applied sciences

MDPI

Article

Face De-Identification Using Convolutional Neural Network (CNN) Models for Visual-Copy Detection

Jinha Song, Juntae Kim and Jongho Nang *

Department of Computer Science and Engineering, Sogang University, Seoul 04107, Republic of Korea; jinhasong@sogang.ac.kr (J.S.); jtkim1211@sogang.ac.kr (J.K.)
* Correspondence: jhnang@sogang.ac.kr

Abstract: The proliferation of media-sharing platforms has led to issues with illegally edited content and the distribution of pornography. To protect personal information, de-identification technologies are being developed to prevent facial identification. Existing de-identification methods directly alter the pixel values in the face region, leading to reduced feature representation and identification accuracy. This study aims to develop a method that minimizes the possibility of personal identification while effectively preserving important features for image- and video-copy-detection tasks, proposing a new deep-learning-based de-identification approach that surpasses traditional pixel-based alteration methods. We introduce two de-identification models using different approaches: one emphasizing the contours of the original face through feature inversion and the other generating a blurred version of the face using D2GAN (Dual Discriminator Generative Adversarial Network). Both models were evaluated on their performance in image- and video-copy-detection tasks before and after de-identification, demonstrating effective feature preservation. This research presents new possibilities for personal-information protection and digital-content security, contributing to digital-rights management and law enforcement.

Keywords: de-identification; CNN; feature inversion; GAN; D2GAN; image-copy detection; video-copy detection

Citation: Song, J.; Kim, J.; Nang, J. Face De-Identification Using Convolutional Neural Network (CNN) Models for Visual-Copy Detection. *Appl. Sci.* **2024**, *14*, 1771. https://doi.org/10.3390/app14051771

Academic Editor: Andrea Prati

Received: 5 January 2024
Revised: 13 February 2024
Accepted: 19 February 2024
Published: 21 February 2024

1. Introduction

The surge in media-sharing platforms has increased the distribution of unauthorized and explicit content, raising privacy concerns. Government agencies and media-sharing platforms create databases of harmful images and videos for content moderation. While current methods use large-scale datasets for content filtering, a challenge arises when using deep learning: using facial data raises ethical issues around privacy. Thus, there is a need for a novel approach that constructs training datasets by eliminating facial information, ensuring privacy preservation while minimizing information loss, and facilitating the development of robust models to address these concerns. To address this challenge, our study focuses on facial de-identification, with the aim of protecting privacy without compromising data utility. Traditional methods [1–5] typically use pixel-level transformations, such as average filters, median filters, or Gaussian filters, to obscure the facial regions. However, these pixel-level transformations often compromise the balance between preserving the utility of data and protecting individual privacy. While they effectively obscure facial features to address privacy concerns, this direct manipulation of pixel values can significantly degrade the quality of the dataset, resulting in a loss of critical-feature representation. This degradation not only diminishes the effectiveness of subsequent analytical models but also raises questions about the adequacy of privacy protection, as overly blurred or distorted images may still leave room for re-identification through advanced techniques or contextual information.

We propose a deep-learning-based de-identification approach that preserves the features of input images while anonymizing facial regions in image- and video-copy-detection tasks. This paper focuses on facial de-identification, aiming for minimal data-information loss and for personal-information protection. To obscure facial regions, we employ (1) Feature Inversion [6,7] and (2) D2GAN [8] techniques: (1) The feature-inversion-based method emphasizes the contours of the original face. It uses features extracted from the intermediate layers of a Convolutional Neural Network (CNN) model, focusing on the facial region of the input image, to reconstruct the original input data. During reconstruction, features are extracted from both the original and de-identified facial images, and their similarity is integrated into the loss function. This allows the model to learn to generate images with similar features after de-identification. The feature-inversion method was selected due to its proficiency in reconstructing images with emphasis on contours and significant features, observed particularly in natural images, where it promotes feature representation based on contours. This aligns with the goal of preserving crucial elements while de-identifying faces. (2) The D2GAN-based method blurs the face region. It uses a generator and two discriminators to de-identify the facial region. One discriminator distinguishes whether the generated face image is real or fake, and the other determines whether the generated face image is original or de-identified. D2GAN was opted for due to its distinctive approach to handling the ambiguity in the direction of image transformation, common in typical GANs when presented with facial images. The dual-discriminator structure of D2GAN effectively addresses this by directing the generation towards producing blurred facial images, which ensures a substantial degree of anonymity and maintains feature similarity. This approach also uses the same loss function as the feature-inversion method, enabling the model to learn to generate images with similar features after de-identification. Both models demonstrate high feature-similarity performance after de-identification, confirming that image features can be preserved without loss even through the de-identification process. Furthermore, the results of the image- and video-copy-detection experiments using our proposed de-identification method showed that, despite significant changes in the face region, the performance before and after de-identification was almost identical. This demonstrates the effectiveness of our proposed method in generating privacy-protecting images and suggests its potential for broad applicability in various tasks, including copy-detection tasks. Building on the foundation laid in the introduction, this paper delves into the specifics of CNN-based de-identification models for image- and video-copy-detection tasks.

Section 2 reviews the current state of de-identification techniques, image-generation methodologies, and task-related research. Section 3 outlines the application scenarios and provides a detailed description of the two proposed models. Section 4 evaluates the performance of de-identification models through metrics, datasets, experimental results, and a comparative analysis of performance in copy-detection tasks. Section 5 discusses the challenges encountered, potential vulnerabilities, and areas for enhancement in the models. Finally, Section 6 concludes the paper by summarizing the research findings and highlighting the contributions of this work.

2. Related Work

In this section, we present the related work. Section 2.1 offers an overview of the existing research on de-identification, Section 2.2 describes the image-generation method for de-identification, and Section 2.3 introduces the research domains, where the model proposed in this paper has been applied.

2.1. De-Identification

De-identification primarily aims at privacy protection. It involves transforming or encrypting the original image to protect it from recognition systems, even if the given image appears identical, thereby minimizing identity information leakage due to the image. In [9], a framework for facial-image de-identification through adversarial perturbations in

