FACEMAE: PRIVACY-PRESERVING FACE RECOGNITION VIA MASKED AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Face recognition, as one of the most successful applications in artificial intelligence, has been widely used in security, administration, advertising, and healthcare. However, the privacy issues of public face datasets have attracted increasing attention in recent years. Previous works simply mask most areas of faces or synthesize samples using generative models to construct privacy-preserving face datasets, which overlooks the trade-off between privacy protection and data utility. In this paper, we propose a novel framework FaceMAE, where the face privacy and recognition performance are considered simultaneously. Firstly, randomly masked face images are used to train the reconstruction module in FaceMAE. We tailor the instance relation matching (IRM) module to minimize the distribution gap between real faces and FaceMAE reconstructed ones. During the deployment phase, we use trained FaceMAE to reconstruct images from masked faces of unseen identities without extra training. The risk of privacy leakage is measured based on face retrieval between reconstructed and original datasets. Experiments prove that the identities of reconstructed images are difficult to be retrieved. We also perform sufficient privacy-preserving face recognition on several public face datasets (*i.e.*CASIA-WebFace and WebFace260M). Compared to previous state of the arts, FaceMAE consistently **reduces at least 50% error rate** on LFW, CFP-FP and AgeDB.

1 INTRODUCTION

In the past decade, face recognition has achieved remarkable and continuous progress in improving recognition accuracy Deng et al. (2019); Wen et al. (2016); An et al. (2021); Zhu et al. (2021); Wang et al. (2022); Yi et al. (2014); Sun et al. (2014) and has been widely used in daily activities such as online payment and security for identification. Advanced face recognition algorithms Deng et al. (2019); Wen et al. (2016); An et al. (2021); Wang et al. (2022; 2019; 2021); Zeng et al. (2020); Zhang et al. (2019); Wang et al. (2018); He et al. (2010); Wu et al. (2018); Zhang et al. (2016); Wen et al. (2016) and large-scale public face datasets Zhu et al. (2021); Yi et al. (2014); Liu et al. (2018) are two key factors of these progresses. Nevertheless, collecting and releasing large-scale face datasets raise increasingly more concerns on the privacy leakage of identity membership Hayes et al. (2019); Wu et al. (2019) and attribute Mirjalili et al. (2020); Bortolato et al. (2020) of training samples in recent years. Generating large-scale privacy-preserving face datasets for downstream tasks is urgent and challenging for face recognition community Wu et al. (2019); Mirjalili et al. (2020); Qiu et al. (2021).

This paper focuses on the membership privacy of a public face dataset instead of inferring specific training samples or model parameters Rigaki & Garcia (2020); De Cristofaro (2020). Specifically, in this scenario, the adversary aims to infer whether target identities are in the training set by retrieving them with query face images when the privacy-preserving face dataset is accessible. Protecting such membership privacy is crucial for practical applications. Our goal is to reduce the face retrieval accuracy, in other words, reduce the risk of membership privacy leakage, while keep the informativeness for training deep models on the privacy-preserving face dataset.

As illustrated in Fig. 1, traditionally distortion such as blurring, noising and masking is applied to face images for reducing the privacy Dufaux & Ebrahimi (2010); Korshunov & Ebrahimi (2013). These naive distortion methods reduce the privacy as well as semantics in a face image, thus pro-

