Machine Vision and Applications (2025) 36:65
https://doi.org/10.1007/s00138-025-01687-8

RESEARCH

Emotion-aware face de-identification with generative adversarial networks

Md Shopon[1] · **Marina L. Gavrilova**[1]

Received: 19 February 2024 / Revised: 20 March 2025 / Accepted: 24 March 2025 / Published online: 9 April 2025
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2025

Abstract

This paper presents a novel end-to-end emotion-preserving face de-identification approach based on Generative Adversarial Networks (GANs), specifically utilizing the StyleGAN architecture. The proposed method generates natural-looking de-identified images by creating a synthetic face dataset and leveraging the DeepFace model for gender classification and representative image selection. To enhance emotion preservation, an improved SimSwap framework is introduced, incorporating a novel loss function designed to maintain emotional expressions. The DeepFace model is further utilized to classify and recognize emotional expressions in both original and de-identified images. Emotion preservation during face swapping is explicitly enforced by minimizing attribute preservation loss. A comprehensive ablation study demonstrates the effectiveness of the proposed GAN components. Experimental results show that the method outperforms recent face de-identification techniques in both accuracy and emotion preservation.

Keywords Facial images · Facial emotions · Privacy · De-identification · Injection-based GANs · Deep learning · Biometrics

1 Introduction

In the modern digital landscape, the widespread availability of camera-equipped mobile devices, along with the rise of surveillance technologies has raised significant concerns about personal privacy. The inadvertent capture of individuals in photographs has amplified the demand for effective de-identification methods to protect anonymity. Privacy regulations, such as the General Data Protection Regulation (GDPR) in the European Union [26], mandate stringent privacy protection measures, highlighting the need for advanced solutions to safeguard individual identities. Integrating privacy-enhancing systems into surveillance and recreational cameras can play a crucial role in ensuring compliance with such regulations while preserving personal privacy in an increasingly interconnected world.

De-identification is emerging as a key technique for protecting personal data, complementing other privacy-preserving approaches such as encryption and cancelability [14]. It involves removing identifiable information from datasets while retaining their utility for analysis [18]. The de-identification of biometric data is essential for safeguarding privacy while ensuring that data remains useful for research and other applications without violating ethical or legal regulations. Facial images, in particular, are highly distinctive and can be used to accurately identify individuals [12]. However, since facial data is often publicly accessible, it is vulnerable to misuse by malicious actors. On the other hand, de-identifying facial images can result in the loss of critical attributes such as gender, age, and ethnicity. Therefore, selectively preserving certain attributes enables greater flexibility while maintaining privacy. In particular, retaining emotional content in biometric data can be more beneficial than simple de-identification, especially in fields such as psychology, psychiatry, and neuroscience [13]. Emotions play a crucial role in these disciplines, offering valuable insights into a patient's mental state. Moreover, emotional content enhances the usability of facial images for training machine learning models, contributing to improved data analysis and applications in affective computing.

Traditional de-identification techniques, such as blurring or pixelation, can help protect privacy, but often result in the

✉ Md Shopon
md.shopon@ucalgary.ca

[1] Department of Computer Science, University of Calgary, Calgary, AB, Canada

loss of vital information, including emotions, age, gender, and facial expressions [25]. This loss is particularly problematic in domains where biometric attributes are essential, such as law enforcement and healthcare. Recently, a few machine learning-based de-identification methods have been proposed [1, 9]; however, these approaches do not preserve emotional content, which is critical in specific applications. This paper introduces a novel GAN-based approach for generating realistic de-identified images that retain the emotional content of the original face while ensuring privacy protection. To the best of our knowledge, this is the first end-to-end emotion-preserving face de-identification method. Here, "end-to-end" refers to a comprehensive and seamless framework that encompasses the entire de-identification process without requiring external interventions.

This research develops a novel de-identification system capable of preserving emotions while safeguarding an individual's privacy. The following are the main contributions:

- The first end-to-end emotion-preserving de-identification system that utilizes deep learning without any intermediary processing steps is proposed.
- A novel method for emotion-preserving de-identification using Generative Adversarial Networks (GANs), which can accurately and seamlessly generate a high-fidelity de-identified version of a face while unequivocally preserving the subject's emotions, is introduced.
- A novel attention-based identity injection module is proposed for producing natural-looking de-identified images.
- The StyleGAN architecture is utilized to generate a synthetic face image dataset for proxy face images, which can increase the diversity of images available for training the GAN.
- A novel attribute-preserving loss function is introduced to preserve the subject's emotions during de-identification.
- The proposed enhanced SimSwap architecture is capable of handling non-frontal views and wearable accessories of subjects.

The developed system is particularly valuable in applications where recognizing and interpreting facial expressions and emotions are crucial, such as healthcare and psychology. The proposed method was trained using the VGGFace2 dataset and validated on the CELEB-A and LFW datasets.

This paper is an extended version of a conference paper originally presented at the 2023 International Symposium on Visual Computing, held in Lake Tahoe, USA [19]. In this extended version, several significant enhancements have been made to further validate and refine the proposed method. First, a comprehensive cross-sectional experiment has been conducted to evaluate the simultaneous effectiveness of the approach in both de-identification and emotion preservation. The results highlight the robustness of the proposed method in maintaining anonymity while preserving emotional expressions. Additionally, an extended ablation study has been performed to assess the contribution of each component to the overall performance. This detailed analysis emphasizes the critical role of the proposed components in achieving optimal results in both de-identification and emotion preservation. Furthermore, new visualizations have been introduced to provide a clearer understanding of the cross-sectional results and the quality of the generated samples. These visual aids offer valuable insights into the effectiveness of the proposed method, demonstrating its ability to produce natural-looking de-identified images while retaining emotional expressions.

2 Literature review

Emotion-preserving face de-identification is an emerging research area that is gaining increasing attention due to the widespread use of facial recognition technology across various domains, from smartphone authentication to public surveillance. With growing concerns over personal privacy and the potential misuse of biometric data, de-identification techniques have been proposed to address these challenges. However, traditional de-identification methods often result in the loss of essential attributes, such as emotions, which play a fundamental role in human behavior and have significant implications in psychology, psychiatry, and neuroscience.

Traditional de-identification methods can be broadly categorized into complete de-identification, soft biometrics de-identification, and traditional biometric-preserving de-identification [20]. Figure 1 illustrates different biometric de-identification methods. In recent years, neural network-based approaches have gained popularity in face de-identification research. These methods leverage machine learning models, such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), to generate de-identified faces by learning from large datasets. CNNs can be trained to predict facial attributes, including emotions, and subsequently used to modify these attributes while generating de-identified faces.

