applied sciences

MDPI

Article

A Novel Generative Model for Face Privacy Protection in Video Surveillance with Utility Maintenance

Yuying Qiu [1], Zhiyi Niu [1], Biao Song [1,*], Tinghuai Ma [1], Abdullah Al-Dhelaan [2] and Mohammed Al-Dhelaan [2]

[1] School of Computer Science, Nanjing University of Information Science & Technology, Nanjing 210044, China; jklqiuyuying@foxmail.com (Y.Q.); nzy0921@foxmail.com (Z.N.); thma@nuist.edu.cn (T.M.)
[2] Computer Science Department, King Saud University, Riyadh 11451, Saudi Arabia; dhelaan@ksu.edu.sa (A.A.-D.); mdhelaan@ksu.edu.sa (M.A.-D.)
* Correspondence: bsong@nuist.edu.cn

Abstract: In recent years, the security and privacy issues of face data in video surveillance have become one of the hotspots. How to protect privacy while maintaining the utility of monitored faces is a challenging problem. At present, most of the mainstream methods are suitable for maintaining data utility with respect to pre-defined criteria such as the structure similarity or shape of the face, which bears the criticism of poor versatility and adaptability. This paper proposes a novel generative framework called Quality Maintenance-Variational AutoEncoder (QM-VAE), which takes full advantage of existing privacy protection technologies. We innovatively add the loss of service quality to the loss function to ensure the generation of de-identified face images with guided quality preservation. The proposed model automatically adjusts the generated image according to the different service quality evaluators, so it is generic and efficient in different service scenarios, even some that have nothing to do with simple visual effects. We take facial expression recognition as an example to present experiments on the dataset CelebA to demonstrate the utility-preservation capabilities of QM-VAE. The experimental data show that QM-VAE has the highest quality retention rate of 86%. Compared with the existing method, QM-VAE generates de-identified face images with significantly improved utility and increases the effect by 6.7%.

Keywords: face de-identification; autoencoders; privacy protection; vector quantization; deep learning; differential privacy

Citation: Qiu, Y.; Niu, Z.; Song, B.; Ma, T.; Al-Dhelaan, A.; Al-Dhelaan, M. A Novel Generative Model for Face Privacy Protection in Video Surveillance with Utility Maintenance. *Appl. Sci.* **2022**, *12*, 6962. https://doi.org/10.3390/app12146962

Academic Editor: David Megías

Received: 21 April 2022
Accepted: 28 June 2022
Published: 9 July 2022

1. Introduction

In recent years, with the advancement of imaging devices and fast-transmission media, high-quality video surveillance has been popularized. People's identity and behavior information are always exposed to video surveillance in public spaces. While the intensive surveillance distribution has contributed towards making people's lives more convenient in many ways, care needs to be taken as a huge amount of image sources is being stored, transmitted, or published around the world, and they also become easily accessible and more likely to be abused.

The face is significant to reveal a person's most distinguishing features in images [1], so the abuse of facial images in surveillance can violate the safety of a person's privacy. In practice, face de-identification cannot be simply described as the complete elimination of private information, but it must also take into account the real utility of the image. If the image is protected by simple methods such as blurring or occlusion, along with identity, other vital non-biometric traits may be buried. Thus, the de-identified image will lose its original value in the video surveillance of different service situations, such as services that are sensitive to facial emotion or facial shape, etc. As a result, protecting the privacy data in a face image and maintaining the application quality of the de-recognized image in the service system at the same time become crucial. For example, privacy protection is

crucial in social settings, but if useful information such as facial expressions can be retained in the meantime, it will considerably contribute to the total expression of information. Nevertheless, the link between the facial privacy features to be removed and the useful information to be retained is often nonlinear and sometimes even contradictory, which is the conundrum that the present facial de-identification research is confronted with. As a result, figuring out how to achieve the aims of protecting privacy while retaining quality is a challenge.

Face image de-identification includes obscuring the identifying information in the face image, on the one hand, while having the resulting images have a similar effect to the original image in the service, on the other hand. Many recent research works have focused on achieving the aforementioned ideal equilibrium. Reference [2] suggested a technique of face de-identification that depends on the detection of the face and key points, followed by variational adaptive filtering, to achieve the face de-identification while preserving expression. Reference [3] devised unique verifying and regulating modules in the process of face de-identification to ensure that de-identified photos maintain structure similarity. Reference [4] fine-tuned the encoding section of the typical autoencoder and completed the de-identification operation in the latent space, allowing the original face pictures to retain their nature and variety while creating new faces with various identity labels. These technologies are oriented toward one specific quality standard (such as facial expressions) [5], and there is not yet a model that can automatically adjust the process of generating images according to the needs of different service evaluations. The need to develop an effective face de-identification model for privacy protection cannot be overstated.

To alleviate the above-mentioned constraints, we build on previous techniques and propose a novel de-identification strategy called Quality Maintenance-Variational AutoEncoder (QM-VAE) in this paper. The proposed strategy makes use of a powerful generative model, Vector Quantized Variational Autoencoder (VQ-VAE) [6], and is capable of producing high-quality de-identified faces. We process the images, respectively, using four typical protection methods, resulting in four sets of images without private information. Then, for specific services, we modify the loss function to establish a service quality evaluation index that guides the merging of these images. This approach ensures that the resulting image retains some service-related quality, and because the model's input does not contain private information, it also ensures that privacy is protected.

The following are the major contributions of this work:

- We present a new framework for facial privacy protection, called QM-VAE, which de-identifies the image first and then reconstructs its utility. We integrate vector quantization into the structure of the generative model, so that the model can generate high-quality face images. QM-VAE takes multiple groups of de-identified images as the input, which can make use of the advantages of existing technology to speed up the training and achieve the goal of maintaining service quality.
- We establish a service-oriented loss function by adding service quality loss to guide the generation of de-identified face images with utility maintenance. The proposed framework treats the service quality evaluator as a black box and can adjust the generated image automatically according to the different evaluation results, with wide applicability.
- We take facial expression service as an example and illustrate the viability of our model on the CelebA dataset. We find the most appropriate method configuration by adjusting the parameters. Experimental results show that our solution results in de-identified face images with significantly improved utility. Compared with the traditional methods of de-identification and AMT-GAN, the QM-VAE model has obvious advantages in maintaining specific service quality and increases the utility retention rate by at least 6.7%.

The remainder of this paper is organized as follows. In Section 2, we review the previous research on face de-identification. In Section 3, we present the preliminary work

of our research and utility evaluation methods that we integrate. In Section 4, we present the detailed information of our proposed model, QM-VAE, which combines a generative module incorporating vector quantization with loss of service quality. In Section 5, we conduct a quantitative evaluation experiment to verify the performance of our proposed method, and the results are presented to demonstrate its effectiveness. We discuss the results and present several recommendations for further research in Section 6 and show a simple conclusion of our research in Section 7.

2. Related Work

In recent years, face image privacy protection has been extensively researched [7–9]. Traditional privacy protection research has focused on some simple methods applicable to standard static images, such as image pixelation, image blurring, color block masking [10], etc. These rudimentary techniques, however, are insufficient for removing all privacy data from the original image [10,11]. Reference [12] took advantage of a multi-task deep learning network to detect the location of privacy-sensitive information and then provided simple protection through image blurring, while the visuals were severely compromised.

k-anonymity [13–15] offers a series of techniques with theoretical anonymity guarantees [16]. In these techniques, the resulting face is represented by the average face of the k nearest faces in the face image set. Because of the alignment error when calculating the average face, the resulting image often suffers from ghosting artifacts. For this problem, Reference [17] proposed the q-far de-identification approach based on the active appearance model [18]. The q-far technique includes an extra pose estimation step to align the faces before calculating the average, which successfully eliminates the ghosting artifact phenomenon in the generated face. K-Same-Net [19] innovatively combines the k-Same algorithm with Generative Neural Networks (GNNs) [20], which can erase the identity of the face and retain the selected features. Reference [21] proposed the k-Same-Select algorithm, where the k-Same algorithm is used independently to segment mutually exclusive subsets from the input sets, and semi-supervised learning [22] and collaborative training technologies [23] are utilized to learn the information to be expressed by the original images. This method was proven to have great potential for utility preservation on the FERET database [24]. Reference [25] divided the face space into subspaces that are sensitive to identity or utility and then used the k-anonymity de-identification algorithm to deal with the latter. In this way, while resisting face identification, it can still meet the goal of image utility preservation. Each of the k-Same-algorithm-based methods has a common drawback of being unable to provide unique de-identified outputs for various images.

