

SUMMARY OF CHANGES

We are immensely pleased to be offered the helpful suggestions, and have revised the manuscript fully according to the reviewers' comments. In this revision, we have taken the opportunity to address reviewers' concerns that were kindly drawn to our attention. The following summarizes the latest revision conducted to the paper.

1. Preliminaries section was merged to Related Work. Problem section and ϵ -Dimension Reduction Privacy section were merged to Methodology.
2. Section 2 (Related Work) was thoroughly revised to update our references with more latest work (around 10 most recent works were added).
3. Section 3.1.1 (Problem Statement) was revised and modified to give more information about the problem and the sample scenario.
4. All formula serial numbers were added. Figures and tables are rearranged to improve the readability.
5. Conducting more experiments with different structure of AutoGAN-DRP (i.e., VGG19, VGG16, basic CNN were applied to Generator and Re-constructor). Results were updated in Section 4.
6. Implementing AutoGAN-DRP on one more color dataset (CelebA). Experiment results and discussion in Section 4 were updated correspondingly.
7. Table 1 (Implementation information) in Section 4 was added to provide precise implementation information of components' structure in Section 4 (Experiment and Discussion)
8. Section 6 was added and new experiments were conducted to compare to other privacy preservation techniques using Differential Privacy (DP) and Principle Component Analysis (PCA).
9. Table 2 (Sample visualization of AutoGAN, DP, PCA over three datasets) in Section 6 was added to show more intuitive results of AutoGAN, DP, PCA based techniques.

I. RESPONSE TO REVIEWER 1

The authors would like to express sincere thanks to the reviewer for the thorough and constructive comments. In what follows, we present detailed comments in response to the individual points raised by the reviewer and elaborate on how the manuscript has been revised.

[Comment 1.1](#) The authors proposed an input-privacy-preserving technique that reduces input images into lower-dimensional signals. The difficulty of image recovery is enhanced by training the encoder ("Generator" in this manuscript) to produce low-dimensional signal that cannot be decoded by another neural network ("Reconstruct" in Figure 3). Their idea is reasonable to some extent, but it is difficult for me to regard it as a serious privacy-preserving mechanism, mainly because the security is validated only heuristically by the reconstructor and discriminator used in the training phase. Since the capacity of reconstructor is finite, it may possible to decode reduced inputs by a more powerful reconstructor which can be easily prepared.

Response:

Thank you for the insightful comment. To the best of our knowledge, in the case of using non-linear methods to reduce number of data dimension especially neural network, the well-known method could be used to recover the original data is to use an auto-encoder. The architecture of an auto-encoder mainly includes two parts, encoder and decoder. The encoder structure varies and could be fully connected network, convolutional network. The decoder structure usually is an inverted structure of the encoder. Instead of using typical fully-connected neural network, we use Deep Convolutional Neural Networks which are more effective for images. To investigate the robustness of AutoGAN-DRP, in this revision we conducted more experiments using most current powerful structure of convolution networks (i.e., VGG16, VGG19) for encoders (Generator) and their inverted structures for decoders (Reconstructor). By examining most recent powerful structure of reconstructor, we aim at evaluating strong adversaries who attempt to reconstruct our original data. Beside the comparison with GAP in section 5, we also conduct more experiments to compare with other privacy preservation techniques (i.e., Differential Privacy, Principle Component Analysis for privacy preservation) in section 6. The result in section 4, 5, 6 show that our methods not only maintain data utility but also preserve data owners' privacy.

”

4. Experiments and Discussion

In this section, we demonstrate our experiments over three popular supervised face image datasets: *the Extended Yale Face Database B* [29], *AT&T* [30] and *CelebFaces Attributes Dataset (CelebA)* [31].

4.1. Experiment Setup

The Extended Yale Face Database B (YaleB) contains 2,470 grayscale images of 38 human subjects under different illumination conditions and their identity label. In this dataset, the image size is 168x192 pixels. The AT&T dataset has 400 face images of 40 subjects. For convenience, we resize each image of these two dataset to 64x64 pixels. CelebA is a color facial image dataset containing 202,599 images of 10,177 subjects. 1,709 images of the first 80 subjects are used for our experiment. Each image is resized to 64x64x3 pixels. All pixel values are scaled to the range of [0,1]. We randomly select 10% of each subject's images for validation and 15% for testing dataset.

The Generator and Re-constructor in Figure 3 are implemented by three different structures. Specifically, we follow the architecture of recent powerful models VGG19, VGG16 [32] and a basic convolutional network (CNN). We modify the models to adapt with our data size (64x64). Discriminator and Classifier are built on fully connected neural network and convolutional network respectively. leaky ReLU is used as activation function for hidden layers. We use linear activation function for Generator output layers and softmax activation functions for other components' output layers. Each component is trained in 5 local iterations (n_r, n_g, n_d, n_c), and the entire system is trained in 500 iterations for global loop (n). The target distribution is drawn from Gaussian distribution (with the covariance value of 0.5 and the mean is the average of the training data). Table 1 provides detail information of neural networks' structures and other implementation information.

To evaluate the reliability of the system we test our system with different level of authentication corresponding to binary classification (single-level) and multi-class classification (multi-level). For the single-level authentication system in the scenario, we consider half of the subjects in the dataset are valid to access the database system while the rest are invalid. We randomly divide the dataset into two groups of subject and labels their images to (1) or (0) depending on their access permission. For the cases of multi-level authentication system experiments, we divide the subjects into four groups and eight groups so that the authentication server becomes four-class and eight-class classifier respectively.

4.2. Utility

We use accuracy metric to evaluate the utility of dimensionally-reduced data. The testing dataset is tested with the classifier extracted from our framework. Different structures of Generator and Re-constructor are applied including VGG19, VGG16, basic CNN on different privilege levels which correspond to multi-class classification. Figure 4 illustrates the accuracies for different dimensions from three to seven over the three facial datasets. Overall, the accuracies improve when the number of dimension increases. The accuracies on the two gray image datasets (AT&T and Yale_B) reaches 90% and higher when using VGG with only seven dimensions. The figure for Celeba is smaller but it still reaches 80%. In general, VGG19 structure performs better than using VGG16 and basic CNN in terms of utility due to the complexity (table 1) and adaptability to image datasets of VGG19. As the dimension number is reduced from 4,096 (64x64) to 7, we can achieve a compression ratio of 585 yet achieve accuracy of 90% for the two gray datasets and 80% for the color dataset. This implies our method could gain a high compression ratio and maintain a high utility in terms of accuracy. During conducting experiments we also observe that the accuracy could be higher if we keep the original resolution of images. However for convenience and reducing the complexity of our structure we resize images to the same size of 64x64 pixels.

4.3. Privacy

In this study, the Euclidean distance is used to measure the distance between original and reconstructed images: $dist(x, \hat{x}) = ||x - \hat{x}||^2$. Figure 5 illustrates the average distances between original images and reconstructed images on testing data with different ϵ constraints for seven dimensions and single-level authentication and VGG19 structure. The achieved distances (red lines) are always larger than the hyper-parameter ϵ (black dotted lines) where ϵ is less than 0.035 for AT&T, 0.052 for YaleB and 0.067 for CelebA. Due to the fact that the Re-constructor is trained using the training dataset (we consider the adversary can reach the model and the training data), our framework can only force the distance within a certain range as shown in 5. Since the mean of the target distribution is set to the mean of training dataset, reconstructed

images will be close to the mean of training dataset which we believe it will enlarge the distance and expose less individual information. Therefore, the range of epsilon can be estimated base on the expectation of the distance between the testing sample and the mean of training data. The intersection between the red line and the dotted black line points out the largest distance our framework can achieve. The first section of table 2 demonstrates samples and their corresponding reconstructions in single-level authentication and seven dimensions with different achieved accuracies and distances. The reconstructed images could reach nearly identical, thus making it visually hard to recognize the identity of an individual. ”