Traditional face de-identification methods utilize techniques such as blurring [6], warping [21], and pixelation [10] to protect individuals' privacy in images and videos. Figure 2 presents examples of these methods. These approaches modify or obscure facial features while maintaining the overall structure of a face, making identification more challenging for unauthorized individuals. However, they often degrade visual quality, resulting in unnatural or

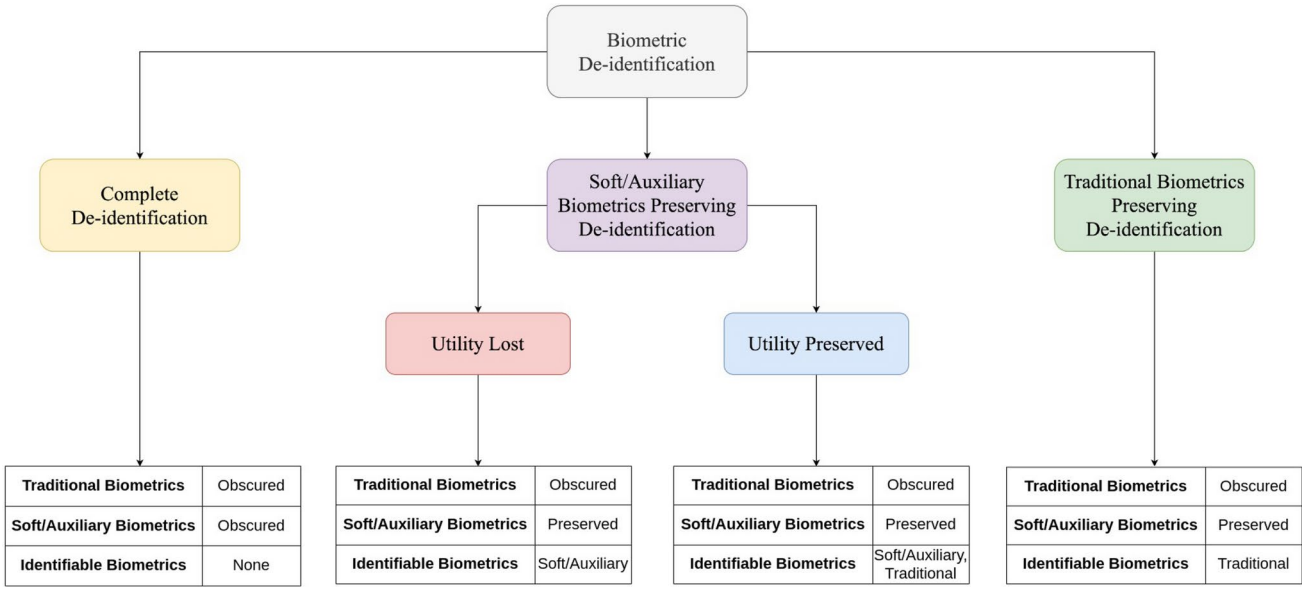


Fig. 1 Categorization of different de-identification methods [20]

Fig. 2 Traditional de-identification method on GTAV dataset [23]



distorted appearances. Additionally, they fail to preserve soft biometric attributes, which are crucial for biometric analysis.

Several recent studies have proposed neural network-based methods for face de-identification, leveraging deep learning to enhance de-identification performance while preserving identity-protecting features. These methods employ various techniques to generate images that obscure identity while maintaining emotional content. In [2], the authors combined a Gaussian Mixture Model (GMM)-based background subtraction algorithm with an improved version of the GrabCut algorithm to detect and segment pedestrians. Faces were detected separately using a standard face detector, and a neural art algorithm was applied to obfuscate them by blending styles from reference images. Sun et al. [22] proposed a head inpainting technique for privacy protection in online photo sharing, using facial landmarks and surrounding image context to generate realistic de-identified faces. Ren et al. [15] introduced an adversarial training-based approach for video face anonymization, which removed privacy-sensitive facial information while maintaining spatial action detection performance. Khorzooghi et al. [9] employed a style-mixing approach with generative adversarial networks (GANs), reporting emotion preservation accuracy of 52 and 60% using CIAGAN and Deep Privacy, respectively. Shahbaz et al. [17] utilized an image processing-based method for emotion-preserving de-identification, generating a proxy set of faces using StyleGAN and later using these proxy faces to obfuscate the target face.

This work introduces a novel GAN-based method for emotion-preserving face de-identification. Existing approaches often struggle to maintain facial emotions, particularly in non-frontal images or in cases of occlusion. The proposed method addresses these limitations by generating a proxy set of anonymous faces using StyleGAN. By mapping an input face to a corresponding proxy face with similar emotions, the method achieves effective de-identification while preserving emotional expressions. Key advantages include the use of StyleGAN to produce a diverse and realistic synthetic face dataset and a novel loss function designed explicitly for emotion preservation. Additionally, this approach demonstrates robustness in handling non-frontal face images and subjects with wearable accessories.

3 Proposed method

This research proposes a novel methodology for emotion-preserving face de-identification using Generative Adversarial Networks (GANs). The proposed method involves

several steps: (1) generating a synthetic face dataset using StyleGAN [8], (2) classifying the gender of the target image and selecting a gender-matched proxy image using DeepFace [16], and (3) applying the Enhanced SimSwap architecture to perform face swapping while preserving emotional expressions. The proposed method is depicted in Fig. 3.

3.1 SimSwap framework and its enhancements

The proposed method builds on the SimSwap [4] architecture, a state-of-the-art face-swapping framework designed for high-fidelity face swapping while preserving target face attributes. To address its limitations and improve performance, we introduce key enhancements, including an attention mechanism and an Attribute Preservation (AP) Loss. Below, we describe the original SimSwap framework and its enhanced version in a unified structure.

3.1.1 Core components of SimSwap

The SimSwap architecture consists of three main components: an encoder, a decoder, and a discriminator. The encoder extracts facial features from the input image using convolutional layers and activation functions. These features are passed to the Identity Injection Module (IIM), which transfers the identity of a source face to a target face while maintaining attribute information. The decoder generates the final swapped image, and the discriminator distinguishes between real and fake images to guide the adversarial training process.