The emergence of Generative Adversarial Networks (GANs) [26] provides a new path for face de-identification research [15,27–30]. In [31], a GAN was used for the inpainting of facial landmarks in the conditioned heads. This preserves some details of the images to a certain extent while generating face images with high visual quality, which is more effective in face de-identification than pure blurring approaches. Wu et al. [3] proposed the CGAN-based PPGAN to tackle the face de-identification issue by using a single face image as the input and producing a de-identified image with the structural similarity. Although the image processed by the PPGAN may keep the original emotion, its identity obfuscation quality is not particularly strong [1]. In order to improve the effect of de-identification, Li et al. [32] put forward the SF-GAN method, which balances the effects of different elements on face de-identification by constructing various external mechanisms. Reference [4] proposed a variety of face de-identification methods based on fully connected and convolutional autoencoders, learning to moving the sample to other samples with desired attributes and away from those with conflicting attributes through the training of an encoder. Reference [33] combined the Variational Autoencoder (VAE) [34] and CGAN [35] into the Variational Generative Adversarial Network (VGAN) and proposed a Privacy-Preserving Representation-Learning method based on it. This method learns the image representation that is explicitly disentangled from the identity information, effectively preserving the privacy while achieving other tasks such as expression morphing.

The principle of Differential Privacy (DP) [36] is adopted by many face de-identification technologies [37]. Reference [38] proposed the SDC-DP algorithm, a novel noise dynamic allocation algorithm based on differential privacy using the standard deviation circle radius, which effectively reduces the relative error and improves the accuracy. Reference [39] introduced a neural network privacy analysis method based on the f-differential [40], which can boost forecast accuracy without going over the privacy budget by tweaking some model parameters. In [41], a Privacy-Preserving Adversarial Protector Network (PPAPNet) was developed as an improved technique for adding noise to the face. Experiments revealed that the PPAPNet performs remarkably in converting original images into high-quality de-identified images and resisting inversion attacks [42]. Reference [43] directly processed face images in the pixel space to realize DP, regardless of the image's distribution characteristics. On this premise, the exponential mechanism proposed can provide superior visual quality for image confusion using the Laplacian mechanism, along with strong universality. On the contrary, Reference [44] proposed Privacy using EigEnface Perturbation (PEEP), which utilizes local DP to add perturbations to image distribution features and stores these perturbed data on third-party servers to avoid privacy attacks such as model memorization attacks [45] or membership inference [46]. In [47], a Privacy-Preserving Semi-Generative Adversarial Network (PPSGAN) was proposed, which uses a self-attention mechanism [48] to add noise to the category-independent features of each image selectively, allowing the resulting image to preserve the original label. The PPSGAN is more practical than many common technologies, such as image filtering, random noise addition, and even, GAN generation.

In addition to modifying faces directly, many studies solve the privacy protection problem in video surveillance using other angles. Reference [49] proposed a Sliding Window Publication (SWP) algorithm for face image publishing and sorting, innovatively converting the privacy security issue of images into that of the data stream, opening up a completely different perspective for image privacy protection. Reference [50] presented a secure video transmission strategy, which uses watermarking and video scrambling technology, making the multimedia stream low quality and unavailable in transmission, until it is decoded at the authorized endpoint. In [51], a cycle Vector-Quantized Variational Autoencoder was proposed to encode and decode the video and complete the task of privacy protection using multiple heterogeneous data sources in the video. This method uses a fusion mechanism to integrate the video and extracted audio and regards the extracted audio as random noise with a non-pattern distribution to realize visual information hiding. It has excellent performance in video compression, video reconstruction, and visual quality maintenance.

The current research mainly focuses on a specific application scenario, and the evaluation standard of image quality was fixed in the process of the model design. Although these techniques have achieved extraordinary results, their application scope is strictly limited and their adaptability is weak to meet different service evaluation metrics. At the same time, the quality requirements of actual service applications are often varied and even abstract (difficult to judge by vision or a simple function). To overcome the restrictions of the approaches mentioned above, it is very valuable to design privacy protection and quality maintenance technology for different services.

This paper proposes a novel de-identification and quality maintenance framework, QM-VAE. We treated the specific service scenario as a black box instead of defining specific calculation formulas, making our methods flexible, adaptability, and highly efficient.

3. Preliminary

3.1. Face De-Identification Methods

To accomplish the mission of removing private information, we selected four classic privacy protection methods to process the images. The first method is blindfold, where we

set the pixel value of the line where the eye area is located in the image to 0. T_{eyes} represents the set containing the eye coordinates. The function for blindfold is as follows:

$$x_{i,j}^1 = \begin{cases} 0, & \text{if } (i,j) \in T_{\text{eyes}} \\ x_{i,j}, & \text{other} \end{cases} \tag{1}$$

The second method is to add a mosaic in the center of the face, which makes color blocks with a side length of 8 pixels based on the value of points in the original image. T_{face} is the set containing all the face coordinates, and $g(x)$ represents the average value of pixels covered by color blocks. The function for adding a mosaic is as follows:

$$x_{i,j}^2 = \begin{cases} g(x), & \text{if } (i,j) \in T_{face} \\ x_{i,j}, & \text{other} \end{cases} \tag{2}$$

The next method is to add the Laplace noise to the probability density function, as indicated in Equation (3) with λ equal to 1 and μ equal to 0 in our study.

$$f(x) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}} \tag{3}$$

The last method is to convert images into anime style. The pre-trained UGATIT network [52] provides a convenient way to generate cartoon images.

We created four sets of images without privacy-sensitive information using the procedures described above. If we try to use these images directly in the application scenarios, we will suffer the adverse consequences of the decline in service quality. As a result, we need to create a fusion algorithm for them to restore quality in various conditions.

3.2. Facial Expression Recognition Module

Human facial expressions contain emotion, mentality, etc., and many service scenarios obtain useful information by recognizing facial expressions. Thus, we took facial expression recognition as an example to complete the work of service quality maintenance [53–55].

We define the expression set L to contain seven elements: "angry", "disgust", "fear", "happy", "sad", "surprise", and "neutral". The essence of facial expression recognition is to realize classification tasks [56,57]. Compared to other kinds of neural networks, Convolutional Neural Networks (CNNs) [56,58] show better performance in this work [59,60]. CNNs are made up of several convolutional layers with pooling layers and fully connected layers. Each convolutional layer extracts features from the input data, and the final target representative features are reflected in the highest convolutional layer. Then, the output enters the fully connected layers to complete the mapping from the input image to the expression tag. The weights of the network were iteratively adjusted through back-propagation and trained and tested on the Kaggle facial expression recognition challenge dataset [61].

The workflow of the expression recognizer is illustrated in Figure 1. For the initial image, the pre-trained Haar cascade classification model [62] is used to detect the facial region in the gray image and return the boundary rectangle of the detected face. This step is very necessary so that we can reduce the dimension of the original image to a form similar to the image of the facial expression recognition network training set, effectively improving the reliability of facial expression recognition. Taking the cropped image x as the input, the output $res(x)$ of the model will obtain the values of seven neurons for facial expression recognition for this image. As shown in Equation (4), we selected the index of the maximum value and took the corresponding expression as the result of facial expression recognition $E(x)$.

$$E(x) = L_k, \quad \text{where } k = \arg\max_i (res(x)_i) \tag{4}$$

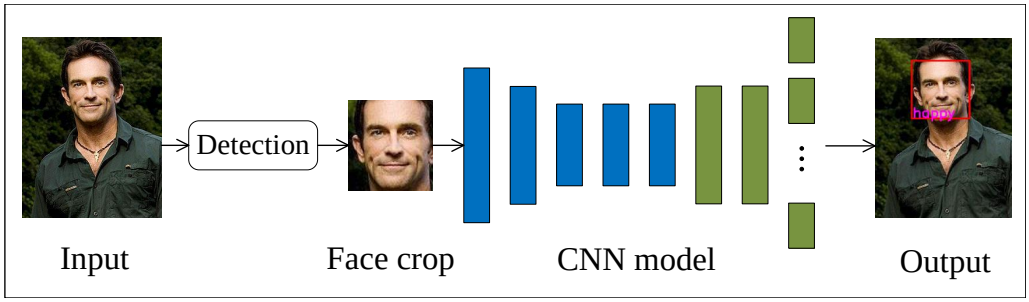

Figure 1. The workflow of the expression recognizer.