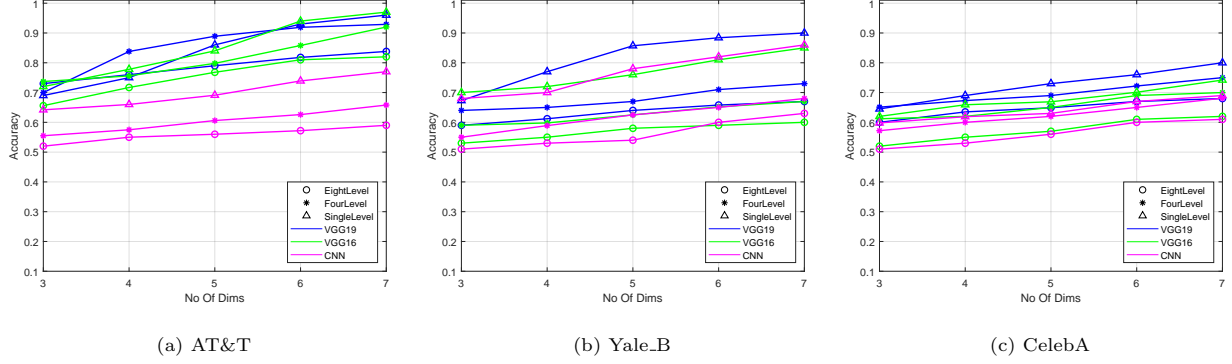


Figure 4: Accuracy for Different Number of Reduced Dimensions

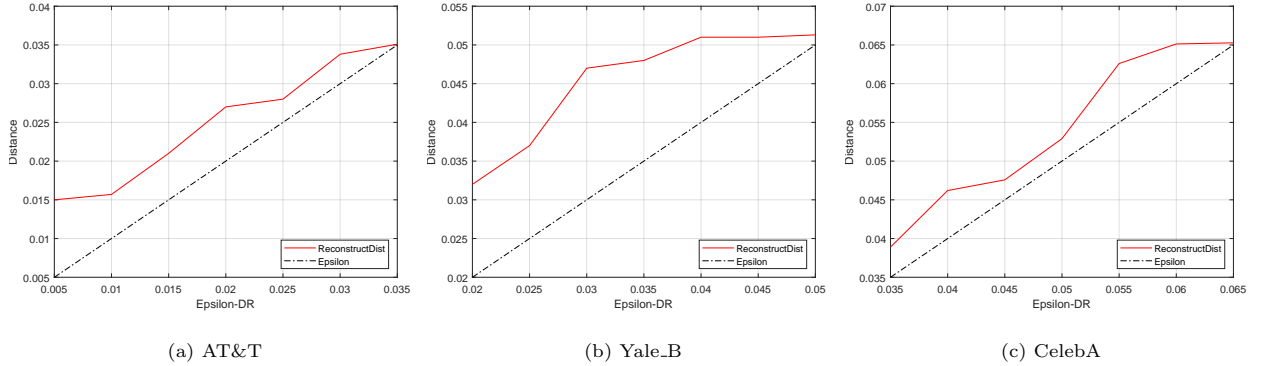


Figure 5: Average Distance Measurement Result { 7 dimensions, Single-Level}

6. Visual comparison to privacy preservation techniques using Differential Privacy (DP) [24] and Principle Component Analysis (PCA) [33]

In this section, we compare AutoGAN-DRP ability to visually protect privacy with other privacy preservation methods. We choose a widely used tool for privacy preserving Differential Privacy (DP) [24] and another privacy preservation method utilizing a dimensionality reduction technique Principle Component Analysis (PCA)[33].

In these experiments, we implement AutoGAN-DRP with VGG19 structure for the Generator and Reconstructor and the parameter setting as shown in table 1. The images are reduced to seven dimensions

for different values of $\epsilon - DR$ to achieve different distances and accuracies. The datasets are grouped into two groups corresponding to binary classifiers. For implementing DP, we first generate a classifier on the authentication server by training the datasets with a VGG19 binary classifier (the structure of hidden layers is similar to our Generator in table 1). The testing images are then perturbed using differential privacy method. Specifically, Laplace noise is added to the images with the sensitivity coefficient of 1 (it is computed by the maximum range value of each pixel [0,1]) and different DP epsilon parameters (this DP epsilon is different from our $\epsilon - DR$). The perturbed images are then sent to the authentication server and fed to the classifier. We visually compare the perturbed images and the accuracy of this method and AutoGAN.

In addition, we follow instruction from FRiPAL [11] in which the clients reduce image dimension using Principle Component Analysis method and send reduced features to the server. FRiPAL claims that by reducing image dimension, their method can be more resistant to reconstruction attacks. The experiments are performed with different number of reduced dimension. The images are reconstructed using *Moore-Penrose inverse* method with assumption that an adversary has access to the model. The classification accuracy is evaluated using a classifier which has similar structure to AutoGAN’s classifier.

Table 2 shows image samples and results over the three datasets. Overall, AutoGAN-DRP is more resilient to reconstruction attacks compared to the other two techniques. For instant, at the accuracy of 79% on AT&T dataset, 80% on YaleB and 73% on CelebA, we cannot distinguish entities from the others. For DP method, the accuracy decreases when the DP epsilon decreases (adding more noise). Thus, the perturbed images are harder to recognize. However, at the accuracy of 57%, we are still able to distinguish identities by human eyes because DP noise does not focus on the important pixels. For PCA, the accuracy also goes down when the number of dimensions reduces and the distances increase. Since PCA transformation is linear and deterministic, the original important information can be significantly reconstructed using the inverse transformation deriving from the model or training data. For example, at the accuracy of 75% on AT&T, 71% on YaleB and 68% on CelebA we still can differentiate individuals in the group. Thus, for security purposes, our proposed method shows the advantage in securing the data. ”




























































































































		AT&T				YaleB				CelebA			
Acc			0.93	0.79	0.65		0.90	0.80	0.69		0.73	0.66	0.59
Dist			0.0116	0.0198	0.0245		0.0184	0.0246	0.0585		0.0513	0.0531	0.06618
AutoGAN-DRP													
													
													
													
	Org	(7)	(7)	(7)		Org	(7)	(7)	(7)	Org	(7)	(7)	(7)
Acc			0.69	0.63	0.57		0.68	0.60	0.58		0.62	0.59	0.56
Dist			0.0164	0.0313	0.0405		0.0149	0.0314	0.0407		0.0200	0.0418	0.0509
Differential Privacy													
													
													
													
	Org	(11)	(8)	(7)		Org	(11)	(8)	(7)	Org	(11)	(8)	(7)
Acc			0.90	0.75	0.60		0.87	0.83	0.71		0.71	0.68	0.57
Dist			0.0197	0.0264	0.0348		0.0228	0.0266	0.0287		0.0362	0.0379	0.0511
PCA													
													
													
													
	Org	(15)	(7)	(5)		Org	(15)	(7)	(5)	Org	(15)	(7)	(5)
Acc : Average accuracy on testing data Dist: Average Euclidean distance between original images and reconstructed/perturbed images Org : Original images () Experiment parameters: epsilon for DP and number of reduced dimensions for PCA and AutoGAN-DRP													

Table 2: Sample visualization of AutoGAN, DP, PCA over three datasets

[Comment 1.2](#) My concern is enhanced by the fact that I could not find the precise information on the neural network architectures used in the study.

Response:

In the revision, we added table 1 and provided more information on neural network architectures in section 4.1 (Experiment setup) as follows.