feature space was proposed, demonstrating the conversion of original images to substitute images. Another study with similar objectives was [10], which proposed a natural way to protect identity through strong 3D prior information and delicate generation design by using a "divide and conquer" strategy to train a GAN with adjusted loss to hide 3D separated identity codes and preserve image utility. And study [11] protected identity information by encrypting and decrypting facial-identity information in latent space based on the StyleGAN2 [12] generator. The advantage of this method is that it provides strong privacy protection by transforming an existing image into an alternative image while remaining cosmetically identical to the original image. The disadvantage is that adversarial transformations may only partially bypass the recognition system in certain situations. This technique benefits image sharing in online spaces where privacy is essential. The authors in [13] proposed a facial de-identification system using Conditional GAN [14] to generate realistic faces, satisfying various facial features (e.g., gender, hair color, and facial shape). A similar study by Li et al. [15], used a facial-attribute-transfer model to change the face while maintaining the natural appearance of the de-identified face through an encoder and decoder neural network. Additionally, study [16] proposed a model that uses a StyleGAN [17] to blend the styles or features of the target face and the auxiliary face into a natural-looking face. It has the advantage of vigorously protecting user identity while allowing re-identification back to the original facial image. However, it can be limited in generating highly controlled features and requires significant computational resources to train the model. Applicable scenarios include protecting personal identities in public databases, such as social media. In [18], a method for generating adversarial identity masks to hide identity from recognition systems was proposed. The approach presented there protects faces from facial recognition systems by overlaying a mask on the image, ensuring no change in the image's appearance. This method is particularly suitable for security-critical environments where the identity of the individual needs to be protected. Similarly, the authors of [19] proposed a method to hide identity from facial-recognition systems through a diffusion model. This technique has the advantage of effectively anonymizing identifying information while maintaining the natural appearance of the image, among other things. Applicable scenarios include various digital platforms that require the protection of an individual's identity, such as photo sharing. De-identification research has also been conducted in the medical field. In [20], a method was presented that combines the local-differential-privacy algorithm with GLOW, a flow-based deep generative model, to protect personal identification information in medical images. This method has the advantage of enhancing patient privacy in healthcare, but consideration should be given to its scalability in other areas that require a high degree of data protection.

2.2. Image-Generation Method

Feature inversion is a method of reconstructing the original input data using features extracted from the intermediate layers of a trained CNN model. This is aimed at understanding the model's representation from various perspectives. Articles [6,7] focused on discerning the visual information encoded in each layer of the CNN and identifying information through feature inversion. They demonstrated an optimization technique that reconstructs the image closest to the activation of the layers while preserving the features of the original image as much as possible. GAN (Generative Adversarial Network) [21] is an image-generation model using two networks, a generator and a discriminator. Due to the diversity in input images, common GANs can suffer from mode collapse, where images are transformed only into certain types. To address this, D2GAN [8] was proposed. Among the two discriminators included in D2GAN, one differentiates between real and fake images like a regular GAN, while the other assigns high values to generated images and low values to real images, thereby solving the mode-collapse issue. In [22], the authors introduced a de-identification method using D2GAN for medical images. That paper presented a method to effectively fuse and emphasize critical information in medical images using ED-D2GAN. In our study, we propose a de-identification method utilizing

feature inversion and D2GAN. The method using feature inversion reconstructs the input facial image to produce an image that is visually unrecognizable, yet retains the necessary features for image- and video-copy-detection tasks. This is implemented by utilizing the intermediate-layer features of feature inversion. The method using D2GAN is trained to blur the image while maintaining important features. This method was inspired by the medical-image-fusion method in [22]. An additional discriminator is introduced to the traditional generator and discriminator structure, distinguishing between the blurred and original images, and learning to generate de-identified blurred images while maintaining the crucial features.

2.3. Visual-Copy Detection

Visual-Copy Detection refers to the process of identifying duplicates or copies within image and video content. This task is crucial for detecting unauthorized use or infringement of digital content and is divided based on the input format into 'Image-Copy Detection' and 'Video-Copy Detection'. These areas play a key role in protecting the originality and copyright of visual content. Image-copy detection is a research field that focuses on detecting the replication or manipulation of original images and identifying subtle differences between the original and altered images or finding illegally manipulated copies. The study in [23] proposes an image copy-move forgery-detection method based on fused features and density clustering. This method offers a novel approach to more precisely identifying replicated or manipulated parts of an image, integrating features of the duplicated image to accurately locate the forged area. The proposed method has proven effective in detecting unauthorized image replication and manipulation, emphasizing the utility of CNNs in this domain. In [24], a Self-Supervised Contrastive Descent (SSCD) model based on self-supervised contrastive training was proposed. This method altered the architecture and training objectives, incorporated pooling operators in the instance-matching domain, and applied contrast learning to augmentation for combining images for image-copy-detection tasks. Video-copy detection focuses on identifying and matching original video content and its various forms of alterations, which can include cropping parts of the video, adding logos, color changes, etc. This field primarily aims to prevent copyright infringement and detect illegally uploaded content. Video-copy detection typically involves stages of feature extraction, video matching, and validation. In [25], significant improvements were made in the processing speed and accuracy for video-copy detection and video-retrieval tasks by extracting crucial information from video data and selecting relevant information for optimal search results. The S^2VS [26] model leverages self-supervised learning to train the model in proxy tasks and transition to target tasks after fine tuning. A single universal model trained using the method proposed in this research achieved the state-of-the-art in video-copy-detection tasks. In this study, we compare and analyze image- and video-copy-detection performance before and after de-identification, based on the research in [24–26]. This comparison demonstrates the effectiveness of the proposed de-identification method in image- and video-copy-detection tasks and how it differs from the existing methodologies.

3. Proposed Methods

In this section, we propose two de-identification models using CNNs. In Section 3.1, we describe the overall scenario in which we want to apply the two models proposed in this paper. Section 3.2 describes the first model using feature inversion, and Section 3.3 discusses the second model using D2GAN. The learning results and performances of these models are discussed in Section 4.

3.1. Application Scenarios of the Proposed Models

In this section, we describe the application scenarios of the two de-identification methods proposed in this study. Figure 1 intuitively illustrates the complete scenario of applying the de-identification method. The figure visually illustrates how personally identifiable information within illegal content can be protected by de-identifying it before

it is shared over the network, especially on the client side, before it is delivered to media-sharing platforms and cloud storage via uploaders on the client side. On the client side, the device uploading the media detects the facial regions of the media through "Face Detection" before uploading the media, and only those regions are de-identified through "De-identification" and uploaded through "Media Uploader". On the server side, when the de-identified media are delivered through the "Media Reciever", features are extracted using the "Feature Extractor" to proceed with the visual-copy-detection process and are compared to the features in the database consisting of existing illegal-content features through the "Copy Detector", and illegal content is classified through the "Illegal Contents Classifier". The classification result is uploaded to the illegal-content database if it is illegal content or to the media-sharing platform or cloud storage where the client wants to upload it if it is legal content. Humans then review the newly uploaded content in the illegal-content database. Features are extracted through the "Feature Extractor" used for visual-copy detection and uploaded to the illegal-content-feature database, which is used to classify the newly uploaded content. This approach has the advantage of preserving personally identifiable information even if the content is intercepted during the upload process, protecting the integrity of the data while minimizing the potential spread of illegal content. Sections 3.2 and 3.3 introduce two models that can be applied to the scenarios described in this section.