Identity injection module (IIM)

The IIM is the core component of SimSwap. It comprises two parts: identity extraction and identity embedding. The identity extraction part uses modified residual blocks with adaptive instance normalization (AdaIN) to extract identity features from the source face. The identity embedding part injects these features into the target face while preserving its attributes, such as emotional expressions. This process ensures that the swapped image retains the identity of the source face while maintaining the attributes of the target face.

Loss functions The SimSwap architecture is trained using five loss functions to ensure high-quality face swapping:

- *Identity Loss* [24] Ensures that the swapped face retains the identity of the source image. It is calculated as the cosine similarity between the identity vectors of the source and generated faces:

$$L_{id} = \frac{1}{N}\sum_{i=1}^{N}\cos(\theta_i) \tag{1}$$

where N is the number of image pairs, and θ_i is the cosine similarity between the i-th pair.

- *Reconstruction Loss* Ensures that the generated face is similar to the target face. It is computed as the L1 distance between the generated and target images:

$$L_{rec} = ||G(x_s, z) - x_t||_1 \tag{2}$$

where $G(x_s, z)$ is the generated image, and x_t is the target image.

- *Adversarial loss* Encourages the generator to produce realistic images that can fool the discriminator. It uses the Hinge loss formulation for stable training.

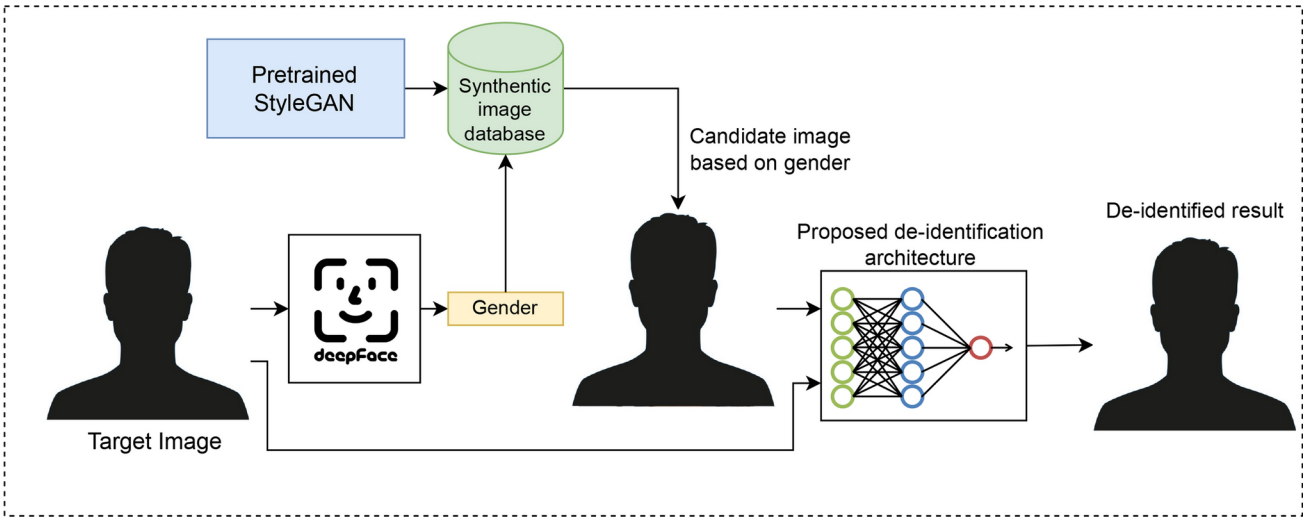


Fig. 3 Overall flow of the proposed emotion-preserving face de-identification system

- *Gradient penalty* Regularizes the discriminator to prevent large gradients, ensuring stable adversarial training.
- *Weak feature matching loss* Ensures that the generated image matches the target image in terms of low-level features. It is calculated as the L1 distance between features extracted by the discriminator from real and fake images.

3.1.2 Enhancements to SimSwap

To improve the performance of SimSwap, we introduce two key enhancements: (1) an attention mechanism in the Identity Injection Module (IIM) for better identity transfer control, and (2) an Attribute Preservation (AP) Loss to preserve non-identity attributes such as emotional expressions.

Attention mechanism in IIM The attention mechanism enhances the IIM by enabling selective focus on crucial facial regions during identity transfer. It generates attention maps that highlight significant regions in both the source and target faces. These maps are used to compute weights for each spatial location in the feature maps, ensuring that important identity-related regions are emphasized during the fusion of source and target features. The attention-guided features are computed as:

$$F_{\text{att}} = F \otimes A \tag{3}$$

where F represents the original facial features, A is the attention map, and $\otimes$ denotes element-wise multiplication. The attention maps A are generated using a softmax function applied to spatial attention weights W, which are obtained through convolutional layers:

$$W = \text{softmax}(W_{\text{conv}}(F)) \tag{4}$$

This mechanism ensures that identity information is injected more accurately, reducing artifacts and distortions in the swapped image.

Attribute preservation loss The Attribute Preservation (AP) Loss is introduced to ensure that non-identity attributes, such as emotional expressions, are preserved in the de-identified image. It is defined as the difference between the classification losses of the original and de-identified images. Both images are passed through a pretrained DeepFace model, and the cross-entropy loss is computed for their predicted emotional expressions. The AP Loss is given by:

$$L_{attr} = \sum_{i=1}^{N_{attr}} [H(y_i, f_i(\hat{x})) - H(y_i, f_i(x))] \tag{5}$$

where N_{attr} is the number of attributes, y_i is the ground truth label for attribute i, x is the original image, $\hat{x}$ is the de-identified image, $f_i(\cdot)$ is the classifier for attribute i, and $H(\cdot)$ is the cross-entropy loss. Minimizing this loss ensures that the de-identified image retains the same emotional expressions as the original.

3.2 Training the enhanced SimSwap architecture

The proposed Enhanced SimSwap architecture was trained on the VGGFace2 dataset using six loss functions (the original five from SimSwap and the new AP Loss). The model was implemented in PyTorch and trained on an Nvidia GeForce GTX-2080Ti GPU. Images smaller than 250 $\times$ 250 were removed to ensure high-quality training data, and the remaining images were aligned and cropped to a size of 224$\times$224. The face recognition model in the IIM was initialized using a pretrained Arcface [5] model. The Adam optimizer was used with beta1 = 0 and beta2 = 0.999. Training was performed for over 500 epochs, with separate batches for image pairs of the same identity and different identities (Figs. 4 and 5).

By incorporating the attention mechanism and the Attribute Preservation Loss, the Enhanced SimSwap architecture achieves improved control over identity transfer and better preservation of non-identity attributes. These enhancements result in high-quality de-identified images that retain emotional expressions while effectively concealing the original identity.