” The Generator and Re-constructor in Figure 3 are implemented by three different structures. Specifically, we follow the architecture of recent powerful models VGG19, VGG16 [32] and a basic convolutional network (CNN). We modify the models to adapt with our data size (64x64). Discriminator and Classifier are built on fully connected neural network and convolutional network respectively. leaky ReLU is used as activation function for hidden layers. We use linear activation function for Generator output layers and softmax

activation functions for other components' output layers. Each component is trained in 5 local iterations (n_r, n_g, n_d, n_c), and the entire system is trained in 500 iterations for global loop (n). The target distribution is drawn from Gaussian distribution (with the covariance value of 0.5 and the mean is the average of the training data). Table 1 provides detail information of neural networks' structures and other implementation information. ”

	VGG16			VGG19			CNN		
	Hidden layers	Units	Parameter	Hidden layers	Units	Parameter	Hidden layers	Units	Parameter
Generator	Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Dense Dense	64x2 128x2 256x3 512x3 512x3 1024 1024	16,295,623	Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Dense Dense	64x2 128x2 256x4 512x4 512x4 1024 1024	21,605,319	Conv BatchNorm Conv BatchNorm Conv BatchNorm Dense	256 512 1024 1024	16,451,847
Reconstructor	Dense Dense Dense Reshape Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T	1024 1024 1024 512x3 512x3 256x3 128x2 64x2	10,184,000	Dense Dense Dense Reshape Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T	1024 1024 1024 512x4 512x4 512x4 256x4 128x2 64x2	13,281,472	Dense BatchNorm Reshape Conv BatchNorm Conv BatchNorm Conv	1024 1024 1024 512 256	18,048,256
Classifier	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048	12,636,168	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048	12,636,168	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048	12,636,168
Discriminator	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737
Shared parameters: optimizer Adam, learning rate 0.0001, 7 dimensions Hardware: GPU Testla T4 16Gb, CPU Xeon Processors @2.3Ghz Software: Tensorflow 2.0 beta. The number of parameters are reported by model.summary() from Keras library									

Table 1: Implementation information

II. RESPONSE TO REVIEWER 2

The authors would like to express sincere thanks to the reviewer for the thorough and constructive comments. In what follows, we present detailed comments in response to the individual points raised by the reviewer and elaborate on how the manuscript has been revised.

[Comment 2.1](#) This paper proposed a dimension reduction-based method for privacy preservation. However, there are some problems.

1. The organizational structure of the article is unreasonable, section 2, section 3 and section 4 should be introduced in the introduction and related work. I hope it can improve readability.

Response:

Thank you for the insightful comment.

1. As recommended, in this revision we merged Section 3 (Preliminary) into Section 2 (Related Work).
2. The problem was introduced in the introduction. However, to clarify the problem and provide a comprehensive scenario for the problem, we introduced again the problem in section 4. For this reason, in the revision Section 4 (Problem) and Section 5 (ϵ -Dimension Reduction Privacy) were merged to the top of Section 6 (Methodology). Thus, we hope the changes can support the readability of our methodology.

[Comment 2.2](#) Please briefly describe the content in Figure 1, which will be more helpful for readers to understand.

Response:

Thank you for your insightful comment. In the revision, we clarify Figure 1 in the problem statement section (Section 3.1.1) as follows. " We introduce the problem through the practical scenario mentioned in Section 1. Figure 1 briefly describes the entire system in which staff members (clients) in a company request access to company resources, such as websites and data servers through a face recognition access control system. For example, if member n requests to web server 2, the local device first takes a facial photo of the member by an attached camera, locally transforms it into lower dimension and send to an authentication center. The authentication server then obtains the reduced dimensional features and determine his/her access eligibility without the clear face of the requesting member using a classifier. We consider that the system has three levels of privilege (i.e., single level, four-level, eight-level) corresponding to three groups of members. We assume the authentication server is semi-honest (it obeys the work procedure but might be used to infer personal information). Once the server is compromised, an adversary in the authentication center can reconstruct the face features to achieve plain-text face images and determine members' identity. "

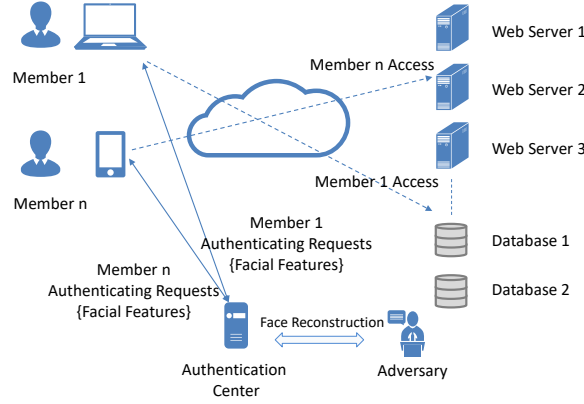


Figure 1: Attack Model

[Comment 2.3](#) Please explain what is the operation of $E[\cdot]$ in equation (1)?

Response:

In the revision we added explanation for equation (1) as follows.


”

$$E[|dist(x, \hat{x})|] \geq \epsilon \quad (1)$$

where $E[\cdot]$ is the expectation, $\epsilon \geq 0$, $x' = F(x)$, $\hat{x} = R(x')$, and $R(\cdot)$ is the Reconstruction Function. ”

[Comment 2.4](#) Please explain which method is used to project original data to low dimensional space?

Response:

In this revision, we clarify and provide more detail on the method used to project data to lower dimension space in Section 5 (Methodology) as follows .

” To tackle the problem, we propose a deep learning framework for transforming face images to lower dimension vectors which are hard to be clearly reconstructed. We leverage the structure of an auto-encoder [19] which contains encoder and decoder (in this work, we called them generator and re-constructor) in order to reduce data dimension. More specifically, the lower dimensional represents are extracted from the middle layer of the auto-encoder (output of the generator). The dimensionally-reduced data can be sent to the authentication server as an authentication request. We consider an adversarial as a re-constructor implemented by the decoder of the Auto-encoder. To prevent fully reconstructing images, the framework utilizes the discriminator in GAN [18] to direct reconstructed data to a desired target distribution. In this work, the target distribution is sampled from Gaussian distribution and the mean is the average of the whole training data. After the transformation projects the data into a lower dimension domain, the re-constructor can only partially reconstruct the data. Therefore, the adversary might not be able to recognize an individual’s identity. To maintain data utility, we also use feedback from a classifier while training the auto-encoder as a part of its loss function. The entire framework can enlarge the distance between original data and its reconstruction to preserve individual privacy and retain significant data information. The dimensionally-reduced transformation model is extracted from the framework and provided to clients for reducing their facial image dimensions. The classification model will be used in an authentication center that classifies whether an authentication request is valid to have access $\{1\}$ or not $\{0\}$. ”

[Comment 2.5](#) Please explain how to estimate the range of values of epsilon ?

Response:

In this study, the distance is defined as the L2 norm distance between the original data and reconstructed data. In this revision we added our observation in order to estimate the range of ϵ to Section 4 (Experiment and Discussion) as follows.

”...It is our view that the reconstructed images will be close to the mean of training dataset so that it will expose less individual information. Therefore, the epsilon can be estimated base on the expectation of the distance between the original data and the mean of training data...”

[Comment 2.6](#) There are few methods for comparing experiments, no convincing. More recent methods should be used in the comparison experiments. In addition, computation complexity or time consumption of the methods should also be given. I hope this will help to improve the quality of the paper.

Response:

In this revision we conducted more experiments using most current powerful structure of convolution networks (i.e., VGG16, VGG19 [32], basic CNN) for encoders (Generator) and their inverted structures for decoders (Reconstructor). By examining most recent powerful structure of reconstructor, we aim at evaluating strong adversaries who attempt to reconstruct our original data. Beside the comparison with GAP in section 5, we also conduct more experiments to compare with other privacy preservation techniques (i.e., Differential Privacy, Principle Component Analysis for privacy preservation) in section 6. In addition, we added table 1 in Section 4 which provides detail information of used neural networks, number of parameters in each component and hardware and software information to improve paper quality. Section 4 (Experiment and Discussion) is updated and Section 6 is added as follows.

”

4. Experiments and Discussion

In this section, we demonstrate our experiments over three popular supervised face image datasets: *the Extended Yale Face Database B* [29], *AT&T* [30] and *CelebFaces Attributes Dataset (CelebA)* [31].

4.1. Experiment Setup

The Extended Yale Face Database B (YaleB) contains 2,470 grayscale images of 38 human subjects under different illumination conditions and their identity label. In this dataset, the image size is 168x192 pixels. The AT&T dataset has 400 face images of 40 subjects. For convenience, we resize each image of these two dataset to 64x64 pixels. CelebA is a color facial image dataset containing 202,599 images of 10,177 subjects. 1,709 images of the first 80 subjects are used for our experiment. Each image is resized to 64x64x3 pixels. All pixel values are scaled to the range of [0,1]. We randomly select 10% of each subject’s images for validation and 15% for testing dataset.