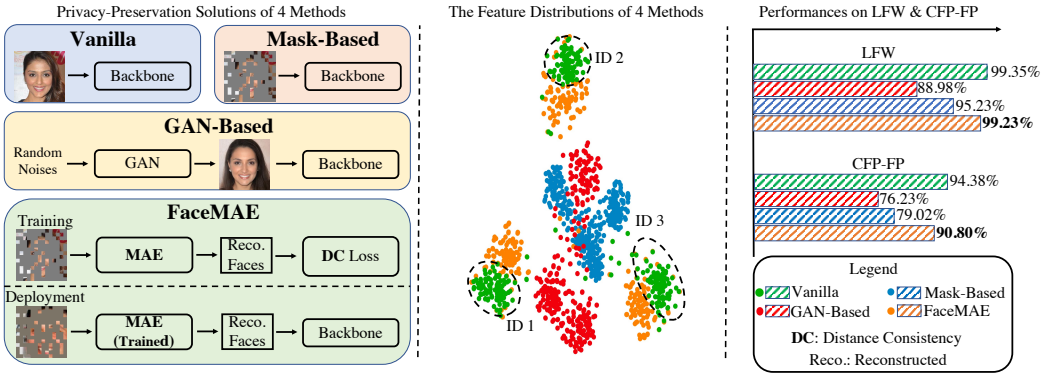

Figure 1: The left part shows four training paradigms of face recognition, namely vanilla, mask-based, GAN-based methods and our FaceMAE. The abbreviation 'Reco.' means 'Reconstructed', and 'DC' means 'Distance Consistency'. We also visualize the feature distributions of face images in central part. It is obviously that the face images learned by FaceMAE have more similar distribution with original faces. The right panel shows the performance comparisons of four methods on LFW and CFP-FP datasets. Compared to training on original and generated faces, mask-based and FaceMAE methods need less (25%) visible information of the whole faces. FaceMAE can capture the original face feature distribution more accurately thus it outperforms GAN-based and mask-based methods.

ducing unsatisfying recognition performance. The recent works try to generate privacy-preserving face dataset by synthesizing identity-ambiguous faces Qiu et al. (2021), de-identification faces Wu et al. (2019); Li & Lin (2019) or attribute-removed faces Mirjalili et al. (2020); Bortolato et al. (2020) with adapted generative adversarial networks (GANs). They can generate as many synthetic faces as required for privacy-preserving face recognition task. Although these GAN-based methods can generate real-looking faces, the utility of these generated faces for training deep models are not guaranteed. As shown in the central part of Fig. 1, there exists remarkable domain gap between generated and original faces, which leads to poor face recognition performance of GAN-based methods Qiu et al. (2021). Another drawback of GAN based methods is that the generators trained on one dataset cannot be deployed on unseen identities. In addition, all raw face images of the target dataset are required to train the generative models, which causes extra privacy leakage risks.

For membership privacy, an intuitive sense is that a raw face contains more privacy than the masked one. As shown in Fig. 1, the features distribution of masked faces are close to each other, which indicates the masked faces only contains little privacy. This naturally raises a question: "Is it feasible to reconstruct informative face images from these masked faces that are good for training deep models?" In this paper, we propose a novel framework FaceMAE, where the face privacy and recognition performance are considered simultaneously. Specifically, FaceMAE consists of two stages, named training and deployment. In the training stage, we adapt masked autoencoders (MAE) He et al. (2021) to reconstruct a new dataset from randomly masked face dataset. Distinct from vanilla MAE, the objective of FaceMAE is to generate the faces those are beneficial for face recognition training. Therefore, we utilize instance relation matching module (IRM) instead of Mean Squared Error loss (MSE-Loss) in vanilla MAE, which aims to minimize the difference between the relation graphs of original and reconstructed faces. To the best of our knowledge, we are the first to tailor a new optimization object rather than inherit the MSE-Loss for MAE. Once the training stage is finished, we apply the trained FaceMAE to generate reconstructed dataset and train face recognition backbone on reconstructed dataset. The trained FaceMAE can be easily deployed on any masked face dataset without any extra training, which shows strong generalization of our proposed method.

In experiments, we verify that our FaceMAE outperforms the state-of-the-art synthetic face dataset generation methods by a large margin in terms of recognition accuracy in multiple large-scale face datasets. As shown in Fig. 1, when training on reconstructed images from 75% masked faces of CASIA-WebFace, FaceMAE consistently **reduces at least 50% error rate** than the runner-up method (SynFace) on LFW, CFP-FP, and AgeDB datasets. It means better utility of the privacy-preserving face images generated by our method. We implement real-to-synthetic face retrieval

experiments namely membership inference, and the results show that our FaceMAE makes membership inference significantly hard.

