared identities to be distinct, which is reflected by a low cosine similarity: anonymization (measuring whether a pseudonym can still be linked back to its subject), diversity (measuring whether different keys map the same subject to distinct pseudonym identities), and differentiation (measuring whether different subjects map to distinct pseudonym identities under the same key). The fourth property, pseudonym consistency, instead requires a high cosine similarity, as it measures whether the same subject maps to the same pseudonym identity across separate frames and videos. Together, these four measured properties correspond exactly to the identity objectives that the Projector is trained on in the proposed framework.

Each property is measured through the cosine similarity between two facial embeddings. Because face recognition models are trained to group embeddings of similar identities together based on cosine similarity, this score serves as a direct indicator of how similar the underlying identities are [34, 35, 41]. The formula is expressed as

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\|_2 \|B\|_2}, \tag{3}$$

where A and B are two facial embedding vectors, $A \cdot B$ is their dot product, and $\|A\|_2$ and $\|B\|_2$ are their respective L2 magnitudes. A value approaching 1 indicates that the embeddings are nearly identical, while lower or negative values represent distinct identities.

A single similarity score on its own does not tell the system whether two embeddings represent the same subject. The decision is controlled by a threshold that acts as a fixed boundary on the similarity score. Any score exceeding the threshold is classified as a match, meaning that the subject has been re-identified, while scores below the threshold are treated as distinct identities. Because the choice of threshold affects both the false positive rate and the false negative rate, its placement dictates the central trade-off that the evaluation metrics are designed to capture.

4.2 Evaluation Metrics AUC and EER

The individual cosine similarity scores between faces of the subjects and pseudonyms must be summarized into statistical metrics that describe the overall behavior of the pseudonymization framework. The two most standard metrics for biometric re-identification are the Equal Error Rate (EER) and the Area Under the Curve (AUC). Both metrics evaluate scores across all possible decision thresholds rather than a single chosen boundary, which makes them objective measures that do not depend on a specific threshold.

The Receiver Operating Characteristic (ROC) curve provides a visual representation of how a system performs across every possible decision boundary [42]. Understanding this curve requires defining the underlying classification outcomes: the True Positive Rate (TPR) measures the frequency of correctly identifying a match, the False Positive Rate (FPR) measures how often the system incorrectly links two different identities, and the False Negative Rate (FNR) measures how often the system incorrectly rejects a true match. By plotting the TPR against the FPR, the ROC curve illustrates the trade-off between correctly recognizing actual matches and incorrectly making false links. Each point along the curve corresponds to a specific cosine similarity threshold, and the diagonal dashed line represents random chance, where a classifier performs no better than guessing.

The Equal Error Rate (EER) is the operating point on the ROC curve where the FPR equals the FNR, marked in Figure 10. A low EER indicates that the system separates matching from non-matching pairs accurately at a balanced operating point and is therefore preferred.

The Area Under the Curve (AUC) quantifies overall system scores by integrating the ROC curve across all thresholds, visualized as the shaded region in Figure 10. Perfect separation

between the two classes means the system can distinguish matching from non-matching pairs at every threshold, whereas a score of random chance indicates the ROC curve coincides with the diagonal line. To align with the unified reporting standards of the proposed framework, all AUC and EER scores are scaled from their traditional 0 to 1 format to a range between 0 and 100. Under this scaling, a perfect anonymization AUC of 100.0 means that subjects and generated pseudonyms are so distinct that no overlap exists between them, while a random classifier yields a score of 50.0.

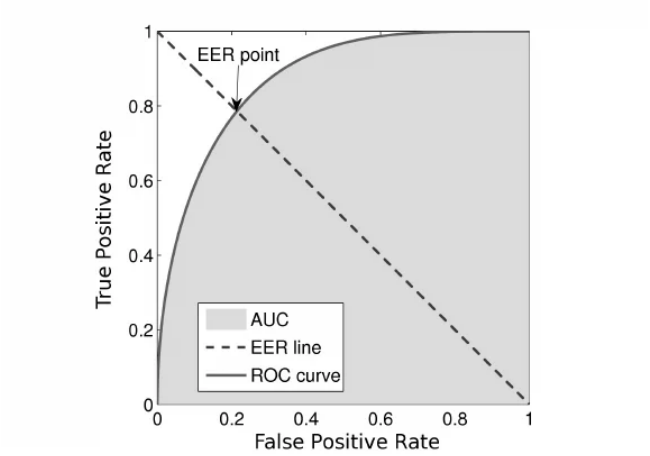

Figure 10. Receiver Operating Characteristic (ROC) curve with marked metrics. The x-axis represents the False Positive Rate (FPR, incorrectly linking two different identities), and the y-axis represents the True Positive Rate (TPR, correctly identifying a match). Each point on the solid ROC curve corresponds to a specific decision threshold for biometric matching. The dashed diagonal line represents the score of a random classifier. The Equal Error Rate (EER) point marks where FPR equals the False Negative Rate (FNR), i.e., where the curve crosses the diagonal going from (0,1) to (1,0). The shaded area under the ROC curve represents the Area Under the Curve (AUC), which quantifies the overall discriminative ability of the system. Note that these metrics are traditionally evaluated on a 0 to 1 scale, while the proposed framework scales them to the 0 to 100 range for readability. Evaluating these threshold-independent metrics is essential for determining the overall Re-ID scores of the generated pseudonyms. Figure adapted from Tronci et al. [43].

4.3 Visual Quality and Contextual Preservation Metrics

Beyond identity-level evaluation, a pseudonymization framework must also be assessed on how it preserves the visual properties of the original frames. Three complementary metrics are used in this thesis to measure this preservation.

The Structural Similarity Index Measure (SSIM) compares two images on the basis of luminance, contrast, and local structure [44]. Rather than measuring raw pixel-level differences, SSIM evaluates whether the spatial patterns and intensity relationships of the original frame are retained in the modified output. It is particularly suited for evaluating whether the background, environment, and composition of a frame remain intact after pseudonymization. While traditionally measured on a scale from -1 to 1, the proposed framework scales SSIM scores to a range between -100 and 100 for consistency, where 100.0 indicates perfect structural agreement.

The Learned Perceptual Image Patch Similarity (LPIPS) metric measures perceptual similarity using deep features extracted by a pre-trained convolutional network [45]. Two images are compared in this learned feature space, and their distance reflects how similar they appear to a human observer rather than how closely their pixels match. LPIPS complements SSIM by capturing high-level appearance information, such as texture and visual realism, that purely structural metrics may miss. Like the other metrics in the proposed framework, LPIPS values are scaled from their default 0 to 1 range to a 0 to 100 scale, where lower values indicate higher perceptual similarity.

Landmark distance evaluates the preservation of pose and facial expression. Facial landmark detectors locate predefined keypoints on the face, such as the corners of the eyes and mouth [46]. The metric is computed as the normalized pixel-level deviation between the aligned landmark positions of the faces of the subject and the pseudonym in the output. A low landmark distance indicates that the output retains the head pose and facial expressions of the subject, which are essential for downstream tasks that depend on facial motion.

4.4 Utility and Probabilistic Evaluation Metrics

Visual quality metrics quantify how a pseudonymization framework affects the appearance of the data, but they do not measure whether the modified data remains useful for downstream analysis. A complementary evaluation strategy is to apply pre-trained downstream models, such as action or emotion classifiers, to the pseudonymized output and compare their score to the unmodified videos. The underlying assumption is that if a downstream model continues to perform reliably on the modified data, then the relevant signal for that task has been preserved.

To evaluate this preserved utility, downstream tasks often rely on specialized pre-trained architectures. Video action recognition, for example, utilizes models like VideoMAE [47]. VideoMAE is a transformer-based architecture that replaces convolutional filters with attention mechanisms, learning relationships between every pair of structural and temporal input regions [48]. This attention mechanism allows the model to capture long-range dependencies across both space and time. Each input region is a spatial and temporal patch covering a few frames and a small portion of the image. Attention across these patches lets the model relate, for example, a patch showing a subject picking up a toothbrush at the start of a video to a patch showing the brushing motion several seconds later. Similarly, facial emotion recognition utilizes models like Emo-AffectNet [49], which is trained on multiple large-scale emotion datasets to predict discrete emotion categories from short videos. Architecturally, it combines a CNN backbone for per-frame facial feature extraction with a temporal module that aggregates these features into a single video-level emotion prediction. Because these models capture complex spatio-temporal dependencies and subtle facial features, VideoMAE and Emo-AffectNet serve as the two pre-trained downstream models used to evaluate the proposed framework in Part 3.

Three metrics quantify downstream task success. Classification accuracy provides an interpretable summary of whether the task remains solvable after pseudonymization by measuring the proportion of videos correctly assigned to their ground-truth class by the classifier, calculated as

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i), \tag{4}$$

where N is the total number of evaluation videos, $\hat{y}_i$ is the predicted class for video i, y_i is its true ground-truth class, and $\mathbb{1}(\cdot)$ is the indicator function that returns 1 if the prediction matches the ground truth and 0 otherwise. However, classification accuracy treats every input equally and can overestimate effectiveness when the evaluation set is imbalanced across classes. To

address class imbalance, the Unweighted Average Recall (UAR) ensures every class contributes equally to the final value regardless of sample count. The UAR computes the recall separately for each class and averages the per-class scores, given by

$$\text{UAR} = \frac{1}{C} \sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}, \tag{5}$$

where C is the total number of classes, TP_c represents the true positives for class c, and FN_c represents the false negatives for that same class.

The Prediction Agreement Rate complements the ground-truth metrics by measuring how often predictions on the pseudonymized video result in the exact same prediction as on the unmodified videos, defined as

$$\text{Agreement Rate} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{y}_i' = \hat{y}_i), \tag{6}$$

where N is the total number of videos, $\hat{y}_i'$ is the model's predicted class for the pseudonymized video i, and $\hat{y}_i$ is the predicted class for the corresponding baseline video. A high agreement rate demonstrates that the pseudonymization process preserves video utility by maintaining the same downstream classifier behavior as the unmodified video. Together, these three metrics describe whether the downstream task succeeds, how success is distributed across classes, and how closely the modified data mimics the behavior of the baseline. To align with the reporting standards of the proposed framework, Accuracy, UAR, and Agreement Rate are all scaled from a standard proportion of 0 to 1 up to a percentage range of 0 to 100.

When comparing the metrics of a classifier on two distinct datasets, it is necessary to determine whether any observed drop in accuracy is statistically significant. The exact two-sided McNemar test evaluates paired nominal data [50]. This paired design fits scenarios where a downstream model classifies the exact same sample under two different conditions. The test isolates discordant pairs: instances where the baseline model is correct but the modified method is incorrect (denoted as b), and instances where the baseline is incorrect but the modified method is correct (denoted as c). Letting $n = b + c$ represent the total number of discordant pairs and $k = \min(b, c)$, the exact p-value is calculated using the binomial distribution:

$$p = \min\left(1, 2 \sum_{i=0}^{k} \binom{n}{i} 0.5^n\right). \tag{7}$$

A low p-value, commonly defined as $p < 0.05$, confirms that a change in classification accuracy is statistically significant rather than an artifact of random chance. Consequently, this exact two-sided McNemar test is used to rigorously evaluate the metric drops caused by the proposed pseudonymization framework.

References for Part 1 & 2

[1] Runfang Guo et al. "Development and application of emotion recognition technology — a systematic literature review". In: *BMC Psychology* 12.1 (2024), pp. 1–25. DOI: 10.1186/s40359-024-01581-4.

[2] Duaa Shehada et al. "An Explainable Framework for Mental Health Monitoring Using Lightweight and Privacy-Preserving Federated Facial Emotion Recognition". In: *Sensors* 25.23 (2025), p. 7320. DOI: 10.3390/s25237320. URL: https://doi.org/10.3390/s25237320.

[3] S. Qiao and H. Liu. "Automatic classification of criminal activities for security surveillance by keyframes detection and advanced inception techniques". In: *Scientific Reports* 16.1 (2026), pp. 1–14. DOI: 10.1038/s41598-025-30199-8.

[4] Kumuda P et al. "A comprehensive review of AI-powered campus surveillance". In: *ITM Web of Conferences* 81 (Jan. 2026). DOI: 10.1051/itmconf/20268101017.

[5] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj.

[6] Centers for Medicare & Medicaid Services. *The Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. Public Law 104-191, 110 Stat. 1936. Accessed: 2026-03-27. 1996. URL: https://www.cms.gov.

[7] Jingyi Cao et al. "Face De-Identification: State-of-the-Art Methods and Comparative Studies". In: *IEEE Transactions on Broadcasting* PP (Jan. 2025), pp. 1–21. DOI: 10.1109/TBC.2025.3639783.

[8] Roland Stenger et al. "Evaluating the Impact of Face Anonymization Methods on Computer Vision Tasks: A Trade-Off Between Privacy and Utility". In: *IEEE Access* PP (Dec. 2024), pp. 1–1. DOI: 10.1109/ACCESS.2024.3519441.

[9] Yu Kong and Yun Fu. "Human Action Recognition and Prediction: A Survey". In: *International Journal of Computer Vision* 130.5 (2022), pp. 1366–1401. DOI: 10.1007/s11263-022-01594-9.

[10] Shan Li and Weihong Deng. "Deep Facial Expression Recognition: A Survey". In: *IEEE Transactions on Affective Computing* 13.3 (2022), pp. 1195–1215. DOI: 10.1109/TAFFC.2020.2981446.

[11] Vishal M. Patel, Nalini K. Ratha, and Rama Chellappa. "Cancelable Biometrics: A review". In: *IEEE Signal Processing Magazine* 32.5 (2015), pp. 54–65. DOI: 10.1109/MSP.2015.2434151.

[12] R. Dhanyalakshmi et al. "A Survey on Face-Swapping Methods for Identity Manipulation in Deepfake Applications". In: *IET Image Processing* 19 (June 2025). DOI: 10.1049/ipr2.70132.

[13] Yue Han et al. *Face Adapter for Pre-Trained Diffusion Models with Fine-Grained ID and Attribute Control*. 2024. arXiv: 2405.12970 [cs.CV]. URL: https://arxiv.org/abs/2405.12970.

[14] Zhuowen Yuan et al. "On Generating Identifiable Virtual Faces". In: *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. ACM, Oct. 2022, pp. 1465–1473. DOI: 10.1145/3503161.3548110. URL: http://dx.doi.org/10.1145/3503161.3548110.

[15] Miaomiao Wang et al. “A Key-Driven Framework for Identity-Preserving Face Anonymization”. In: *Proceedings of the 32nd Annual Network and Distributed System Security Symposium (NDSS)*. Internet Society, Jan. 2025. DOI: 10.14722/ndss.2025.230729.

[16] Anil K. Jain, Arun Ross, and Salil Prabhakar. “An Introduction to Biometric Recognition”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 14 (Feb. 2004), pp. 4–20. DOI: 10.1109/TCSVT.2003.818349.

[17] Article 29 Data Protection Working Party. *Opinion 05/2014 on Anonymisation Techniques*. Tech. rep. 0829/14/EN WP216. European Commission, 2014. URL: https://europa.eu.

[18] Jonathan Katz and Yehuda Lindell. *Introduction to Modern Cryptography*. Chapman & Hall/CRC Cryptography and Network Security Series. CRC Press, 2007. ISBN: 978-1-58488-551-1. DOI: 10.1201/9781420010756.

[19] Anil K. Jain, Karthik Nandakumar, and Abhishek Nagar. “Biometric Template Security”. In: *EURASIP Journal on Advances in Signal Processing* 2008 (Mar. 2008), p. 579416. DOI: 10.1155/2008/579416.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[21] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning Representations by Back-Propagating Errors”. In: *Neurocomputing: Foundations of Research*. MIT Press, 2002, pp. 213–222. ISBN: 9780262281744. DOI: 10.7551/mitpress/1888.003.0013.

[22] Feng Wang et al. “NormFace: L2 Hypersphere Embedding for Face Verification”. In: *Proceedings of the 25th ACM International Conference on Multimedia*. 2017, pp. 173–181. DOI: 10.1145/3123266.3123359.

[23] Rich Caruana. “Multitask Learning”. In: *Machine Learning* 28.1 (July 1997), pp. 41–75. DOI: 10.1023/A:1007379606734.

[24] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer, 2006. ISBN: 978-0387310732.

[25] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[26] Xiang Li et al. “Predicting the effective mechanical property of heterogeneous materials by image based modeling and deep learning”. In: *Computer Methods in Applied Mechanics and Engineering* 347 (2019), pp. 735–753. DOI: 10.1016/j.cma.2019.01.005.

[27] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2014, pp. 2672–2680. URL: https://papers.nips.cc/paper/5423-generative-adversarial-nets.

[28] Zhengwei Wang, Qi She, and Tomas E. Ward. “Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy”. In: *ACM Computing Surveys* 54.2 (2021), pp. 1–38. DOI: 10.1145/3439723.

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 6840–6851. DOI: 10.48550/arXiv.2006.11239.

[30] Arash Vahdat and Karsten Kreis. “Improving Diffusion Models as an Alternative to GANs”. In: *NVIDIA Developer Blog* (Apr. 2022). Accessed: June 1, 2026. URL: https://developer.nvidia.com/blog/improving-diffusion-models-as-an-alternative-to-gans-part-1/.

[31] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 10674–10685. DOI: 10.1109/CVPR52688.2022.01042.

[32] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

[33] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 27. 2014, pp. 3320–3328.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[35] Jiankang Deng et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4690–4699. DOI: 10.1109/CVPR.2019.00482.

[36] Tero Karras, Samuli Laine, and Timo Aila. “A Style-Based Generator Architecture for Generative Adversarial Networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4401–4410. DOI: 10.1109/CVPR.2019.00453.