4. Method

4.1. System Model

Face de-identification in the context of image utility maintenance is complicated work. In the process of pure privacy information removal, the possible loss of image utility is ignored, so the service quality will also suffer as a result. To solve this problem, we present the QM-VAE framework, a unique generative model of image utility maintenance, and establish a loss function including the loss of service quality to guide the model's training. Figure 2 depicts the general structure of the de-identification module and QM-VAE. Firstly, we applied four protection methods (blindfold, cartoon, Laplace, and mosaic) for face images to construct de-identified datasets. Following that, we input these datasets into QM-VAE, a generative model guided by service quality evaluation. During the training phase, we calculated whether the utility of the output images in the service scenario was consistent with the original data and took this calculation result as a part of the loss function. Backpropagation plays a significant role in updating the output continuously in order to maintain service quality.

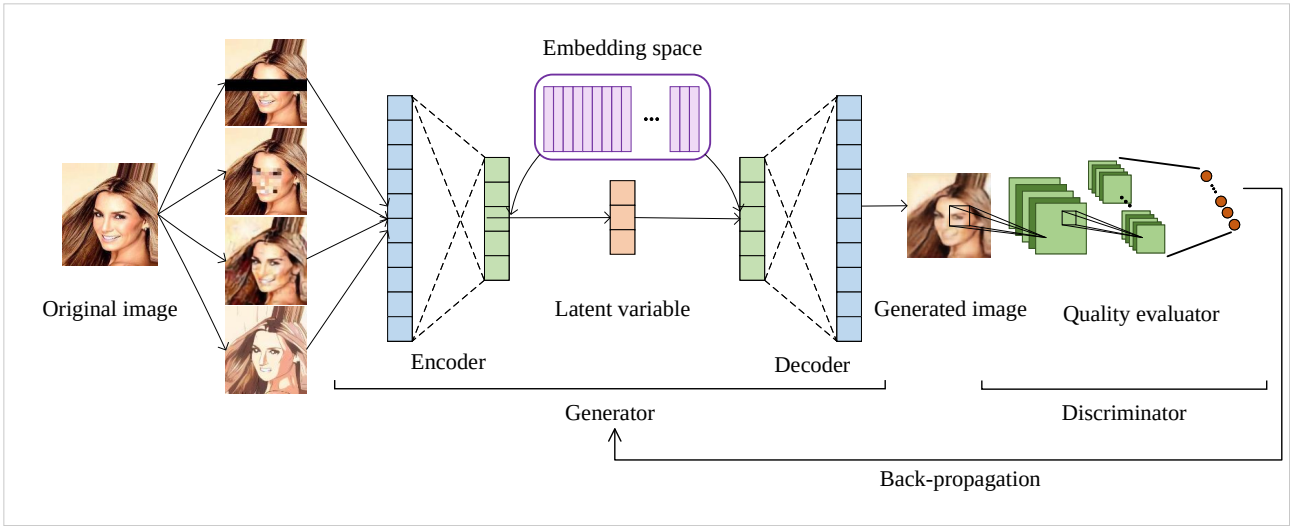

Figure 2. The overall structure of the de-identification module and QM-VAE.

4.2. Architecture and Working

Our research aims to develop a multi-input generative model that uses several face images with privacy information removed to synthesize one image with quality maintenance. QM-VAE offers an inspiring framework for efficiently generating de-identified images with utility preservation through neural network training.

QM-VAE is inspired by the concept of Vector Quantization (VQ), initially proposed for VQ-VAE, which integrates discrete latent variables into the structure of the model. Compared with common autoencoders, VQ-VAE creates a latent embedding space e between the model input x and the low-dimensional variables z. The model's whole parameter set may be split into three sections: the encoder, the decoder, and the embedding space e. The

model's prior distribution is used to learn to generate discrete variables. The output of the encoder will be mapped to the embedding space, and the mapped embedding vector will be fed to the decoder.

The input x of the model is formed by splicing images processed by four privacy protection methods. The encoder is fed x and outputs $ze(x)$. Different from the ordinary autoencoder model, $ze(x)$ does not go straight to the decoder, but searches for the closest embedding in the embedding space e. This embedding will replace $ze(x)$ and enter the decoder as the latent variable z. The probability of the posterior classification distribution $q(z \mid x)$ is defined as Equation (5), which means that the index of the vector nearest $ze(x)$ will be set to 1, and the rest will be set to 0.

$$q(z = k \mid x) = \begin{cases} 1, & \text{if } k = \arg\min_j \mid\mid z_e(x) - e_j \mid\mid_2 \\ 0, & \text{other} \end{cases} \tag{5}$$

$ze(x)$ finds the element closest to itself in the embedding space and transmits it to the decoder as its own mapping in the embedding space, thus realizing the process of discretization, as shown in Equation (6).

$$z_q(x) = e_k, \quad \text{where } k = \arg\min_j \mid\mid z_e(x) - e_j \mid\mid_2 \tag{6}$$

Our model extracts the latent feature space of image data. In the nonlinear process of mapping the latent vector to the embedding vector, the model simultaneously copies the gradient from the decoder's input $zq(x)$ to the encoder's output $ze(x)$. The gradient promotes the update of assignment in Equation (5) and, in turn, influences the result of the encoder's discretization. Since the encoder's output and decoder's input share the same D-dimensional space, the gradient will instruct the encoder to adjust its output to reduce the reconstruction loss of image fusion.

Equation (7) presents the model's total loss function. The L_e in the first term measures the service quality loss. In this work, we took the scenario of facial expression recognition as an example, and the specific solution process is described in detail in Equation (8). The second term describes the image reconstruction loss of the decoder and encoder, that is the Mean-Squared Error (MSE) between the original images and the model outputs. The third term of the function is designed to optimize the vector-quantized embedding space and constantly update the dictionary in the process of model training. The last term is commitment loss, which constrains the encoder, preventing it from updating too quickly, so that its output deviates from the embedding vector. To sum up, the complete loss function for model training is as follows:

$$L = \alpha L_e + (1 - \alpha) \log p(x \mid z_q(x)) + \mid\mid sg[z_e(x)] - e \mid\mid_2^2 + \beta \mid\mid z_e(x) - sg[e] \mid\mid_2^2, \tag{7}$$

where α is used to assign the proportional weight of the loss of service and the loss of image fusion. The value of α ranges from 0 to 1. In our experiment, we adjusted the value of α and analyzed the corresponding quality of service. The specific results are presented in Section 4.3. sg is the stop gradient operator with zero partial derivatives and fixed operands when computing forward. The first two terms of the loss function are optimized by both the encoder and decoder. The third term is optimized by embedding space e. The last one is only optimized by the encoder.

According to the method described above, in each iteration of QM-VAE training, we record the original images' and the output images' expressions, respectively. With the same result as the goal, the service quality loss function L_e is determined, as shown in Equation (8). When the result of facial expression recognition of the output is the same as

that of the initial image, it is considered that the quality of service is maintained, and L_e is set to 0; otherwise, it is set to 1.

$$L_e = \begin{cases} 0, & \text{if } E(x) = E(z(x)) \\ 1, & \text{if } E(x) \neq E(z(x)) \end{cases} \tag{8}$$

4.3. Method Execution Process

We drew on the idea of destroying the image utility first and then reconstructing it to formulate the process of our method. Figure 3 vividly shows the execution process of QM-VAE.

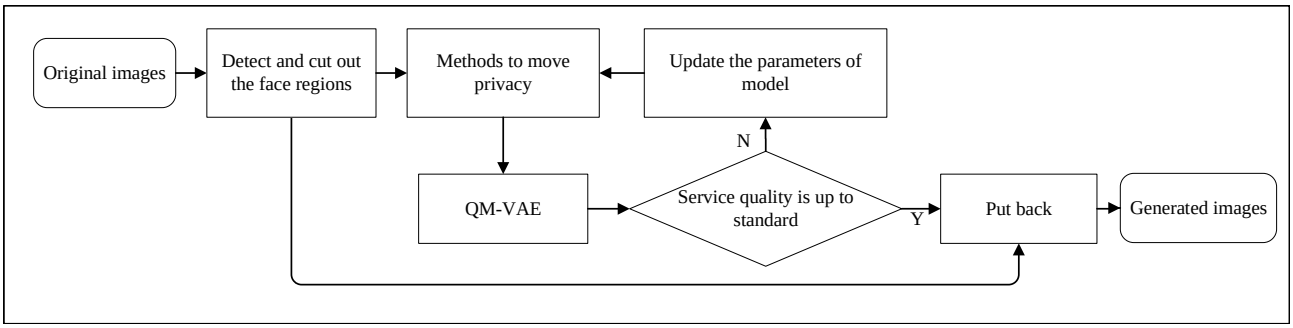

Figure 3. The flowchart of QM-VAE.