We summarize our contributions as:

- We propose a new paradigm for privacy-preserving face recognition, named FaceMAE, that can be easily used on any face dataset to reduce the privacy leakage risk.
- Our FaceMAE jointly considers the privacy-persevering and recognition performance by adopting IRM module to minimize the distribution gap of original and reconstructed faces.
- Sufficient experiments verify that our FaceMAE outperforms the state-of-the-art methods, especially we **reduce at least 50% error rate** than previous best method on LFW, CFP-FP, and AgeDB datasets. In addition, the risk of privacy leakage decreases around **20%** in the face retrieval experiment.

2 Related Work

2.1 Face Dataset and Privacy

Large-scale public face datasets Parkhi et al. (2015); Guo et al. (2016); Nech & Kemelmacher-Shlizerman (2017); Liu et al. (2018); Zhu et al. (2021) are typically collected by searching online images based on a name list of celebrities. Instead, those tech giants may obtain much more private face data Taigman et al. (2015); Schroff et al. (2015) for training their models. Both of them cause many privacy concerns, as they collect and store the original high-resolution personal images.

Although differential privacy (DP) Dwork et al. (2006); Abadi et al. (2016) based methods have theoretical guarantees of privacy leakage and have been applied to easy datasets like MNIST LeCun et al. (1998), it is not applicable to generate large-scale high-resolution face dataset, because of low utility of the generated samples Xie et al. (2018); Cao et al. (2021). The recent works on generating privacy-preserving face dataset try to model the collect real faces with advanced GAN models and then synthesize fake face images with modifications Wu et al. (2019); Li & Lin (2019); Mirjalili et al. (2020); Bortolato et al. (2020); Qiu et al. (2021). For example, Qiu et al. (2021) propose to synthesize identity-ambiguous faces by mixing two label vectors. Different from de-identification methods that aim to remove the identity of face images Wu et al. (2019), we expect the privacy-preserving dataset can be used to train downstream identification models. Besides the utility problem, the above methods all have to access the raw face images of the target identities for training generative models.

Encoder-decoder models are learned in You et al. (2021); Yang et al. (2022a) that can convert private faces into privacy-preserving images and then invert them back to original ones. Although our FaceMAE also has encoder-decoder structure, our method is different from them in three main aspects: 1) Their protected images are hardly useful for downstream tasks, while ours can be used for training downstream models directly. 2) Their protected images can be inverted to original ones, while ours cannot be inverted. 3) They have to access raw face images, while our method only needs to access desensitized face images. An orthogonal technique to relieve face privacy concerns can be federated face recognition Meng et al. (2022); Bai et al. (2021); Aggarwal et al. (2021), while it is still challenging to achieve comparable performance to that of centralized training.

2.2 Membership Inference Attack

Membership inference Brickell & Shmatikov (2008); Li et al. (2010); Shokri et al. (2017) is a fundamental privacy problem in machine learning applications and the attack against membership is extremely destructive for medical and finical applications. In the popular membership inference attack protocol, a predictor is learned to infer the membership of specific sample in the training set given access to the black-box model Shokri et al. (2017); Sablayrolles et al. (2019). The averaged prediction accuracy or some other metrics can be derived based on the input and output of shadow models Sablayrolles et al. (2019); Carlini et al. (2021); Rezaei & Liu (2021).

However, this protocol is designed for small-scale classification datasets and it is not applicable to large-scale face datasets with millions of identities. Note that the representation model trained on

a set of face identities will include only the embedding module and the identity classifier will be dropped. Thus, it is not suitable to predict the existence of a training face image based on its output, i.e. feature, in face recognition task. As aforemotioned, we propose to measure the membership privacy leakage based on real-synthetic face image retrieval.