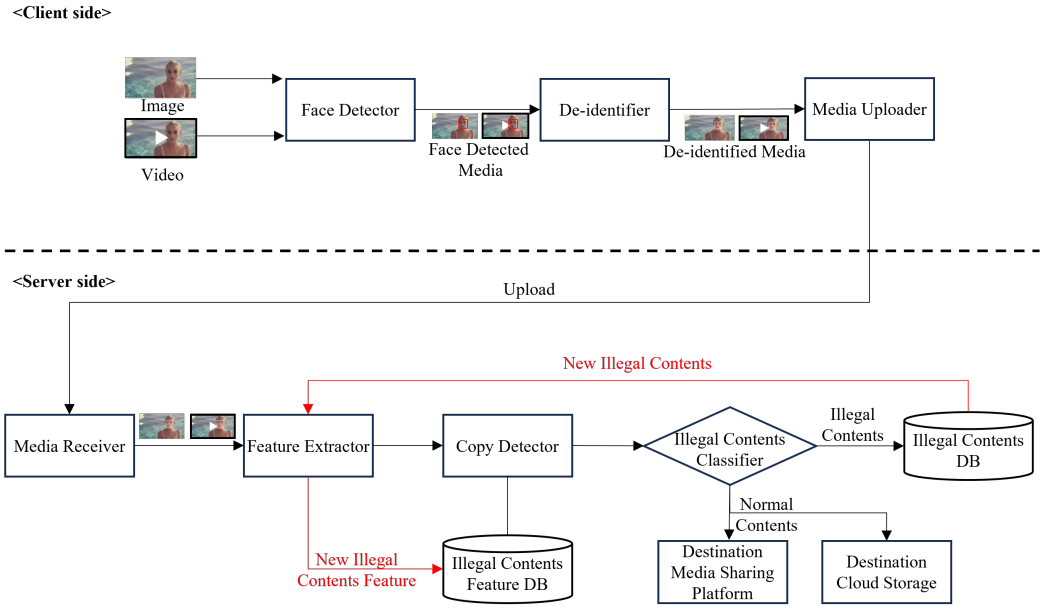


Figure 1. Overview of the de-identification-process scenario. This figure illustrates the steps whereby content is processed through a de-identification program before the user uploads media. The de-identified content is validated through an illegal-content-copy-detection model before being transmitted to the media-sharing platform, and if identified as illegal content the upload is blocked.

3.2. Face De-Identification Model Using Feature Inversion

In this section, we present our proposed de-identification model that employs feature inversion, as depicted in Figure 2. Feature inversion involves reconstructing the original input data from features extracted from the intermediate layer of a CNN model. Leveraging this concept, we build a de-identification model as follows. Initially, an image is selected from the dataset, and the face region is detected and cropped. Face detection employs the yolov7-face [27] model, an adaptation of yolov7 [28] specifically trained on the WiderFace [29] dataset. The cropped-face-region image undergoes feature inversion to reconstruct the de-identified face image. Following this step, the modified ResNet50 architecture, similar to the feature extractor, is utilized to implement the feature inversion model, effectively creating the de-identified face image. The architecture ensures the maintenance

of essential features from the original image while generating high-resolution de-identified images through an enhanced upsampling process. This module adjusts the channels of the skip connection using a 1×1 convolution and combines it with the upsampled feature map, thus facilitating the restoration of the original image size utilizing features extracted from the deep layers of the network. This process aims to preserve the feature similarity between the original and de-identified images, ultimately producing an image that maintains the crucial visual information of the original while effectively removing personal identification information. To guarantee that the reconstructed face-region-images differ, the mean square error (MSE) loss measures the similarity, which is integrated into the overall loss function. In addition, the similarity is assessed by extracting features from the original full image, encompassing the face region, and the de-identified full image with the replaced face region. For the image-copy-detection task, we employ the ResNet50 feature extractor as outlined in [24], and for video-copy detection, we use the ResNet50 feature extractor from [25,26]. The loss applied to the feature-inversion training of the face region is defined as follows:

$$L_{\text{total}} = \lambda_{RS} L_{RS} + \lambda_{FS} L_{FS} \tag{1}$$

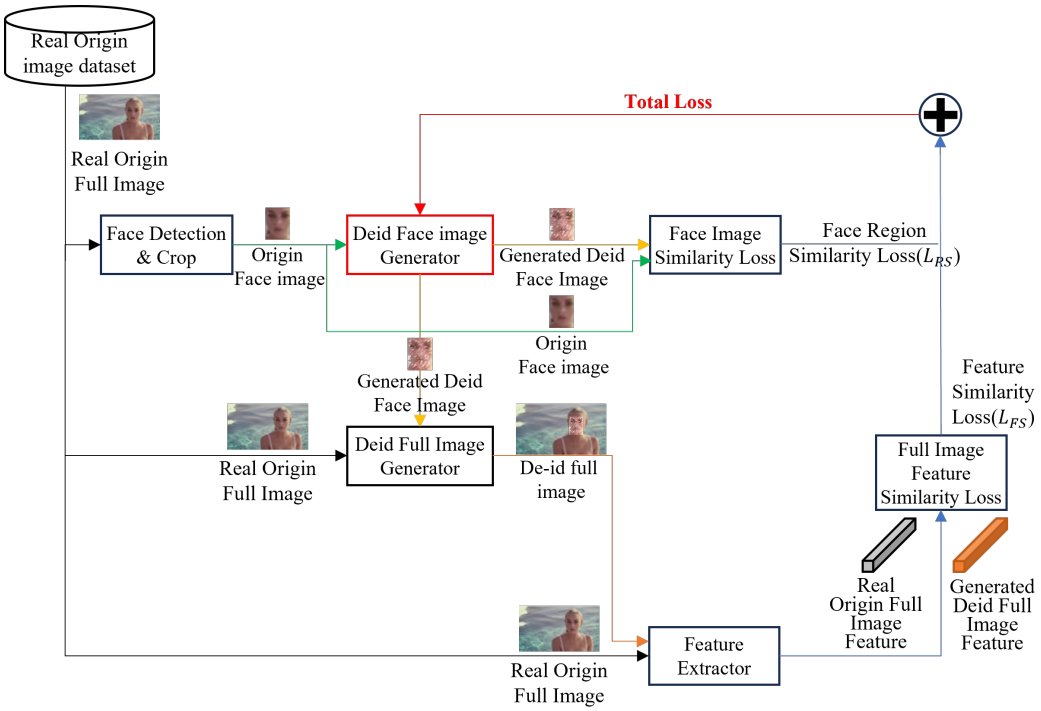


Figure 2. Model structure of face de-identification using feature inversion. After detecting and extracting facial regions from the original image, the de-identification generator de-identifies the face image and the entire image. The generated de-identified face image and the whole image are optimized through several loss functions to maintain similarity to the original, and this process finally minimizes the total loss, including the face-image-similarity loss, the face-region-similarity loss, and the whole-image-feature-similarity loss.