The Generator and Re-constructor in Figure 3 are implemented by three different structures. Specifically, we follow the architecture of recent powerful models VGG19, VGG16 [32] and a basic convolutional network (CNN). We modify the models to adapt with our data size (64x64). Discriminator and Classifier are built on fully connected neural network and convolutional network respectively. leaky ReLU is used as activation function for hidden layers. We use linear activation function for Generator output layers and softmax activation functions for other components’ output layers. Each component is trained in 5 local iterations

(n_r, n_g, n_d, n_c) , and the entire system is trained in 500 iterations for global loop (n). The target distribution is drawn from Gaussian distribution (with the covariance value of 0.5 and the mean is the average of the training data). Table 1 provides detail information of neural networks' structures and other implementation information.

	VGG16			VGG19			CNN		
	Hidden layers	Units	Parameter	Hidden layers	Units	Parameter	Hidden layers	Units	Parameter
Generator	Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Dense Dense	64x2 128x2 256x3 512x3 512x3 1024 1024	16,295,623	Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Dense Dense	64x2 128x2 256x4 512x4 512x4 1024 1024	21,605,319	Conv BatchNorm Conv BatchNorm Conv BatchNorm Dense	256 512 1024 1024	16,451,847
Reconstructor	Dense Dense Dense Reshape Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T	1024 1024 1024 512x3 512x3 256x3 128x2 64x2	10,184,000	Dense Dense Dense Reshape Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T	1024 1024 1024 512x4 512x4 256x4 128x2 64x2	13,281,472	Dense BatchNorm Reshape Conv BatchNorm Conv BatchNorm Conv	1024 1024 1024 512 512 256	18,048,256
Classifier	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048 2048	12,636,168	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048 2048	12,636,168	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048 2048	12,636,168
Discriminator	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737
Shared parameters: optimizer Adam, learning rate 0.0001, 7 dimensions Hardware: GPU Testla T4 16Gb, CPU Xeon Processors @2.3Ghz Software: Tensorflow 2.0 beta. The number of parameters are reported by model.summary() from Keras library									

Table 1: Implementation information

To evaluate the reliability of the system we test our system with different level of authentication corresponding to binary classification (single-level) and multi-class classification (multi-level). For the single-level authentication system in the scenario, we consider half of the subjects in the dataset are valid to access the database system while the rest are invalid. We randomly divide the dataset into two groups of subject

and labels their images to (1) or (0) depending on their access permission. For the cases of multi-level authentication system experiments, we divide the subjects into four groups and eight groups so that the authentication server becomes four-class and eight-class classifier respectively.

4.2. Utility

We use accuracy metric to evaluate the utility of dimensionally-reduced data. The testing dataset is tested with the classifier extracted from our framework. Different structures of Generator and Re-constructor are applied including VGG19, VGG16, basic CNN on different privilege levels which correspond to multi-class classification. Figure 4 illustrates the accuracies for different dimensions from three to seven over the three facial datasets. Overall, the accuracies improve when the number of dimension increases. The accuracies on the two gray image datasets (AT&T and Yale_B) reaches 90% and higher when using VGG with only seven dimensions. The figure for Celeba is smaller but it still reaches 80%. In general, VGG19 structure performs better than using VGG16 and basic CNN in terms of utility due to the complexity (table 1) and adaptability to image datasets of VGG19. As the dimension number is reduced from 4,096 (64x64) to 7, we can achieve a compression ratio of 585 yet achieve accuracy of 90% for the two gray datasets and 80% for the color dataset. This implies our method could gain a high compression ratio and maintain a high utility in terms of accuracy. During conducting experiments we also observe that the accuracy could be higher if we keep the original resolution of images. However for convenience and reducing the complexity of our structure we resize images to the same size of 64x64 pixels.

4.3. Privacy

In this study, the Euclidean distance is used to measure the distance between original and reconstructed images: $dist(x, \hat{x}) = ||x - \hat{x}||^2$. Figure 5 illustrates the average distances between original images and reconstructed images on testing data with different ϵ constraints for seven dimensions and single-level authentication and VGG19 structure. The achieved distances (red lines) are always larger than the hyper-parameter ϵ (black dotted lines) where ϵ is less than 0.035 for AT&T, 0.052 for YaleB and 0.067 for CelebA . Due to the fact that the Re-constructor is trained using the training dataset (we consider the adversary can reach the model and the training data), our framework can only force the distance within a certain range as shown in 5. Since the mean of the target distribution is set to the mean of training dataset, reconstructed images will be close to the mean of training dataset which we believe it will enlarge the distance and expose less individual information. Therefore, the range of epsilon can be estimated base on the expectation of the distance between the testing sample and the mean of training data. The intersection between the red line and the dotted black line points out the largest distance our framework can achieve. The first section of table 2 demonstrates samples and their corresponding reconstructions in single-level authentication and seven dimensions with different achieved accuracies and distances. The reconstructed images could reach nearly identical, thus making it visually hard to recognize the identity of an individual. ”

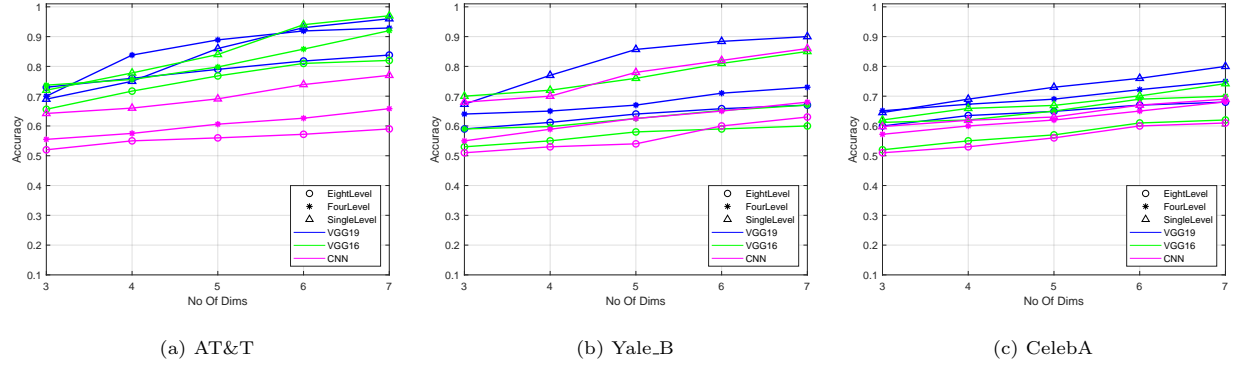


Figure 4: Accuracy for Different Number of Reduced Dimensions

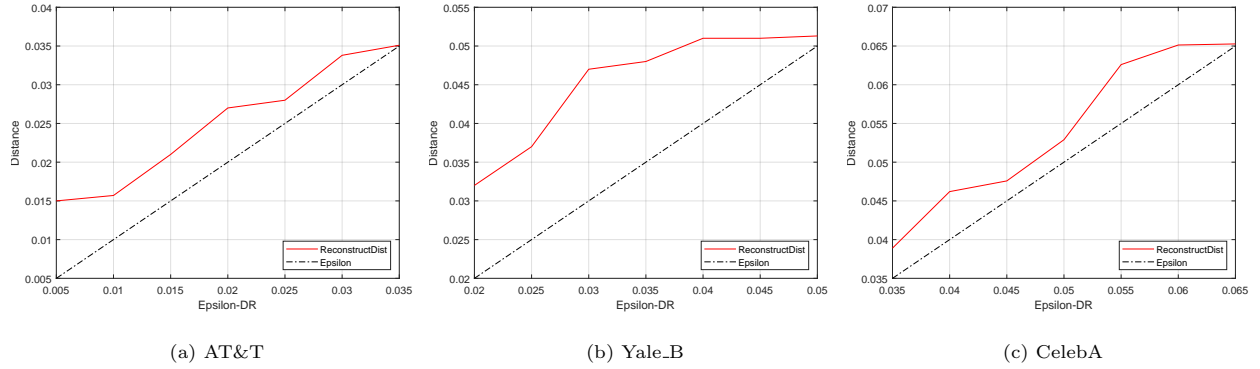


Figure 5: Average Distance Measurement Result { 7 dimensions, Single-Level}

6. Visual comparison to privacy preservation techniques using Differential Privacy (DP) [24] and Principle Component Analysis (PCA) [33]

In this section, we compare AutoGAN-DRP ability to visually protect privacy with other privacy preservation methods. We choose a widely used tool for privacy preserving Differential Privacy (DP) [24] and another privacy preservation method utilizing a dimensionality reduction technique Principle Component Analysis (PCA) [33].