4 Experimental results

4.1 Dataset

Three different datasets were used for training and validation. The VGGFace2 dataset [3] was used to train the proposed architecture. It is a large-scale face recognition dataset containing over 3.31 million images of 9,131 subjects, covering a diverse range of poses, ages, and ethnicities. The Celeb-A dataset [11] was used to evaluate the proposed model. This dataset includes over 200,000 images of celebrities, each annotated with 40 different facial attributes such as hair color, facial hair, age, and gender. Additionally, the Labeled Faces in the Wild (LFW) dataset [7] was used for further evaluation. LFW is a widely used benchmark dataset for face recognition, containing over 13,000 face images collected from the web, each labeled with the corresponding person's name.

The CELEB-A and Labeled Faces in the Wild (LFW) datasets were used to evaluate the proposed architecture's performance. The evaluation process was carried out in two

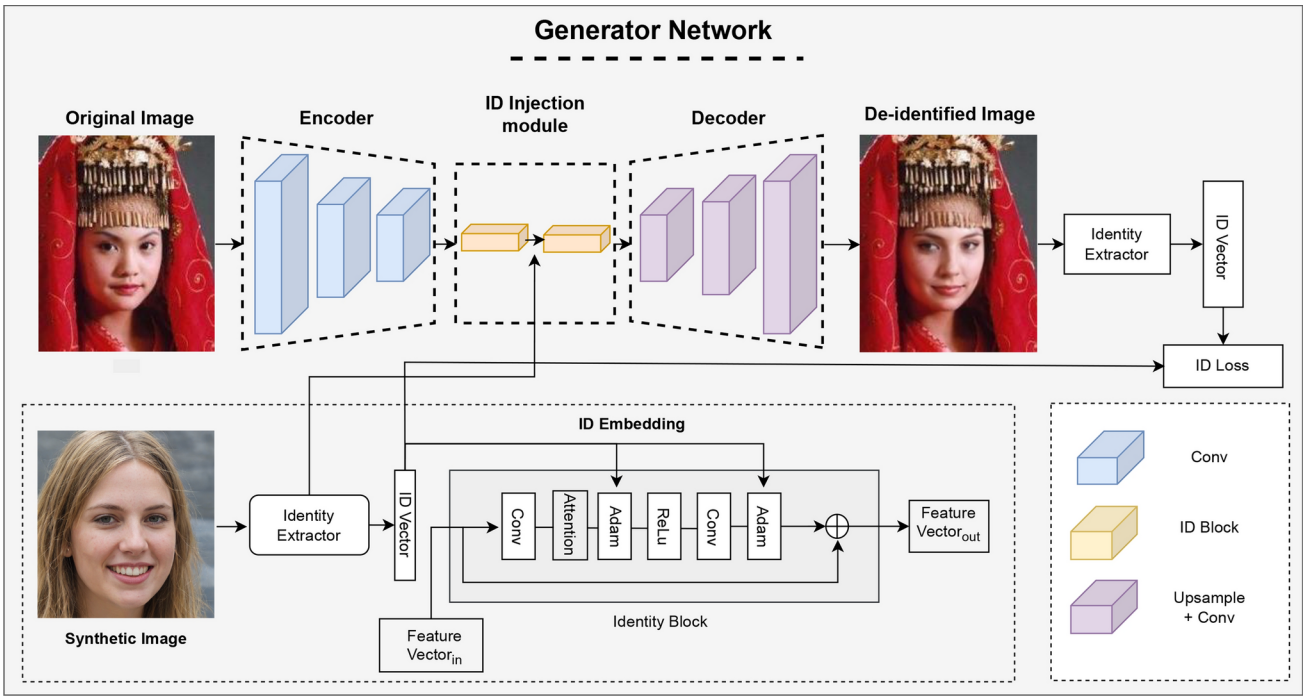


Fig. 4 The generator architecture of the proposed system

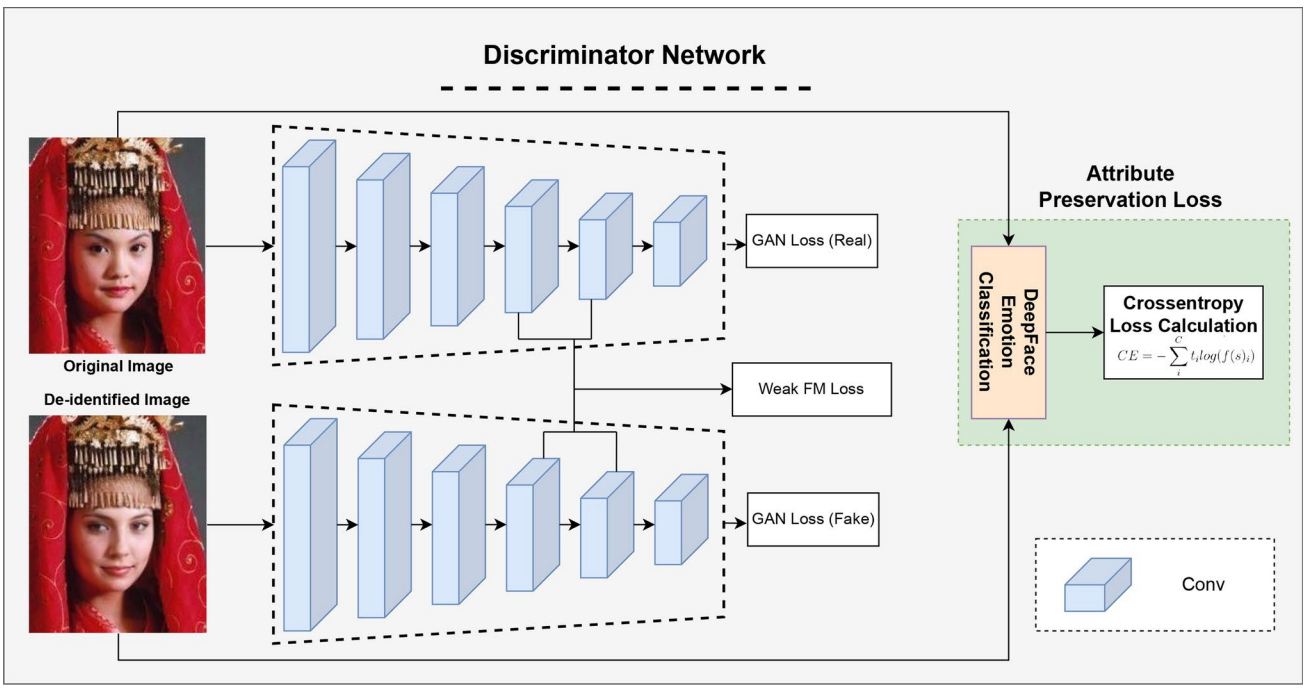


Fig. 5 The discriminator architecture of the proposed GAN

parts: first, evaluating the de-identification accuracy, and second, assessing the emotion preservation accuracy. In the first part, a candidate image was passed through the trained model to generate a de-identified result, and the de-identification performance was measured using accuracy as the evaluation metric. In the second part, the DeepFace library was employed to extract facial features containing identity and attribute information, which were then used to assess emotion preservation accuracy. The produced results were used to evaluate the emotion preservation performance. The proposed method was compared with the baseline SimSwap architecture. The results demonstrate the effectiveness of