Equation (1) illustrates the overall formula, where L_{RS} represents the similarity to the face region and L_{FS} denotes the similarity of the features extracted from both the original and de-identified full images. Our model targets the dissimilarity in the face region after de-identification while seeking to maintain the feature similarity from the original full image, leading to the calculation of the final loss as described above. The expressions for each loss, L_{RS} and L_{FS}, are described in Equations (2) and (3). Equation (2) calculates the MSE loss between the original and the de-identified face in an image, where N represents the number of face regions in the image, O_i signifies the pixel value of the original face region, and D_i denotes the pixel value of the de-identified-face region:

$$L_{\mathrm{RS}} = \frac{1}{N}\sum_{i=1}^{N}(O_i - D_i)^2 \tag{2}$$

Equation (3) computes the cosine similarity of the features extracted from the original and de-identified full images, where M represents the number of dimensions of the feature vector, F_{O_j} is the feature vector of the original full image, and F_{D_j} is the feature vector of the de-identified image:

$$L_{FS} = 1 - \left(\frac{1}{M}\sum_{j=1}^{M}\frac{F_{O_j}\cdot F_{D_j}}{\left\|F_{O_j}\right\|\left\|F_{D_j}\right\|}\right) \tag{3}$$

A lower value of L_{RS} indicates a higher similarity between the original and the de-identified face. In our proposed model, to avoid similarity between the two faces, we set the value of λ_{RS} to be low during the learning process. Consequently, we aim for the features extracted from the original and de-identified images to exhibit similarity. Therefore, we set the value of λ_{FS} relatively high. Owing to this structure, the model's loss function enables the Deid Face Image Generator to train in a manner that ensures the feature vectors generated from the entire image and the de-identified face image are similar. At the same time, it reflects the similarity between the original and de-identified facial regions. Our proposed face de-identification using the feature-inversion model showcases its ability to maintain the utility of de-identified images while ensuring strong privacy protection. This method opens up promising avenues in the advancement of image-based privacy and security technologies. For the evaluation of the proposed model, we use several evaluation metrics to assess the performance of de-identified images and videos, including Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and cosine similarity. SSIM and PSNR are important for evaluating the visual similarity of facial regions before and after de-identification, while cosine similarity measures the similarity of features between the transformed image and the original image, to assess how effectively the model transformed the image to preserve privacy. For comparative analysis, this study compares performance using query images and videos to de-identification processing applied and performance using the original dataset without such processing. This allows us to quantitatively evaluate the impact of de-identification on image- and video-clone-detection tasks. Through experiments with the DISC21 and VCDB datasets, we show that even after de-identification, the model maintains its ability to detect clones while effectively protecting privacy. These evaluation methods and comparison criteria will help to provide a comprehensive understanding of how effectively models can protect privacy while maintaining the usefulness of the data.

3.3. Face-De-Identification Model using D2GAN

This section describes the de-identification model that employs the D2GAN model proposed in this study, as illustrated in Figure 3. In contrast to the feature-inversion model, which requires two sets of data—images and corresponding bounding-box labels—for training, face de-identification using the D2GAN model requires three components: the original image, the bounding-box labels, and images with the face regions blurred using a Gaussian filter. The original images are represented by x_{orin}, serving as a basis for learning how to generate de-identified versions and ensuring the model accurately captures and retains the essential visual information of each face. The generator G utilizes an input vector, which is denoted by z, from the latent space, to produce de-identified facial images. This allows for a diverse range of outputs from a single-input image through the manipulation of z. The D2GAN model is designed to mitigate mode collapse by using two discriminators. The first discriminator discerns whether the image is real or synthetic, whereas the second distinguishes between the original and the de-identified image, similar to a conventional GAN. The generator used in this context is the renowned ResNet generator network. The ResNet-based generator network is structured to transform input images into de-identified

versions while preserving essential features of the original image. It employs a series of downsampling, residual blocks, and upsampling layers. This architecture ensures the preservation of crucial visual information and maintains the integrity of the original image during the de-identification process. Equation (4) represents the loss function of the discriminator D_{RF} that determines whether an input image is real or fake. In this equation, G denotes a generator that creates de-identified face images, while D_{RF} signifies a discriminator that discerns whether an image is real or fake. The loss is updated by computing the operation as outlined in Equation (4):

$$L_{D_{RF}} = -(\mathbf{E}_{x_{orin}}[\log D_{RF}(x) + \mathbf{E}_z \log(1 - D_{RF}(G(z)))]) \quad (4)$$

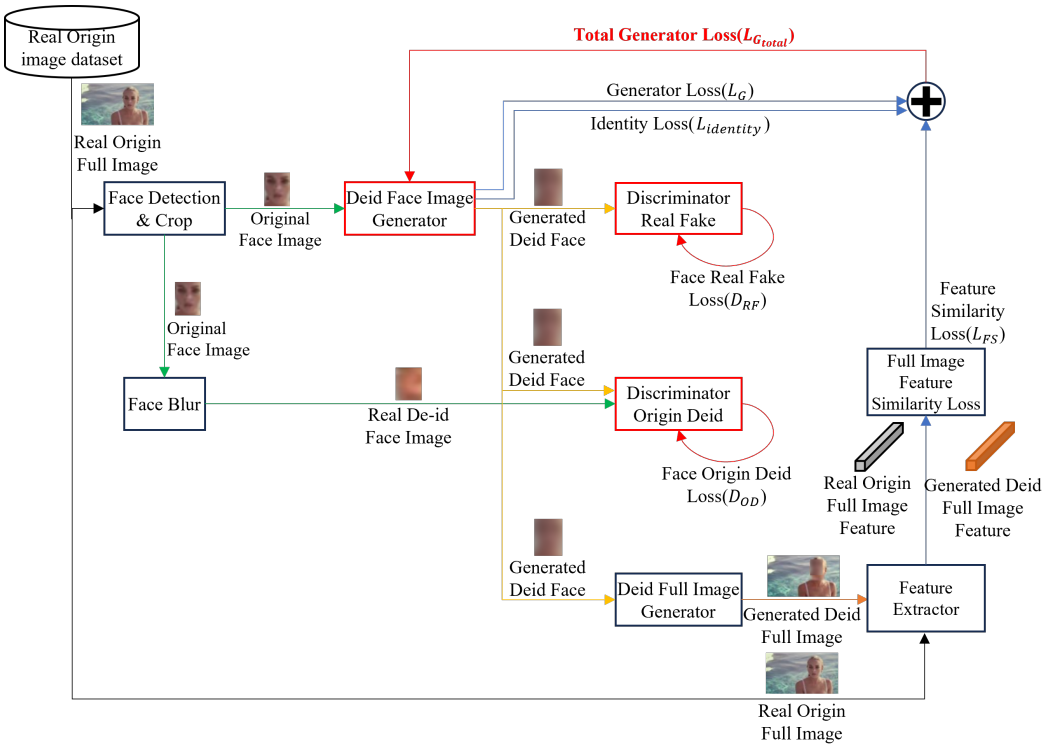


Figure 3. Model structure of face de-identification using D2GAN. This structure illustrates the stages of detecting and cropping faces from the original image, followed by a de-identification process that generates transformed facial images. The generated de-identified facial images are then overlaid onto the original full image to create the final de-identified image. The de-identification process utilizes two discriminators to distinguish between real and fake images and to assess the authenticity of the face region de-identification. The model is optimized to minimize a comprehensive loss function, which includes the losses from the two discriminators, the feature similarity between the original and de-identified full images, and the generator's loss.