[37] Tero Karras et al. “Analyzing and Improving the Image Quality of StyleGAN”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8110–8119. DOI: 10.1109/CVPR42600.2020.00813.

[38] Amit H. Bermano et al. “State-of-the-Art in the Architecture, Methods and Applications of StyleGAN”. In: *Computer Graphics Forum* 41.2 (2022), pp. 591–611. DOI: 10.1111/cgf.14502.

[39] Iryna Korshunova et al. “Fast Face-swap Using Convolutional Neural Networks”. In: *arXiv preprint arXiv:1611.09577* (Nov. 2016). DOI: 10.48550/arXiv.1611.09577.

[40] Chong Mou et al. *T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models*. 2023. arXiv: 2302.08453 [cs.CV]. URL: https://arxiv.org/abs/2302.08453.

[41] Yaniv Taigman et al. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220.

[42] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (June 2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[43] Roberto Tronci, Giorgio Giacinto, and Fabio Roli. “Dynamic Score Combination: A Supervised and Unsupervised Score Combination Method”. In: *Multiple Classifier Systems*. Vol. 5519. Lecture Notes in Computer Science. Springer, 2009, pp. 163–177. DOI: 10.1007/978-3-642-03070-3_13.

[44] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[45] Richard Zhang et al. “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068.

[46] Adrian Bulat and Georgios Tzimiropoulos. "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 (ICCV)*. 2017, pp. 1021–1030. DOI: 10.1109/ICCV.2017.116.

[47] Zhan Tong et al. "VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 10078–10093.

[48] Ashish Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 5998–6008.

[49] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. "In Search of a Robust Facial Expressions Recognition Model: A Large-Scale Visual Cross-Corpus Study". In: *Neurocomputing* 514 (2022), pp. 435–448. DOI: 10.1016/j.neucom.2022.10.013. URL: https://www.sciencedirect.com/science/article/pii/S0925231222012656.

[50] Alan Agresti. *Categorical Data Analysis*. 3rd. Hoboken: John Wiley & Sons, 2013. ISBN: 978-0-470-46363-5. DOI: 10.1002/9781118357590.

PART 3

Scientific Article

Preserving Video Utility via Consistent Subject- and Key-derived Pseudonyms combined with Face Swapping[1]

Timo van Hoorn
Computer Vision Lab
Delft University of Technology

Jan van Gemert
Computer Vision Lab
Delft University of Technology

P. Benschop
Signal Processing Systems (SPS)
Dept. of Microelectronics, TU Delft

Abstract

The face and its surrounding context are a strong signal for video analysis in sensitive domains, powering action recognition in forensics and longitudinal emotion analysis in medicine. However, faces are biometric data that privacy regulations such as the GDPR and HIPAA protect, forbidding their storage without protective measures. Pseudonymization solves this problem by replacing each face with a generated one, called a pseudonym. To remain useful, a pseudonymization method must satisfy three requirements: preserving the context around the face, mapping the same subject to the same pseudonym across separate videos, and avoiding any biometric database that links subjects to their pseudonyms. No existing method satisfies all three. Face swapping preserves context but depends on a biometric database to stay consistent, while subject- and key-conditioned pseudonym generators remove that database but discard the original frame along with its context. This thesis closes the gap with SKPG-Swap: a hybrid framework in which a lightweight Subject- and Key-conditioned Pseudonym Generator (SKPG) derives a consistent pseudonym from a subject's face and a secret key, combined with a face-swap model that blends that pseudonym back into the original frame. Evaluated against bounding-box rendering strategies built on the same SKPG backbone, SKPG-Swap retains nearly all of the action-recognition accuracy of unmodified videos on UCF101 and outperforms the other pseudonymization methods on RAVDESS emotion recognition. A controlled experiment further shows that assigning a subject a consistent pseudonym identity, rather than an inconsistent one, results in more stable predictions across videos, motivating the consistency requirement.

Keywords: Video Pseudonymization, Face Swapping, Action Recognition, Emotion Recognition.

[1]The source code and models used for the implementation and evaluation of the proposed framework are publicly available at: `https://github.com/tfrvanhoorn/msc-thesis-utility-preserving-video-pseudonymization`

1 Introduction

Computer vision is increasingly used to analyze video in sensitive domains such as forensics [4, 5] and medicine [6, 7], where the processing and storage of facial data are strictly controlled by privacy regulations like the GDPR [8] and HIPAA [9]. These regulations forbid the storage of biometric identifiers without strong protective measures [8, Art. 5(1)(c)]. To analyze this data compliantly, every face must be removed through anonymization or replaced through pseudonymization. Anonymization permanently destroys all links to the subject, while pseudonymization masks the identity using a generated substitute that can be reversed [10]. Destructive anonymization techniques such as blurring and masking degrade the accuracy of downstream models that rely on facial detail and context [11]. Such downstream tasks include action recognition in forensic surveillance [12, 13] and the longitudinal tracking of patient emotional states through emotion recognition [14, 15]. These tasks demand a pseudonymization method that removes the original identity without losing the visual cues the downstream models depend on.

Preserving the analytical utility of pseudonymized video for these tasks places three requirements on the pseudonymization system. The first is **consistent pseudonymization**, defined as the mapping of a subject to the same pseudonym across continuous frames and separate videos. Without consistency, variance can be introduced into the predictions of downstream longitudinal tasks. The second requirement is **context preservation**. The pseudonym must be rendered back into the original frame, changing the identity while keeping other context intact. This context includes head pose, facial expression, and environment. Here, environment denotes the background surrounding the face, together with the objects and lighting interacting with it. Action recognition and emotion recognition both depend on some of these cues [16, 17], and discarding them alongside the original identity degrades the utility of the pseudonymized data [18]. The third requirement is that the pseudonymization must be **biometric database-free**. The system should avoid storing a database that connects subjects to their assigned pseudonyms to achieve consistency. Such a database represents a single point of failure that exposes

Figure 1. Qualitative comparison of pseudonymization methods on one RAVDESS [1] (left) and two VoxCeleb2 [2] (middle and right) videos. Rows show, from top to bottom, the original frames, DeepPrivacy2 [3], two baseline rendering strategies using the shared generation backbone (SKPG-BB and SKPG-Reenact), and the proposed face-swap rendering (SKPG-Swap). Each block of three columns shows three frames sampled from the same video for one subject. SKPG-Swap visually preserves more surrounding context of the original frames than SKPG-BB and SKPG-Reenact.

every protected identity if breached [19]. Unlike a leaked password, biometric identifiers cannot be revoked or reissued once exposed [20], so a database breach permanently compromises every subject it contains. The 2019 BioStar 2 incident [21] illustrates this risk. A single misconfigured biometric platform exposed more than one million unencrypted fingerprint and face recognition records.

No existing pseudonymization method satisfies all three requirements. Face-swap techniques such as Face-Adapter [22] preserve context by blending the pseudonym's identity into the context of the original frame [23, 24]. However, these techniques only achieve consistent pseudonymization when links between subjects and pseudonyms are stored, as face swapping requires a reference image of the pseudonym. This forces pipelines to maintain a database containing these links, violating the biometric database-free requirement. A distinct family of frameworks relies on subject- and key-conditioned pseudonymization. These methods train a neural network to map a subject-key pairing to a consistent pseudonym, achieving consistency through training objectives rather than storing vulnerable data. However, existing methods in this family stop after generating the pseudonym and never render it back into the original frame, losing its context.

To address the trade-off between context preservation and database-free security, this thesis proposes a hybrid framework centered on a lightweight **Subject- and Key-conditioned Pseudonym Generator (SKPG)**. The SKPG generates consistent pseudonyms from a subject and a binary cryptographic key. Then a face-swap renderer, based on Face-Adapter, blends the pseudonym back into the original frame while preserving its context.

To evaluate face swapping as a rendering choice, the experiments compare three rendering strategies on the shared SKPG backbone. These strategies include bounding-box overwrite, reenactment-aligned bounding-box overwrite, and the proposed face swap. Bounding-box overwrite crops the generated face and pastes it directly onto the subject's bounding box in the original frame. The reenactment-aligned variant first aligns the pseudonym's pose and expression to the subject, then overwrites that same region. Figure 1 illustrates the three rendering strategies on three sample videos. The bounding-box strategies visibly discard the context around the face, while the proposed face-swap strategy blends the pseudonym into the original frame and preserves context.

In summary, this thesis makes three primary contributions:

- The integration of a Subject- and Key-conditioned Pseudonym Generator with a face-swap renderer based on Face-Adapter, producing a consistent pseudonymization pipeline (SKPG-Swap) that preserves context while remaining biometric database-free.
- A comparison of rendering strategies built on

the SKPG backbone, benchmarking a bounding-box overwrite (SKPG-BB), a reenactment-aligned bounding-box overwrite (SKPG-Reenact), and the proposed face swap (SKPG-Swap), evaluated on action recognition and emotion recognition.

- A demonstration that, across all tested rendering strategies on the SKPG backbone, mapping a subject to a consistent pseudonym identity produces more stable downstream predictions than mapping the same subject to inconsistent pseudonym identities.

2 Related Works

Classical Anonymization Techniques Early video anonymization methods rely on destructive techniques such as blurring, pixelation, and solid masking [25, 26, 27]. These techniques are computationally efficient but remove the facial detail that downstream tasks such as emotion recognition, action recognition, pose estimation, and anomaly detection rely on [17, 28, 29, 30, 31, 32]. This motivates identity generation approaches that replace rather than destroy these facial cues.

Pseudonym Generation Modern deep learning approaches overcome the utility limitations of classical techniques by replacing faces with generated ones, preserving realistic cues that computer vision models trained on real faces can understand. Methods based on Generative Adversarial Networks (GAN) like DeepPrivacy2 [3] replace the subject with a generated pseudonym that is sampled at random, using facial landmarks to preserve the original head pose. Because the pseudonym is sampled at random, the method cannot deliberately map a subject to one fixed pseudonym. Consistent pseudonymization instead requires a repeatable mapping from a subject to a pseudonym. CIAGAN [33] addresses this consistency problem by conditioning a GAN on an identity embedding. Conditioning on identity ensures that the generated pseudonym matches this identity and can be reproduced consistently, an approach also adopted by diffusion-based generators such as DCFace [34]. Identity-conditioned generation therefore provides the foundation for consistent pseudonymization.

The existing literature motivates this consistency mainly through within-video utility and cross-video re-identification linking [11], where the goal is to recognize the same subject across separate videos. However, the role of consistency in stabilizing downstream predictions across videos remains unaddressed.

Face Swapping and Context Preservation Building on identity-conditioned generation, face swapping blends the pseudonym into the subject's environment, preserving context like lighting, hair, skin tone, texture, and background occlusion. GAN-based methods such as SimSwap [35] established the technique with lightweight architectures suitable for fast rendering, but tend to produce coarse detail and blending artifacts around the swapped face. Diffusion-based methods such as DiffSwap [36] and Face-Adapter [22] produce sharper and more realistic detail at the expense of higher computational costs [11]. To satisfy the consistency requirement, pseudonymization pipelines that only use face swapping require references of pseudonyms to be linked to the subject. However, these links need to be stored in a biometric identity database that violates the biometric database-free requirement.

Subject- and Key-Conditioned Pseudonymization Subject- and key-conditioned pseudonymization methods are neural networks trained to generate a pseudonym directly from the subject's facial embedding and a secret cryptographic key. Frameworks such as IVFG [37] and KFAAR [38] feed the embedding and key into a StyleGAN-based generator [39, 40], training the network using a multi-task objective designed so that four key properties emerge from the mapping itself:

- **Anonymity:** The pseudonym's identity cannot be linked to the subject.
- **Consistency:** A specific subject-key pairing maps to the same pseudonym identity.
- **Differentiation:** Different subjects map to distinct pseudonym identities, even when the same key is used.
- **Diversity:** Different keys on the same subject map to a distinct pseudonym identity.

As the network is trained on consistency, it replaces the vulnerable biometric database that face-swap pipelines depend on to achieve consistent pseudonymization. Another advantage of IVFG and KFAAR is that they are keyed. To link pseudonyms back to the subjects, both the trained pipeline and the key need to be compromised, adding an extra layer of security. However, the published pipelines for these methods stop after generating the pseudonym. Their output is a standalone image of a pseudonym carrying its own head pose, expression, and environment. While KFAAR appends a face-reenactment component to align the pose and expression of the generated pseudonym to the subject, it remains an image-generation pipeline that does not render the pseudonym back into the original frame. Furthermore, neither framework has been evaluated on video utility preservation. Subject- and key-conditioned methods therefore secure database-free, consistent pseudonymization but have not been combined with a renderer that integrates the pseudonym into the original frame while preserving context, leaving a gap that this thesis closes.

Table 1. Comparison of privacy-preserving video analysis methods across the requirements for utility-preserving pseudonymization. Consistency indicates if the method achieves within or cross-video consistent pseudonymization, where *Not Applicable* implies that the method does not generate a pseudonym. Database-free refers to the absence of a stored biometric database. Keyed indicates that pseudonym generation is bound to a secret cryptographic key, strengthening the security of the method. Context preservation lists which non-identity cues survive. Every prior method fails at least one requirement. The bottom row shows that the proposed framework (SKPG-Swap) is the only method satisfying all requirements.

Method	Consistency	Database-free	Keyed	Context Preservation
Masking	Not applicable	✓	×	None
Blurring	Not applicable	✓	×	None
DeepPrivacy2	Within video*	✓	×	Pose, environment
Face swapping	Cross video†	×	×	Pose, expression, environment
IVFG	Cross video	✓	✓	None
KFAAR	Cross video	✓	✓	Pose, expression
SKPG-Swap (Ours)	Cross video	✓	✓	Pose, expression, environment

*Can achieve weak cross-video consistency when the subject's appearance is similar across videos.
†Achieved via a biometric database, violating the database-free requirement.

Table 1 summarizes the trade-offs for existing methods and illustrates how the proposed SKPG-Swap framework satisfies all requirements.

3 Proposed Method

The proposed framework, **SKPG-Swap**, uses two components to address separate requirements. The Subject- and Key-conditioned Pseudonym Generator (SKPG) combines a lightweight projector with a StyleGAN2 generator to produce a pseudonym from the subject's facial embedding and a secret cryptographic key. This backbone is designed to satisfy the consistency and biometric database-free requirements. A diffusion-based face-swap renderer, based on Face-Adapter [22], then blends this pseudonym back into the original frame, restoring the context that the SKPG output alone discards. Figure 2 illustrates the full pipeline.

3.1 System Pipeline

The pipeline processes a subject through several components. First, the original frame containing the subject's face x is fed into the Encoder (E), which extracts an identity embedding with dimensionality d_z,

$$z = E(x), \qquad z \in \mathbb{R}^{d_z}. \tag{1}$$

This embedding is then passed to the Projector (P), which first preprocesses z and a binary cryptographic key $k \in \{0,1\}^{d_k}$ with dimensionality d_k, then concatenates the preprocessed vectors, and finally feeds them into a Multi-Layer Perceptron (MLP) to produce a projected vector z'. Directly concatenating z and k was considered problematic as identity embeddings from pre-trained face recognition models are typically L2-normalized, while the cryptographic key is an unnormalized binary vector that can have a different dimensionality. As a result, the two inputs differ in both dimensionality ($d_z \gg d_k$) and numerical magnitude. A raw concatenation forces the downstream MLP to learn both differences before it can balance the contributions of the two inputs, which complicates the optimization of the learning objectives. To remove this burden from the optimizer, the Projector preprocesses each input separately before concatenation. The latent vector z is L2-normalized to align with the convention used by face recognition models,

$$\hat{z} = \frac{z}{\|z\|_2}. \tag{2}$$

In parallel, the cryptographic key k is preprocessed to align its scale and dimensionality with the normalized identity embedding $\hat{z}$. First, the key is L2-normalized to ensure its numerical magnitude is balanced with the identity vector,

$$\hat{k} = \frac{k}{\|k\|_2}. \tag{3}$$

Next, this normalized key is transformed into the target dimensionality d_z through a learned linear layer followed by a ReLU activation,

$$k' = \mathrm{ReLU}(W_k \hat{k} + b_k), \qquad W_k \in \mathbb{R}^{d_z \times d_k}, b_k \in \mathbb{R}^{d_z}, \tag{4}$$

where W_k represents the learnable weight matrix of the linear projection layer, and b_k is the corresponding learnable bias vector. The projection produces a key-based vector k' that matches $\hat{z}$ in dimensionality, and whose magnitudes adjust during training to fall in a range comparable to that of $\hat{z}$. The MLP then combines the identity-based vector $\hat{z}$ with the key-based vector k' into the projected vector

$$z' = P(z,k) = \mathrm{MLP}\big(\hat{z} \,\|\, k'\big), \tag{5}$$

where $\|$ denotes concatenation along the feature axis. By resolving the dimensional and scale mismatches in

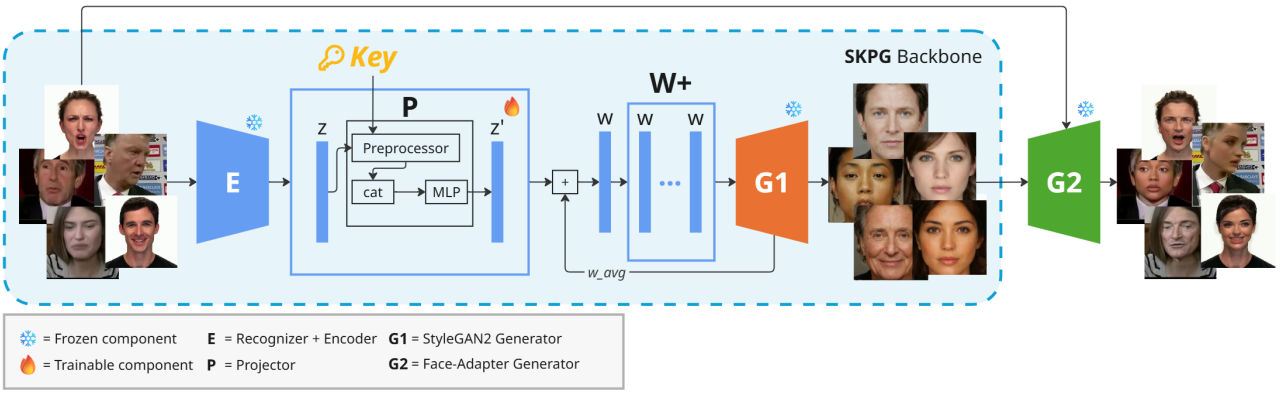

Figure 2. A simplified overview of the proposed hybrid framework. The Subject- and Key-conditioned Pseudonym Generator (SKPG) first generates a pseudonym based on the subject and a binary cryptographic key: the subject's face is encoded into a latent vector z by the Encoder E, preprocessed together and concatenated with the key in the Projector P, and mapped through the StyleGAN2 generator G_1 to generate a consistent pseudonym. After the SKPG has generated a pseudonym, the Face-Adapter generator G_2 face swaps the pseudonym back into the original frame to preserve context. Only the lightweight Projector P is trained, while the Encoder, StyleGAN2 generator, and Face-Adapter generator are frozen. This design keeps the framework biometric database-free and consistent through the SKPG backbone while preserving context through the Face-Adapter generator.

this dedicated projection step, the network frees the downstream MLP to focus on its primary objective of merging the identity and the key.