In these experiments, we implement AutoGAN-DRP with VGG19 structure for the Generator and Reconstructor and the parameter setting as shown in table 1. The images are reduced to seven dimensions for different values of $\epsilon - DR$ to achieve different distances and accuracies. The datasets are grouped into two groups corresponding to binary classifiers. For implementing DP, we first generate a classifier on the authentication server by training the datasets with a VGG19 binary classifier (the structure of hidden layers is similar to our Generator in table 1). The testing images are then perturbed using differential privacy method. Specifically, Laplace noise is added to the images with the sensitivity coefficient of 1 (it is computed by the maximum range value of each pixel $[0,1]$) and different DP epsilon parameters (this DP epsilon is different from our $\epsilon - DR$). The perturbed images are then sent to the authentication server and fed to the classifier. We visually compare the perturbed images and the accuracy of this method and AutoGAN.

In addition, we follow instruction from FRiPAL [11] in which the clients reduce image dimension using Principle Component Analysis method and send reduced features to the server. FRiPAL claims that by reducing image dimension, their method can be more resistant to reconstruction attacks. The experiments are performed with different number of reduced dimension. The images are reconstructed using *Moore–Penrose inverse* method with assumption that an adversary has access to the model. The classification accuracy is evaluated using a classifier which has similar structure to AutoGAN’s classifier.

Table 2 shows image samples and results over the three datasets. Overall, AutoGAN-DRP is more resilient to reconstruction attacks compared to the other two techniques. For instance, at the accuracy of 79% on AT&T dataset, 80% on YaleB and 73% on CelebA, we cannot distinguish entities from the others. For DP method, the accuracy decreases when the DP epsilon decreases (adding more noise). Thus, the perturbed images are harder to recognize. However, at the accuracy of 57%, we are still able to distinguish identities by human eyes because DP noise does not focus on the important pixels. For PCA, the accuracy also goes down when the number of dimensions reduces and the distances increase. Since PCA transformation is linear and deterministic, the original important information can be significantly reconstructed using the inverse transformation deriving from the model or training data. For example, at the accuracy of 75% on AT&T, 71% on YaleB and 68% on CelebA we still can differentiate individuals in the group. Thus, for security purposes, our proposed method shows the advantage in securing the data.”

































































































		AT&T				YaleB				CelebA		
Acc		0.93	0.79	0.65		0.90	0.80	0.69		0.73	0.66	0.59
Dist		0.0116	0.0198	0.0245		0.0184	0.0246	0.0585		0.0513	0.0531	0.06618
AutoGAN-DRP												
												
												
												
	Org	(7)	(7)	(7)	Org	(7)	(7)	(7)	Org	(7)	(7)	(7)
Acc		0.69	0.63	0.57		0.68	0.60	0.58		0.62	0.59	0.56
Dist		0.0164	0.0313	0.0405		0.0149	0.0314	0.0407		0.0200	0.0418	0.0509
Differential Privacy												
												
												
												
	Org	(11)	(8)	(7)	Org	(11)	(8)	(7)	Org	(11)	(8)	(7)
Acc		0.90	0.75	0.60		0.87	0.83	0.71		0.71	0.68	0.57
Dist		0.0197	0.0264	0.0348		0.0228	0.0266	0.0287		0.0362	0.0379	0.0511
PCA												
												
												
												
	Org	(15)	(7)	(5)	Org	(15)	(7)	(5)	Org	(15)	(7)	(5)
Acc : Average accuracy on testing data Dist: Average Euclidean distance between original images and reconstructed/perturbed images Org : Original images () Experiment parameters: epsilon for DP and number of reduced dimensions for PCA and AutoGAN-DRP												

Table 2: Sample visualization of AutoGAN, DP, PCA over three datasets

[Comment 2.7](#) What is the advantage of AutoGAN-DRP in terms of privacy protection compared to GAP? The advice I would like to give is that the authors could give more details to explain the advantages of this method further.

Response:

AutoGAN-DRP is visually protecting the images themselves by reducing image dimension and extracting the most important information of images which maintains utility of data. As shown in Figure 6, at the same distortion level AutoGAN provides higher accuracy which implies our method can add more noise than GAP but still maintain high data utility.

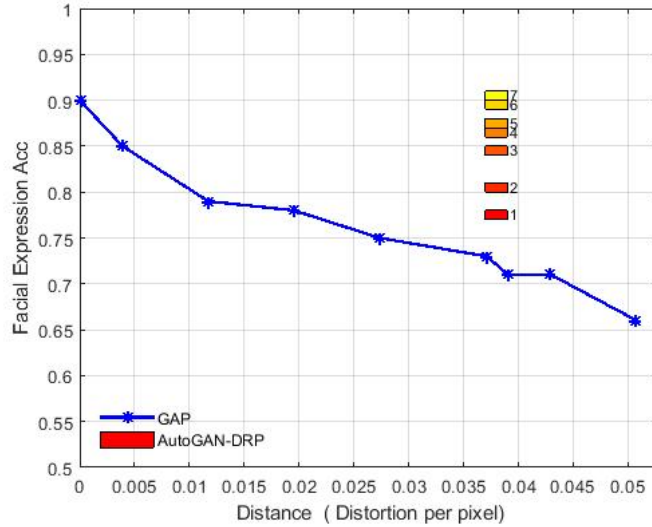


Figure 6: GENKI Facial Expression Accuracy Vs Distance using GAP and AutoGAN-DRP

To provide readers intuitive information about our method privacy preservation capacity, we conducted more comparison experiments as presented in Section 6 Visual comparison to privacy preservation techniques using Differential Privacy (DP) [24] and Principle Component Analysis (PCA) [33] as follows.

6. Visual comparison to privacy preservation techniques using Differential Privacy (DP) [24] and Principle Component Analysis (PCA) [33]

In this section, we compare AutoGAN-DRP ability to visually protect privacy with that using two other techniques Differential Privacy and Principle Component Analysis. First, we implement AutoGAN-DRP with VGG19 structure for the Generator and Re-constructor and the parameter setting as shown in table 1. The images are reduced to seven dimensions for different values of $\epsilon - DR$ to achieve different distances and accuracies. The datasets are grouped into two groups corresponding to binary classifiers. We also implement DP [24] method on our datasets. We generate a classifier on the authentication server by training the datasets with a VGG19 binary classifier (the structure of hidden layers is similar to our Generator in table 1). The testing images are then perturbed using differential privacy method. Specifically, Laplace noise is added to the images with sensitivity coefficient of 1 (it is computed by the maximum range value of each pixel [0,1]) and different DP epsilon parameters (this DP epsilon is different from our $\epsilon - DR$). The perturbed images are then sent to the authentication server and fed to the classifier. We visually compare the perturbed images and the accuracy of this method and AutoGAN. In addition, we implement another experiment following FRiPAL [11] in which the clients reduce image dimension using Principle Component Analysis method and send reduced features to the server. FRiPAL claims that by reducing image dimension, their method can be more resistant to reconstruction attacks. The experiments are performed with different number of reduced dimension. The images are reconstructed using *Moore-Penrose inverse* method with assumption that an adversary has access to the model. The classification accuracy is evaluated using a classifier which has similar structure to AutoGAN's classifier.