Equation (5) depicts the loss function of the discriminator D_{OD} that distinguishes between the original and de-identified images. In this equation, similar to Equation (4), G denotes a generator that creates de-identified face images, while D_{OD} signifies a discriminator that discerns whether an image is original or de-identified. The de-identified images produced by the generator G are represented by z, and are used alongside original images to train D_{OD} in distinguishing between original and altered facial features. The loss is updated by computing the operation as outlined in Equation (5):

$$L_{D_{OD}} = -(\mathbf{E}_{x_{deid}}[\log D_{OD}(x) + \mathbf{E}_z \log(1 - D_{OD}(G(z)))]) \quad (5)$$

The Real vs. Fake Discriminator (D_{RF}) is tasked with differentiating between real face images and those generated by the generator. It plays a crucial role in assessing

the authenticity of the images produced during the training process. Additionally, the Original vs. De-identified Discriminator (D_{OD}) is trained to accurately distinguish between original and de-identified face images. This discriminator ensures that the de-identification process effectively obscures personal identifying features while retaining the essential characteristics of the original image.

Finally, Equation (6) represents the overall loss of the generator responsible for producing the de-identified image. An input image is represented by x_i, and y_i represents the input image blurred with a Gaussian filter. This loss utilizes a combination of Binary Cross-Entropy (BCE) Loss and Logistic Loss for the generator. The BCE Loss measures the discrepancy between the discriminator's prediction and the actual label, indicating how effectively the generator deceives the discriminator into perceiving the generated images as real. Meanwhile, the Logistic Loss, applied through a sigmoid activation function (σ), further refines the model's predictions, minimizing the difference between the predicted probabilities and the actual binary outcomes. Together, these two losses enhance the generator's ability to produce realistic and convincing de-identified images. Additionally, Identity Loss is considered, which calculates the L1 distance between the generated de-identified face and the original face, thus preserving essential characteristics while enhancing privacy. The final loss is calculated by incorporating each λ ratio. The feature-similarity loss is denoted by L_{FS} and uses Equation (3), as used in the feature-inversion model.

$$L_{G_{total}} = \lambda_{GAN} \times \left(-\sum_i [y_i \times \log \sigma(G(x_i)) + (1 - y_i) \times \log(1 - \sigma(G(x_i)))] \right) + \lambda_{Identity} \times \frac{1}{N} \sum_i |G(x_i) - x_i| + \lambda_{FS} L_{FS} \tag{6}$$

The design of our model generates images in a manner akin to Gaussian blurring, thereby maintaining the feature similarity of the overall image without significant compromise. This balanced approach ensures enhanced privacy while preserving the image's intended utility. The primary rationale for this configuration is to achieve an optimal balance between practicality and the necessity for privacy protection. We anticipate that this strategy will deliver substantial value in real-world applications. Similar to the models proposed in Section 3.2, we use several evaluation metrics to evaluate the performance of de-identified images and videos, including Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and cosine similarity. For comparative analysis, we also compare the performance using query images and videos with de-identification applied to them to the performance using the original dataset without such processing. This comparative analysis provides a deeper understanding of how the proposed models effectively preserve the key characteristics of the data required for the clone-detection task while enhancing privacy.

4. Experiments

This section presents both the quantitative and qualitative performance of the models described in Section 3, focusing on their application to both image- and video-copy detection.

4.1. Dataset

To train and evaluate the proposed models, we utilized datasets such as Widerface [29], VGGFace2-HQ [30], FFHQ [17], and CelebA-HQ [31], with detailed information presented in Table 1. According to Table 1, it is noted that the images in these datasets often contain multiple faces and not just a single one, suggesting that reconstructed datasets using these sources could enable training and evaluation for scenarios with both single and multiple faces. Recognizing the significance of the face area's proportion in images for feature representation post-de-identification, we found that larger face regions could significantly

alter the features. Table 2 shows the distribution of images based on the size of the face region across these datasets, indicating a bias in the proportion of faces within images. By reconstructing the training dataset to have a uniform distribution across various face-area ratios, we ensured consistent and effective model training regardless of the face region's ratio in the images. Advanced face-bounding-box detection was performed, using the yolov7-face [27] model, specifically the yolov7-w6+TTA variant, achieving 84fps on a V100 GPU. The dataset now includes a minimum of 15,000 images for each face-area ratio, providing a robust basis for evaluation.

Table 1. Result of the proposed models in de-identification and feature-similarity experiments.

Dataset	Widerface [29]	VGGFace2 -HQ [30]	FFHQ [17]	CelebA-HQ [31]
Number of Images	30,250	113,758	70,000	202,599
Number of Identities	-	903	-	10,177
Number of Faces	140,040	151,108	139,000	231,560
Avg Number of Faces per Image	4.63	1.33	1.99	1.14

Table 2. Training-dataset composition and proportion of face region in images.

Proportion of Face Region in Images	Dataset				
	Widerface [29]	VGGFace2 -HQ [30]	FFHQ [17]	CelebA -HQ [31]	Ours
0–10%	27,035	20	32,134	72,357	20,534
10–30%	3129	2566	37,038	110,611	20,511
30–50%	72	50,104	811	19,120	22,009
50–70%	4	46,077	17	497	22,007
70–100%	0	14,989	0	14	15,003
Total	25,237	113,756	70,000	202,599	100,064

For evaluating the performance before and after de-identification in image-copy-detection tasks, we employed the DISC21 [24] and CopyDays [32] datasets, and for video-copy detection the VCDB dataset was utilized. The composition of the datasets for image-copy detection is detailed in Table 3, which includes analyses of the total number of images and the quantity containing faces, showcasing the diversity and complexity of the datasets. The presence of faces in images plays a crucial role in image- and video-copy-detection tasks, particularly in assessing the efficiency of the proposed model's de-identification process. Quantifying the number of instances with visible faces allows for a better measurement of the de-identification impact applied by the model. This information is vital in evaluating how well the model distinguishes between the manipulated and original images under various conditions. Similarly, Table 4 presents the VCDB dataset composition, including the total number of video clips and the proportion containing faces. This metric is essential as it directly correlates with the model's ability to effectively obfuscate facial identities while preserving the integrity of non-target features, thus serving as a crucial indicator of model efficiency. The model's performance is measured not only in terms of its overall accuracy but also in terms of its precision in handling images and videos containing faces, ensuring that the data utility for copy-detection purposes is not compromised by de-identification processing. By including a large and diverse set of video clips in VCDB, the model's performance in consistently detecting and tracking de-identified faces within video streams can be comprehensively evaluated. Utilizing these datasets for evaluation allows for a thorough analysis of the model's performance across various scenarios, with a particular focus on the accuracy and reliability of face de-identification.