2.3 Training Sample Synthesis

Training deep neural networks with synthetic samples is a promising solution when collecting real training data is expensive or infeasible for privacy issues. As the most successful generative model, Generative Adversarial Networks (GANs) Goodfellow et al. (2014); Brock et al. (2018) are leveraged or adapted to synthesize training images for downstream tasks. For example, Antoniou et al. (2017) use GANs to augment training samples for few-shot target domain. Zhang et al. (2021); Yang et al. (2022b) learn to synthesize pixel-level image-annotation pairs, in order to minimize the manual annotation efforts. However, it is known that the synthetic and real data would have serious domain gap, and thus models trained on synthetic data perform badly on real testing data Ravuri & Vinyals (2019); Zhao & Bilen (2022). To narrow the domain gap between synthetic and real face images, Qiu et al. (2021) propose to mixup synthetic and real faces. Zhao & Bilen (2022) learn informative latent vectors of a pre-trained GAN model corresponding to informative synthetic training images by explicitly matching synthetic and real data distribution in many embedding spaces. Instead, we learn FaceMAE which can convert the largely masked faces to informative training samples for training high-performance recognition models.

3 Proposed Method

In this section, we first overview the pipeline of FaceMAE. Then, we briefly revisit the masked autoencoders (MAE) He et al. (2021) and formally introduce masked face reconstruction. After that, a carefully designed instance relation matching (IRM) module is presented. Finally, we give the measurement of the membership privacy leakage of face datasets based on face image retrieval.

3.1 Overview of FaceMAE

An overview of the FaceMAE is illustrated in Fig. 2. FaceMAE includes two stages, named training and deployment. In training stage, we first randomly mask the original faces with (by default) 75% ratio from a public dataset. Then, the masked faces are fed into FaceMAE to obtain the reconstructed ones. Instead of minimizing the Mean Squared Error loss (MSE-Loss) between original and reconstructed faces in pixel space, we tailor the instance relation matching (IRM) module to minimize distribution gap between real faces and FaceMAE reconstructed ones in feature space. In deployment stage, we apply the trained FaceMAE to reconstruct the masked faces from another dataset with unseen identities and assemble these reconstructed faces as a privacy-preserving dataset for face recognition tasks. Note that FaceMAE is trained on one dataset with raw images and then deployed to masked faces from unseen identities in order to construct the privacy-preserving dataset for these unseen identities.

3.2 Masked Face Reconstruction

Revisiting Masked Autoencoders. Masked Autoencoders (MAE) He et al. (2021) is a simple but efficient self-supervised pretraining strategy for image classification and its downstream tasks, such as object detection, instance segmentation, and semantic segmentation. MAE utilizes an asymmetric encoder-decoder architecture to reconstruct the unseen patches from masked image. To train MAE, the training images are randomly masked as the input, and the MSE-Loss between the original image and MAE reconstructed ones is minimized. With MAE pre-training, data-hungry models like ViT-Large/-Huge Dosovitskiy et al. (2020) can be trained well with improved generalization performance.

Thanks to MAE's image completion ability, we adapt it to reconstruct privacy-preserving faces from the masked ones. Specifically, given N raw face images $\mathcal{I} = \{I_i\}|_{i=1}^{|\mathcal{I}|}$, we resize them to 224 $\times$ 224 size. The resized faces are divided into non-overlapping 16 $\times$ 16 patches. Then, we randomly select a subset (*e.g.* 25%) of patches and mask (*i.e.*, remove) the rest. Each patch represents a *token*