Table 2 shows image samples and results over the three datasets. Overall, AutoGAN-DRP is more resilient to reconstruction attacks compared to the other two techniques. At the accuracy of 79% on AT&T dataset, 80% on YaleB and 73% on CelebA, we cannot distinguish entities from the others. For DP method, the accuracy decreases when the DP epsilon decreases (adding more noise). Thus, the perturbed images are harder to recognize. However, at the accuracy of 57%, we are still able to distinguish identities by human eyes. For PCA, the accuracy also goes down when the number of dimensions reduces and the distances increase. At the accuracy of 75% on AT&T, 71% on YaleB and 68% on CelebA we still can differentiate individuals in the group. ”








































































	AT&T					YaleB					CelebA			
Acc		0.93	0.79	0.65		0.90	0.80	0.69		0.73	0.66	0.59		
Dist		0.0116	0.0198	0.0245		0.0184	0.0246	0.0585		0.0513	0.0531	0.06618		
AutoGAN-DRP														
														
														
														
	Org	(7)	(7)	(7)	Org	(7)	(7)	(7)	Org	(7)	(7)	(7)		
Acc		0.69	0.63	0.57		0.68	0.60	0.58		0.62	0.59	0.56		
Dist		0.0164	0.0313	0.0405		0.0149	0.0314	0.0407		0.0200	0.0418	0.0509		
Differential Privacy														
														
														
														
	Org	(11)	(8)	(7)	Org	(11)	(8)	(7)	Org	(11)	(8)	(7)		
Acc		0.90	0.75	0.60		0.87	0.83	0.71		0.71	0.68	0.57		
Dist		0.0197	0.0264	0.0348		0.0228	0.0266	0.0287		0.0362	0.0379	0.0511		
PCA														
														
														
														
	Org	(15)	(7)	(5)	Org	(15)	(7)	(5)	Org	(15)	(7)	(5)		
Acc : Average accuracy on testing data Dist: Average Euclidean distance between original images and reconstructed/perturbed images Org : Original images () Experiment parameters: epsilon for DP and number of reduced dimensions for PCA and AutoGAN-DRP														

Table 2: Sample visualization of AutoGAN, DP, PCA over three datasets

[Comment 2.8](#) Please rearrange the chart in paper, I hope it can help the paper to improve the readability.

Response:

Thank you for your suggestion. In this revision, we rearranged figures and tables by putting them closed to their related sections.

[Comment 2.9](#) In section 6.2, there is a spelling mistake, please correct it.

Response:

Thank you for your comment. In this revision, we fixed that spelling mistake

[Comment 2.10](#) In the paper, some formulas have no serial numbers, please correct it.

Response:

Thank you for your comment. In this revision, we added serial numbers on all of the formulas.

[Comment 2.11](#) The number of references is insufficient, please add the corresponding references, I hope this will help to improve the quality of the paper.

Response:

Thank you for the comment. According to your recommendation, we updated our references for the most recent years in the related work (10 references added).

III. RESPONSE TO REVIEWER 3

Overall, this paper proposed a GAN-based dimension reduction method for privacy preservation. The motivation is practical while the method is well presented. However the novelty is still limited and the authors need to explain more clearer about their contribution.

Comment 3.1 The main optimization formulation needs more explanation, especially from the perspective of theoretical analysis.

Response:

Thank you for the comment. In this revision, we revised to clarify the main optimization and updated section 3.3 (AutoGAN-based Dimension Reduction for Privacy Preserving) as follows. " ...Our proposed framework can learn a DR function $F(\cdot)$ that preserves privacy at certain value of ϵ evaluated by $\epsilon - DR$ privacy. The larger distance implies higher level of privacy. Figure-3 represents our learning system in which X' is generated from a Generator G . X' can be classified by a classifier C and can resist data reconstruction of an aggressive attack implemented by a trainable Re-constructor R . We use a binary classifier for single-level authentication system and multi-class classifier for multi-level authentication system. The Discriminator D is a neural network working with G as a minimax game. The Discriminator aims to differentiate the reconstructed data from a target distribution and send feedback to Generator. The Generator is trained so that it directs reconstructed data distribution close to the target distribution to ensure a distance between reconstructed and original data. To enlarge the distance, the selected target distribution should be different from the original data distribution. The problem becomes finding an optimal solution for the Generator. At the optimal solution, the Generator plays a role in projecting original images to lower dimension space in which only information for the classifier is retained and reducing perception information for the reconstruction. The optimal problem can be formulated as equation (4):

$$\theta^* = \theta \arg \min (\alpha \phi \arg \min \mathcal{L}_C - \beta \omega \arg \min \mathcal{L}_D - \gamma \varphi \arg \min \mathcal{L}_R + \mathcal{C}(\epsilon)) \quad (4)$$

Where α, β, γ are weights of components in the objective function and freely tuned. $\mathcal{C}(\epsilon)$ is a constraint function with respect to hyper-parameter ϵ .

The re-constructor plays its role as an aggressive adversary attempting to reconstruct original data by training R using known data. The loss function of R is the mean square error of original training data and reconstructed data, as displayed in (5):

$$\mathcal{L}_R = \sum_i^n (x_i - \hat{x}_i)^2 \quad (5)$$

The classifier C keeps the performance of the classification task in a lower dimension domain and sends feedback to G . The classifier loss function (6) is defined by a cross entropy of the class target y and predicted class \hat{y} . \mathcal{L}_D (7) is the cross-entropy loss of the Discriminator.

$$\mathcal{L}_C = - \sum_{i=1}^n \sum_{j=1}^m y_{ij} \log(\hat{y}_{ij}) \quad y, \hat{y} \in \{0, 1, \dots, m-1\} \quad (6)$$

$$\mathcal{L}_D = - \sum_i^n (t_i \log(\hat{t}_i) + (1 - t_i) \log(1 - \hat{t}_i)) \quad t, \hat{t} \in \{0, 1\} \quad (7)$$

Where m denotes the number of classes and n denotes the number of samples.

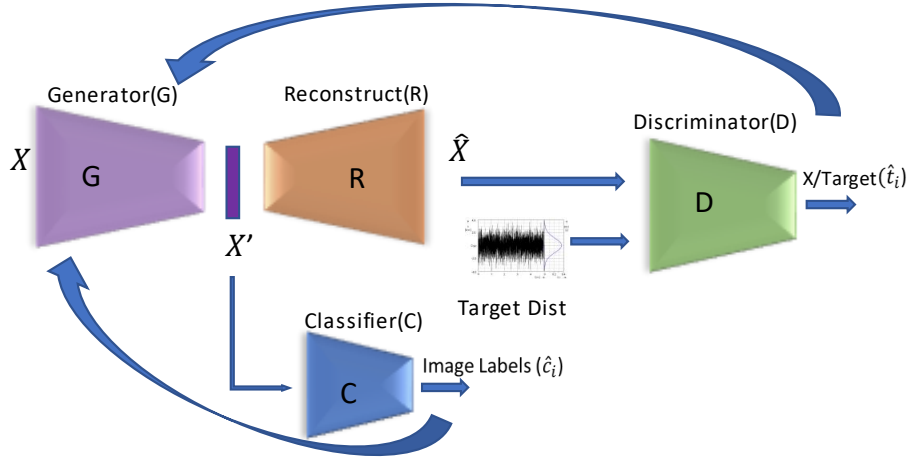


Figure 3: AutoGAN-DRP

”

[Comment 3.2](#) Not enough comparison in the experiments section, especially evaluation standard is not clear

Response:

In this revision we conducted more experiments using most current powerful structure of convolution networks (i.e., VGG16, VGG19 [32], basic CNN) for encoders (Generator) and their inverted structures for decoders (Reconstructor). By examining most recent powerful structure of reconstructor, we aim at evaluating strong adversaries who attempt to reconstruct our original data. Beside the comparison with GAP in section 5, we also conduct more experiments to compare with other privacy preservation techniques (i.e., Differential Privacy, Principle Component Analysis for privacy preservation) in section 6.