Table 3. Overview of the image-copy-detection datasets.

Category	Detail	Dataset	
		DISC21 [24]	CopyDays [32]
Number of Images	Dev Query	50,000	1000
	Test Query	50,000	-
	Reference	1,000,000	3212
	Train	1,000,000	1000
Number of Images Containing Faces	Dev Query	3369	58
	Test Query	2925	-
	Reference	86,539	332
	Train	80,645	65
Resolution	Range	32 × 32–1024 × 1024	56 × 72–3008 × 2000
	# of Varieties	10,860 distinct	1071 distinct

Table 4. Overview of the VCDB dataset.

Category	Detail	Value
Number of Videos	Core Videos	528
	Background Videos	100,000
Number of Videos Containing Faces	Core Videos	528
	Background Videos	97,429
Length	Max Length of Video	2662 s
	Min Length of Video	3 s
	Avg Length of Video	183.63 s
Resolution	Range	320 × 214–1920 × 1080
	Number of Varieties	43 distinct

4.2. Evaluation Metrics

In this study, we evaluated the proposed de-identification methods by comparing the face-region similarity, using SSIM and PSNR against the existing research [9,18,19]. SSIM assesses structural similarity, measuring how de-identification minimizes identifiable information in comparison to the original image. PSNR indicates pixel-level differences between the original and de-identified images. Additionally, we utilized cosine similarity to compare features extracted from the entire image, including faces before and after de-identification, observing the impact of the face-region ratio on the overall performance. For image-copy-detection tasks, evaluation followed the de-identification procedure described in Section 3.1, compared to the existing research [24], using metrics such as the mAP (mean average precision), μAP (micro average precision), Acc@1 (accuracy@1), and $Rcl@p90$ (recall@p90). The μAP represents the average precision across all copy-detection results, while the Acc@1 measures the accuracy of the highest-probability-class prediction, and the Rcl@p90 indicates the accuracy rate of the model for the top-90%-probability classes. For the video-copy-detection task, we applied the μAP metric to evaluate the performance after de-identification, to compare the performance before and after de-identification with the study [25,26].

4.3. Face Verification and Feature Similarity

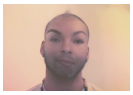
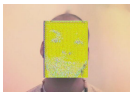
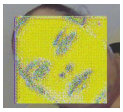
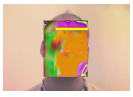
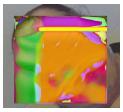
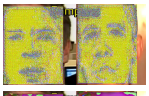
Table 5 in this study categorizes the performance of the proposed models based on the face-region ratio within images. The Structural Similarity Index (SSIM) indicates that values closer to 1 signify higher similarity between the face regions before and after de-identification. The Peak Signal-to-Noise Ratio (PSNR) suggests that values closer to 0 mean higher similarity between the face regions before and after de-identification. Additionally, cosine similarity, with values near 1, denotes a high similarity between two features, demonstrating the model's efficiency in generating privacy-preserving images, even when face regions vary and occupy a significant portion of the image. Table 6 shows the de-identified images from each model, categorized by the size ratio of face regions within each

dataset type, clearly demonstrating the models' efficiency in creating privacy-preserving images by showing the dissimilarity to the original faces.

Table 5. Result of the proposed models in the de-identification and feature-similarity experiments.

Task (Feature Extractor)	Metric	SSIM		PSNR		Feature Similarity	
		Feature Inversion	D2GAN	Feature Inversion	D2GAN	Feature Inversion	D2GAN
Image-Copy Detection (SSCD ResNet50)	0–10%	0.0950	0.4266	7.6847	9.4259	0.9995	0.9994
	10–30%	0.0881	0.3860	7.3589	9.4587	0.9997	0.9965
	30–50%	0.0866	0.3708	6.9805	9.2420	0.9977	0.9937
	50–70%	0.0736	0.3830	6.1259	9.1317	0.9947	0.9931
	70–100%	0.0810	0.3831	6.0036	7.2015	0.9937	0.9998
	Total	0.0941	0.4218	7.6439	9.4282	0.9976	0.9991
Video-Copy Detection (S²VS ResNet50)	0–10%	0.0259	0.4273	7.7064	9.3651	0.9993	0.9996
	10–30%	0.0237	0.4156	7.3543	9.3787	0.9908	0.9973
	30–50%	0.0229	0.4070	7.2664	9.3218	0.9733	0.9937
	50–70%	0.0232	0.4164	7.1750	9.3070	0.9668	0.9923
	70–100%	0.0264	0.4377	7.6666	9.8981	0.9890	0.9874
	Total	0.0255	0.4253	7.6492	9.3663	0.9876	0.9892

Table 6. Result of the generated face regions with the proposed methods based on the percentage of face regions in the image.

Dataset	Type	0–10%	10–30%	30–50%	50–70%	70–100%
Widerface [29]	Original Image					
	Feature Inversion					
	D2GAN					
DISC21 [24]	Original Image					
	Feature Inversion					
	D2GAN					
VCDB [33]	Original Image					
	Feature Inversion					
	D2GAN					

4.4. Evaluation on Image- and Video-Copy-Detection Tasks

In this study, only query images and videos from the entire dataset were de-identified according to the scenario described in Section 3.1. Therefore, the performance evaluated using the unaltered dataset was compared to the performance evaluated using the de-

identified query images and videos alongside the non-de-identified reference images and videos, to assess the before-and-after effects of de-identification. In this context, the evaluation of the image-copy-detection task specifically focused on applying the de-identification processes to query images and videos. In the case of image-copy detection, the performance labeled 'Total' in Table 7 was evaluated using the entire DISC21 dataset, which corresponds to the 'Dev Query' and 'Reference' categories in Table 3's '# of Images', while 'Only Face' was evaluated using only the 'Dev Query' and 'Reference' images from the '# of Images Containing Faces' category. Similarly, for video-copy detection, the 'Total' performance in Table 8 was assessed using all the videos in the VCDB dataset, corresponding to the 'Core Videos' in Table 4's '# of Videos', and 'Only Face' was assessed using only the 'Core Videos' from the '# of Videos Containing Faces' category. We performed a comparative analysis of the performance before and after the application of de-identification, using the evaluation methodologies described in [24,26]. Presented in Tables 7 and 8, the results indicate no significant alteration in performance for both the 'Total' and 'Only Face' categories following de-identification, which demonstrates our model's effectiveness in preserving privacy without compromising its ability to detect copies. This underscores the utility of our approach in protecting personal information, as it maintains robust copy-detection capabilities in various conditions, whether faces are present or not in the media.