the attribute preservation loss function introduced in this work (Fig. 6).

The goal was to measure the ability of the proposed architecture to correctly de-identify and preserve emotions. A test set was created using all the images in the CELEB-A and Labeled Faces in the Wild dataset.

4.2 Evaluation metrics

To evaluate the effectiveness of the proposed emotion-preserving face de-identification method, we employ a face matching-based approach to assess de-identification accuracy. This approach involves comparing the original and de-identified images using a face matcher to determine whether the de-identification has been successfully achieved. The evaluation metrics and methodology are outlined in detail below.

4.2.1 De-identification accuracy

De-identification accuracy is measured by evaluating whether the de-identified image can still be recognized as the original identity. This is done using a face matcher, which compares the original and de-identified images. If the face matcher fails to identify the de-identified image as the same person as the original image, the de-identification is considered successful. The de-identification accuracy is then calculated as the ratio of successfully de-identified samples to the total number of samples:

$$\text{De-Id Accuracy} = \frac{\#\text{ of Successfully De-Id Samples}}{\text{Total Number of Samples}} \times 100\% \tag{6}$$

This metric provides a quantitative measure of the system's ability to conceal the original identity while preserving non-identity attributes such as emotional expressions.



Fig. 6 Generated de-identified faces preserving original emotions

4.2.2 Face matching process

The face matching process is performed using a pretrained face recognition model, such as ArcFace [5] or DeepFace [16]. The steps are as follows:

1. *Feature Extraction* The face recognition model extracts feature vectors from both the original and de-identified images.
2. *Similarity Calculation* The cosine similarity between the feature vectors of the original and de-identified images is computed. The cosine similarity ranges from −1 to 1, where 1 indicates identical faces and −1 indicates completely dissimilar faces.
3. *Thresholding* A predefined similarity threshold τ is used to determine whether the de-identified image is recognized as the same person as the original image. If the cosine similarity is below τ, the de-identification is considered successful.

4.2.3 Emotion preservation evaluation

In addition to de-identification accuracy, we evaluate the system's ability to preserve emotional expressions in the de-identified images. This is done using a pretrained emotion recognition model, such as DeepFace. The emotion preservation score is calculated as the percentage of de-identified images that retain the same emotional expression as the original image:

$$\text{Emotion Preservation Score} = \frac{\#\text{ of Images with Preserved Emotions}}{\text{Total Number of Samples}} \times 100\% \quad (7)$$

4.2.4 Overall evaluation framework

The overall evaluation framework combines de-identification accuracy and emotion preservation score to provide a comprehensive assessment of the proposed method. A successful de-identification system must achieve high de-identification accuracy while maintaining a high emotion preservation score. These metrics are reported for both the original SimSwap architecture and the proposed Enhanced SimSwap architecture to demonstrate the improvements brought by the attention mechanism and Attribute Preservation Loss.

4.3 Experimentation and analysis

De-identification accuracy refers to the effectiveness of a method in removing personally identifiable information from an image. The higher the accuracy, the better the method is. Table 1 demonstrates the de-identification accuracy of two popular face recognition datasets, CELEB-A and LFW. CELEB-A dataset has 202,599 images with a de-identification accuracy of 93.37%, whereas LFW dataset has 13,000 images with a de-identification accuracy of 96.89

Table 1 De-identification accuracy of CELEB-A and LFW datasets

Dataset	Number of images	De-identification accuracy (%)
CELEB-A	202,599	93.37
LFW	13,000	96.89

Table 2 Emotion preservation accuracy of CELEB-A and LFW datasets

Dataset	Number of images	Emotion preservation accuracy (Rank-1) (%)	Emotion preservation accuracy (Rank-2) (%)
CELEB-A	202,599	72.32	86.44
LFW	13,000	78.18	90.31

Table 3 Cross-section of De-identification and emotion preservation accuracy

Dataset	Number of images	Cross-section result (%)
CELEB-A	202,599	87.21
LFW	13,000	88.92

Emotion preservation accuracy measures how well a system can maintain the emotional expression of a person's face, even when subjected to various transformations or distortions. The higher the accuracy, the better the system is at preserving the original emotional expression of the individual in the image. Table 2 shows the emotion preservation accuracy of the CELEB-A dataset, which contains 202,599 facial images. The emotion preservation accuracy is reported for the top two ranks, with the accuracy for Rank-1 being 72.32% and for Rank-2 being 86.44%.

We further conducted an additional cross-sectional study to highlight that our proposed method effectively manages both de-identification and emotion preservation simultaneously. This extra experiment underscores our approach's ability to safeguard personal identity while also maintaining the emotional content in images. Table 3 provides a concise summary of the cross-section results for de-identification and emotion preservation accuracy in the CELEB-A and LFW datasets. In the CELEB-A dataset, consisting of 202,599 images, the cross-section result is 87.21%, representing an aggregated measure of performance in both de-identification and emotion preservation tasks. Similarly, for the LFW dataset with 13,000 images, the cross-section result is 88.92%, indicating the overall effectiveness of the applied methods in simultaneously achieving high accuracy in de-identification and emotion preservation. These percentages serve as comprehensive metrics that encapsulate the combined success of the proposed method in maintaining

privacy and preserving emotional content within the given image datasets. Figure 7 shows some of the sample de-identified images where both de-identification was performed and emotion was preserved.

Figure 6 displays three images side by side: the original image, a synthetic image generated using StyleGAN to replace the person's face, and the de-identified image with identifying features removed. The de-identified images look natural and effectively preserve the original face's structure while concealing the person's identity. The results demonstrate the approach's robustness in preserving emotional expressions and privacy in challenging scenarios. This versatility makes it suitable for real-world applications, enhancing privacy protection without compromising emotional expression preservation.