Regarding utility, the evaluation is evaluated by the classification accuracy. For privacy, we consider the distance between the reconstructed images using an aggressive re-constructor and the original images. The distance in this case is defined as L₂ norm distance. In this revision, we clarify this in section 4.3 (Privacy) as follows.

” In this study, the Euclidean distance is used to measure the distance between original and reconstructed images: $dist(x, \hat{x}) = ||x - \hat{x}||^2$. ”

Section 4 (Experiment and Discussion) is updated and Section 6 is added as follows.

”

4. Experiments and Discussion

In this section, we demonstrate our experiments over three popular supervised face image datasets: *the Extended Yale Face Database B* [29], *AT&T* [30] and *CelebFaces Attributes Dataset (CelebA)* [31].

4.1. Experiment Setup

The Extended Yale Face Database B (YaleB) contains 2,470 grayscale images of 38 human subjects under different illumination conditions and their identity label. In this dataset, the image size is 168x192 pixels.

The AT&T dataset has 400 face images of 40 subjects. For convenience, we resize each image of these two dataset to 64x64 pixels. CelebA is a color facial image dataset containing 202,599 images of 10,177 subjects. 1,709 images of the first 80 subjects are used for our experiment. Each image is resized to 64x64x3 pixels. All pixel values are scaled to the range of $[0,1]$. We randomly select 10% of each subject’s images for validation and 15% for testing dataset.

The Generator and Re-constructor in Figure 3 are implemented by three different structures. Specifically, we follow the architecture of recent powerful models VGG19, VGG16 [32] and a basic convolutional network (CNN). We modify the models to adapt with our data size (64x64). Discriminator and Classifier are built on fully connected neural network and convolutional network respectively. leaky ReLU is used as activation function for hidden layers. We use linear activation function for Generator output layers and softmax activation functions for other components’ output layers. Each component is trained in 5 local iterations (n_r, n_g, n_d, n_c), and the entire system is trained in 500 iterations for global loop (n). The target distribution is drawn from Gaussian distribution (with the covariance value of 0.5 and the mean is the average of the training data). Table 1 provides detail information of neural networks’ structures and other implementation information.

	VGG16			VGG19			CNN		
	Hidden layers	Units	Parameter	Hidden layers	Units	Parameter	Hidden layers	Units	Parameter
Generator	Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Dense Dense	64x2 128x2 256x3 512x3 512x3 1024 1024	16,295,623	Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Conv_block Max_pooling Dense Dense	64x2 128x2 256x4 512x4 512x4 1024 1024	21,605,319	Conv BatchNorm Conv BatchNorm Conv BatchNorm Dense	256 512 1024 1024	16,451,847
Reconstructor	Dense Dense Dense Reshape Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T	1024 1024 1024 512x3 512x3 256x3 128x2 64x2	10,184,000	Dense Dense Dense Reshape Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T Up_sampling Conv_block-T	1024 1024 1024 512x4 512x4 512x4 256x4 128x2 64x2	13,281,472	Dense BatchNorm Reshape Conv BatchNorm Conv BatchNorm Conv	1024 1024 1024 512 256	18,048,256
Classifier	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048	12,636,168	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048	12,636,168	Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout Dense BatchNorm Dropout	2048 2048 2048 2048	12,636,168
Discriminator	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737	Conv Dropout Conv Dropout Flatten Dense Dense	128 256 64 1024	5,084,737
Shared parameters: optimizer Adam, learning rate 0.0001, 7 dimensions Hardware: GPU Testla T4 16Gb, CPU Xeon Processors @2.3Ghz Software: Tensorflow 2.0 beta. The number of parameters are reported by model.summary() from Keras library									

Table 1: Implementation information

To evaluate the reliability of the system we test our system with different level of authentication corresponding to binary classification (single-level) and multi-class classification (multi-level). For the single-level authentication system in the scenario, we consider half of the subjects in the dataset are valid to access the database system while the rest are invalid. We randomly divide the dataset into two groups of subject and labels their images to (1) or (0) depending on their access permission. For the cases of multi-level authentication system experiments, we divide the subjects into four groups and eight groups so that the authentication server becomes four-class and eight-class classifier respectively.

4.2. Utility

We use accuracy metric to evaluate the utility of dimensionally-reduced data. The testing dataset is tested with the classifier extracted from our framework. Different structures of Generator and Re-constructor are applied including VGG19, VGG16, basic CNN on different privilege levels which correspond to multi-class classification. Figure 4 illustrates the accuracies for different dimensions from three to seven over the three facial datasets. Overall, the accuracies improve when the number of dimension increases. The accuracies on the two gray image datasets (AT&T and Yale_B) reaches 90% and higher when using VGG with only seven dimensions. The figure for Celeba is smaller but it still reaches 80%. In general, VGG19 structure performs better than using VGG16 and basic CNN in terms of utility due to the complexity (table 1) and adaptability to image datasets of VGG19. As the dimension number is reduced from 4,096 (64x64) to 7, we can achieve a compression ratio of 585 yet achieve accuracy of 90% for the two gray datasets and 80% for the color dataset. This implies our method could gain a high compression ratio and maintain a high utility in terms of accuracy. During conducting experiments we also observe that the accuracy could be higher if we keep the original resolution of images. However for convenience and reducing the complexity of our structure we resize images to the same size of 64x64 pixels.

4.3. Privacy

In this study, the Euclidean distance is used to measure the distance between original and reconstructed images: $dist(x, \hat{x}) = ||x - \hat{x}||^2$. Figure 5 illustrates the average distances between original images and reconstructed images on testing data with different ϵ constraints for seven dimensions and single-level authentication and VGG19 structure. The achieved distances (red lines) are always larger than the hyper-parameter ϵ (black dotted lines) where ϵ is less than 0.035 for AT&T, 0.052 for YaleB and 0.067 for CelebA . Due to the fact that the Re-constructor is trained using the training dataset (we consider the adversary can reach the model and the training data), our framework can only force the distance within a certain range as shown in 5. Since the mean of the target distribution is set to the mean of training dataset, reconstructed images will be close to the mean of training dataset which we believe it will enlarge the distance and expose less individual information. Therefore, the range of epsilon can be estimated base on the expectation of the distance between the testing sample and the mean of training data. The intersection between the red line and the dotted black line points out the largest distance our framework can achieve. The first section of table 2 demonstrates samples and their corresponding reconstructions in single-level authentication and seven dimensions with different achieved accuracies and distances. The reconstructed images could reach nearly identical, thus making it visually hard to recognize the identity of an individual. ”