After projection, the projected vector z' is added to the pre-calculated average latent vector w_{avg} of the StyleGAN2 [40] generator,

$$w = w_{\text{avg}} + z'. \tag{6}$$

By anchoring the latent at w_{avg}, the model starts from a realistic face. w_{avg} is the average latent vector of the StyleGAN2 W space and therefore lies in its densest, most realistic region. Starting from this point lets the network learn small, identity-driven offsets rather than discover realistic faces from scratch. To match the required input format of the StyleGAN2 generator, the vector w is then replicated across 18 style layers to form the extended latent representation $W^+ \in \mathbb{R}^{18\times d_z}$. This step is denoted by a tiling operator $\mathcal{T} : \mathbb{R}^{d_z} \to \mathbb{R}^{18\times d_z}$, defined as

$$W^+ = \mathcal{T}(w) = (\underbrace{w, w, \ldots, w}_{18 \text{ times}}). \tag{7}$$

The StyleGAN2 generator (G_1) then maps this extended latent to generate a pseudonym,

$$x' = G_1(W^+) = G_1(\mathcal{T}(w)). \tag{8}$$

In the final stage, the pseudonym x' is combined with the original frame x in the Face-Adapter generator (G_2), which performs a face swap to produce the final pseudonymized frame $\hat{x}$,

$$\hat{x} = G_2(x, x'). \tag{9}$$

This generator preserves the original context by blending the generated pseudonym into the subject in the original video frame.

The swap replaces only the face and keeps the subject's hair from the original frame. The hair is left in place because Face-Adapter blends the new face into the existing head, and regenerating the hairline would introduce artifacts.

3.2 Training Strategy

Only the SKPG backbone is utilized during training and the Face-Adapter generator (G_2) does not participate in the training loop. Training optimizes the Projector (P), including its key projection parameters W_k and b_k, while the Encoder (E) and StyleGAN2 generator (G_1) are frozen. Face-Adapter is excluded due to its high parameter count and large diffusion-based architecture. Including it in the training process would increase the computational cost because of the added volume of gradients required for backpropagation. Its inclusion is unnecessary for the core optimization, as the loss functions are specifically designed to refine the identities generated by StyleGAN2. Furthermore, Face-Adapter is already pretrained to preserve these pseudonyms' identities during swapping. Further training would only be fine-tuning the model on the identities in the training set, introducing the risk of overfitting.

The optimization is controlled by a multi-task learning objective built on a Cosine Embedding Loss ($L_{\cos}$). The loss compares two identity embeddings f_1 and f_2 through their cosine similarity $\cos(f_1, f_2)$, and a label $l \in \{1, -1\}$ selects how the pair is treated. With $l = 1$ the loss decreases as the two embeddings align, pulling them

toward the same identity. With $l = -1$ the loss instead penalizes similarity and pushes the embeddings apart, until their cosine similarity falls to the margin m, beyond which the penalty is zero. The margin m therefore sets the minimum separation enforced in the push-apart case. The margin is a tunable hyperparameter held fixed during training. Combining both cases gives

$$L_{\cos}(f_1, f_2, l, m) = \begin{cases} 1 - \cos(f_1, f_2), & l = 1, \\ \max\big(0,\ \cos(f_1, f_2) - m\big), & l = -1. \end{cases} \tag{10}$$

In each of the loss terms that follow, l is fixed to a single value, $+1$ when the two embeddings should match and -1 when they should differ, so l acts as a fixed switch per term rather than a quantity that varies during training.

For simplicity, the SKPG backbone is denoted by the shorthand

$$G(x, k) = G_1\big(\mathcal{T}(w_{\text{avg}} + P(E(x), k))\big), \tag{11}$$

which generates a pseudonym from a subject's face x and a secret key k.

The five primary loss components are defined as follows, where the symbol R denotes a pre-trained face recognition model used to extract identity embeddings, x_1 and x_2 are two samples of one subject, y_1 is a sample of a different subject, and k_1 and k_2 are two distinct keys.

- **Anonymity loss (L_{ano}):** ensures the pseudonym identity is different from the identity of the subject,

$$L_{\text{ano}} = L_{\cos}\big(R(G(x_1, k_1)),\ R(x_1),\ -1,\ m\big). \tag{12}$$

- **Consistency loss (L_{con}):** ensures that different samples of the same subject generate the same pseudonym identity when using the same key,

$$L_{\text{con}} = L_{\cos}\big(R(G(x_1, k_1)),\ R(G(x_2, k_1)),\ 1,\ m\big). \tag{13}$$

- **Diversity loss (L_{div}):** ensures that a subject generates different pseudonym identities when using different keys,

$$L_{\text{div}} = L_{\cos}\big(R(G(x_1, k_1)),\ R(G(x_1, k_2)),\ -1,\ m\big). \tag{14}$$

- **Differentiation loss (L_{dif}):** ensures that distinct subjects generate different pseudonym identities even if the same key is applied,

$$L_{\text{dif}} = L_{\cos}\big(R(G(x_1, k_1)),\ R(G(y_1, k_1)),\ -1,\ m\big). \tag{15}$$

- **Regularization loss (L_{reg}):** ensures the pseudonyms are realistic,

$$L_{\text{reg}} = \|w - w_{\text{avg}}\|_2^2, \tag{16}$$

where w_{avg} represents the average latent vector of the StyleGAN2 generator. By penalizing large deviations from w_{avg}, L_{reg} constrains w to remain within the realistic regions of the StyleGAN2 latent space.

Algorithm 1 Training the Projector P

Require: Training set $\mathcal{D}$. Encoder E, StyleGAN2 generator G_1, Recognizer R. Mean latent w_{avg}. Weights $\boldsymbol{\lambda}$. Margin m. Learning rate η. Epochs N.
Ensure: Trained Projector P

```
Initialize Projector parameters θ_P
for epoch = 1, ..., N do
    for all mini-batches in D do
        Given identities x and y, sample x_1, x_2 and y_1
        Sample binary keys k_1, k_2
        S ← {(1,1), (1,2), (2,1)}
        x'_ab ← G(x_a, k_b) for (a, b) ∈ S
        x'_{y_1} ← G(y_1, k_1)
        f_ab ← R(x'_ab) for (a, b) ∈ S
        f_{y_1} ← R(x'_{y_1})
        f_1 ← R(x_1)
        L_ano ← L_cos(f_11, f_1, −1, m)
        L_con ← L_cos(f_11, f_21, +1, m)
        L_div ← L_cos(f_11, f_12, −1, m)
        L_dif ← L_cos(f_11, f_{y_1}, −1, m)
        L_reg ← ‖P(E(x_1), k_1)‖_2^2
        L_tot ← λ_ano L_ano + λ_con L_con + λ_div L_div + λ_dif L_dif + λ_reg L_reg
        θ_P ← θ_P − η ∇_{θ_P} L_tot
    end for
end for
return P
```

These five components are combined into a single multi-task objective, where each weight λ_i controls the relative importance of its corresponding term,

$$\begin{aligned} L_{tot} = \lambda_{\text{ano}} L_{\text{ano}} + \lambda_{\text{con}} L_{\text{con}} + \lambda_{\text{div}} L_{\text{div}} \\ + \lambda_{\text{dif}} L_{\text{dif}} + \lambda_{\text{reg}} L_{\text{reg}}. \end{aligned} \tag{17}$$

Algorithm 1 summarizes the full training procedure. Each mini-batch draws two samples from one identity and one sample from another identity, runs them through the Encoder, Projector, and StyleGAN2 generator, and updates only the Projector's parameters from the total loss L_{tot}.

4 Experimental Results

The proposed framework is evaluated by establishing the shared experimental setup, followed by three primary experiments assessing privacy, action recognition, and emotion recognition. Then two ablation studies quantify two specific design choices for the Projector's architecture.

4.1 Experimental Settings

Training Datasets Two datasets were used during training. CelebA [41], consisting of 202,599 facial images of 10,177 unique identities, was used to train the SKPG backbone. CelebA-HQ [42], a high-quality subset of CelebA, was used to train the StyleGAN2 model from scratch. To ensure that the generated pseudonyms are clearly visible without biometric-irrelevant features, CelebA-HQ was filtered to exclude images containing accessories.

Training Configuration The framework incorporates several core architectures. The Encoder uses FaceNet (Inception-ResNet-v1) [43] pre-trained on the VGGFace2 dataset [44] to extract identity embeddings. The StyleGAN2 generator uses the official PyTorch ADA release [45], trained from scratch on the filtered CelebA-HQ dataset. The Face-Adapter [22] generator uses the Stable Diffusion v1-5 base model [46] with the official Face-Adapter checkpoints on HuggingFace.

Training was conducted on a single GPU with at least 12GB of VRAM. The Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, the learning rate of 3×10^{-4}, and the Cosine Embedding Loss margin of 0.4 were all adapted from IVFG to avoid computationally prohibitive hyperparameter tuning. The Projector's Multi-Layer Perceptron (MLP) was configured with hidden dimensions of (1024, 512) and trained using an 80/20 training-validation split over 10 epochs. The binary cryptographic key had a dimensionality of 128 to provide a sufficiently large key space . Finally, the loss weights were set to $\lambda_{\text{ano}} = 0.1$, $\lambda_{\text{con}} = 1$, $\lambda_{\text{div}} = 1$, $\lambda_{\text{dif}} = 1.25$, and $\lambda_{\text{reg}} = 1.5$.

The low λ_{ano} value matches the IVFG configuration. Generated pseudonyms inherently differ from the real identity distribution and easily satisfy anonymity. Therefore, heavy optimization on this objective is unnecessary. Furthermore, λ_{dif} was weighted heavier than the consistency and diversity parameters because the model initializes the Projector's output to w_{avg}, resulting in similar generated pseudonyms for every subject, requiring a stronger penalty to successfully differentiate them.

The complete training configuration was shared across Experiments 1, 2, and 3. The two ablations reuse this configuration and only vary the Projector's architecture by disabling input normalization and not anchoring to w_{avg} to isolate each design choice.

Rendering Strategies Under Evaluation The proposed SKPG backbone builds upon the subject- and key-conditioned generation principles introduced by IVFG [37] and KFAAR [38]. To evaluate how different downstream rendering strategies affect the final pseudonymized video, the SKPG backbone was kept fixed while only the rendering strategy varied. This isolates the impact of the rendering step and establishes a direct comparison between the proposed swap-based approach and the methodologies of these existing frameworks.

The rendering strategies were implemented as follows:

- **SKPG-BB (Inspired by IVFG):** IVFG generates a standalone pseudonym but does not specify a mechanism for integrating it back into the original video frame. To evaluate this approach on video data, SKPG-BB adapts the method by applying a naive bounding-box overwrite. The face region from the SKPG output is cropped and pasted directly over the bounding box of the subject's face.
- **SKPG-Reenact (Inspired by KFAAR):** KFAAR improves upon static generation by using face reenactment to align the generated head pose and expression with the subject before overwriting the region. Because the specific *FaceVid2Vid* model [47] used in KFAAR is private, SKPG-Reenact replicates this methodology by using Face-Adapter in its *face-reenactment* mode. This step inherently aligns the generated pseudonym's pose with the target frame, eliminating the need for manual cropping before pasting it onto the subject's face.
- **SKPG-Swap (Ours):** Represents the proposed context-preserving hybrid framework. This strategy uses diffusion-based face swapping via Face-Adapter operating in *face-swap* mode. Rather than overwriting the region with a cropped image, it takes the complete, uncropped pseudonym generated by the SKPG backbone and blends its identity into the bounding box of the subject's face.

Alongside SKPG-BB and SKPG-Reenact, the proposed framework was also compared to DeepPrivacy2 [3]. DeepPrivacy2 is database-free, preserves head pose and environment, and is cross-video consistent under similar conditions in different videos. These properties make the method a relevant external baseline. The SKPG-BB/Reenact variants isolate the rendering choice, while DeepPrivacy2 represents a strong prior method that shares the database-free constraint.

4.2 Experiment 1: Privacy and Identity Preservation

The first experiment answers the question: Does the proposed framework anonymize subjects while preserving consistency, diversity, differentiation, and context?

Dataset The experiment used VoxCeleb2 [2], a dataset of celebrity videos with associated identity labels, to test the framework on real-world video footage. All clips were processed to 5 fps. The dataset was filtered to 400 identities with 2 samples per identity to save computational cost.

Pre-trained Models and Configuration Identity embeddings for the calculation of Re-identification (Re-ID) metrics were extracted using the ArcFace model [48] provided by the InsightFace Buffalo_L package. The proposed framework was compared against DeepPrivacy2 and the SKPG-BB and SKPG-Reenact baselines.

Metrics The following Re-ID metrics quantify four properties of the framework using the Area Under the Curve (AUC) and Equal Error Rate (EER) of the cosine similarities between ArcFace embeddings. Regardless of whether a metric aims for a high or low similarity score, the evaluation is structured such that a higher AUC and a lower EER always indicate better results. Three of the metrics target a low cosine similarity to ensure identities are distinctly separated. Anonymization (Anon) compares the subject with its pseudonym, Diversity (Div) compares pseudonyms generated from the same subject using different keys, and Differentiation (Dif) compares pseudonyms from different subjects under the same key. In contrast, Consistency (Con) targets a high cosine similarity, as it compares pseudonyms generated from different video samples of the exact same subject using the same key.

Context preservation was measured through three complementary metrics. Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) jointly quantify whether the environment remains intact, where higher SSIM and lower LPIPS indicate better preservation. Landmark Distance penalizes frameworks that fail to preserve the head pose and expressions of the subject.

To ensure every metric could be calculated, the evaluation was performed using a batch-based setup. Each batch consisted of two distinct identities to calculate differentiation, two sample videos per identity to calculate consistency, and two different cryptographic keys to calculate diversity.

Results Table 2 presents the context preservation metrics. SKPG-Swap (Ours) achieved the lowest landmark distance (0.050), and SSIM and LPIPS scores close to DeepPrivacy2, indicating that it preserves context comparably to DeepPrivacy2. While SKPG-Reenact improved landmark distance compared to SKPG-BB due to its face-reenactment step, both rendering strategies were substantially worse in preserving context than SKPG-Swap. SKPG-Swap is therefore the best SKPG variant at preserving context.

Table 3 shows the anonymization results. The proposed framework reached an Anon AUC of 78.4, ranking below the bounding-box rendering strategies and above DeepPrivacy2. Notably, SKPG-BB achieved a near-perfect score (AUC 99.9), effectively eliminating any link back to the subject. SKPG-Reenact ranks just below SKPG-BB. The proposed framework therefore anonymizes less strongly than the bounding-box strategies, but more strongly than DeepPrivacy2.

Table 4 reports consistency, both within and across videos. For within-video consistency, SKPG-Swap (Ours) demonstrated competitive results (AUC 92.1), scoring similarly to both SKPG-BB and SKPG-Reenact. For cross-video consistency, SKPG-BB maintained the highest consistency, followed by SKPG-Reenact. In comparison, both SKPG-Swap (AUC 64.9) and DeepPrivacy2 experienced a larger drop than the bounding-box strategies across separate videos, scoring similarly to each other. SKPG-Swap therefore preserves pseudonym consistency within a video but is less consistent across separate videos.

Finally, Table 5 reports the diversity and differentiation metrics. SKPG-Swap matched DeepPrivacy2 on differentiation (AUC 81.4) and outscored the bounding-box rendering strategies by a small margin. In terms of diversity, while SKPG-Swap demonstrates moderate key-driven variation (AUC 74.8), it does not achieve the same level of distinctness as the bounding-box rendering strategies SKPG-BB (AUC 99.8) and SKPG-Reenact.

In summary, the proposed framework successfully preserves strong context, differentiation, and diversity, while achieving moderate anonymization. Furthermore, it maintains reliable within-video consistency, although it does not provide full cross-video consistency.

Table 2. Context preservation metrics on VoxCeleb2. SKPG-Swap (Ours) achieved the lowest landmark distance, and an SSIM and LPIPS close to DeepPrivacy2, substantially outperforming the bounding-box rendering strategies SKPG-BB and SKPG-Reenact on context preservation.