At the beginning of the process, we perform face recognition on the original image before privacy protection, in which the face area is cut out for subsequent processing. Then, we utilize the face de-identification methods mentioned in Section 3.1 to modify the cropped area. In the utility reconstruction stage, de-identified images will be fed into the QM-VAE.

The service quality evaluator quantifies the image utility loss by using the difference of attributes in the service scene before and after processing. It influences the training process of the model, so as to make the model generate images that meet both privacy removal and utility preservation, and, finally, provides reliable and safe faces for application scenes.

5. Experimentation and Results

5.1. Experimental Setup

We implemented QM-VAE on the CelebA dataset [63], an open dataset containing sufficient face images after alignment and cropping. We selected the first 1000 images in the dataset for the experiments, of which, the first 80% were used for training and the remaining 20% for testing. To verify the wide applicability of QM-VAE, we divided the datasets into six categories according to expressions, including angry, fear, happy, sad, surprise, and neutral. The expression distribution of the whole dataset is shown in Figure 4.

We coded QM-VAE using the Tensorflow 1.8 and Keras 2.24 framework developed by Google, America, and trained it on GTX 2070. Paper [6] proved that when β in Equation (7) takes a value between 0.1 and 2.0, there is no significant change in the result, so we set the value of β to 0.25 in the experiment. During the whole system training, we employed the ADAM optimizer [64] with a learning rate of 1×10^{-3} during the whole model training. The simulation parameters are listed in Table 1.

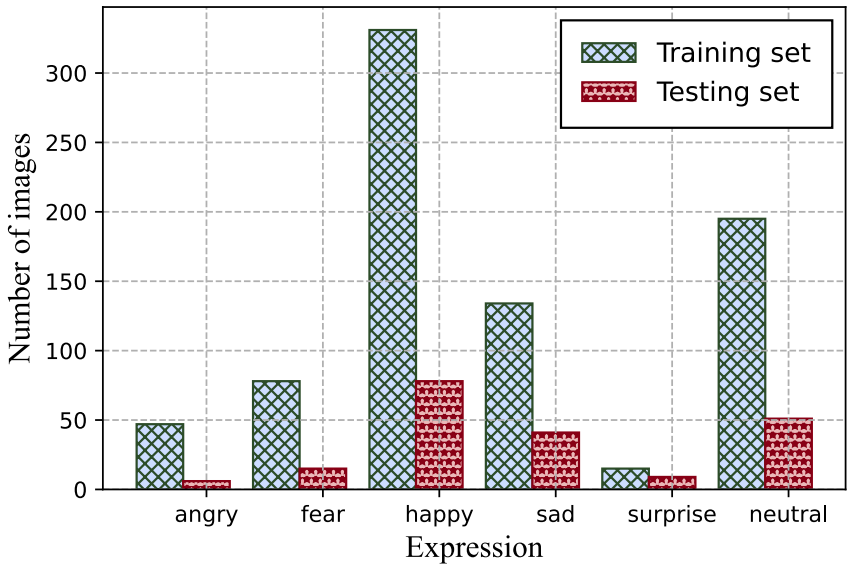

Figure 4. The distribution of the dataset.

Table 1. The simulation parameters.

Methods	Epochs	α
QM-VAE	10	from 0 to 1 in steps of 0.1
QM-VAE	20	from 0 to 1 in steps of 0.1
QM-VAE	30	from 0 to 1 in steps of 0.1
QM-VAE	40	from 0 to 1 in steps of 0.1
Blindfold	\	\
Mosaic	\	\
Cartoon	\	\

5.2. Face Image De-Identification

We adopted four de-identification methods to process 1000 face images. According to the chosen method, we can divide the processed images into four data groups: blindfold, mosaic, Laplace, and cartoon. Examples of the original images and processed images are illustrated in Figure 5.

In order to demonstrate that our proposed framework has a good effect in specific service evaluation, we need to prove that the generated image has the same result as the original one after passing through the facial expression recognizer. We calculated the expression consistency between the generated and original image. The expression recognition neural network was utilized to predict the expression of the original image (E_1) and generated image (E_2).

First, we used the cv2.CascadeClassifier face classifier for face detection. Then, we input the recognized region into the trained expression recognition model to obtain the values of E_1 and E_2. As shown in Equation (8), if the values of E_1 and E_2 are the same, this means that effective data privacy protection is achieved, so the returned value of the function is 0. Otherwise, it is 1. As the number of training rounds grows, we calculate the service quality loss rate by dividing the sum of the return values by the number of pictures. If the service quality loss rate continues to decrease, this indicates that the effect of the model on image utility maintenance is improving.

We measured the service quality loss rate for the four de-identification methods. As indicated in Table 2, there was a significant decrease in the service quality of facial expressions, up to 46.5%. It can be concluded that the utility of these images is obviously affected after these methods.

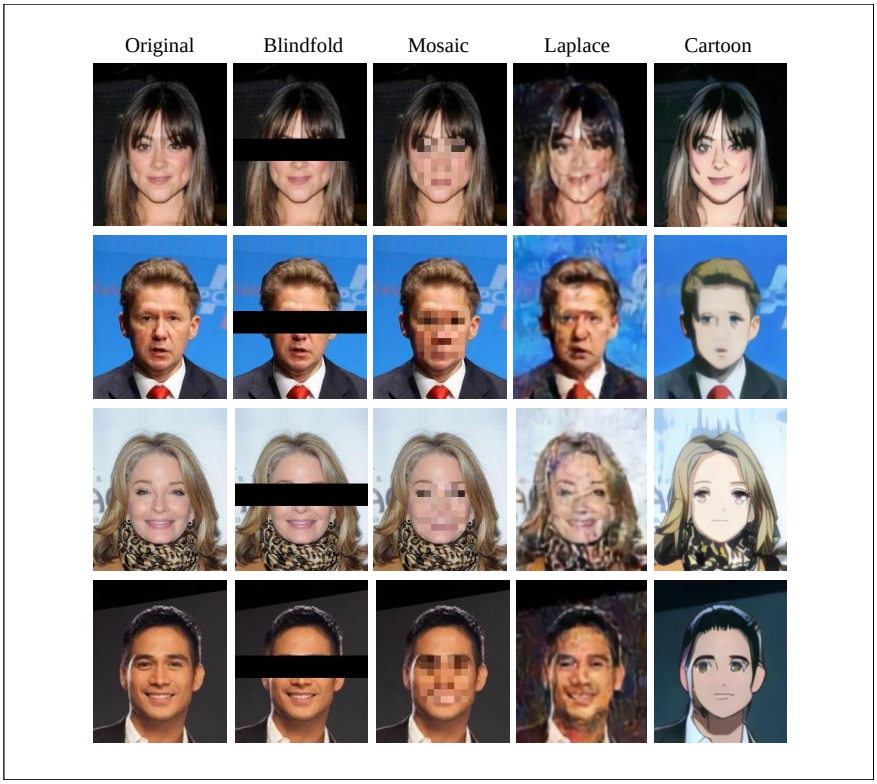

Figure 5. The results of the de-identified image.

Table 2. Service quality loss rate under the four de-identification modes.

Type	Blindfold	Mosaic	Laplace	Cartoon
Quality loss rate	0.462	0.207	0.207	0.465

5.3. Images' Utility Maintenance

We put one original image and four corresponding de-identified images together as a set of data for QM-VAE training. We used a set of training data as the input of the model and obtained the generated image corresponding to each set of data. In the process of training, Equation (7) is used as the loss function. In order to obtain the minimum loss rate, we conducted many experiments. We changed the parameter α that determines the weights of the MSE and service quality loss in the loss function and then observed the model effect of service quality maintenance, respectively, to obtain the most appropriate composition of the loss function.