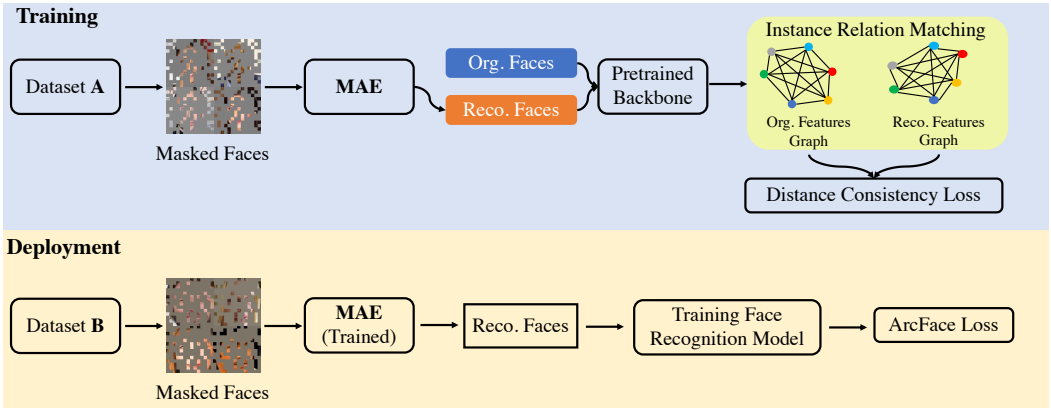

Figure 2: Illustration of the pipeline of FaceMAE. Org. and Reco. denote original and reconstructed, respectively. In training stage, FaceMAE learns to generate reconstructed faces from masked faces. Instance relation matching (IRM) aims to ensure reconstructed faces that have similar distance graph with original ones in feature space. Distance consistency loss is used to optimize FaceMAE. In deployment stage, we directly apply trained FaceMAE on unseen masked dataset to obtain reconstructed faces for face recognition training.

embedding. Only the visible patches are fed into transformer encoder E to obtain visible tokens. After that, a shallow decoder D is designed to reconstruct the input image from visible mask tokens. The whole reconstruction progress can be formulated as

$$\hat{I} = D(E[\gamma(I)]), \tag{1}$$

where $\gamma(\cdot)$ represents the random mask operation.

3.3 Instance Relation Matching

We expect the reconstructed face images are informative as the training samples for downstream tasks rather than high visual fidelity. The vanilla MAE adopts MSE-Loss to minimizes differences between reconstructed and original faces in pixel space for visual fidelity, which may ignore the distribution gap between original faces and reconstructed ones. In order to make reconstructed faces more informative as training samples, we design the instance relation matching (IRM) module to learn FaceMAE. The IRM consists of two optimization components, named instance matching (IM) and relation matching (RM) losses.

With the raw and corresponding reconstructed images, we extract the features using an extra pretrained embedding model $\phi(\cdot)$. We combine the features of the raw and reconstructed image set as two matrix $\mathbf{F} = [\phi(\mathbf{I_1}), ..., \phi(\mathbf{I_N})\}]$ and $\hat{\mathbf{F}} = [\phi(\hat{I}_1), ..., \phi(\hat{I}_N)]$ respectively, where each row is a feature vector. We construct two graphs in the feature space with the data points as nodes and similarity as edges respectively. Then, the distribution consistency loss $\mathcal{L}_{dc}$ is calculated based on the matching of both nodes and edges of the two graphs, in other words, the instance matching and relation matching respectively. We formulate it as

$$\mathcal{L}_{dc} = \underbrace{\Delta(\mathbf{F}, \hat{\mathbf{F}})}_{\text{Instance Matching}} + \beta \underbrace{\Delta(\langle \mathbf{F}\ ,\ \mathbf{F}^{\mathrm{T}} \rangle, \langle \hat{\mathbf{F}}\ ,\ \hat{\mathbf{F}}^{\mathrm{T}} \rangle)}_{\text{Relation Matching}}, \tag{2}$$

where Δ measures the averaged pairwise Euclidean distance, $\langle \cdot\ ,\ \cdot \rangle$ is the inner production, and β is a trade-off coefficient between instance matching and relation matching.