Method	SSIM ↑	LPIPS ↓	Landmark Dist ↓
DeepPrivacy2	81.2	9.8	0.053
SKPG-BB	53.7	40.9	0.113
SKPG-Reenact	54.5	45.8	0.064
SKPG-Swap (Ours)	80.0	10.9	0.050

Table 3. Anonymization results on VoxCeleb2. SKPG-Swap (Ours) reached an Anon AUC of 78.4, ranking above DeepPrivacy2 and below the bounding-box rendering strategies, with SKPG-BB achieving near-perfect anonymization (AUC 99.9).

Method	Anon AUC ↑	Anon EER ↓
DeepPrivacy2	62.5	41.8
SKPG-BB	99.9	0.2
SKPG-Reenact	96.7	8.6
SKPG-Swap (Ours)	78.4	26.3

Table 4. Consistency results on VoxCeleb2, measured both within and across videos. Original indicates the consistency scores on the unmodified videos. SKPG-Swap (Ours) matched the bounding-box strategies on within-video consistency (AUC 92.1) but dropped on cross-video consistency (AUC 64.9).

Method	Within AUC ↑	Within EER ↓	Cross AUC ↑	Cross EER ↓
Original	96.1	5.0	88.0	16.4
DeepPrivacy2	96.0	9.8	65.4	38.7
SKPG-BB	93.1	14.7	86.6	21.6
SKPG-Reenact	92.7	14.5	77.0	27.9
SKPG-Swap (Ours)	92.1	14.3	64.9	38.5

Table 5. Diversity and Differentiation results on VoxCeleb2. Original shows scores on the unmodified videos. The N/A values in the original and DeepPrivacy2 rows indicate that the diversity score cannot be calculated as their resulting videos are not keyed. SKPG-Swap (Ours) achieved the highest differentiation among the pseudonymization methods while matching DeepPrivacy2, but a lower diversity score (AUC 74.8) than SKPG-BB (AUC 99.8).

Method	Div AUC ↑	Div EER ↓	Dif AUC ↑	Dif EER ↓
Original	N/A	N/A	95.9	9.9
DeepPrivacy2	N/A	N/A	81.2	26.8
SKPG-BB	99.8	0.9	79.1	28.5
SKPG-Reenact	91.9	13.8	76.8	29.9
SKPG-Swap (Ours)	74.8	29.3	81.4	26.2

4.3 Experiment 2: Utility Preservation Through Action Recognition

The second experiment answers the question: How does the proposed framework impact action recognition on face-dependent tasks?

Dataset The experiment used UCF101 [49], a human action recognition dataset consisting of realistic action videos. All clips were processed to 5 fps. To ensure an unbiased evaluation, only the official testing split was used, as the evaluation model that was used is trained on the UCF101 training split. The dataset was filtered to include only five classes, namely *ApplyEyeMakeup*, *ApplyLipstick*, *BrushingTeeth*, *PlayingFlute*, and *ShavingBeard*, leaving 92 test videos. This selection is a deliberately challenging subset because these classes depend on head pose, expressions, and how objects interact with the face. Therefore, maintaining high accuracy on this subset demonstrates that pseudonymization preserves the context necessary for face-dependent action recognition.

Pre-trained Models and Configuration To evaluate the downstream utility of the pseudonymized videos, this experiment used an action recognition model based on the VideoMAE architecture [50]. The *nateraw/videomae-base-finetuned-ucf101* model was used, implemented via Hugging Face. This model was fine-tuned on the 101 action classes of the UCF101 dataset, ensuring the model is familiar with the chosen subset while having a broader range of possible predicted classes to capture misclassifications.

Metrics Utility preservation was quantified through three metrics computed on the predictions of the fine-tuned VideoMAE model. Classification Accuracy is the ratio of videos correctly classified into their original action categories after the pseudonymization process. The Prediction Agreement Rate complements accuracy by measuring how often the model assigns the pseudonymized video to the same class it predicted on the corresponding unmodified video, regardless of whether that prediction is correct. A high agreement rate indicates that the method preserves the downstream behavior of the model, not only the fraction of clips that land on the correct label. Because the downstream classifier evaluates the exact same clips across all methods, the evaluation uses the exact two-sided McNemar test [51] to determine whether the resulting differences in accuracy compared to the unmodified baseline are statistically significant. Appendix A, Table A.1 reports the discordant clip counts, the McNemar χ^2, and the exact p-value for every method. Per-class confusion matrices in Appendix B complement these aggregate metrics by showing which classes each method correctly classifies and which it fails on.

Results Table 6 presents the action recognition results, comparing SKPG-Swap against DeepPrivacy2, SKPG-BB, and SKPG-Reenact, alongside the baseline results on the unmodified videos. The unmodified videos established an upper bound for classification accuracy (91.3). SKPG-Swap retained nearly all of the analytical utility (90.2), scored similarly to DeepPrivacy2, and outscored the two bounding-box rendering strategies, SKPG-BB and SKPG-Reenact. The face-reenactment component in SKPG-Reenact provided no measurable advantage over SKPG-BB in classification accuracy. McNemar's exact

test on the correct/incorrect outcomes confirms that the accuracy gap to the original baseline is not statistically significant for SKPG-Swap ($p = 1.0$) or DeepPrivacy2, while SKPG-BB and SKPG-Reenact are significantly worse than the unmodified baseline. Table A.1 lists all per-method significance scores. In terms of prediction agreement with the unmodified baseline, SKPG-Swap and DeepPrivacy2 each match predictions on a large fraction of clips (94.6), while SKPG-BB and SKPG-Reenact drop substantially. This trend directly supports the established ranking in classification accuracy.

The per-class confusion matrices in Appendix B expose the source of the accuracy gap. Both bounding-box renderers fail on the two classes that depend on facial detail, *ApplyEyeMakeup* and *ApplyLipstick*, and frequently misclassify them as unrelated classes such as *BlowDryHair* under SKPG-BB and *Haircut* under SKPG-Reenact. The three classes that involve face-interacting objects without depending on facial detail, *BrushingTeeth*, *PlayingFlute*, and *ShavingBeard*, stay high under every method. SKPG-Swap matches the unmodified baseline closely on every class, including the two detail-dependent ones.

Overall, the proposed framework preserves strong utility for face-dependent action recognition. Unlike bounding-box methods, it maintains high accuracy and prediction agreement even on tasks requiring facial details, matching the unmodified baseline with no statistically significant drop.

4.4 Experiment 3: Cross-Video Utility Through Emotion Recognition

This experiment evaluates the framework on facial emotion recognition. This is expected to be a harder setting than action recognition as modifying the face can distort the small visual cues classifiers rely on. The experiment is organized around three questions:

1. How does each pseudonymization framework impact per-clip emotion recognition?
2. How much does the prediction change when only the SKPG generated pseudonym is varied, with the original frame held fixed?
3. Does a consistent pseudonym identity produce more stable predictions across separate videos than an inconsistent one?

Dataset and Pre-trained Models The experiment used RAVDESS [1], the Ryerson Audio-Visual Database of Emotional Speech and Song. The Emo-AffectNet model [52] was used for emotion classification. The specific checkpoint used for this model excluded RAVDESS from training such that no clip in the evaluation set had been seen by the pre-trained model. Emo-AffectNet recognizes seven emotions: *Neutral*, *Happiness*, *Sadness*, *Anger*, *Fear*, *Disgust*, and *Surprise*. The *Calm* emotion present in the RAVDESS dataset was excluded as the classification model does not recognize it. From the remaining clips, the first statement was selected for all 24 actors, all intensity levels, and both repetitions, resulting in 624 evaluation clips. All clips were processed at 10 fps. A higher fps compared to the videos in the previous experiments was chosen because emotion recognition depends on more temporal and structural detail.

Metrics To answer the first question, the evaluation measures whether each pseudonymized clip is classified correctly in isolation. The **Unweighted Average Recall (UAR)** reports the mean per-class recall and matches the primary metric used in the original Emo-AffectNet paper. A high UAR indicates that the classifier remains accurate on the pseudonymized clips. The **Prediction Agreement Rate** complements the UAR by measuring how often the classifier assigns the pseudonymized clip to the same class as the prediction on the corresponding unmodified clip, regardless of whether that prediction is correct. A high agreement rate indicates that the method preserves the per-clip behavior of the downstream classifier rather than only the fraction of correctly classified clips. Per-class confusion matrices in Appendix E complement both metrics by showing which emotions each method classifies correctly and which it confuses. Because the downstream classifier scores the same clips across all methods, the evaluation applies the two-sided exact McNemar test [51] to the per-clip correct/incorrect outcomes to test whether each method classifies significantly fewer clips correctly than the unmodified baseline, with the per-method scores reported in Appendix D, Table A.2.

To answer the second question, the evaluation quantifies the prediction variance directly. The experiment ran a controlled comparison with two cryptographic keys. For every original clip i in the evaluation set, two pseudonymized versions were generated, $V_{i,A}$ using Key A and $V_{i,B}$ using Key B. Since the original frame was locked across $V_{i,A}$ and $V_{i,B}$, the pseudonym blended into the frame was the only intended difference between the two versions. The **Label Flip Rate** compares these two versions and reports the fraction of the N evaluation clips for which the predicted label $\hat{y}$ changes between the two keys, where $\mathbb{1}(\cdot)$ is the indicator function that returns 1 when the predictions do not match and 0 otherwise,

$$\text{Flip Rate} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_{i,A} \neq \hat{y}_{i,B}). \tag{18}$$

A high flip rate shows that changing the pseudonym alone can push the prediction across the decision boundary between emotions. One caveat is that a different key gives the same subject a different pseudonym identity, but the pseudonyms can differ in more than identity alone. They

Table 6. Utility Preservation on UCF101 Action Recognition. Original indicates results on the unmodified videos as reference. SKPG-Swap (Ours) matches the unmodified baseline's prediction on 94.6 of clips, the same rate as DeepPrivacy2, and McNemar's exact test finds no significant difference in accuracy from the unmodified baseline ($p = 1.0$). The two bounding-box rendering strategies drop substantially and are significantly worse than the unmodified baseline ($p < 10^{-5}$).

Method	Classification Accuracy ↑	Pred. Agreement ↑
Original	91.3	N/A
DeepPrivacy2	88.0	94.6
SKPG-BB	68.5	71.7
SKPG-Reenact	68.5	71.7
SKPG-Swap (Ours)	90.2	94.6

can also vary in visual context that is not controlled, like pose, expression, and environment. Therefore, any change in prediction when the key changes cannot be attributed to the change in pseudonym identity alone. Figure A.6 in Appendix C shows a schematic example of a prediction flip caused by changing the key and also illustrates how a pseudonym's context can differ in addition to its identity.

Finally, to answer the third question, two metrics test whether a consistent pseudonym identity produces more stable predictions when the same subject is recorded across separate videos. The setting mimics longitudinal tracking by treating pairs of clips that share the same subject and the same target emotion as repeated videos of the same subject. The pairs were collected under the same studio conditions, but they still capture the multi-video structure that longitudinal analysis has to handle. The **Pairwise Agreement Rate** is the fraction of the M clip pairs for which the classifier predicts the same emotion category on both clips, where $\hat{y}_{\text{clip 1}}$ and $\hat{y}_{\text{clip 2}}$ are the predicted labels on the two clips of a pair, and $\mathbb{1}(\cdot)$ is the indicator function that returns 1 if the predictions match and 0 otherwise,

$$\text{Agreement} = \frac{1}{M} \sum_{\text{pairs}} \mathbb{1}(\hat{y}_{\text{clip 1}} = \hat{y}_{\text{clip 2}}). \qquad (19)$$

The metric captures stability of the predicted label across the pair regardless of whether that label is correct. Per-clip classification accuracy was not used because the *same-key* and *different-key* conditions can reach the same per-clip accuracy while still differing sharply in how often two clips in a pair receive the same prediction, hiding the within-pair stability that longitudinal tracking depends on. The **Conditional Accuracy** is the probability that the classifier correctly classifies the second clip given that it correctly classified the first one, where Correct_1 and Correct_2 denote correct classification on the first and second clip of a pair and $\text{Count}(\cdot)$ counts the pairs satisfying its condition,

$$P(\text{Correct}_2 \mid \text{Correct}_1) = \frac{\text{Count(Both Correct)}}{\text{Count(Clip 1 Correct)}}. \qquad (20)$$

The conditional accuracy removes the penalty of the classifier's baseline failure rate and therefore shows stability on already correctly classified clips.

Both metrics are reported under two conditions. The *same-key* condition assigned the same key to both clips in a pair, simulating ideal consistent pseudonymization. The *different-key* condition assigned different keys to the two clips, simulating a subject receiving two different pseudonym identities across videos. The *different-key* condition was evaluated in both orderings of every pair, such that each key appeared once as the first clip and once as the second. Figure A.7 in Appendix C gives a schematic worked example of the two conditions and how they can have a different output. This schematic also illustrates how a pseudonym's context can differ even when its identity is fixed in the *same-key* condition. A wider gap between the two conditions indicates a larger longitudinal instability caused by inconsistent pseudonym identities.

Table 7 verifies that the same key gives a consistent pseudonym identity and different keys give inconsistent ones on the subjects in the RAVDESS dataset, using the consistency and diversity metrics from Experiment 1. Every method exceeded 99.0 on all metrics. The gaps in Tables 9 and 10 therefore reflect a change in pseudonym identity when changing the key.

Results Table 8 answers the first question on per-clip emotion recognition utility. Among the pseudonymization methods, the proposed framework achieved the highest UAR (24.0), outperforming SKPG-Reenact, SKPG-BB, and DeepPrivacy2. Prediction Agreement follows the same ordering, with SKPG-Swap matching the unmodified baseline on the largest fraction of clips and DeepPrivacy2 the smallest. Nevertheless, agreement is low for every method, consistent with the sharp UAR drop. No method preserves the per-clip behavior of the unmodified baseline on this task, and the column is meaningful here only as a relative ranking rather than as evidence of preserved behavior. DeepPrivacy2 sits lowest because it collapses every clip onto *Happiness* and coincides with the unmodified baseline only where the unmodified baseline already predicted *Happiness*. However, the ranking matches the action recognition result from Experiment 2 and confirms that the context preserved by face swapping benefits per-clip downstream emotion classification

Table 7. Validation of the *same-key* and *different-key* interpretation on RAVDESS. Within-video consistency, cross-video consistency, and diversity AUCs are computed exactly as in Experiment 1. Every evaluated method exceeded 99.0 on all three, confirming that the same key maps to a consistent pseudonym identity and different keys map to inconsistent pseudonym identities at near-perfect rates on this dataset.

Method	Within Con AUC ↑	Cross Con AUC ↑	Div AUC ↑
SKPG-BB	99.9	99.9	99.9
SKPG-Reenact	99.4	99.4	99.4
SKPG-Swap (Ours)	99.8	99.7	99.7

relative to the other methods. The McNemar test on the per-clip outcomes confirms that every pseudonymization method, including SKPG-Swap, classifies significantly fewer clips correctly than the unmodified baseline, as detailed in Appendix D, Table A.2.

Table 8. Per-clip emotion recognition on RAVDESS. SKPG-Swap (Ours) reached the highest UAR (24.0) and the highest Prediction Agreement with the unmodified baseline among the pseudonymization methods. Agreement is low for every method, consistent with the sharp UAR drop, so the column ranks the methods by how little they diverge from the unmodified baseline rather than showing that any method preserves its per-clip behavior. SKPG-Swap diverges least, confirming that the context preserved by face swapping benefits per-clip emotion classification relative to the alternatives.

Method	UAR ↑	Pred. Agreement ↑
Original	53.6	100.0
DeepPrivacy2	14.3	16.3
SKPG-BB	14.6	20.8
SKPG-Reenact	20.7	21.9
SKPG-Swap (Ours)	24.0	23.8

The per-class confusion matrices in Appendix E show where each method succeeds and fails per class. DeepPrivacy2 predicts *Happiness* for every clip, including clips whose true emotion is *Neutral*, which the unmodified baseline recognized perfectly. SKPG-BB only ever predicts *Happiness* or *Neutral*, never the other five classes. SKPG-Reenact predicts *Neutral* for most clips but classifies *Happiness* correctly on under half of its clips. SKPG-Swap classifies *Happiness* correctly on the majority of its clips and also produces correct predictions on *Anger*, *Sadness*, and *Surprise*, which explains its higher UAR. Even SKPG-Reenact and SKPG-Swap, the two methods that preserve facial expression, recover only *Happiness* and *Neutral* with any reliability. These are the two classes the unmodified baseline already recognizes most easily. Harder emotions such as *Anger*, *Fear*, and *Surprise* remain difficult after pseudonymization.

Table 9 answers the second question. Changing the pseudonym while holding the original clip fixed produced a substantial label flip rate across all methods. SKPG-BB recorded the highest flip rate, consistent with its bounding-box overwrite that generates an entire new face without aligning the subject's expression. SKPG-Reenact recorded the lowest flip rate and the proposed framework ranked between them (38.1). Therefore, because of the high flip rates, a varying pseudonym is a measurable trigger of prediction variance.

Table 9. Label flip rate caused by changing pseudonyms on RAVDESS. This metric compares predictions on $V_{i,A}$ and $V_{i,B}$, which share the same clip and differ only in the pseudonym. SKPG-BB recorded the highest flip rate, SKPG-Reenact the lowest, and the proposed framework sat between them. These results indicate that changing the pseudonym can introduce variance in predictions

Method	Label Flip Rate ↓
SKPG-BB	72.0
SKPG-Reenact	28.7
SKPG-Swap (Ours)	38.1

Table 10 answers the third question by testing whether a consistent pseudonym identity reduces this variance across pairs of videos. For every evaluated method, both Pairwise Agreement and Conditional Accuracy were substantially higher in the *same-key* condition than in the *different-key* condition. The proposed framework reached higher *same-key* scores than its *different-key* condition on both Pairwise Agreement and Conditional Accuracy (90.5 vs. 53.6 on Conditional Accuracy), and SKPG-BB and SKPG-Reenact followed the same pattern. Across every method, holding the pseudonym identity consistent across the pair (*same-key*) produced more stable predictions than letting it change (*different-key*), demonstrating in this controlled setting that consistent pseudonymization reduces prediction variance across videos.