The test set was used to calculate the expression loss rate for various α values and training epochs. We obtained the changing trend of the expression loss rate as indicated in Figure 6.

From the above results, we can consider that when α is about 0.4, the effect of the model is the best. when training for 40 rounds, the loss rate of facial expression recognition is only 0.14, which is less than that of the four protected images, and the loss rate is reduced by at least 6.7%. This demonstrates that our approach has a remarkable effect on preserving the quality of face expressions. Compared with image fusion, which only considers the mean-squared error, the result of $\alpha = 0.4$ has more advantages in the number of training rounds, which reflects that the loss of the quality of service plays a significant role in the overall loss function. The resulting images are shown in Figure 7.

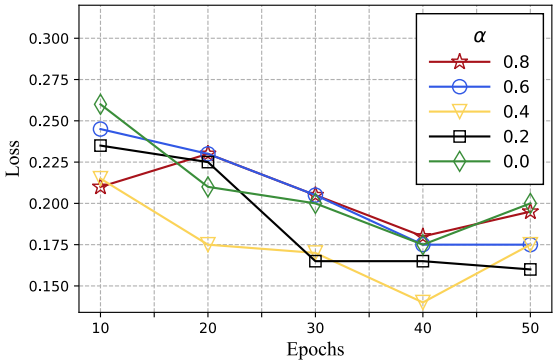

Figure 6. The changing trend of the expression loss rate.

Compared with the method without considering the specific loss of service quality, QM-VAE has advantages in maintaining service quality. We used AMT-GAN [6], one of the latest privacy protection technologies, to process the dataset images, and the image results are shown in Figure 8. The loss rate of the service image quality of AMT-GAN is 0.395, which is about 0.255 more than that of QM-VAE, which obviously shows that QM-VAE performs better in maintaining the quality of service.

We used the model trained for 40 epochs when α is 0.4 for the following analysis. We measured the loss rate of six kinds of facial expressions under different methods. The results are shown in Table 3.

Table 3. Facial expression loss rate under different methods.

Loss Rate	Angry	Fear	Happy	Sad	Surprise	Neutral
QM-VAE	0.167	0.4	0.064	0.171	0.33	0.117
Blindfold	0.396	0.701	0.039	0.726	0.75	0.870
Mosaic	0.226	0.290	0.196	0.149	0.208	0.232
Laplace	0.321	0.204	0.178	0.246	0.333	0.191
Cartoon	0.755	0.452	0.504	0.343	0.542	0.423
AMT-GAN	0.551	0.776	0.144	0.677	0.826	0.476

The service quality of each expression was maintained to some extent under the QM-VAE treatment, and the loss rate of the fear group and surprise group was slightly higher because of the small sample size.

We calculated the difference of the values corresponding to the facial expression recognition results before and after being processed by the four selected methods, AMT-GAN and QM-VAE, and drew error scatter plots and error bar diagrams of data, as shown in Figures 9 and 10. Compared with the other five methods, QM-VAE shows a more stable and excellent utility maintenance ability when the results of facial expression recognition are diverse.

In the problem of face recognition, we also hope to make the smallest possible changes to the image in the process, that is to erase only the private information and maintain its similarity to the original image as much as possible, which is beneficial to maintaining the effectiveness of the image. We used the MSE to calculate the degree of changes to the image. If you set the coefficient of the MSE in the loss function to 0 (set the value of α to 1), the resulting image is as shown in Figure 11. From the resulting images represented by these three images, we cannot recognize faces, let alone extract their expression information. It can be considered that the maintenance of image utility is ineffective in this case.

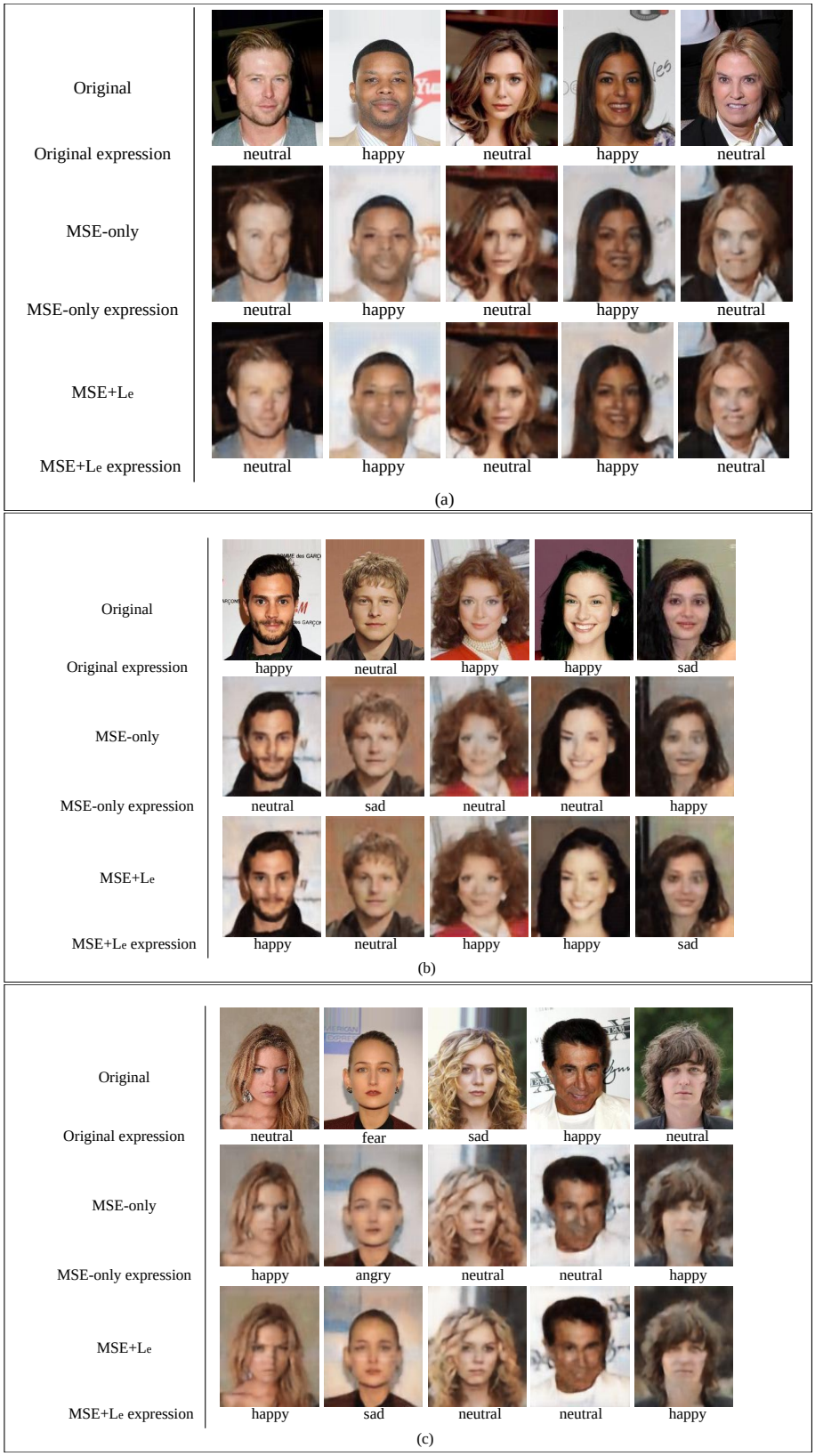

Figure 7. The result of service quality maintenance (**a**) Image utility was maintained without expression loss. (**b**) Image utility was maintained only after expression loss was added. (**c**) Image utility was damaged even after adding expression loss.