3.4 Privacy Leakage Risk

Our method is used to generate privacy-preserving face dataset. The potential attack to privacy is that the adversary may try to infer (retrieve) his/her interested identities in the face dataset by retrieving them with some other real faces. We measure the risk of this membership privacy leakage

based on face image retrieval experiment. Specifically, we assume that the reconstructed set $\mathcal{S} = \{(\boldsymbol{s}_i, y_i)\}|_{i=1}^{|\mathcal{S}|}$ is released and attacker has a set of real faces $\mathcal{A} = \{(\boldsymbol{a}_i, z_i)\}|_{i=1}^{|\mathcal{A}|}$ of his/her interested identities. He/she can leverage well-trained models to extract features and retrieve the set $\mathcal{R}$ based on $\mathcal{A}$. If the top-K retrieved faces involve the ground-truth identity, we count it as one leakage event. Then, the averaged leakage risk $R(\mathcal{A}, \mathcal{S}, K)$ can be obtained by averaging multiple attacks:

$$R(\mathcal{A}, \mathcal{S}, K) = \frac{1}{\mathcal{A}} \Sigma_{i=1}^{|\mathcal{A}|} \mathbf{1}(z_i \in T(\boldsymbol{a}_i, \mathcal{S}, K)) \tag{3}$$

where $T(\boldsymbol{a}_i, \mathcal{S}, K)$ is a retrieval function that returns the label set of the K reconstructed faces closest to the real query face $\boldsymbol{a}_i$ in embedding space.

4 Experiments

4.1 Datasets

We first introduce the datasets used in training and deployment stages of FaceMAE and the three datasets for face verification.

WebFace260M for FaceMAE Training. By default, we use WebFace260M Zhu et al. (2021) to train the FaceMAE. WebFace260M is the latest million-scale face benchmark, which is constructed for the research community. The high-quality training set of WebFace260M contains 42 millions faces of 2 millions identities. We use 10% data for FaceMAE training.

CASIA-WebFace for FaceMAE Deployment. After training FaceMAE on WebFace260M, we deploy the trained FaceMAE on CASIA-WebFace Yi et al. (2014). The dataset includes 494, 414 face faces of 10, 575 real identities collected from the Internet.

Face Verification Datasets. We mainly validate our face recognition model on 3 face recognition benchmarks, including LFW Huang et al. (2008), CFP-FP Sengupta et al. (2016), and AgeDB Moschoglou et al. (2017). LFW is collected from the Internet which contains 13, 233 faces with 5, 749 IDs. CFP-FP collects 7, 000 celebrities' faces with 500 identities. Each identity contains 10 frontal view and 4 profile views. AgeDB contains 16, 488 images of various famous people, such as actors/actresses, writers, scientists, politicians, etc. Every image is annotated with respect to the identity, age and gender attribute. There exist a total of 568 distinct subjects.

4.2 Implementation Details

Our FaceMAE is implemented with Pytorch Paszke et al. (2019). **Training Stage:** We follow the default hyper-parameters of original MAE. The default patch size is 16, the size of input image is 224 $\times$ 224, the batch size is 256 per GPU, and mask ratio is 75%. For masked face reconstruction, we use eight A100 80G GPUs to train ViT-Base for 200 epochs (40 epochs for warm up). The pretrained backbone (in Fig. 2) is ResNet-50 and its parameters is downloaded from InsightFace toolbox[1]. The default β is 1, the base learning rate is 1.5e-4, and the weight decay is 0.05. **Deployment Stage:** We apply FaceMAE to reconstruct faces from masked CASIA-WebFace and train these reconstructed faces using ResNet-50 backbone from InsightFace. To make fair comparison, we set all hyper-parameters as same as InisghtFace. More details can be found in Supplementary Material.