To demonstrate the effectiveness of the attribute preservation loss in the proposed method, we compared our work with the baseline architecture. Figure 8 presents a comparison of de-identification, emotion preservation, and cross-sectional accuracy. The results show that the proposed method achieved higher de-identification accuracy on the CELEB-A and LFW datasets. Specifically, the proposed method attained 9.44% improvement in Rank-1 accuracy and 16.29% improvement in Rank-2 accuracy for the CELEB-A dataset. For the LFW dataset, the proposed method outperformed the baseline with 11.84% higher Rank-1 accuracy and 11.34% higher Rank-2 accuracy. These results demonstrate the superior performance of the proposed attribute preservation loss.

We also compared our method with the state-of-the-art facial de-identification system EPIC [17] (see Table 4). Both methods were evaluated using the LightFace pretrained model. For de-identification accuracy, our proposed method achieved a slightly higher accuracy of 97.54%. However, our method significantly outperformed EPIC in terms of emotion preservation accuracy, achieving 78.32%, compared to 57% reported by their method. This substantial improvement in emotion preservation highlights the effectiveness of our attention-guided Identity Injection Module (IIM) and attribute preservation loss function, both of which contribute to better emotion retention during the de-identification process. These results indicate that our proposed

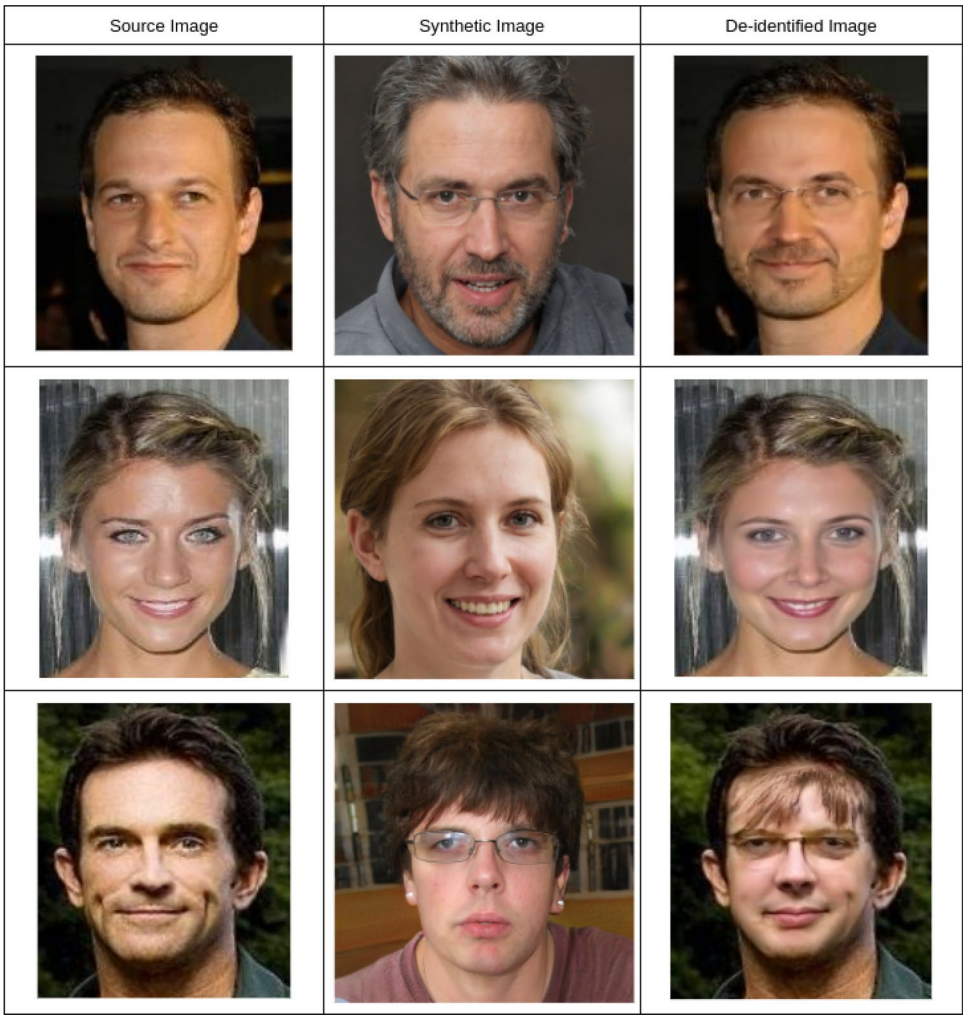
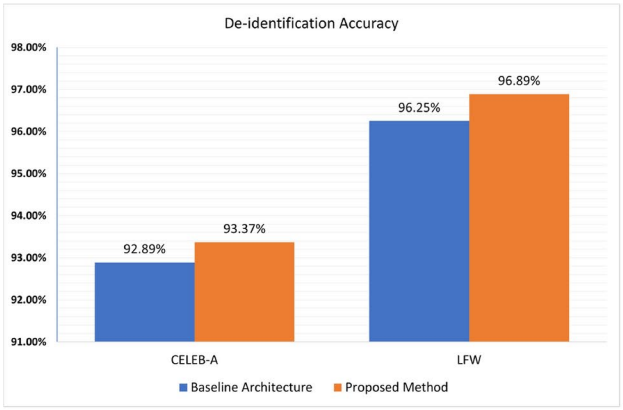
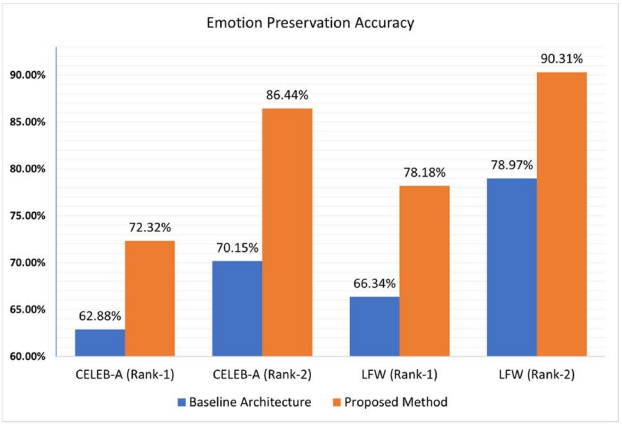