4.5 Experiment 4: Ablation of Input Preprocessing in the Projector

The fourth experiment answers the question: How does the input preprocessor in the Projector affect its capacity to optimize the learning objectives?

Section 3.1 argues that the magnitude mismatch between the unnormalized binary key k and the L2-normalized latent vector z complicates the optimization of the learning objectives as the Projector's MLP has to learn to balance the inputs. This ablation tests this

Table 10. Pairwise agreement rate and conditional accuracy on RAVDESS on every pair of clips that share the same subject and target emotion. The *different-key* condition is evaluated in both orderings of every pair. For every method, the *same-key* condition resulted in higher agreement and conditional accuracy, indicating that a consistent pseudonym identity gives more stable predictions across videos.

	Pairwise Agreement Rate ↑		**Conditional Accuracy ↑**	
Method	*same-key*	*different-key*	*same-key*	*different-key*
SKPG-BB	97.4	27.5	99.2	14.0
SKPG-Reenact	90.9	70.7	85.9	44.2
SKPG-Swap (Ours)	85.9	60.3	90.5	53.6

argument by training a second Projector without the input preprocessor and comparing its behavior to that of the Projector used in Experiment 1.

Configurations Two Projectors were trained under settings identical to Experiment 1. They differ only in the preprocessing steps applied before concatenation.

- **Raw Projector**. z and k are concatenated without L2 normalization and without rescaling of k.
- **Full Projector**. Both z and k are L2-normalized, and k is additionally rescaled to a magnitude comparable to z through the learned projection, as described in Section 3.1. This is the same trained Projector evaluated in Experiment 1.

Each Projector was then evaluated through the SKPG-BB, SKPG-Reenact, and SKPG-Swap rendering strategies, resulting in six evaluations. Evaluations used the same VoxCeleb2 dataset from Experiment 1.

Metrics The ablation reports five metrics from Experiment 1 that together measure how well the Projector optimizes its learning objectives: Anonymization (Anon), Diversity (Div), Differentiation (Dif), within video Consistency (Within Con) and Consistency across videos (Cross Con). A well-balanced Projector optimizes all five properties and achieves high Area Under the Curve (AUC) scores on all of them.

Results Table 11 shows the results for the ablation of the input preprocessor. Adding the preprocessing steps improves every metric across all three rendering strategies, with the single exception of the already-saturated Anon score for SKPG-BB. The consistency of this direction indicates that the magnitude mismatch does not only restrict the Projector's use of the key but broadly degrades its capacity to optimize the learning objectives. Without the preprocessor, the downstream MLP must resolve the scaling differences internally while simultaneously learning the multi-objective mappings.

The face-swap strategy amplifies this improvement. On SKPG-Swap, adding the preprocessor causes substantial rises in cross-video consistency (+6.2 AUC) and key-driven diversity (+7.8 AUC), alongside smaller increases in anonymization and within-video consistency. SKPG-Reenact shows a similar increase in cross-video consistency and diversity.

The preprocessor is therefore an important architectural choice. By resolving the dimensional and scale mismatches using L2-normalization and a dedicated key projection layer, the downstream MLP better optimizes the Projector's learning objectives.

4.6 Experiment 5: Ablation of the w_{avg} Anchor

The fifth experiment answers the question: Is the w_{avg} anchor necessary, given that the regularization loss already pulls the latent vector w toward w_{avg}?

The anchor introduced in Section 3.1 adds the Projector's output z' to the StyleGAN2 average latent vector w_{avg} before mapping it to W^+. This addition initializes the vector inside a realistic region of the StyleGAN2 latent space W, ensuring the Projector can focus on learning small identity-driven offsets in this region. The regularization loss (Eq. (16)) pulls w towards w_{avg} during training, which makes the anchor look redundant. However, the anchor and the regularization loss act at different moments: the anchor fixes the initialization, while the regularization loss applies a soft penalty during optimization. Removing the anchor therefore leaves the training objective unchanged and instead shifts the initialization to the unanchored Projector output. This ablation tests whether the regularization loss alone can recover realistic pseudonyms from this unanchored state, and how removing the anchor affects the optimization of the learning objectives.

Configurations Two Projectors were trained under settings identical to Experiment 1, differing only in how the latent vector w is constructed before replication to W^+.

- **No-Anchor**: the latent w is the projected vector, $w = z'$.
- **Full**: the latent w is the addition of the projected vector to the anchor, $w = w_{\text{avg}} + z'$. This is the same Projector evaluated in Experiment 1.

Table 11. Ablation of the input preprocessor in the Projector on VoxCeleb2. Parentheses next to the *Full Projector* scores show the gain over the *Raw Projector*. The *Full Projector* scores higher than the *Raw Projector* on all five Re-ID metrics for every rendering strategy, with the sole exception of the already-saturated anonymization score for SKPG-BB. The preprocessor therefore consistently improves the Projector's ability to satisfy all learning objectives, with the strongest benefit on key-driven diversity and cross-video consistency under SKPG-Swap.

Method	Anon AUC ↑	Within Con AUC ↑	Cross Con AUC ↑	Div AUC ↑	Dif AUC ↑
Raw Projector (no preprocessor, ablated)					
SKPG-BB	99.9	92.3	84.6	99.2	76.9
SKPG-Reenact	94.9	90.5	70.7	84.4	75.5
SKPG-Swap	74.4	89.8	58.7	67.0	80.7
Full Projector (L2 normalization and key rescaling)					
SKPG-BB	99.9 (+0.0)	93.1 (+0.8)	86.6 (+2.0)	99.8 (+0.6)	79.1 (+2.2)
SKPG-Reenact	96.7 (+1.8)	92.7 (+2.2)	77.0 (+6.3)	91.9 (+7.5)	76.8 (+1.3)
SKPG-Swap	78.4 (+4.0)	92.1 (+2.3)	64.9 (+6.2)	74.8 (+7.8)	81.4 (+0.7)

Both Projectors were evaluated through the SKPG-BB, SKPG-Reenact, and SKPG-Swap rendering strategies on the VoxCeleb2 dataset from Experiment 1.

Metrics The ablation reports pseudonym realism qualitatively in Figure 3 and learning behavior in the loss curves of Figure 4. The five Re-ID metrics from Experiment 1 measure how well the learning objectives are optimized: Anonymization (Anon), Diversity (Div), Differentiation (Dif), within-video Consistency (Within Con), and cross-video Consistency (Cross Con), each reported as AUC across all three rendering strategies.

Results The generated pseudonyms in Figure 3 demonstrate that the anchor primarily provides a realistic initialization. The *Full* Projector produces recognizable pseudonyms from the first epoch, whereas the *No-Anchor* Projector produces faces that are unrealistic and barely recognizable. Within the same training capacity, the *No-Anchor* variant never fully recovers, and the generated pseudonyms still show visual distortions at epoch 10.

The loss curves in Figure 4 confirm the initialization argument quantitatively. The regularization loss stays low for the *Full* configuration from the first epoch, while for the *No-Anchor* variant, it starts much higher and never closes the gap. The *Full* configuration also holds a lower re-identification (Re-ID) loss through the early epochs, although the Re-ID losses of both configurations eventually converge. For context, this Re-ID loss is the weighted sum of the anonymization, diversity, consistency, and differentiation losses used to optimize the re-identification metrics. While applying a higher regularization weight (λ_{reg}) could theoretically force a lower regularization loss for the *No-Anchor* variant, doing so would likely make the identity objectives harder to optimize, as the regularization objective would become more dominant. Therefore, the initialization provided by the w_{avg} anchor makes it easier to optimize the regularization objective, while maintaining at least as much capacity to optimize the other identity objectives.

The Re-ID metrics in Table 12 show that the anchor's downstream value depends on the rendering strategy. The addition of the anchor hardly changes the scores for SKPG-BB, which exposes the SKPG output directly. Therefore, the initial structural distortions barely affect the pseudonym's Re-ID scores directly. Instead, these distortions become costly only after Face-Adapter blends the faces back into the original frames. Adding the anchor substantially increases cross-video consistency and diversity for SKPG-Reenact. SKPG-Swap amplifies these gains even further for these two metrics (up to +14.4 AUC on diversity), while additionally achieving an increase in anonymization (+11.1 AUC). Both strategies render the output using Face-Adapter, which struggles to integrate a distorted pseudonym. The anchor therefore matters at the rendering step, not in the raw SKPG output.

In conclusion, the w_{avg} anchor remains necessary despite the regularization loss, because its realistic initialization ensures the structural realism of the generated pseudonyms. This directly translates into downstream improvements when those faces are processed through advanced face-reenactment or face-swapping rendering steps like in SKPG-Swap and SKPG-Reenact.

5 Discussion

This thesis evaluated three rendering strategies on the shared SKPG backbone, plus an external baseline, against Re-ID, context preservation, and downstream-utility metrics. Some useful insights emerged from training and evaluation.

Identity leakage in SKPG-Swap A single mechanism explains the Re-ID metrics for SKPG-Swap: because it blends the subject into the output, part of the subject's identity remains in the pseudonymized frame. This leakage moves the pseudonym closer to the subject's original face, lowering anonymization (Anon AUC 78.4 against 99.9 for SKPG-BB). It pulls every pseudonym of a subject toward that subject's identity regardless of the key,

Table 12. Ablation of the w_{avg} anchor on VoxCeleb2. Re-ID metrics are reported as AUC for each rendering strategy. Parentheses next to the *Full* scores show the gain over *No-Anchor*. Removing the anchor changes the Re-ID metrics by at most 1.3 percentage points on the bounding-box SKPG-BB strategy, but substantially degrades them on SKPG-Reenact and SKPG-Swap. Both of these run the pseudonym through an extra rendering step, and that step is where the anchor's downstream value emerges rather than in the SKPG generated pseudonym.

Method	Anon AUC ↑	Within Con AUC ↑	Cross Con AUC ↑	Div AUC ↑	Dif AUC ↑
No-Anchor (ablated)					
SKPG-BB	100.0	93.8	86.3	99.1	80.4
SKPG-Reenact	95.2	90.9	69.3	82.1	76.8
SKPG-Swap	67.3	88.0	55.3	60.4	82.3
Full (with w_{avg} anchor)					
SKPG-BB	99.9 (-0.1)	93.1 (-0.7)	86.6 (+0.3)	99.8 (+0.7)	79.1 (-1.3)
SKPG-Reenact	96.7 (+1.5)	92.7 (+1.8)	77.0 (+7.7)	91.9 (+9.8)	76.8 (+0.0)
SKPG-Swap	78.4 (+11.1)	92.1 (+4.1)	64.9 (+9.6)	74.8 (+14.4)	81.4 (-0.9)

lowering diversity (Div AUC 74.8 against 99.8). Because each output retains part of the subject's identity, two different subjects stay separable, raising differentiation (Dif AUC 81.4 against 79.1). Diversity and differentiation moving in opposite directions is the signature of leaked identity rather than noise added by the rendering step. A general loss of pseudonym quality would lower both at once.

The drop in cross-video consistency for SKPG-Swap (Cross Con AUC 64.9) has two separate causes. The first is the leakage described above. Not only the subject's identity is leaked by the blend but also video-specific appearance such as lighting and head pose. This video-specific appearance causes the cross-video consistency of the unmodified videos to drop to 88.0 from a within-video consistency of 96.1. Therefore, leakage of video-specific appearance inherently lowers cross-video consistency for SKPG-Swap as well. The second cause is the SKPG backbone itself, which is not perfectly consistent across separate videos even before any rendering into the original frame. SKPG-BB, which exposes the generator output directly, only reaches a cross-video consistency of 86.6 against a score of 93.1 within a video. Thus, part of the cross-video drop for SKPG-Swap originates in the generated pseudonym and is then amplified by the video-specific leakage of the blend.

Trading Re-ID scores for downstream utility A second pattern runs across the rendering strategies: the stronger the Re-ID scores, the weaker the downstream utility. SKPG-BB and SKPG-Reenact score highest on nearly all Re-ID metrics but score worst on both downstream tasks. In contrast, SKPG-Swap trades some Re-ID scores for context preservation, allowing it to retain almost all of the unmodified baseline action recognition accuracy and reach the highest emotion recognition UAR among the pseudonymization frameworks. Comparing the two bounding-box strategies isolates the contribution of face reenactment. Aligning head pose and expression helps the expression-dependent emotion recognition task, but gives no benefit to action recognition. The action recognition tasks depended mainly on the objects interacting with the face, which the bounding-box overwrite removes regardless of how well the face is aligned.

Same-key consistency stabilizes predictions The longitudinal experiment shows a consistent pattern across every rendering strategy: pairs of videos assigned the same key agree more often than pairs assigned different keys. The *same-key* versus *different-key* comparison cannot isolate the pseudonym identity as the sole cause of this gap, because two *same-key* SKPG outputs can still differ in context, so non-identity attributes vary between them as well. Even the identity-driven part of the gap is not inherent to the pseudonym identity: with the subject's expression held fixed, an ideal renderer and classifier would return the same emotion for any pseudonym, so the gap instead reflects that the renderer and the classifier are imperfect and respond in part to the pseudonym identity rather than to the subject's expression. Because the renderer and the classifier are held fixed across the *same-key* and *different-key* conditions, the difference in stability between the two conditions is attributable to pseudonym consistency rather than to the renderer or the classifier. What the experiment does show is that keeping the pseudonym identity consistent across videos produces more stable predictions than letting it vary. Therefore, the consistency objective on which the SKPG backbone is trained is a motivated property that stabilizes downstream predictions.

The *different-key* gap is not a practical weakness of the framework, since the framework is designed to assign the same key to the same subject and therefore always operates in the *same-key* setting. Prediction instability only appears when predictions are compared across separate videos of the same subject under different keys, and the instability is removed once a consistent key keeps the pseudonym identity fixed. The *different-key* setting therefore functions only as a control that illustrates what would happen if consistent pseudonymization were not

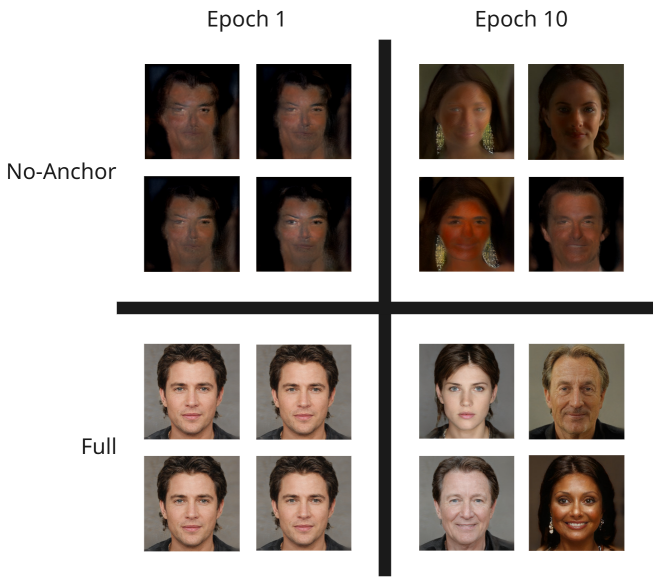

Figure 3. Intermediate pseudonyms produced by the SKPG backbone at epoch 1 (left column) and epoch 10 (right column) for the *No-Anchor* (top row) and *Full* (bottom row) configurations. With the anchor, the Projector starts from a realistic face and produces recognizable pseudonyms from the first epoch. Without the anchor, the outputs at epoch 1 barely resemble human faces, and structural distortions remain visible at epoch 10 despite the regularization loss pulling the latent toward w_{avg}. The anchor, and not the regularization loss, is what supplies the realistic initialization of latent vector w.

enforced.

How rendering amplifies pseudonym instabilities The two ablations point to one further property of the rendering choice: removing the input preprocessor or the w_{avg} anchor hardly changes the SKPG-BB scores but has a clear negative effect on SKPG-Swap. Because SKPG-BB exposes the SKPG output directly, that output stays relatively resilient to both design choices, maintaining high scores across most Re-ID metrics. In contrast, SKPG-Reenact and SKPG-Swap both rely on Face-Adapter, a complex model primarily trained on realistic samples. This model amplifies the effect of minor changes or distortions in the pseudonym. Without input preprocessing and the w_{avg} anchor, the Projector struggles to balance its multi-task objectives and maintain structural realism. Consequently, SKPG-Reenact and SKPG-Swap amplify the small SKPG-level instabilities seen in SKPG-BB into substantial downstream metric drops. The added value of both the preprocessing step and the w_{avg} anchor therefore only fully emerges after face reenactment and face swapping.

Neither IVFG [37] nor KFAAR [38] discusses input preprocessing or anchoring to w_{avg} as design choices, and neither ablates them. This omission probably originates from differences in their pipelines. Because IVFG evaluates only the raw StyleGAN2 output, downstream degradation is barely noticeable. In contrast, KFAAR's reenactment step would only partially expose the severe metric drops that become fully apparent after face-swap rendering.

Accessories as a training shortcut One observed behavior in early training led to a key design choice: the SKPG can satisfy its identity objectives by changing accessories rather than the pseudonym's face. During the initial development phase, the SKPG backbone used a StyleGAN2 generator pre-trained on the full CelebA-HQ dataset. An unexpected behavior emerged during optimization, in which the network discovered that accessories could help satisfy the diversity, differentiation, and consistency objectives. The model exploited accessories in two conflicting ways. In some cases, it added prominent items such as heavy glasses to increase the distance between identity embeddings without modifying the underlying faces. In other cases, it added accessories that were barely noticeable, such as thin wire-frame glasses. Because the identity embeddings of the pseudonyms with these subtle accessories stayed close to the embeddings of pseudonyms without them, the network could reach a low consistency loss while alternating accessories between pseudonyms of the same subject. This instability produced a flickering effect across frames, in which the generated pseudonym alternated between wearing and not wearing an accessory. Another issue was that the Face-Adapter renderer struggled to blend accessories into the original frames without introducing visual artifacts.