Table 7. Comparison results before and after applying de-identification to the evaluation of the image-copy-detection task.

Dataset	Method	Total				Only Face			
		mAP	*µAP*	*Acc@1*	*Rcl@p90*	*mAP*	*µAP*	*Acc@1*	*Rcl@p90*
DISC21 [24]	SSCD	-	72.5	78.2	63.1	-	54.7	68.3	37.1
	Ours (Feature Inversion)	-	71.2	77.8	62.9	-	54.3	67.9	36.9
	Ours (D2GAN)	-	71.3	78.1	62.7	-	54.5	68.0	36.5
CopyDays [32]	SSCD	86.6	98.1	-	-	90.9	97.9	-	-
	Ours (Feature Inversion)	86.1	97.5	-	-	89.8	96.8	-	-
	Ours (D2GAN)	85.9	97.9	-	-	90.1	97.2	-	-

Table 8. Comparison results before and after applying de-identification to the evaluation of the video-copy-detection task.

Dataset	Method	Total		Only Face	
		mAP	*µAP*	*mAP*	*µAP*
VCDB [33]	S^2VS	87.9	73.0	87.9	73.0
	Ours (Feature Inversion)	87.5	72.2	87.5	72.2
	Ours (D2GAN)	87.3	72.4	87.3	72.4
	DnS	87.9	74.0	87.9	74.0
	Ours (Feature Inversion)	86.5	71.8	86.5	71.8
	Ours (D2GAN)	87.1	72.1	87.1	72.1

5. Discussion

The de-identification methods explored in this research offer significant benefits from the perspectives of personal-data protection and image- and video-copy-detection tasks. However, they also come with inherent limitations and challenges encountered during the research process. The face-image datasets provided were limited to images containing faces, without precise face-location information. This necessitated the use of deep-learning-based face-detection models to accurately detect and extract the face regions. However, this process encountered errors and incorrect detection, necessitating additional manual filtering and intervention. This underscores the need for more effort in the data-preprocessing phase and highlights the necessity for future research to develop automated error-detection and correction mechanisms. Moreover, the datasets used were not diverse in terms of the proportion of the face region within the images, often presenting faces that were too large or too small. This imbalance led to normalization challenges, which were addressed by

combining various datasets and restructuring images to distribute the data more evenly, based on the proportion of the face region. This accentuates the need for datasets with a diverse range of face-region sizes and an even distribution of images across different face-region ratios, emphasizing the importance of meticulous data preprocessing and selection in future research.

The two de-identification methods proposed—namely, the feature inversion and D2GAN methods—also presented some limitations. The feature-inversion method, while generating images with emphasized contours, was found to allow for recognition upon close inspection. This underscores the importance of carefully balancing the need for personal-data protection and data utility in the de-identification process. By contrast, the D2GAN method effectively provided de-identification by generating blurred facial images, making it difficult to recognize faces. However, this strong de-identification could reduce the utility of the image, necessitating a careful balance in situations where content utility is a priority. For the feature-inversion model, if we focus more on face-image-similarity loss by adjusting the loss function, the face image can appear more transparent while effectively protecting personally identifiable information. Conversely, if we focus more on feature-similarity loss, the image will be trained to be less identifiable but more similar in important features. This suggests a methodology that satisfies the need for privacy while maintaining the usefulness of the data. On the other hand, by adjusting the amount of facial blur, the D2GAN model can be trained with a face shape for strong de-identification, as shown in our experimental results, or it can be trained with a sharper face shape. This tunability allows for setting the privacy level flexibly according to the requirements of a specific application.

Personal-data protection is a core issue in the digital era, with facial data being among the most sensitive and crucial identifiers. There is a potential conflict between the proliferation of illegal content and the protection of personal information. The proposed models—namely, the feature inversion and D2GAN methods—can play a significant role in resolving this conflict. The feature-inversion method generates images based on facial contours, useful in situations where maintaining the natural appearance of the image is essential. However, caution is required in scenarios demanding high levels of personal-data protection, due to the potential risk of re-identification. On the other hand, the D2GAN method enhances personal-data protection by effectively concealing facial information through blurring. While this strong de-identification is beneficial in highly sensitive situations, it is important to find the right balance in contexts where the utility of the content is valued. Considering these factors, finding a balance between personal-data protection and data utility is crucial. Additionally, in-depth research into how these de-identification methods respond to potential re-identification attacks or vulnerabilities is necessary. The feature-inversion method, due to its emphasis on contours, might be vulnerable to deep-learning-based learning models trained to restore original images from pairs of de-identified and original images. Conversely, the D2GAN method, due to its generation of blurred images, may pose less risk, but it is not entirely risk free. Such analyses will help identify vulnerabilities in de-identification technologies and will lead to the development of more robust personal-data protection mechanisms.

6. Conclusions

In this paper, we explored the viability of de-identification techniques in the context of image- and video-copy-detection tasks. Our proposed methods, leveraging feature inversion and D2GAN, demonstrated consistent performance in preserving essential image features even after substantial modifications in the facial regions. This balance between protecting privacy and maintaining content integrity is pivotal in today's digital era, where data security and personal privacy are paramount. Our approach provides a novel pathway for media forensics, digital-rights management, and privacy-aware computing, marking a substantial advancement in the field.

Looking ahead, further research could be directed towards enhancing the robustness and versatility of these de-identification methods. This includes refining the facial-detection process to reduce manual intervention and exploring the integration of more sophisticated neural-network architectures. Furthermore, the application of these methods could be broadened to encompass a wider range of media-analysis tasks, offering a comprehensive solution for privacy protection in various digital domains.

Moreover, the application of these methods could be broadened to encompass a wider range of media-analysis tasks, offering a comprehensive solution for privacy protection in various digital domains. The implications of our research extend beyond technical innovation. They also contribute to the discourse on ethical AI, emphasizing the critical need for technologies that respect and protect user privacy. By pushing the boundaries of what is possible in media analysis and privacy preservation, we pave the way for future breakthroughs in creating more secure and privacy-conscious digital environments.

Author Contributions: Conceptualization, J.S.; methodology, J.S. and J.K.; software, J.S. and J.K.; validation, J.S. and J.K.; formal analysis, J.S. and J.K.; investigation, J.S. and J.K.; resources, J.S.; data curation, J.S.; writing—original draft preparation, J.S.; writing—review and editing, J.S. and J.K.; visualization, J.S. and J.K.; supervision, J.N.; project administration, J.N.; funding acquisition, J.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00224740, development of technology to prevent and track the distribution of illegally filmed content).