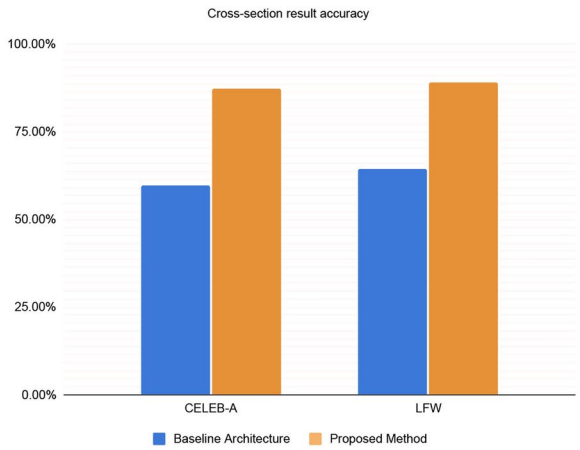
Fig. 7 Sample de-identified images where identification was properly obfuscated and emotion was preserved



(a) De-identification Accuracy.



(b) Emotion Preservation Accuracy.



(c) Cross-sectional Accuracy.

Fig. 8 De-identification and emotion preservation accuracy compared to the baseline architecture

Table 4 Comparison of results using LightFace pre-trained model

Method	De-identification accuracy	Emotion preservation accuracy
EPIC [17]	97.36	57.00
Proposed architecture	97.54	78.32

method excels in emotion preservation, making it well-suited for privacy-sensitive applications where maintaining emotional expressions in de-identified images is crucial (Table 5).

Table 5 Ablation study to demonstrate the effectiveness of attention IIM and attribute preservation loss function for de-identification and emotion-preservation

Dataset	Method	De-identification accuracy			Emotion-preservation accuracy		
		Components			Components		
CELEB-A	SimSwap	✓	✓	✓	✓	✓	✓
	Attention IIM	x	✓	✓	x	✓	✓
	AP loss	x	x	✓	x	x	✓
	Accuracy	81.26%	93.21%	93.37%	48.41%	52.17%	72.32%
LFW	SimSwap	✓	✓	✓	✓	✓	✓
	Attention IIM	x	✓	✓	x	✓	✓
	AP loss	x	x	✓	x	x	✓
	Accuracy	83.95%	95.75%	96.89%	51.63%	54.87%	78.18%

Table 6 Ablation study to demonstrate the effectiveness of attention IIM and attribute preservation loss function for de-identification and emotion-preservation

Dataset	Method	Cross-sectional results		
		Setup 1	Setup 2	Setup 3
CELEB-A	SimSwap	✓	✓	✓
	Attention IIM	x	✓	✓
	AP loss	x	x	✓
	Cross-sectional result	59.73%	75.31%	87.21%
LFW	SimSwap	✓	✓	✓
	Attention IIM	x	✓	✓
	AP loss	x	x	✓
	Cross-sectional result	64.29%	72.44%	88.92%

4.4 Ablation study

Two novel components proposed in this work are the attention-based Identity Injection Module (IIM) and the attribute preservation (AP) loss function. To demonstrate the effectiveness of these components, we performed an ablation study to evaluate de-identification and emotion preservation accuracy. Table 6 presents the results for de-identification and emotion preservation. When evaluating the performance on the CELEB-A dataset, the proposed method without the attention IIM and AP loss achieved an accuracy of 81.26%, which serves as the baseline performance. By incorporating the attention IIM into the SimSwap architecture while keeping the AP loss disabled, the accuracy improved to 93.21%. This highlights the crucial role of the attention mechanism in enhancing de-identification performance, particularly in handling non-frontal face views and facial accessories. Furthermore, when both the attention IIM and the AP loss function were incorporated, the accuracy further increased to 93.37%, demonstrating the complementary effect of those components. The attention mechanism improves the handling of non-frontal face views and accessories, while the AP loss function helps preserve attribute information during the de-identification process.

From Table 6, it can be observed for the CELEB-A dataset, the baseline performance of the proposed method achieved an accuracy of 48.41%. Incorporating the attention IIM into the SimSwap architecture led to an improvement to 52.17%. When both the attention IIM and AP loss function were included, the accuracy increased to 72.32%. This demonstrates that the combination of the attention mechanism and the AP loss function results in a substantial improvement in emotion preservation. The attention mechanism captures essential facial cues related to emotions, while the AP loss function ensures the retention of emotional expressions in the de-identified images. A similar pattern was observed for the LFW dataset.

Additionally, the provided ablation study in Table 6 assesses the impact of the attention IIM and attribute preservation loss function on the cross-sectional performance of the proposed method in de-identification and emotion preservation tasks. For the CELEB-A dataset, the baseline architecture, exhibits the best cross-sectional performance, achieving 59.73% accuracy. Attention IIM and AP loss progressively improve results, with attention IIM reaching 75.31 and 87.21% accuracy in Setups 2 and 3. When all components are included, the highest accuracy is achieved. For the LFW dataset, the SimSwap architecture attained 64.29% accuracy. The attention IIM shows a steady performance increase, reaching 72.44%, and with the AP loss function, similar to the CELEB-A dataset, the accuracy increases to 88.92%. These results collectively demonstrate the effectiveness of the proposed method and the positive influence of the attention IIM and AP loss function on cross-sectional performance.

5 Conclusion

This paper proposes a novel method for emotion-preserving face de-identification using Generative Adversarial Networks (GANs). The method leverages the StyleGAN architecture to generate synthetic face images for proxy face images and the DeepFace model to classify the gender of the target image for face swapping. The Enhanced SimSwap architecture is then employed to de-identify the original image while preserving the subject's emotions. A new loss function introduced within the SimSwap framework improves the emotion preservation quality during the face-swapping process, ensuring the retention of emotional expressions. Experimental results on various datasets demonstrate that the proposed method is effective for both frontal and side views of faces, handles facial occlusions, and achieves a high degree of de-identification while generating realistic de-identified facial images that retain the original emotions. The proposed architecture has potential applications in areas such as law enforcement, medical research, and social media, where face de-identification with emotional expression preservation is crucial. While the focus of this work is on preserving emotions, future research will explore the preservation of other critical identity-related features such as age and gender.

Acknowledgements The authors acknowledge the Natural Sciences and Engineering Research Council (NSERC) Discovery Grant funding, the Innovation for Defense Excellence and Security Network (IDEaS) and the UCalgary Research Excellence Chair Program for the partial funding of this project.