Existing frameworks such as IVFG and KFAAR rely on similar StyleGAN-based architectures and train their generators from scratch on the Labeled Faces in the Wild (LFW) dataset [53], which contains images of subjects wearing accessories. Neither work reports occlusions being exploited to satisfy training objectives nor flickering accessories across frames. However, these issues motivated a different strategy for this framework. Specifically, the StyleGAN2 [40] generator was trained from scratch on a CelebA-HQ subset that explicitly excluded images containing accessories. While this consideration is absent in previous literature, filtering the training data is a crucial step to prevent the pipeline from exploiting occlusions, resulting in visually stable pseudonyms.

5.1 Limitations

Computational cost The face-swap step in SKPG-Swap relies on Face-Adapter, which runs a large Stable Diffusion model through an iterative denoising loop that is computationally expensive. Until this loop is accelerated or replaced with a lighter face-swap component, the framework cannot be used in latency-critical settings

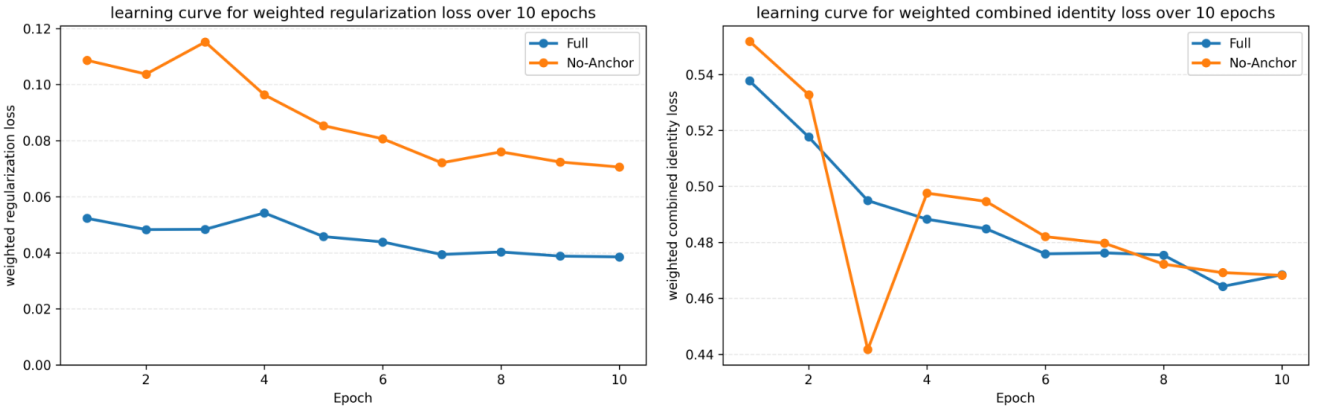

Figure 4. Weighted regularization loss (left) and weighted sum of the four re-identification (Re-ID) losses (right), both per epoch, for the *Full* and *No-Anchor* configurations. The *Full* configuration keeps the regularization loss low from the first epoch while the *No-Anchor* variant never closes the gap. The *Full* Re-ID loss is also lower in the early epochs, while Re-ID losses of the two configurations converge at later epochs. Notably, the *No-Anchor* variant shows a temporary instability at epoch 3, characterized by a sharp drop in Re-ID loss and a corresponding surge in regularization loss. This indicates the optimizer temporarily pushed the latent into an unrealistic region to artificially satisfy the identity objective before the regularization penalty corrected it.

such as live surveillance.

Temporal instability The framework processes each frame independently, which produces visible flickering from one frame to the next. The flickering originates mainly from the Face-Adapter component, a limitation the Face-Adapter authors also acknowledge [22]. Without a dedicated temporal smoothing step, this flickering could degrade downstream models that rely on smooth transitions between frames [11].

Privacy risk from identity leakage Although the SKPG backbone successfully eliminates the need for a vulnerable biometric database, the residual identity leakage from the face-swapping step creates new privacy risks. The measured Anon EER of 26.3 shows that this risk is real. One primary source of this residual leakage is the preservation of the subject's hair. Since the FaceNet [43] and ArcFace [48] embeddings used in the evaluations carry information about hairstyle and hair color [54], the hair preserved by the face swap pulls the pseudonym's embedding back toward that of the subject.

The privacy risk from identity leakage is only partially captured by the current evaluation, which relies on Re-ID metrics that measure the mathematical separation between the identity embeddings of subjects and pseudonyms. These metrics are useful for comparing methods and revealing the trade-offs between them, but they do not capture how a real adversary would attack the system. Three attacks are relevant for future evaluation. The first is a re-identification stress test against a large real-world identity database, in which the adversary tries to match a pseudonym back to the subject. The second is a model inversion attack, in which an adversary with access to the trained models in the pipeline attempts to reconstruct a subject from its pseudonym. The third is a feature reconstruction attack, in which the leaked biometric identifiers are used to link a pseudonym back to the subject. These three attack models capture real-world threats that embedding separation alone cannot reveal.

Within- and cross-subject scope of the consistency benefit The *same-key* consistency reduces prediction variance only partly within a single subject, and not at all across different subjects. The SKPG is trained on a consistency objective that pushes the same subject-key pair toward the same pseudonym identity. This objective, however, is not perfectly achieved, as shown by the cross-video consistency (Cross Con) results from Experiment 1. Consequently, some variation in pseudonym identity remains across videos even under the same key, limiting the variance reduction within the same subject. Across subjects, the property does not apply at all, because different subjects receive different pseudonym identities by the differentiation objective. Comparisons across subjects thus carry pseudonym-driven prediction variance that no choice of key can remove. The *same-key* consistency that stabilizes longitudinal tracking is therefore strictly a within-subject property and does not extend to comparisons between different subjects.

Scope and validity of the longitudinal experiment Beyond the within-subject scope of the consistency claim addressed above, the longitudinal experiment is itself limited by the controlled conditions under which its pairs were collected. Specifically, the pairs used in Experiment 3 are RAVDESS clips of the same subject performing the

same scripted emotion under identical studio conditions and within the same session. While this setting captures the multi-video structure that longitudinal analysis must handle, it does not account for the appearance drift in real longitudinal tracking, such as varying clothing, hairstyles, or lighting conditions. The experiment should therefore be read as a controlled proof-of-mechanism showing that pseudonym consistency stabilizes predictions, rather than as evidence of stability under real-world longitudinal variation.

A second caveat is that the stability metrics rest on a low-accuracy classifier. Because the emotion model reaches a UAR of only 53.6 even on the unmodified baseline, the predictions behind the Pairwise Agreement and Label Flip Rate are often wrong. Consequently, they measure how consistent the predictions are, not whether they are correct. This is problematic because a stable prediction can simply mean the classifier makes the same mistake on both clips rather than recognizing the emotion correctly. However, the conditional accuracy metric only counts pairs whose first clip is classified correctly and it shows a large *same-key* advantage for SKPG-Swap (90.5), therefore supporting the finding among correct predictions as well.

6 Conclusion and Future Work

This thesis developed a hybrid framework that integrates a Subject- and Key-conditioned Pseudonym Generator (SKPG) backbone with a Face-Adapter face-swap renderer to achieve secure, consistent pseudonymization while preserving context and removing the need for a vulnerable biometric database. The integration was evaluated by comparing three rendering strategies on the shared SKPG backbone, namely SKPG-BB, SKPG-Reenact, and the proposed face swap (SKPG-Swap), against the database-free DeepPrivacy2 baseline. In the action recognition evaluation, SKPG-Swap and DeepPrivacy2 both retained the action recognition utility of the unmodified videos, with neither showing a statistically significant drop from the unmodified baseline. The two bounding-box rendering strategies dropped substantially on the same task. SKPG-Swap also achieved the highest emotion recognition score among the evaluated pseudonymization methods. The emotion recognition evaluation additionally motivates consistent pseudonymization, demonstrating that mapping a subject to a consistent pseudonym identity results in more stable predictions across videos than an inconsistent one.

The evaluation also highlights a clear trade-off among the rendering strategies built on the shared SKPG backbone. The bounding-box strategies SKPG-BB and SKPG-Reenact overwrite the face region entirely and therefore reach the strongest anonymization, key-driven diversity, and cross-video consistency. The proposed SKPG-Swap instead routes the pseudonym through a face-swap renderer, which blends it into the original frame and preserves the context. This blending exchanges part of that anonymization, diversity, and cross-video consistency for the downstream utility that a bounding-box overwrite discards. Using face swapping as a rendering strategy is thus what moves the shared backbone from optimizing Re-ID metrics toward maximal data utility, and the appropriate choice along this trade-off depends on the priorities of the target application.

Future research should address the current generation and evaluation constraints along four directions. First, the residual identity leakage introduced by the face-swap renderer should be both reduced and measured more realistically. The leakage can be lowered at its source by extending the swapped region to also replace the subject's hair, because the preserved hair is the main cue that pulls the pseudonym embedding back toward the subject. The leakage should then be measured with adversarial attack simulations, such as a re-identification stress test against a large real-world database, a model inversion attack, and a feature reconstruction attack, since the embedding separation reported by the current Re-ID metrics does not reflect how a real adversary would attack the system. Second, the longitudinal evaluation should be strengthened in two complementary ways. A strictly controlled experiment that holds every non-identity attribute of the generated pseudonym constant across videos would isolate the effect of the pseudonym identity alone. A complementary realistic experiment on separately recorded videos with changes in the appearance of the subjects, together with a downstream task that retains a higher baseline accuracy, would then test whether the stabilizing effect of consistency survives real-world longitudinal variation. Third, the pipeline could be trained end-to-end so that the pseudonym integrates more cleanly into the frame. Folding the face-swap renderer into the multi-task optimization, ideally with a lighter model, could jointly improve the Re-ID scores, context preservation, and visual quality, while also reducing the computational cost that currently rules out latency-critical use. Finally, the flickering introduced by the current per-frame architecture could be resolved by replacing the frame-independent MLP projector with a bidirectional Long Short-Term Memory (LSTM) network and adding a temporal-consistency loss during training. Both changes would let the model share information across neighboring frames instead of processing each frame in isolation. Together, these directions would tighten the privacy guarantee, broaden the evidence for the pseudonym consistency claim, lower the computational cost toward latency-critical use, and reduce the temporal flickering that remains.

Appendix

A Significance Testing for Experiment 2

Table A.1. McNemar significance testing for action recognition on UCF101, with each pseudonymization method compared against the unmodified baseline. Here b is the number of clips the unmodified baseline classifies correctly while the method classifies wrongly, and c is the number of clips the unmodified baseline classifies wrongly while the method classifies correctly. The McNemar χ^2 uses the continuity correction, and significance is read from the two-sided exact test because the discordant counts are small for DeepPrivacy2 and SKPG-Swap. DeepPrivacy2 and SKPG-Swap show no significant difference from the unmodified baseline, while SKPG-BB and SKPG-Reenact are significantly worse, confirming that only the face-swap and DeepPrivacy2 renderers retain baseline action-recognition accuracy.

Method	b	c	McNemar χ^2	Exact p
DeepPrivacy2	4	1	0.80	0.375
SKPG-BB	21	0	19.05	9.5×10^{-7}
SKPG-Reenact	22	1	17.39	5.7×10^{-6}
SKPG-Swap (Ours)	2	1	0.00	1.000

B Confusion Matrices for Experiment 2

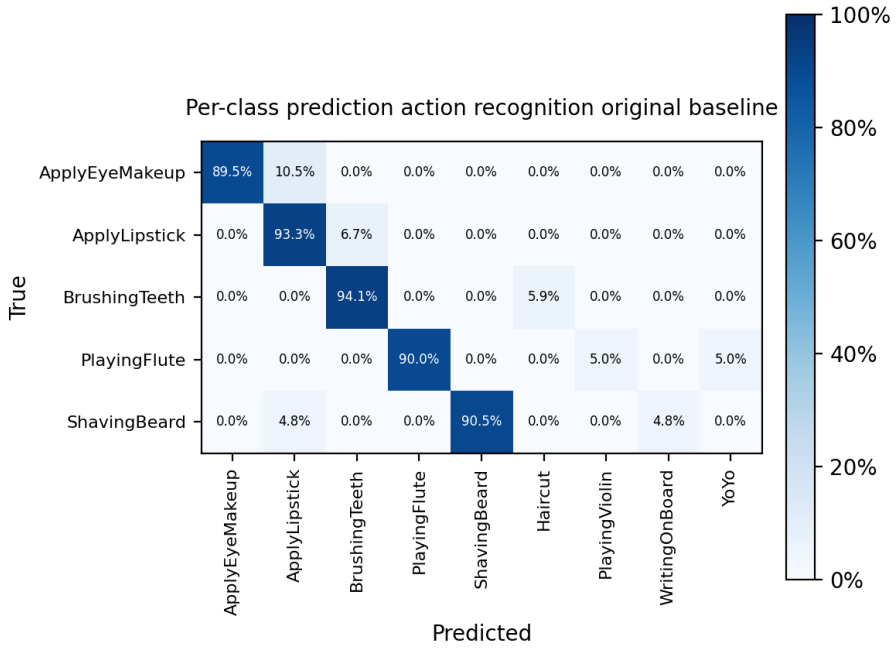

Figure A.1. Per-class confusion matrix for action recognition on the unmodified UCF101 testing split. Rows are true classes, columns are predicted classes, and cells give the percentage of clips of each true class assigned to each predicted class. Diagonal entries above 89% on all five classes confirm that the VideoMAE classifier is a strong baseline on this subset, so any subsequent drop reflects what the pseudonymization removes rather than weakness in the classifier.

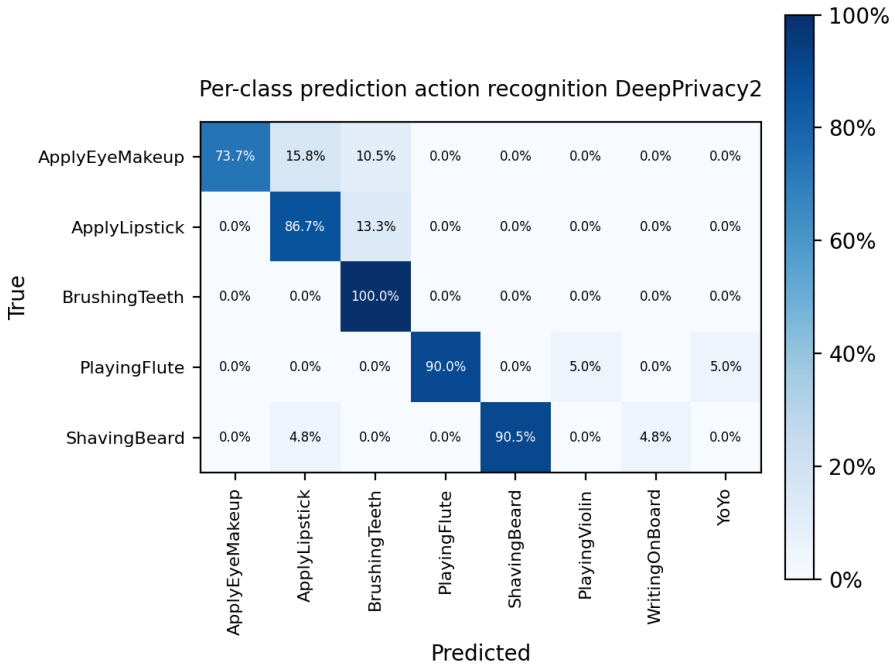

Figure A.2. Per-class confusion matrix for action recognition on UCF101 after DeepPrivacy2 pseudonymization. Diagonal entries stay above 73% on every class, and the largest off-diagonal mass falls between visually adjacent face-region activities (ApplyEyeMakeup → ApplyLipstick, ApplyLipstick → BrushingTeeth). The per-class behavior therefore stays close to the unmodified baseline.

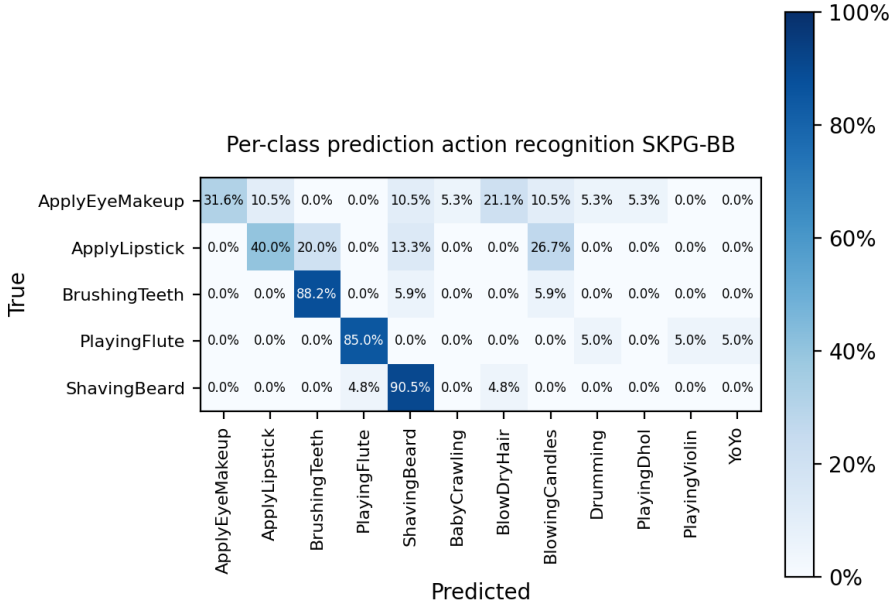

Figure A.3. Per-class confusion matrix for action recognition on UCF101 after SKPG-BB pseudonymization. ApplyEyeMakeup and ApplyLipstick drop to 31.6% and 40.0% and are frequently misclassified as unrelated classes such as BlowDryHair and BlowingCandles, while BrushingTeeth, PlayingFlute, and ShavingBeard remain above 85%. The bounding-box overwrite therefore harms the two classes that depend most on fine facial detail and leaves the object-interaction classes largely intact.