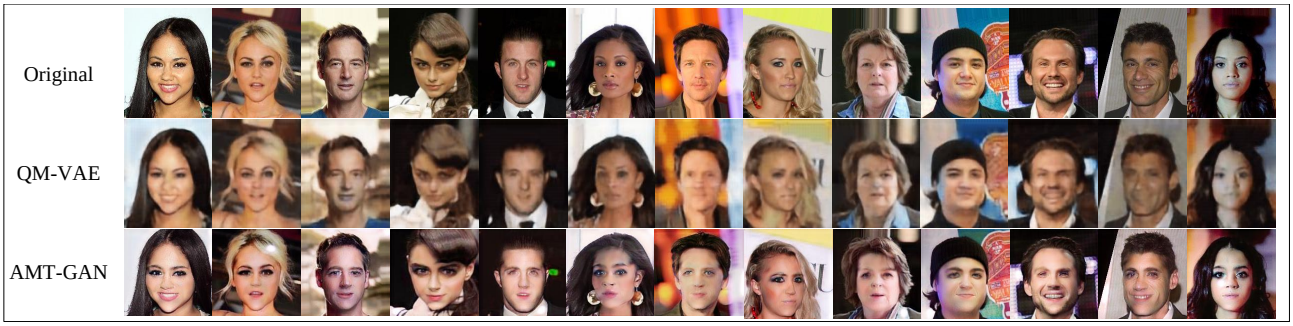

Figure 8. Comparison of the resulting images of QM-VAE and AMT-GAN.

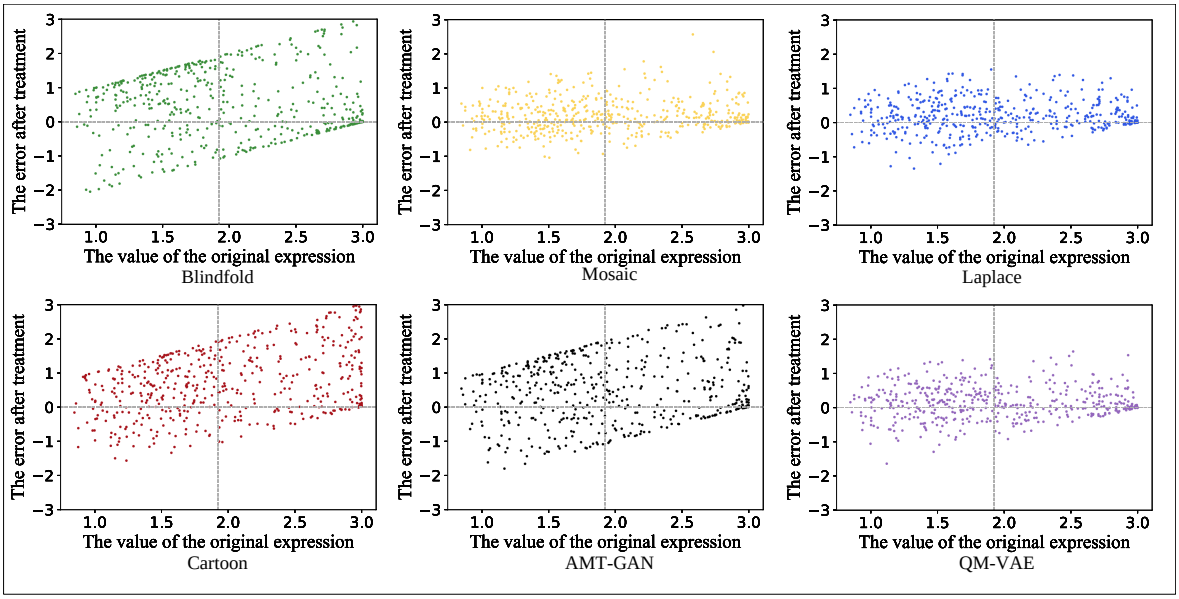

Figure 9. The scatter distribution diagram of the difference between the result values of facial expression recognition before and after processing by different methods.

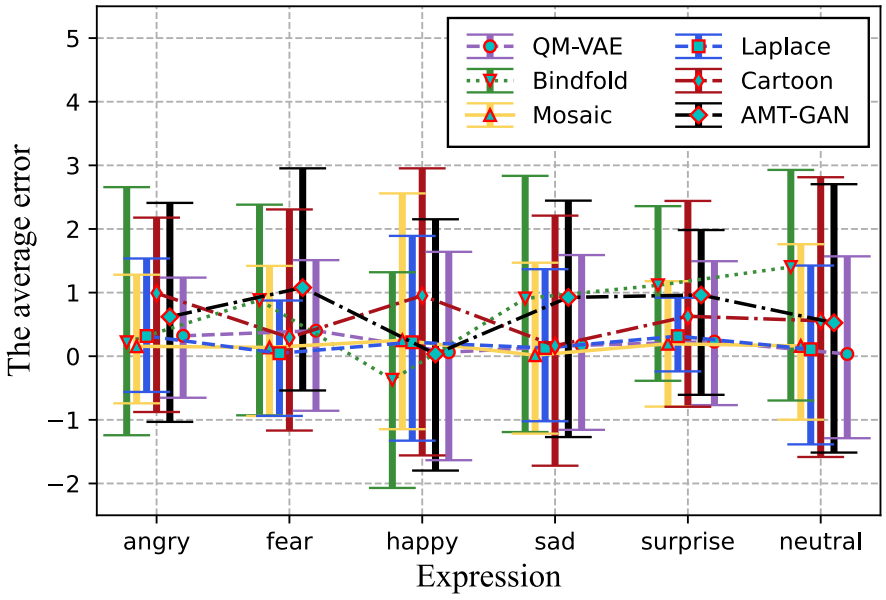

Figure 10. The error bar diagrams of de-identified image by different methods.

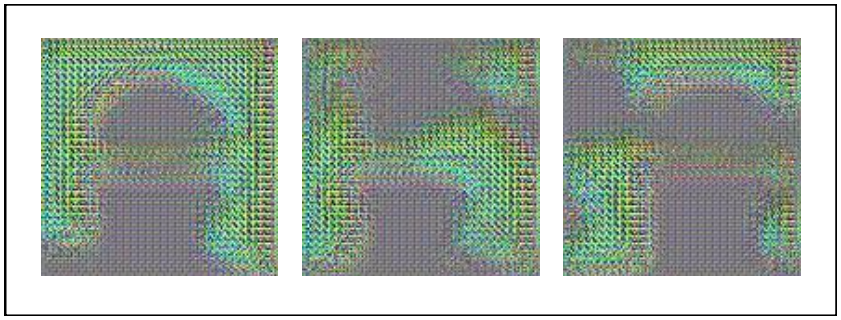

Figure 11. The result of ignoring the MSE.

We can find that if we do not consider the MSE at all, the resulting image will almost collapse. We designed experiments to compare the expression loss of images generated with and without considering the loss function. The experimental data are shown in Table 4.

Table 4. Service quality loss rate of different expressions.

The Coefficient of MSE	Epochs = 20	Epochs = 40
0	0.740	0.825
0.2	0.23	0.18
0.5	0.205	0.150
0.8	0.225	0.165
1	0.210	0.175

The experimental results reveal that the effect of the loss function without the MSE is obviously different from that with the MSE. Within a certain range, the MSE plays a positive role in maintaining the quality of service.

Through the comparative test shown in Figure 8, we can find that the AMT-GAN has excellent visual effect. Therefore, in the service scenario where the image can only be slightly changed, we can add the AMT-GAN as an input of QM-VAE, so that QM-VAE has five inputs. Figure 12 shows the changing trend of model loss when the input number of QM-VAE is 3 (removing blindfold), 4 (used above), and 5 (adding AMT-GAN).

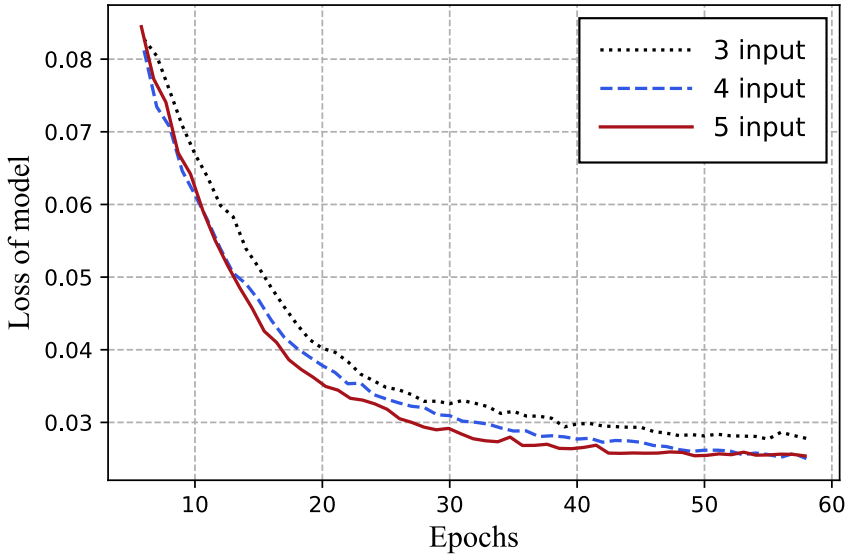

Figure 12. Change of model loss under different inputs.