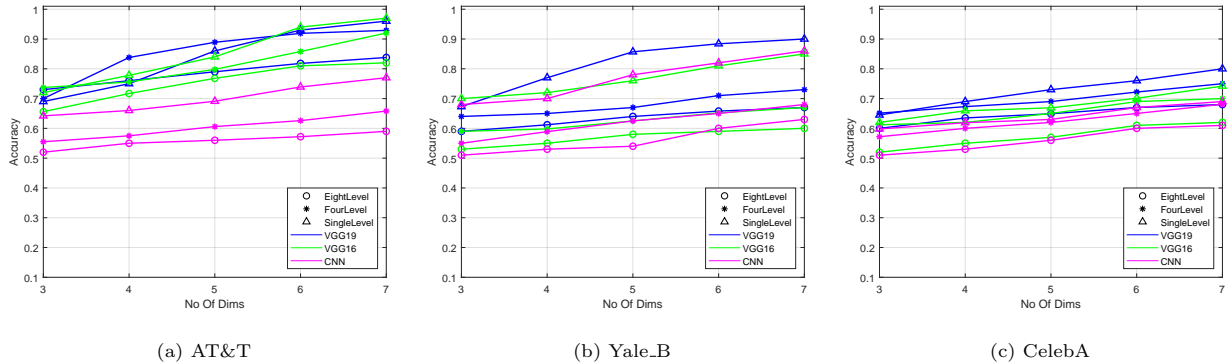


Figure 4: Accuracy for Different Number of Reduced Dimensions

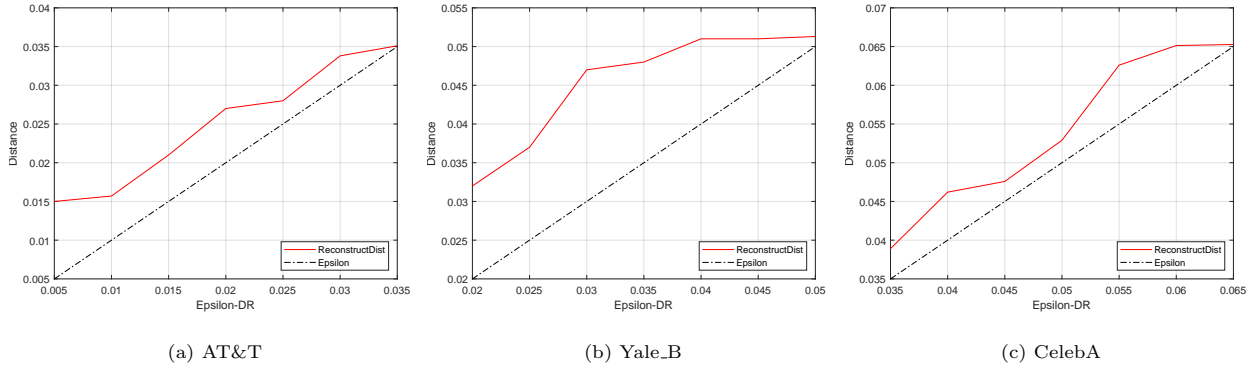


Figure 5: Average Distance Measurement Result { 7 dimensions, Single-Level}

”

6. Visual comparison to privacy preservation techniques using Differential Privacy (DP) [24] and Principle Component Analysis (PCA) [33]

In this section, we compare AutoGAN-DRP ability to visually protect privacy with other privacy preservation methods. We choose a widely used tool for privacy preserving Differential Privacy (DP) [24] and another privacy preservation method utilizing a dimensionality reduction technique Principle Component Analysis (PCA) [33].

In these experiments, we implement AutoGAN-DRP with VGG19 structure for the Generator and Reconstructor and the parameter setting as shown in table 1. The images are reduced to seven dimensions for different values of $\epsilon - DR$ to achieve different distances and accuracies. The datasets are grouped into two groups corresponding to binary classifiers. For implementing DP, we first generate a classifier on the authentication server by training the datasets with a VGG19 binary classifier (the structure of hidden layers is similar to our Generator in table 1). The testing images are then perturbed using differential privacy method. Specifically, Laplace noise is added to the images with the sensitivity coefficient of 1 (it is computed by the maximum range value of each pixel [0,1]) and different DP epsilon parameters (this DP epsilon is different from our $\epsilon - DR$). The perturbed images are then sent to the authentication server and fed to the classifier. We visually compare the perturbed images and the accuracy of this method and AutoGAN.

In addition, we follow instruction from FRiPAL [11] in which the clients reduce image dimension using Principle Component Analysis method and send reduced features to the server. FRiPAL claims that by reducing image dimension, their method can be more resistant to reconstruction attacks. The experiments are performed with different number of reduced dimension. The images are reconstructed using *Moore-Penrose inverse* method with assumption that an adversary has access to the model. The classification accuracy is evaluated using a classifier which has similar structure to AutoGAN’s classifier.

Table 2 shows image samples and results over the three datasets. Overall, AutoGAN-DRP is more resilient to reconstruction attacks compared to the other two techniques. For instant, at the accuracy of 79% on AT&T dataset, 80% on YaleB and 73% on CelebA, we cannot distinguish entities from the others. For DP method, the accuracy decreases when the DP epsilon decreases (adding more noise). Thus, the perturbed images are harder to recognize. However, at the accuracy of 57%, we are still able to distinguish identities by human eyes because DP noise does not focus on the important pixels. For PCA, the accuracy also goes

down when the number of dimensions reduces and the distances increase. Since PCA transformation is linear and deterministic, the original important information can be significantly reconstructed using the inverse transformation deriving from the model or training data. For example, at the accuracy of 75% on AT&T, 71% on YaleB and 68% on CelebA we still can differentiate individuals in the group. Thus, for security purposes, our proposed method shows the advantage in securing the data.”


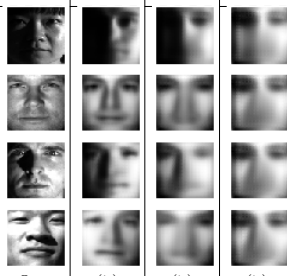
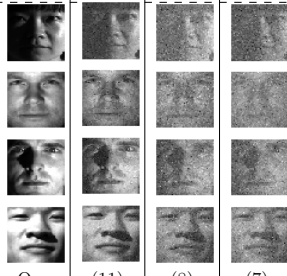


	AT&T				YaleB				CelebA			
	Acc	Dist	Dist	Dist	Acc	Dist	Dist	Dist	Acc	Dist	Dist	Dist
AutoGAN-DRP	---	0.93	0.79	0.65	---	0.90	0.80	0.69	---	0.73	0.66	0.59
	---	0.0116	0.0198	0.0245	---	0.0184	0.0246	0.0585	---	0.0513	0.0531	0.06618
												
	Org	(7)	(7)	(7)	Org	(7)	(7)	(7)	Org	(7)	(7)	(7)
	---	0.69	0.63	0.57	---	0.68	0.60	0.58	---	0.62	0.59	0.56
Differential Privacy	---	0.0164	0.0313	0.0405	---	0.0149	0.0314	0.0407	---	0.0200	0.0418	0.0509
												
	Org	(11)	(8)	(7)	Org	(11)	(8)	(7)	Org	(11)	(8)	(7)
	---	0.90	0.75	0.60	---	0.87	0.83	0.71	---	0.71	0.68	0.57
	---	0.0197	0.0264	0.0348	---	0.0228	0.0266	0.0287	---	0.0362	0.0379	0.0511
PCA	---	0.90	0.75	0.60	---	0.87	0.83	0.71	---	0.71	0.68	0.57
	---	0.0197	0.0264	0.0348	---	0.0228	0.0266	0.0287	---	0.0362	0.0379	0.0511
												
	Org	(15)	(7)	(5)	Org	(15)	(7)	(5)	Org	(15)	(7)	(5)
	---	0.90	0.75	0.60	---	0.87	0.83	0.71	---	0.71	0.68	0.57
Acc : Average accuracy on testing data Dist: Average Euclidean distance between original images and reconstructed/perturbed images Org : Original images () Experiment parameters: epsilon for DP and number of reduced dimensions for PCA and AutoGAN-DRP												

Table 2: Sample visualization of AutoGAN, DP, PCA over three datasets

[Comment 3.3](#) Overall, the paper is not well written, please propose a comprehensive revision

Response:

Thank you for the suggestion. In this revision, we thoroughly revised the manuscript and made several

changes to improve the quality of the paper. The main changes can be listed as follows.

1. Preliminaries section was merged to Related Work. Problem section and ϵ -Dimension Reduction Privacy section were merged to Methodology.
2. Section 2 (Related Work) was thoroughly revised to update our references with more latest work (around 10 most recent works were added).
3. Section 3.1.1 (Problem Statement) was revised and modified to give more information about the problem and the sample scenario.
4. All formula serial numbers were added. Figures and tables are rearranged to improve the readability.
5. Conducting more experiments with different structure of AutoGAN-DRP (i.e., VGG19, VGG16, basic CNN were applied to Generator and Re-constructor). Results were updated in Section 4.
6. Implementing AutoGAN-DRP on one more color dataset (CelebA). Experiment results and discussion in Section 4 were updated correspondingly.
7. Table 1 (Implementation information) in Section 4 was added to provide precise implementation information of components' structure in Section 4 (Experiment and Discussion)
8. Section 6 was added and new experiments were conducted to compare to other privacy preservation techniques using Differential Privacy (DP) and Principle Component Analysis (PCA).
9. Table 2 (Sample visualization of AutoGAN, DP, PCA over three datasets) in Section 6 was added to show more intuitive results of AutoGAN, DP, PCA based techniques.