4.3 Comparison with State-of-the-art Methods

We compare FaceMAE with the state-of-the-art methods, including SynFace Qiu et al. (2021), FaceCrowd Kou et al. (2022) and ArcFace Deng et al. (2019). To investigate the effect of our proposed method, we also compare FaceMAE with vanilla MAE He et al. (2021). The face verification performances on on LFW, CFP-FP, and AgeDB are shown in Tab. 1. These evaluation datasets contain most of face verification cases, including the Internet faces verification, frontal and profile views faces verification, and cross-age faces verification. ArcFace Deng et al. (2019) is a popular margin-based loss for face recognition and we train a ResNet-50 using ArcFace loss as upper-bond results in first row of Tab. 1. SynFace utilizes Disco-GAN Deng et al. (2020) to generate non-existing faces

[1] https://insightface.ai/

Method (Backbone)	Type	Training Dataset	Evaluation Datasets		
			LFW	CFP-FP	AgeDB
ArcFace (R50)* Deng et al. (2019)	No Protection	CASIA-WebFace	99.35	94.38	94.73
SynFace (R50)Qiu et al. (2021)	GAN-Based	Syn-10K-50	88.98	76.23	72.17
FaceCrowd (R18)Kou et al. (2022)	Distortion	Partial Area of CelebA	83.58	-	-
ArcFace (R50)Deng et al. (2019)	Distortion	75% Masked CASIA-WebFace	95.23	79.02	74.13
MAE + ArcFace (R50)Deng et al. (2019)	MAE-Based	Reconstructed CASIA-WebFace	97.62	87.14	85.68
FaceMAE + ArcFace (R18)	MAE-Based	Reconstructed CASIA-WebFace	98.56	89.31	89.13
FaceMAE + ArcFace (R50)	MAE-Based	Reconstructed CASIA-WebFace	**99.23**	**90.80**	**90.25**

Table 1: Comparison with state-of-the-art methods using the metric of face verification accuracy (%). * denotes the upper bond method that is reproduce by official InsighFace. Besides the upper bond results, the best entry of the rest results is marked in **bold**.

using two existing face identities and train ResNet-50 on these generated faces. FaceCrowd chooses partial areas from existing faces, such as eyes and month corners, to train face recognition using ResNet-18. As shown in Tab. 1, FaceMAE outperforms previous state-of-the-art methods on all the three evaluation datasets with a large margin. Specifically, the error rate is reduced **at least 50%** compared to previously best method SynFace Qiu et al. (2021). On the one hand, this demonstrates the effectiveness of our method that can reconstruct more discriminative faces for training. On the other hand, these significant improvements also prove the proposed FaceMAE has better utility for privacy-preserving face recognition.

4.4 Ablation Studies

We perform extensive ablation studies to illustrate the effects of our method. For better evaluation, we conduct experiments where we train FaceMAE on 75% masked WebFace260M dataset and deploy it on CASIA-WebFace (unless otherwise specified).

Evaluation of MSE-Loss and distribution consistency loss $\mathcal{L}_{dc}$. We first analyze the effect of $\mathcal{L}_{dc}$ of FaceMAE. Note that, to better understand the effect of two components of $\mathcal{L}_{dc}$, we separately consider instance matching (IM) loss and relation matching (RM) loss. Experimental results are shown in Tab. 2. As one can see, the best results are achieved when jointly using IM and RM, which shows the compatibility of the two losses. We also investigate the effect of the MSE-Loss. Applying either IM or RM individually also brings non-trivial improvement than using MSE-Loss. Adding MSE-Loss with $\mathcal{L}_{dc}$ cannot further improve the face verification performance, which implies the $\mathcal{L}_{dc}$ can provide sufficient supervision for reconstruction. We visualize the reconstructed faces using MSE-Loss and $\mathcal{L}_{dc}$ respectively in Supplementary Material.