It can be seen that the model loss reduction rate is improved after the addition of blindfold and AMT-GAN, which proves that our model can use existing technologies

according to the service scenarios to improve training efficiency, as well as having wide applicability and flexibility.

Figure 13 demonstrates an example process of privacy protection and utility maintenance by QM-VAE. We tested the utility of the generated image and found that it produces the same result as the original image when it passes through the expression recognition machine, as shown in Figure 14. This proves that QM-VAE is effective in maintaining the work side of utility.

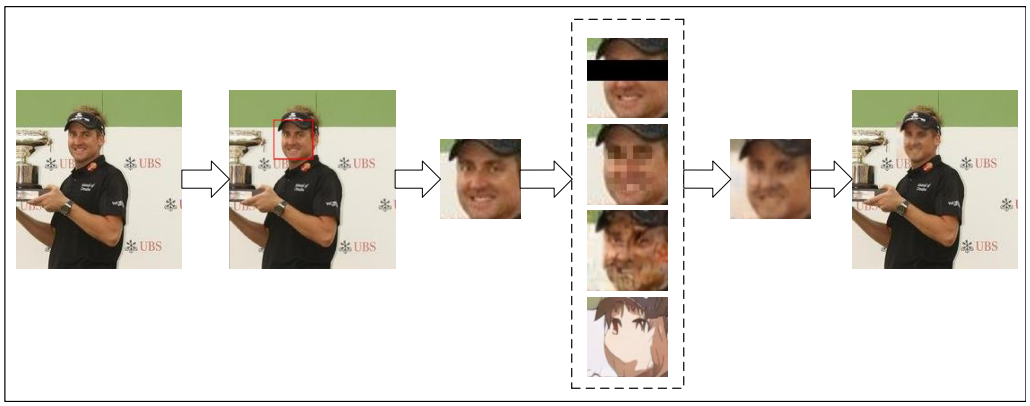

Figure 13. A case process of QM-VAE.

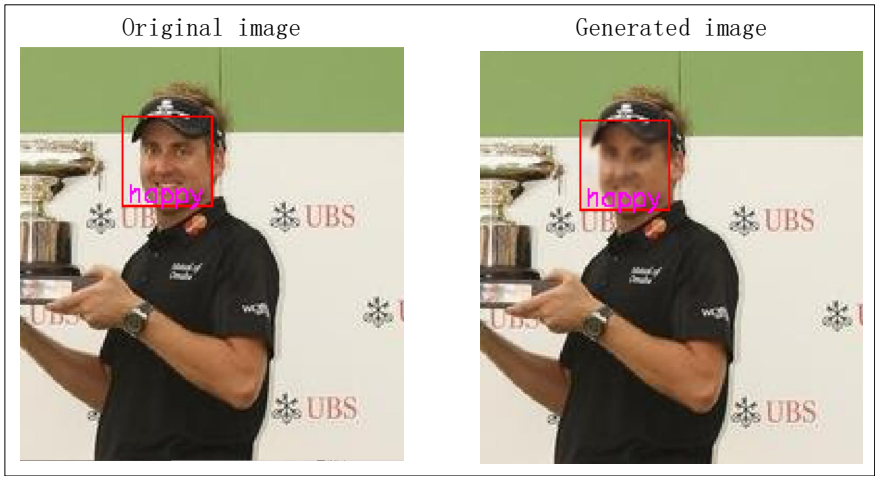

Figure 14. Image utility test.

6. Discussion

Based on the experimental data, we can find that QM-VAE has a stable and superior utility retention ability on the premise of removing privacy. This can be attributed to the multi-input model, which fully integrates a number of existing technologies to realize the protection of privacy. The generative network is used to solve the nonlinear relationship between facial privacy features and the image utility that needs to be preserved. The use of the autoencoder compresses the data in a low-dimensional space, avoiding the explosion of computing requirements, thus efficiently achieving the optimization of privacy protection and quality preservation.

In our work, we quantitatively demonstrated the effectiveness of QM-VAE in maintaining the quality of service, but we lacked a quantitative method to measure the credibility of privacy information removal in QM-VAE. In a future study, it would be meaningful to formulate an appropriate quantitative model to measure the degree of privacy protection. Our future research can provide a breakthrough in the field of face images and be applied to more aspects of video surveillance, such as privacy-sensitive information such as body movements or license plates.

7. Conclusions

In this paper, a novel face de-identification framework named QM-VAE was proposed, integrating various privacy protection methods. We innovatively considered service reviews as a black box and used the results to guide the generation of images that maintain utility. In the experiments, we quantitatively proved that QM-VAE performs admirably in terms of maintaining image utility. We are the first to develop a universal method for privacy protection with great flexibility and efficiency for various service situations.

Author Contributions: Conceptualization, B.S.; methodology, Y.Q., Z.N. and B.S.; software, Y.Q. and Z.N.; validation, Y.Q. and Z.N.; formal analysis, Y.Q.; investigation, Z.N.; resources, B.S.; data curation, Y.Q. and Z.N.; writing—original draft preparation, Y.Q. and Z.N.; writing—review and editing, Y.Q., B.S., T.M., A.A.-D. and M.A.-D.; visualization, Y.Q.; supervision, B.S.; project administration, B.S. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the National Key Research and Development Program of China (International Technology Cooperation Project No. 2021YFE014400) and the National Science Foundation of China (No. 42175194) for funding this work. This work was supported in part by the Deanship of Scientific Research at King Saud University through Research Group No. RGP-264.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Agarwal, A.; Chattopadhyay, P.; Wang, L. Privacy preservation through facial de-identification with simultaneous emotion preservation. *Signal Image Video Process.* **2021**, *15*, 951–958. [CrossRef]
2. Letournel, G.; Bugeau, A.; Ta, V.T.; Domenger, J.P. Face de-identification with expressions preservation. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4366–4370.
3. Wu, Y.; Yang, F.; Xu, Y.; Ling, H. Privacy-protective-GAN for privacy preserving face de-identification. *J. Comput. Sci. Technol.* **2019**, *34*, 47–60. [CrossRef]
4. Nousi, P.; Papadopoulos, S.; Tefas, A.; Pitas, I. Deep autoencoders for attribute preserving face de-identification. *Signal Process. Image Commun.* **2020**, *81*, 115699. [CrossRef]
5. Brkić, K.; Hrkać, T.; Kalafatić, Z.; Sikirić, I. Face, hairstyle and clothing colour de-identification in video sequences. *IET Signal Process.* **2017**, *11*, 1062–1068. [CrossRef]
6. Oord, A.v.d.; Vinyals, O.; Kavukcuoglu, K. Neural discrete representation learning. *arXiv* **2017**, arXiv:1711.00937.
7. Pan, Z.; Yu, W.; Lei, J.; Ling, N.; Kwong, S. TSAN: Synthesized view quality enhancement via two-stream attention network for 3D-HEVC. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 345–358. [CrossRef]
8. Peng, B.; Lei, J.; Fu, H.; Jia, Y.; Zhang, Z.; Li, Y. Deep video action clustering via spatio-temporal feature learning. *Neurocomputing* **2021**, *456*, 519–527. [CrossRef]
9. Lei, J.; Li, X.; Peng, B.; Fang, L.; Ling, N.; Huang, Q. Deep spatial-spectral subspace clustering for hyperspectral image. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 2686–2697. [CrossRef]
10. 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]
11. Neustaedter, C.; Greenberg, S.; Boyle, M. Blur filtration fails to preserve privacy for home-based video conferencing. *Acm Trans. -Comput.-Hum. Interact. (Tochi)* **2006**, *13*, 1–36. [CrossRef]
12. Yu, J.; Zhang, B.; Kuang, Z.; Lin, D.; Fan, J. iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 1005–1016. [CrossRef]
13. Liu, J.; Yin, S.; Li, H.; Teng, L. A Density-based Clustering Method for K-anonymity Privacy Protection. *J. Inf. Hiding Multim. Signal Process.* **2017**, *8*, 12–18. [CrossRef]
14. Gross, R.; Sweeney, L.; De La Torre, F.; Baker, S. Semi-supervised learning of multi-factor models for face de-identification. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.