Data Availability Statement: The datasets used for training and evaluating our models including the VGGFace2-HQ, FFHQ, CelebA-HQ, MS-Celeb-1M, Widerface, DISC21, and VCDB are publicly accessible. The VGGFace2-HQ dataset is available at https://github.com/NNNNAI/VGGFace2-HQ (accessed on 4 January 2024); the FFHQ dataset can be found at https://github.com/NVlabs/ffhq-dataset (accessed on 4 January 2024); the CelebA-HQ dataset is accessible at https://github.com/tkarras/progressive_growing_of_gans (accessed on 4 January 2024); and the MS-Celeb-1M dataset can be accessed at https://exposing.ai/msceleb/ (accessed on 4 January 2024). The Widerface dataset is available at http://shuoyang1213.me/WIDERFACE/ (accessed on 4 January 2024). The DISC21 dataset is available at https://ai.meta.com/datasets/disc21-dataset/ (accessed on 4 January 2024). The VCDB dataset is available at https://fvl.fudan.edu.cn/dataset/vcdb/list.htm (accessed on 4 January 2024). Each dataset's usage in this study has been in accordance with the terms set by their respective providers. The datasets used in this research are publicly accessible and each dataset provider has strictly managed them in accordance with their set privacy protection and usage terms. Therefore, we have confirmed that the data used in this study do not present any issues related to personal-information protection.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
GAN	Generative Adversarial Network
D2GAN	Dual Discriminator Generative Adversarial Network
StyleGAN	Style Generative Adversarial Network
ResNet	Residual Network
SSIM	Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise Ratio
BCE	Binary Cross-Entropy
mAP	Mean Average Precision
μAP	Micro Average Precision
Acc@1	Accuracy at 1
$Rcl@p$90	Recall at Precision 90

References

1. Ribaric, S.; Ariyaeeinia, A.; Pavesic, N. De-identification for privacy protection in multimedia content: A survey. *Signal Process. Image Commun.* **2016**, *47*, 131–151. [CrossRef]
2. Agrawal, P.; Narayanan, P. Person de-identification in videos. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 299–310. [CrossRef]
3. Ivasic-Kos, M.; Iosifidis, A.; Tefas, A.; Pitas, I. Person de-identification in activity videos. In Proceedings of the 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 26–30 May 2014; pp. 1294–1299.
4. Dufaux, F.; Ebrahimi, T. Scrambling for privacy protection in video surveillance systems. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1168–1174. [CrossRef]
5. Dufaux, F.; Ebrahimi, T. A framework for the validation of privacy protection solutions in video surveillance. In Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, Singapore, 19–23 July 2010; pp. 66–71.
6. Mahendran, A.; Vedaldi, A. Understanding deep image representations by inverting them. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5188–5196.
7. Dosovitskiy, A.; Brox, T. Inverting visual representations with convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4829–4837.
8. Nguyen, T.; Le, T.; Vu, H.; Phung, D. Dual discriminator generative adversarial nets. *arXiv* **2017**, arXiv:1709.03831.
9. Xue, H.; Liu, B.; Yuan, X.; Ding, M.; Zhu, T. Face image de-identification by feature space adversarial perturbation. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7554. [CrossRef]
10. Wen, Y.; Liu, B.; Cao, J.; Xie, R.; Song, L. Divide and Conquer: A Two-Step Method for High Quality Face De-identification with Model Explainability. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 5148–5157.
11. Li, D.; Wang, W.; Zhao, K.; Dong, J.; Tan, T. RiDDLE: Reversible and Diversified De-identification with Latent Encryptor. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8093–8102.
12. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
13. Pan, Y.L.; Chen, J.C.; Wu, J.L. Towards a Controllable and Reversible Privacy Protection System for Facial Images through Enhanced Multi-Factor Modifier Networks. *Entropy* **2023**, *25*, 272. [CrossRef] [PubMed]
14. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
15. Li, Y.; Lyu, S. De-identification without losing faces. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Paris, France, 3–5 July 2019; pp. 83–88.
16. Khorzooghi, S.M.S.M.; Nilizadeh, S. StyleGAN as a Utility-Preserving Face De-identification Method. *arXiv* **2022**, arXiv:2212.02611.
17. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
18. Yang, X.; Dong, Y.; Pang, T.; Su, H.; Zhu, J.; Chen, Y.; Xue, H. Towards face encryption by generating adversarial identity masks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 3897–3907.
19. Uchida, H.; Abe, N.; Yamada, S. DeDiM: De-identification using a diffusion model. In Proceedings of the 2022 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 14–16 September 2022; pp. 1–5.
20. Shibata, H.; Hanaoka, S.; Cao, Y.; Yoshikawa, M.; Takenaga, T.; Nomura, Y.; Hayashi, N.; Abe, O. Local differential privacy image generation using flow-based deep generative models. *Appl. Sci.* **2023**, *13*, 10132. [CrossRef]
21. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
22. Zhou, T.; Li, Q.; Lu, H.; Zhang, X.; Cheng, Q. Hybrid multimodal medical image fusion method based on LatLRR and ED-D2GAN. *Appl. Sci.* **2022**, *12*, 12758. [CrossRef]
23. Fu, G.; Zhang, Y.; Wang, Y. Image Copy-Move Forgery Detection Based on Fused Features and Density Clustering. *Appl. Sci.* **2023**, *13*, 7528. [CrossRef]
24. Pizzi, E.; Roy, S.D.; Ravindra, S.N.; Goyal, P.; Douze, M. A self-supervised descriptor for image copy detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14532–14542.
25. Kordopatis-Zilos, G.; Tzelepis, C.; Papadopoulos, S.; Kompatsiaris, I.; Patras, I. DnS: Distill-and-select for efficient and accurate video indexing and retrieval. *Int. J. Comput. Vis.* **2022**, *130*, 2385–2407. [CrossRef]
26. Kordopatis-Zilos, G.; Tolias, G.; Tzelepis, C.; Kompatsiaris, I.; Patras, I.; Papadopoulos, S. Self-Supervised Video Similarity Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 4755–4765.
27. Qi, D. yolov7-Face. Available online: https://github.com/derronqi/yolov7-face (accessed on 7 November 2023).

28. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
29. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. Wider face: A face detection benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5525–5533.
30. Cao, Q.; Shen, L.; Xie, W.; Parkhi, O.M.; Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 67–74.
31. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
32. Douze, M.; Jégou, H.; Sandhawalia, H.; Amsaleg, L.; Schmid, C. Evaluation of gist descriptors for web-scale image search. In Proceedings of the ACM International Conference on Image and Video Retrieval, Fira, Greece, 8–10 July 2009; pp. 1–8.
33. Jiang, Y.G.; Jiang, Y.; Wang, J. VCDB: A large-scale database for partial copy detection in videos. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part IV 13; Springer: Cham, Switzerland, 2014; pp. 357–371.