Author contributions Md Shopon and Marina L. Gavrilova jointly contributed to the conception, design, and execution of the research project leading to the presented manuscript. Md Shopon conducted the initial experiments, implemented the proposed method, and drafted

the manuscript. Throughout the project, Marina L. Gavrilova provided invaluable guidance, advising on experimental design, methodology, and data analysis techniques. Her expertise in the field of machine learning and computer vision significantly influenced the direction of the research and ensured its scientific rigor. Subsequently, Marina L. Gavrilova critically revised the manuscript, offering insightful feedback and suggestions for improvement. Both authors approved the final version of the manuscript and agreed to be accountable for all aspects of the work, ensuring its accuracy and integrity.

Data availability No datasets were generated or analysed during the current study.

Declarations

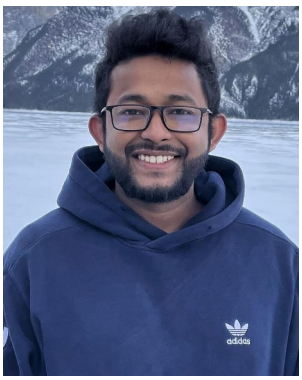
Conflict of interest The authors declare no competing interests.

References

1. Agarwal, A., Chattopadhyay, P., Wang, L.: Privacy preservation through facial de-identification with simultaneous emotion preservation. Signal, Image Video Process. **15**, 951–958 (2021)
2. Brkić, K., Hrkać, T., Sikirić, I., Kalafatić, Z.: Towards neural art-based face de-identification in video data. In: 2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE). pp. 1–5. IEEE (2016)
3. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
4. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: an efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2003–2011 (2020)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
6. Frome, A., Cheung, G., Abdulkader, A., Zennaro, M., Wu, B., Bissacco, A., Adam, H., Neven, H., Vincent, L.: Large-scale privacy protection in google street view. In: IEEE 12th International Conference on Computer Vision, pp. 2373–2380 (2009)
7. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: Detection, Alignment, and Recognition (2008)
8. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
9. Khorzooghi, S.M.S.M., Nilizadeh, S.: StyleGan as a utility-preserving face de-identification method. arXiv preprint arXiv:2212.02611 (2022)
10. Korshunov, P., Cai, S., Ebrahimi, T.: Crowdsourcing approach for evaluation of privacy filters in video surveillance. In: Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia, pp. 35–40 (2012)
11. Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celeb faces attributes (CELEBA) dataset. Retrieved August **15**(2018), 11 (2018)
12. Luo, Y., Gavrilova, M.L., Wang, P.S.: Facial metamorphosis using geometrical methods for biometric applications. Int. J. Pattern Recognit. Artif. Intell. **22**(03), 555–584 (2008)
13. Marcó-García, S., Ferrer-Quintero, M., Usall, J., Ochoa, S., Del Cacho, N., Huerta-Ramos, E.: Facial emotion recognition in neurological disorders: a narrative review. Revue Neurologique **69**(5), 207–219 (2019)
14. Paul, P.P., Gavrilova, M., Klimenko, S.: Situation awareness of cancelable biometric system. Vis. Comput. **30**, 1059–1067 (2014)
15. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 620–636 (2018)
16. Serengil, S.I., Ozpinar, A.: Hyperextended lightface: a facial attribute analysis framework. In: 2021 International Conference on Engineering and Emerging Technologies (ICEET), pp. 1–4. IEEE (2021)
17. Shahbaz Nejad, B., Roch, P., Handte, M., Marrón, P.J.: Enhancing privacy in computer vision applications: an emotion preserving approach to obfuscate faces. In: International Symposium on Visual Computing, pp. 80–90. Springer (2022)
18. Shopon, M., Bari, H., Bhatia, Y., Narayanaswamy, P.K., Tumpa, S.N., Sieu, B., Gavrilova, M.: Biometric system de-identification: concepts, applications, and open problems. Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects pp. 393–422 (2022)
19. Shopon, M., Gavrilova, M.L.: I got your emotion: Emotion preserving face de-identification using injection-based generative adversarial networks. In: International Symposium on Visual Computing, pp. 417–430. Springer (2023)
20. Shopon, M., Tumpa, S.N., Bhatia, Y., Kumar, K.P., Gavrilova, M.L.: Biometric systems de-identification: current advancements and future directions. J. Cybersecur. Priv. **1**(3), 470–495 (2021)
21. Sohn, H., De Neve, W., Ro, Y.M.: Privacy protection in video surveillance systems: analysis of subband-adaptive scrambling in JPEG XR. IEEE Trans. Circuits Syst. Video Technol. **21**(2), 170–177 (2011)
22. Sun, Q., Ma, L., Oh, S.J., Van Gool, L., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5050–5059 (2018)
23. Tarrés, F.: Gtav face database. http:gps-tsc.upc.es/GTAV/ResearchAreas/GTAVDatabase.htm
24. Wang, Z., Tang, X., Luo, W., Gao, S.: Face aging with identity-preserved conditional generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7939–7947 (2018)
25. Wu, Y., Yang, F., Xu, Y., Ling, H.: Privacy-protective-gan for privacy preserving face de-identification. Comput. Sci. Technol. **34**, 47–60 (2019)
26. Zarsky, T.Z.: Incompatible: the GDPR in the age of big data. Seton Hall Law Rev. **47**, 995 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Md Shopon received his B.Sc. in Computer Science and Engineering from the University of Asia Pacific in 2018. He worked as a Lecturer at the University of Asia Pacific from September 2018 to September 2020. He completed his M.Sc. in Computer Science at the University of Calgary in 2022 and is currently pursuing a Ph.D. in Computer Science at the University of Calgary under the supervision of Professor Marina L. Gavrilova. His research interests include computer vision, deep learning, and biometric privacy. He has authored more than 26 international publications in journals and conferences.



Marina L. Gavrilova (Senior Member, IEEE) is currently a Full Professor with the University of Calgary. She is an international expert in the areas of biometric security, machine learning, pattern recognition, and information fusion. She directs the Biometric Technologies Laboratory. She has published more than 300 books, conference proceedings, and peer-reviewed articles. Her professional excellence was recognized by Canada Foundation for Innovation, the Killiam Foundation, the U Make a Difference Award, and the Order of the University of Calgary. She is the Founding Editor-in-Chief of Transactions on Computational Sciences (Springer) and serves on the editorial boards for IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SCIENCES, IEEE ACCESS, The Visual Computer, Sensors, and International Journal of Biometrics.