15. Sun, Z.; Meng, L.; Ariyaeeinia, A. Distinguishable de-identified faces. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; Volume 4, pp. 1–6.
16. Newton, E.M.; Sweeney, L.; Malin, B. Preserving privacy by de-identifying face images. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 232–243. [CrossRef]
17. Samarzija, B.; Ribaric, S. An approach to the de-identification of faces in different poses. In Proceedings of the 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 26–30 May 2014; pp. 1246–1251.
18. Gross, R.; Sweeney, L.; De la Torre, F.; Baker, S. Model-based face de-identification. In Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), New York, NY, USA, 17–22 June 2006; p. 161.
19. Meden, B.; Emeršič, Ž.; Štruc, V.; Peer, P. k-Same-Net: k-Anonymity with generative deep neural networks for face deidentification. *Entropy* **2018**, *20*, 60. [CrossRef] [PubMed]
20. Dosovitskiy, A.; Springenberg, J.T.; Tatarchenko, M.; Brox, T. Learning to generate chairs, tables and cars with convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 692–705. [CrossRef] [PubMed]
21. Gross, R.; Airoldi, E.; Malin, B.; Sweeney, L. Integrating utility into face de-identification. In Proceedings of the 5th International Workshop on Privacy Enhancing Technologies, PET, Cavtat, Croatia, 30 May–1 June 2005; pp. 227–242.
22. Xu, Y.; Shang, L.; Ye, J.; Qian, Q.; Li, Y.F.; Sun, B.; Li, H.; Jin, R. Dash: Semi-supervised learning with dynamic thresholding. In Proceedings of the International Conference on Machine Learning, Virtual Event, Vienna, Austria, 18–24 July 2021; pp. 11525–11536.
23. Ning, X.; Wang, X.; Xu, S.; Cai, W.; Zhang, L.; Yu, L.; Li, W. A review of research on co-training. *Concurr. Comput. Pract. Exp.* **2021**, *21*, e6276. [CrossRef]
24. Phillips, P.J.; Wechsler, H.; Huang, J. The FERET database and evaluation procedure for face–recognition algorithms. *Image Vis. Comput.* **1998**, *16*, 295–306. [CrossRef]
25. Liu, C.; Wang, Y.; Chi, H.; Wang, S. Utility Preserved Facial Image De-identification Using Appearance Subspace Decomposition. *Chin. J. Electron.* **2021**, *30*, 413–418.
26. Ian, G.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27* , 2672–2680.
27. Cai, Z.; Xiong, Z.; Xu, H.; Wang, P.; Li, W.; Pan, Y. Generative adversarial networks: a survey towards and secure applications. *arXiv* **2021**, arXiv:2106.03785.
28. Han, C.; Xue, R. Differentially private GANs by adding noise to Discriminator's loss. *Comput. Secur.* **2021**, *107*, 102322. [CrossRef]
29. Yang, R.; Ma, X.; Bai, X.; Su, X. Differential Privacy Images Protection Based on Generative Adversarial Network. In Proceedings of the 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Guangzhou, China, 9 February 2021; pp. 1688–1695.
30. Hukkelås, H.; Mester, R.; Lindseth, F. Deepprivacy: A generative adversarial network for face anonymization. In Proceedings of the International Symposium on Visual Computing, Lake Tahoe, NV, USA, 7–9 October 2019; Springer: Berlin/Heidelberg, Germany , 2019; pp. 565–578.
31. 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, Salt Lake City, UT, USA, 18–23 June 2018, pp. 5050–5059.
32. Li, Y.; Lu, Q.; Tao, Q.; Zhao, X.; Yu, Y. SF-GAN: Face De-identification Method without Losing Facial Attribute Information. *IEEE Signal Process. Lett.* **2021**, *28*, 1345–1349. [CrossRef]
33. Chen, J.; Konrad, J.; Ishwar, P. Vgan-based image representation learning for privacy-preserving facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1570–1579.
34. Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; Zemel, R. The variational fair autoencoder. *arXiv* **2015**, arXiv:1511.00830.
35. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784 .
36. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
37. Yu, J.; Xue, H.; Liu, B.; Wang, Y.; Zhu, S.; Ding, M. GAN-Based Differential Private Image Privacy Protection Framework for the Internet of Multimedia Things. *Sensors* **2021**, *21*, 58. [CrossRef]
38. Zhou, G.; Qin, S.; Zhou, H.; Cheng, D. A differential privacy noise dynamic allocation algorithm for big multimedia data. *Multimed. Tools Appl.* **2019**, *78*, 3747–3765. [CrossRef]
39. Bu, Z.; Dong, J.; Long, Q.; Su, W.J. Deep learning with Gaussian differential privacy. *Harv. Data Sci. Rev.* **2020**, *2* , 3747–3765. [CrossRef]
40. Dong, J.; Roth, A.; Su, W.J. Gaussian differential privacy. *arXiv* **2019**, arXiv:1905.02383.
41. Kim, T.; Yang, J. Latent-Space-Level Image Anonymization With Adversarial Protector Networks. *IEEE Access* **2019**, *7*, 84992–84999. [CrossRef]
42. Fredrikson, M.; Jha, S.; Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1322–1333.

43. Croft, W.L.; Sack, J.R.; Shi, W. Obfuscation of images via differential privacy: from facial images to general images. *Peer-to-Peer Netw. Appl.* **2021**, *14*, 1705–1733. [CrossRef]
44. Chamikara, M.A.P.; Bertók, P.; Khalil, I.; Liu, D.; Camtepe, S. Privacy preserving face recognition utilizing differential privacy. *Comput. Secur.* **2020**, *97*, 101951. [CrossRef]
45. Nasr, M.; Shokri, R.; Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2019; pp. 739–753.
46. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017, pp. 3–18.
47. Kim, T.; Yang, J. Selective feature anonymization for privacy-preserving image data publishing. *Electronics* **2020**, *9*, 874. [CrossRef]
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
49. Liu, C.; Yang, J.; Zhao, W.; Zhang, Y.; Li, J.; Mu, C. Face Image Publication Based on Differential Privacy. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 6680701. [CrossRef]
50. Carpentieri, B.; Castiglione, A.; De Santis, A.; Palmieri, F.; Pizzolante, R. Privacy-preserving Secure Media Streaming for Multi-user Smart Environments. *ACM Trans. Internet Technol. (Toit)* **2021**, *22*, 1–21. [CrossRef]
51. Xu, H.; Cai, Z.; Takabi, D.; Li, W. Audio-visual autoencoding for privacy-preserving video streaming. *IEEE Internet Things J.* **2021**, *9*, 1749–1761. [CrossRef]
52. Kim, J.; Kim, M.; Kang, H.; Lee, K. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv* **2019**, arXiv:1907.10830.
53. Meng, L.; Sun, Z.; Ariyaeeinia, A.; Bennett, K.L. Retaining expressions on de-identified faces. In Proceedings of the 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 26–30 May 2014.
54. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015 .
55. Ma, T.; Rong, H.; Hao, Y.; Cao, J.; Tian, Y.; Al-Rodhaan, M.A. A novel sentiment polarity detection framework for Chinese. *IEEE Trans. Affect. Comput.* **2019**, *13*, 60–74. [CrossRef]
56. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
57. Ma, T.; Wang, H.; Zhang, L.; Tian, Y.; Al-Nabhan, N. Graph classification based on structural features of significant nodes and spatial convolutional neural networks. *Neurocomputing* **2021**, *423*, 639–650. [CrossRef]
58. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014 .
59. Jeon, J.; Park, J.C.; Jo, Y.; Nam, C.; Bae, K.H.; Hwang, Y.; Kim, D.S. A real-time facial expression recognizer using deep neural network. In Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, Da Nang, Vietnam, 4–6 January 2016; pp. 1–4.
60. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020** , *11*, 1. [CrossRef]
61. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Korea, 3–7 November 2013 .
62. Viola, P.A.; Jones, M.J. Rapid Object Detection using a Boosted Cascade of Simple Features. In Proceedings of the 2001 IEEE Computer Society Conference on Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001.
63. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Large-Scale CelebFaces Attributes (CelebA) Dataset. Available online: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 7 March 2001).
64. Bock, S.; Weiß, M. A proof of local convergence for the Adam optimizer. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.