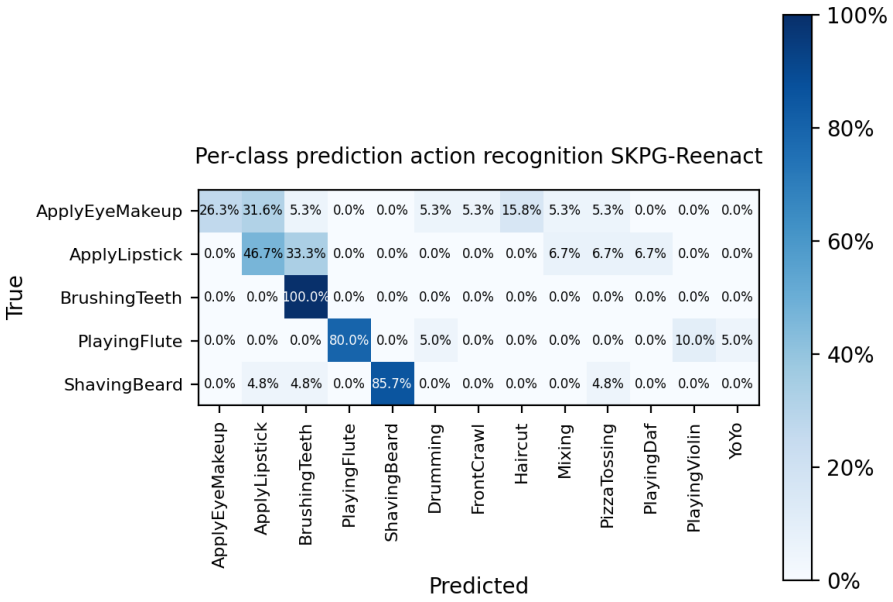

Figure A.4. Per-class confusion matrix for action recognition on UCF101 after SKPG-Reenact pseudonymization. The pattern matches SKPG-BB: ApplyEyeMakeup and ApplyLipstick drop to 26.3% and 46.7% and are frequently misclassified as unrelated classes such as Haircut, while BrushingTeeth, PlayingFlute, and ShavingBeard remain above 80%. Aligning the head pose through reenactment therefore does not recover the detail-dependent classes lost by the bounding-box overwrite.

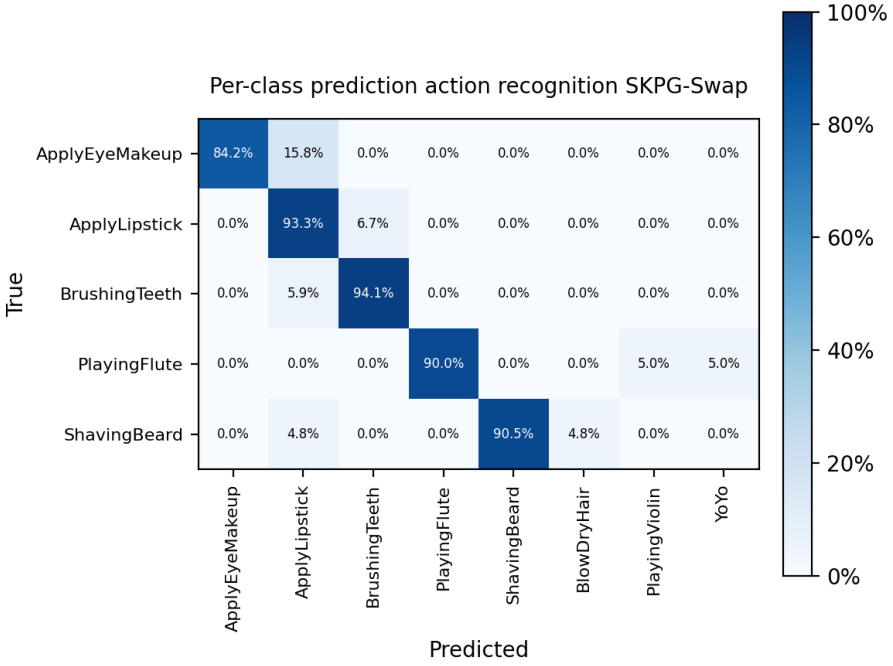

Figure A.5. Per-class confusion matrix for action recognition on UCF101 after SKPG-Swap pseudonymization. Diagonal entries stay within six percentage points of the unmodified baseline on every class, including the detail-dependent ApplyEyeMakeup and ApplyLipstick classes that collapsed under the bounding-box renderers. The face-swap renderer therefore preserves the per-class behavior of the unmodified data.

C Schematic Illustrations for Experiment 3

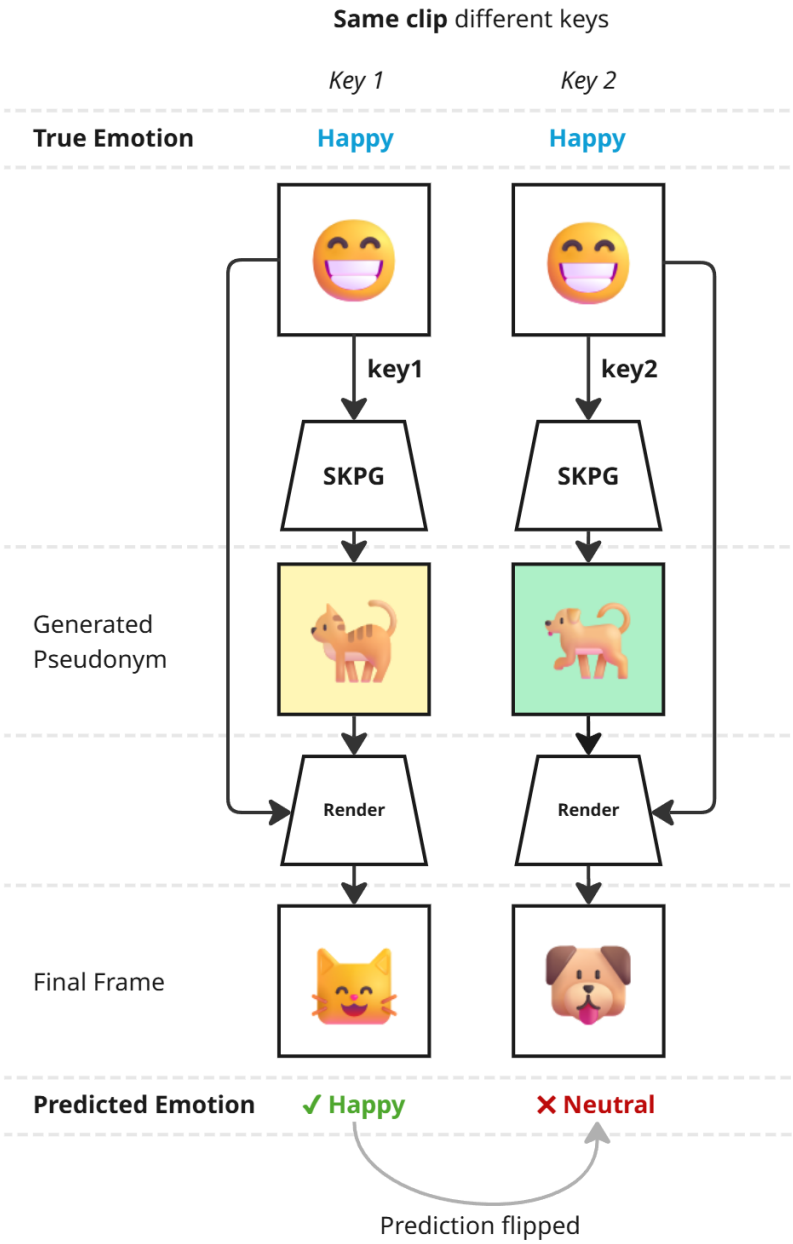

Figure A.6. Schematic illustration of key-induced prediction variance for a single clip (Section 4.4, Eq.(18)). A single source clip with true emotion *Happy* is pseudonymized twice with two different keys. The SKPG backbone maps each key to a different pseudonym identity (a cat for key 1, a dog for key 2), which the renderer blends back onto the same original frame. Although the source clip is held fixed, the downstream classifier predicts *Happy* for the first pseudonym and *Neutral* for the second, illustrating how changing the key alone can push the prediction across the decision boundary. The animal denotes the pseudonym identity and the background color the other, uncontrolled attributes the pseudonym carries. Both change with the key, consistent with the observation that the two pseudonyms differ in more than identity.

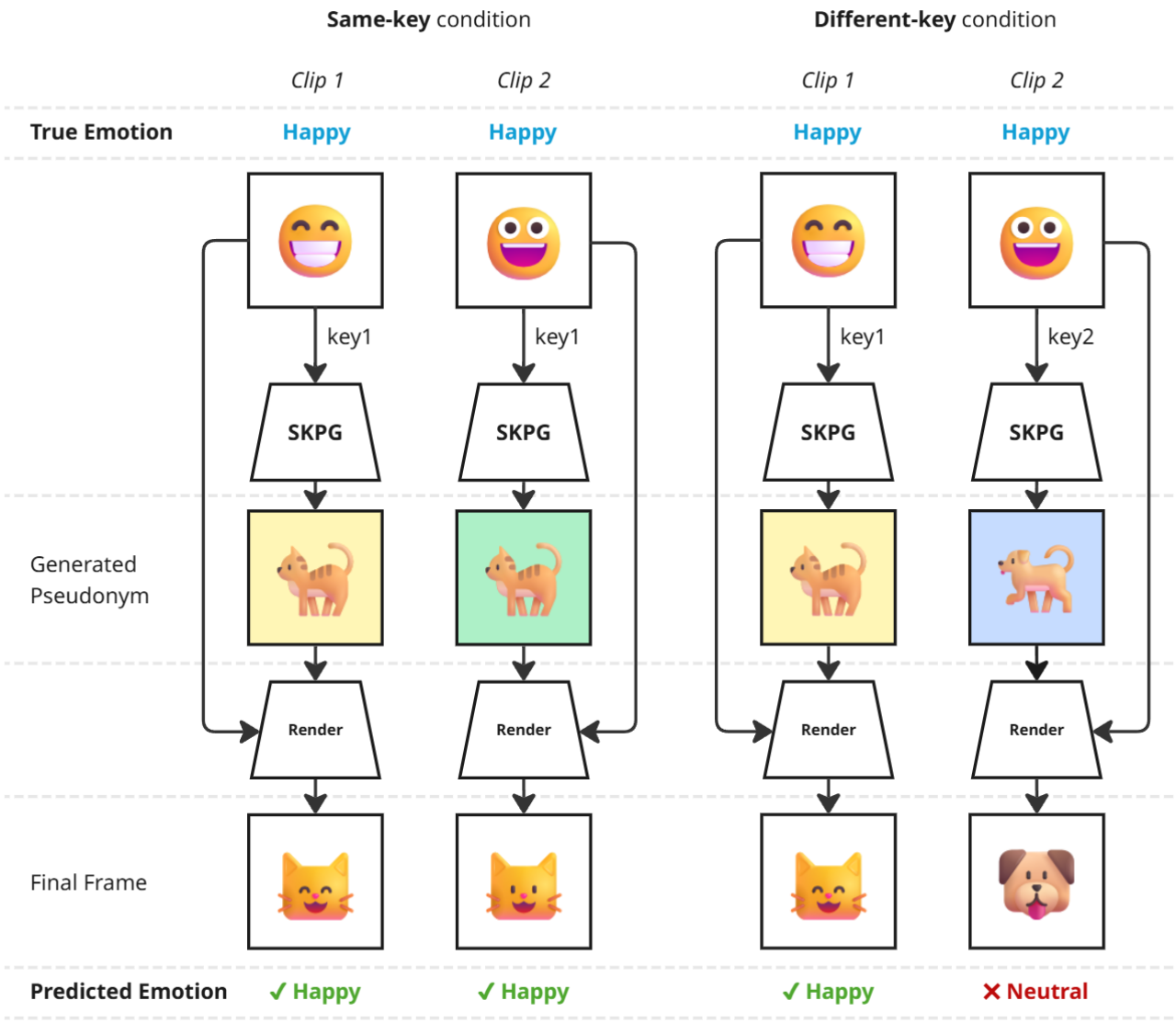

Figure A.7. Schematic illustration of the *same-key* and *different-key* conditions used in the longitudinal evaluation (Section 4.4, Eqs. (19)–(20), Table 10). Two clips of the same subject with the same true emotion (*Happy*) are treated as repeated videos. In the *same-key* condition both clips receive the same key and therefore the same pseudonym identity (cat), and the classifier predicts *Happy* on both, so the pair agrees. In the *different-key* condition the two clips receive different keys and therefore different pseudonym identities (cat and dog), and the prediction on the second clip flips to *Neutral*, so the pair disagrees. The animal denotes the pseudonym identity and the background color the context, which can differ between clips even under the same key. The higher Pairwise Agreement and Conditional Accuracy in the *same-key* condition reflect the stabilizing effect of a consistent pseudonym identity.

D Significance Testing for Experiment 3

Table A.2. McNemar significance testing for emotion recognition on RAVDESS, with each pseudonymization method compared against the unmodified baseline. Here b is the number of clips the unmodified baseline classifies correctly while the method classifies wrongly, and c is the number of clips the unmodified baseline classifies wrongly while the method classifies correctly. The McNemar χ^2 uses the continuity correction. Every method differs significantly from the unmodified baseline ($p < 10^{-58}$), confirming that all pseudonymization methods, including SKPG-Swap, reduce emotion-recognition accuracy significantly below the unmodified baseline on this task.

Method	b	c	McNemar χ^2	Exact p
DeepPrivacy2	222	5	205.53	4.6×10^{-59}
SKPG-BB	264	0	262.00	6.7×10^{-80}
SKPG-Reenact	250	1	245.04	1.4×10^{-73}
SKPG-Swap (Ours)	229	0	227.00	2.3×10^{-69}

E Confusion Matrices for Experiment 3

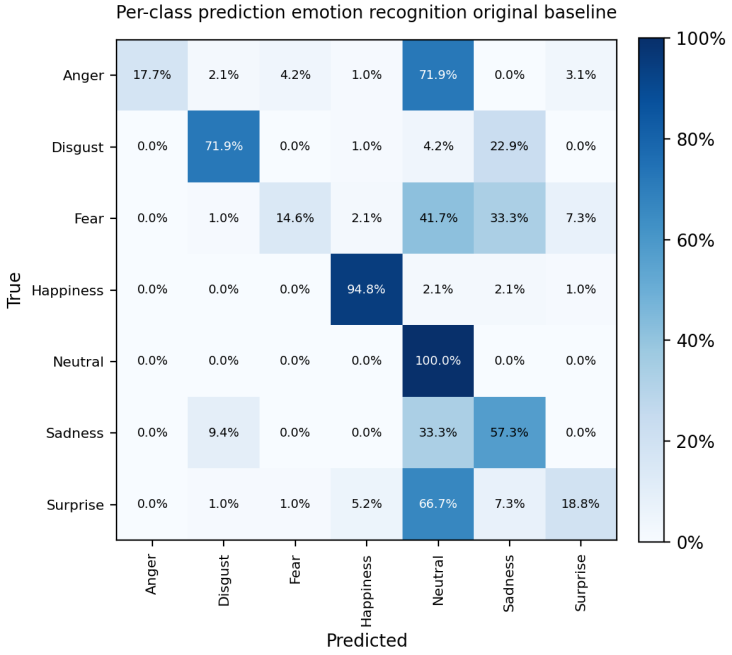

Figure A.8. Per-class confusion matrix for emotion recognition on the unmodified RAVDESS evaluation set. Rows are true classes, columns are predicted classes, and cells give the percentage of clips of each true class assigned to each predicted class. Happiness and Neutral are recognized at 94.8% and 100%, while Anger, Fear, and Surprise are predominantly misclassified as Neutral even on the unmodified data. The unmodified baseline classifier therefore already concentrates predictions on a subset of classes, which sets the reference against which the pseudonymized matrices must be read.

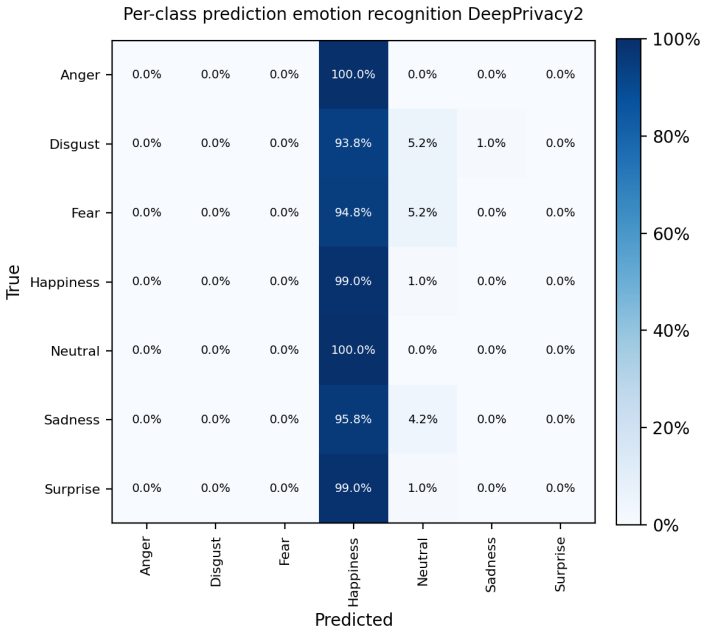

Figure A.9. Per-class confusion matrix for emotion recognition on RAVDESS after DeepPrivacy2 pseudonymization. Every true class collapses onto Happiness with above 93% probability, including the Neutral class that the unmodified baseline recognized perfectly. The classifier therefore loses access to the per-class signal almost entirely under this method.

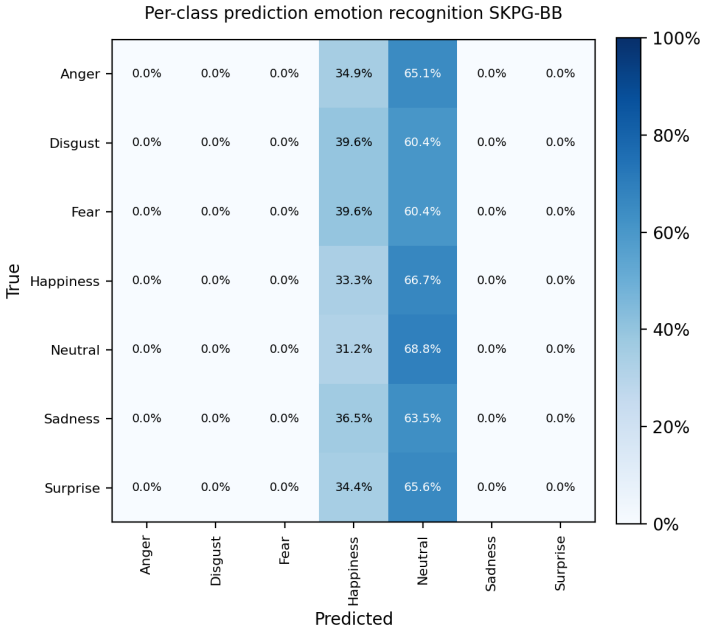

Figure A.10. Per-class confusion matrix for emotion recognition on RAVDESS after SKPG-BB pseudonymization. Every clip is predicted as either Happiness or Neutral, and the remaining five classes never appear in the predictions. The bounding-box overwrite therefore reduces the classifier to a binary decision between two emotion categories.

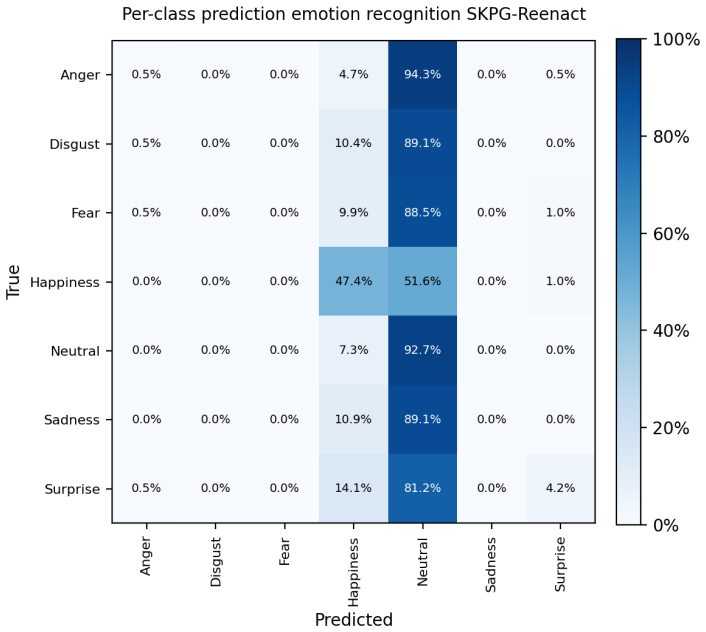

Figure A.11. Per-class confusion matrix for emotion recognition on RAVDESS after SKPG-Reenact pseudonymization. Predictions concentrate on Neutral for every true class except Happiness, which retains 47.4% on its own diagonal. Aligning the head pose through reenactment therefore partly recovers the Happiness class but leaves the remaining classes collapsed onto Neutral.

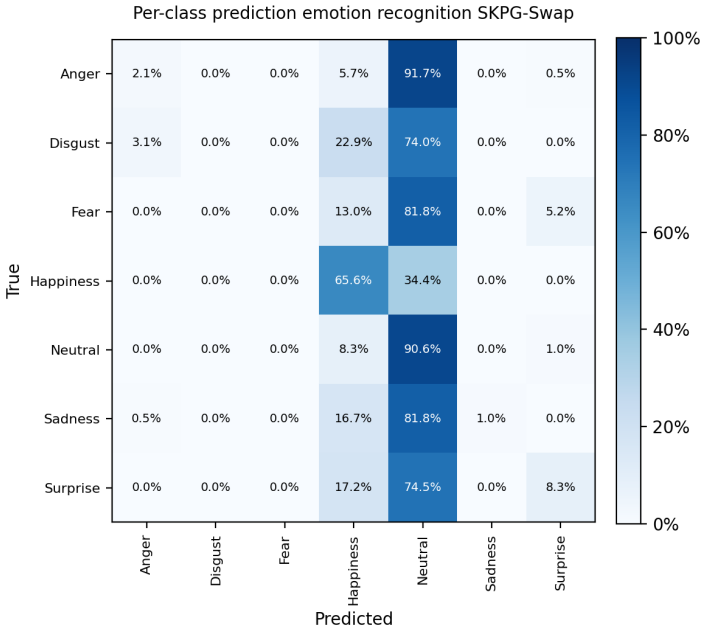

Figure A.12. Per-class confusion matrix for emotion recognition on RAVDESS after SKPG-Swap pseudonymization. Happiness is classified correctly on 65.6% of its clips, and small numbers of clips return to the correct class for Anger, Sadness, and Surprise, while the remaining classes are still predicted as Neutral. The face-swap renderer therefore retains more of the original per-class behavior than the other pseudonymization methods, even though it still drops sharply from the unmodified baseline.

References

[1] Steven R. Livingstone and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A validated multimodal database of emotional speech and song". In: *PLOS ONE* 13.5 (2018), e0196391. DOI: 10.1371/journal.pone.0196391. URL: https://doi.org/10.1371/journal.pone.0196391.

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *Interspeech 2018*. ISCA, Sept. 2018, pp. 1086–1090. DOI: 10.21437/Interspeech.2018-1929.

[3] Håkon Hukkelås and Frank Lindseth. *DeepPrivacy2: Towards Realistic Full-Body Anonymization*. 2022. arXiv: 2211.09454 [cs.CV]. URL: https://arxiv.org/abs/2211.09454.

[4] Jemal Abawajy, Shamsul Huda, and Imran Rao. "A Comprehensive Survey of Machine Learning Methods for Surveillance Videos Anomaly Detection". In: *IEEE Access* 11 (2023), pp. 110986–111011. DOI: 10.1109/ACCESS.2023.3321800.

[5] Muna Almasawa, Lamiaa Elrefaei, and Kawthar Moria. "A Survey on Deep Learning Based Person Re-Identification Systems". In: *IEEE Access* 7 (2019), pp. 175291–175344. DOI: 10.1109/ACCESS.2019.2957336.

[6] Andre Esteva et al. "Deep learning-enabled medical computer vision". In: *npj Digital Medicine* 4.1 (2021), pp. 1–14. DOI: 10.1038/s41746-020-00376-2.

[7] Michal Kolarik et al. "Explainability of deep learning models in medical video analysis: a survey". In: *PeerJ Computer Science* 9 (2023), e1253. DOI: 10.7717/peerj-cs.1253. URL: https://doi.org/10.7717/peerj-cs.1253.

[8] European Parliament and Council of the European Union. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj.

[9] Centers for Medicare & Medicaid Services. *The Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. Public Law 104-191, 110 Stat. 1936. Accessed: 2026-03-27. 1996. URL: https://www.cms.gov.

[10] Roland Stenger et al. "Evaluating the Impact of Face Anonymization Methods on Computer Vision Tasks: A Trade-Off Between Privacy and Utility". In: *IEEE Access* PP (Dec. 2024), pp. 1–1. DOI: 10.1109/ACCESS.2024.3519441.

[11] Jingyi Cao et al. "Face De-Identification: State-of-the-Art Methods and Comparative Studies". In: *IEEE Transactions on Broadcasting* PP (Jan. 2025), pp. 1–21. DOI: 10.1109/TBC.2025.3639783.

[12] S. Qiao and H. Liu. "Automatic classification of criminal activities for security surveillance by keyframes detection and advanced inception techniques". In: *Scientific Reports* 16.1 (2026), pp. 1–14. DOI: 10.1038/s41598-025-30199-8.

[13] Kumuda P et al. "A comprehensive review of AI-powered campus surveillance". In: *ITM Web of Conferences* 81 (Jan. 2026). DOI: 10.1051/itmconf/20268101017.

[14] Runfang Guo et al. "Development and application of emotion recognition technology — a systematic literature review". In: *BMC Psychology* 12.1 (2024), pp. 1–25. DOI: 10.1186/s40359-024-01581-4.

[15] Duaa Shehada et al. "An Explainable Framework for Mental Health Monitoring Using Lightweight and Privacy-Preserving Federated Facial Emotion Recognition". In: *Sensors* 25.23 (2025), p. 7320. DOI: 10.3390/s25237320. URL: https://doi.org/10.3390/s25237320.

[16] Yu Kong and Yun Fu. "Human Action Recognition and Prediction: A Survey". In: *International Journal of Computer Vision* 130.5 (2022), pp. 1366–1401. DOI: 10.1007/s11263-022-01594-9.

[17] Shan Li and Weihong Deng. "Deep Facial Expression Recognition: A Survey". In: *IEEE Transactions on Affective Computing* 13.3 (2022), pp. 1195–1215. DOI: 10.1109/TAFFC.2020.2981446.

[18] Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. "Learning to Anonymize Faces for Privacy Preserving Action Detection". In: *arXiv preprint arXiv:1803.11556* (Mar. 2018). DOI: 10.48550/arXiv.1803.11556.

[19] Nalini Ratha, Jonathan Connell, and Ruud Bolle. "Enhancing Security and Privacy in Biometrics-Based Authentication Systems". In: *IBM Systems Journal* 40 (Jan. 2001), pp. 614–634. DOI: 10.1147/sj.403.0614.

[20] Vishal M. Patel, Nalini K. Ratha, and Rama Chellappa. "Cancelable Biometrics: A review". In: *IEEE Signal Processing Magazine* 32.5 (2015), pp. 54–65. DOI: 10.1109/MSP.2015.2434151.

[21] Josh Taylor. "Major breach found in biometrics system used by banks, UK police and defence firms". In: *The Guardian* (Aug. 2019). URL: https://www.theguardian.com/technology/2019/aug/14/major-breach-found-in-biometrics-system-used-by-banks-uk-police-and-defence-firms.

[22] Yue Han et al. *Face Adapter for Pre-Trained Diffusion Models with Fine-Grained ID and Attribute Control*. 2024. arXiv: 2405.12970 [cs.CV]. URL: https://arxiv.org/abs/2405.12970.

[23] R. Dhanyalakshmi et al. "A Survey on Face-Swapping Methods for Identity Manipulation in Deepfake Applications". In: *IET Image Processing* 19 (June 2025). DOI: 10.1049/ipr2.70132.

[24] Gan Pei et al. *Deepfake Generation and Detection: A Benchmark and Survey*. 2026. arXiv: 2403.17881 [cs.CV]. URL: https://arxiv.org/abs/2403.17881.

[25] Michael Boyle, Christopher Edwards, and Saul Greenberg. "The Effects of Filtered Video on Awareness and Privacy". In: *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW)*. ACM, Dec. 2000, pp. 1–10. DOI: 10.1145/358916.358935.

[26] Khairi Ahmed and Abdulhakim Baroud. "Protecting Privacy with Image Processing: How Blurring and Masking Help Keep Us Safe Online". In: *Journal for Research in Applied Sciences and Biotechnology* 4 (Apr. 2022), pp. 132–138. DOI: 10.55544/jrasb.1.1.18.

[27] Andrea Frome et al. "Large-scale privacy protection in Google Street View". In: *2009 IEEE 12th International Conference on Computer Vision*. 2009, pp. 2373–2380. DOI: 10.1109/ICCV.2009.5459413.

[28] Yifang Li et al. "Blur vs. Block: Investigating the Effectiveness of Privacy-Enhancing Obfuscation for Images". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. July 2017, pp. 1343–1351. DOI: 10.1109/CVPRW.2017.176.

[29] Ralph Gross et al. "Face De-identification". In: *Protecting Privacy in Video Surveillance*. Springer, July 2009, pp. 129–146. ISBN: 978-1-84882-300-6. DOI: 10.1007/978-1-84882-301-3_8.

[30] Miao Xin, Shentong Mo, and Yuanze Lin. "EVA-GCN: Head Pose Estimation Based on Graph Convolutional Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, June 2021, pp. 1462–1471. DOI: 10.1109/CVPRW53098.2021.00162.

[31] Haosong Zhang et al. "PeVL: Pose-Enhanced Vision-Language Model for Fine-Grained Human Action Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2024, pp. 18857–18867. DOI: 10.1109/CVPR52733.2024.01784.

[32] Dan Hendrycks and Thomas G. Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations". In: *International Conference on Learning Representations (ICLR)*. 2019. arXiv: 1903.12261. URL: https://openreview.net/forum?id=S1mu4K197.

[33] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. "CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020, pp. 5446–5455. DOI: 10.1109/CVPR42600.2020.00549. URL: http://dx.doi.org/10.1109/CVPR42600.2020.00549.

[34] Minchul Kim et al. "DCFace: Synthetic Face Generation with Dual Condition Diffusion Model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 12715–12725. DOI: 10.1109/CVPR52729.2023.01223.

[35] Renwang Chen et al. "SimSwap: An Efficient Framework For High Fidelity Face Swapping". In: *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. ACM, Oct. 2020, pp. 2003–2011. DOI: 10.1145/3394171.3413630. URL: http://dx.doi.org/10.1145/3394171.3413630.

[36] Wenliang Zhao et al. "DiffSwap: High-Fidelity and Controllable Face Swapping via 3D-Aware Masked Diffusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023, pp. 8568–8577. DOI: 10.1109/CVPR52729.2023.00828.

[37] Zhuowen Yuan et al. "On Generating Identifiable Virtual Faces". In: *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. ACM, Oct. 2022, pp. 1465–1473. DOI: 10.1145/3503161.3548110. URL: http://dx.doi.org/10.1145/3503161.3548110.

[38] Miaomiao Wang et al. "A Key-Driven Framework for Identity-Preserving Face Anonymization". In: *Proceedings of the 32nd Annual Network and Distributed System Security Symposium (NDSS)*. Internet Society, Jan. 2025. DOI: 10.14722/ndss.2025.230729.

[39] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4401–4410. DOI: 10.1109/CVPR.2019.00453.

[40] Tero Karras et al. "Analyzing and Improving the Image Quality of StyleGAN". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 8110–8119. DOI: 10.1109/CVPR42600.2020.00813.

[41] Ziwei Liu et al. "Deep Learning Face Attributes in the Wild". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015, pp. 3730–3738. DOI: 10.1109/ICCV.2015.425.

[42] Tero Karras et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *International Conference on Learning Representations (ICLR)*. 2018. URL: https://openreview.net/forum?id=Hk99zCeAb.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[44] Qiong Cao et al. "VGGFace2: A dataset for recognising faces across pose and age". In: *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. 2018, pp. 67–74. DOI: 10.1109/FG.2018.00020.

[45] Tero Karras et al. "Training Generative Adversarial Networks with Limited Data". In: *Proc. NeurIPS*. 2020.

[46] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV]. URL: https://arxiv.org/abs/2112.10752.

[47] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. "One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021.

[48] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4690–4699. DOI: 10.1109/CVPR.2019.00482.

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild". In: *arXiv preprint arXiv:1212.0402* (2012). arXiv: 1212.0402 [cs.CV]. URL: https://arxiv.org/abs/1212.0402.

[50] Zhan Tong et al. "VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 10078–10093.

[51] Alan Agresti. *Categorical Data Analysis*. 3rd. Hoboken: John Wiley & Sons, 2013. ISBN: 978-0-470-46363-5. DOI: 10.1002/9781118357590.

[52] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. "In Search of a Robust Facial Expressions Recognition Model: A Large-Scale Visual Cross-Corpus Study". In: *Neurocomputing* 514 (2022), pp. 435–448. DOI: 10.1016/j.neucom.2022.10.013. URL: https://www.sciencedirect.com/science/article/pii/S0925231222012656.

[53] Gary B. Huang et al. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. UM-CS-2008-005. University of Massachusetts, Amherst, Oct. 2008.

[54] Philipp Terhörst et al. "Beyond Identity: What Information Is Stored in Biometric Face Templates?" In: *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*. 2020, pp. 1–10. DOI: 10.1109/IJCB48548.2020.9304901.

Acknowledgment of AI Assistance

During this project, I used AI tools as assistants to speed up my workflow. This gave me more time to focus on the complex, meaningful, and rigorous parts of my research. My approach involved writing detailed prompts, carefully reviewing the responses, and iterating back and forth to refine the output. Ultimately, I guided the models and did the actual work myself. I used the following tools:

- **Gemini:** `https://gemini.google.com/`
- **Claude:** `https://claude.ai/`

I specifically used them to help with:

- **Understanding concepts:** Breaking down difficult math and summarizing research papers.
- **Brainstorming:** Talking through ideas to help organize my thoughts.
- **Coding:** Writing small pieces of code, which I always reviewed and tested myself.
- **Writing:** Making my writing sound more academic and concise.