Exploring the mask ratio. Higher mask ratio means the less information of face is accessible, so the privacy could be protected better. Therefore, in order to minimize the privacy concerns and keep the verification performance simultaneously, we explore the max threshold of mask ratio that can achieve comparable results as original faces. We first randomly mask WebFace260M with four different mask ratios (30%, 50%, 75%, and 90%). Then, we train the FaceMAE models on these masked datasets, respectively. Finally, we deploy these trained models to reconstruct faces from masked CASIA-WebFace and report the verification performances on evaluation datasets. As shown in Tab. 3, the performance degrades slightly when mask ratio increase from 30% to 75%. However, the verification accuracy drops obviously when the mask ratio is set to 90%. It shows that masking too many patches of faces cannot guarantee the face recognition performance. Therefore, considering both performance and privacy, we set mask ratio as 75% by default.

Evaluating the generalization on variant datasets. By default, we train FaceMAE on WebFace260M and apply it to reconstruct faces from masked CASIA-WebFace. To investigate the generalization of training and applying FaceMAE on variant datasets, we exchange the usage of WebFace260M and CASIA-WebFace, *i.e.* WebFace260M as deploying dataset while CASIA-WebFace as training dataset. Another concern is that the identities of the two datasets may overlap. To eliminate the influence of overlapping identities in training and deploying, we use two different 10% subsets of WebFace260M, namely W.$_{sub1}$ and W.$_{sub2}$, where W. is the abbreviation of WebFace260M. We show the experimental results in Tab. 4. Several findings can be concluded in the following.

Operation			Evaluation Datasets		
MSE	IM	RM	LFW	CFP-FP	AgeDB
✓			97.62	86.14	85.68
	✓		98.93	89.69	89.78
		✓	98.92	87.10	86.68
	✓	✓	**99.23**	**90.80**	**90.25**
✓	✓	✓	99.13	90.51	90.19

Table 2: Evaluation of MSE-Loss, IM and RM in the proposed *FaceMAE*. **Bold entries** are best results.

Mask Ratio	Evaluation Datasets		
	LFW	CFP-FP	AgeDB
30%	99.41	93.03	92.87
50%	99.33	91.98	91.56
75%	99.23	90.80	90.25
90%	67.33	57.38	54.01

Table 3: Evaluation of mask ratio of input faces for *FaceMAE* training. All the mask patches are generated by random seed.

Train	Deploy	Evaluation Datasets		
		LFW	CFP-FP	AgeDB
W.	C.	99.23	90.80	90.25
C.	W.	99.37	93.87	94.14
W.$_{sub1}$	W.$_{sub2}$	99.65	94.00	94.36
W.$_{sub2}$	W.$_{sub1}$	99.47	94.12	94.51

Table 4: Evaluation of the generalization of *FaceMAE*. W. and C. represent WebFace260M and CASIA-WebFace.

Arch.	Evaluation Datasets		
	LFW	CFP-FP	AgeDB
R18	98.56	89.31	89.13
R100	99.53	91.91	90.51
MBF	98.40	88.94	88.56
ViT-S	98.45	90.12	90.18

Table 5: Exploration of cross-architecture generalization of FaceMAE. R, MBF and ViT-S denote ResNet, MobileFaceNet, and Vision Transformer.

β	Evaluation Datasets		
	LFW	CFP-FP	AgeDB
0.1	98.94	88.21	87.13
0.5	99.02	89.01	88.96
1.0	99.23	90.80	90.25
10	98.97	87.65	86.96

Table 6: Exploration of different β from 0.1 to 10. β is a trade-off parameter between instance matching loss and relation matching loss.

Mask Area	Evaluation Datasets		
	LFW	CFP-FP	AgeDB
Seed 0	99.23	90.80	90.25
Seed 1	99.10	90.67	90.21
Eye	98.45	89.98	89.76
Mouth	98.77	90.14	89.93

Table 7: Exploring which facial area is more related to the face recognition performance. The first and second rows use different random seeds for masking.

First, deploying the trained FaceMAE on larger datasets can obtain better verification performances. Second, our proposed FaceMAE has good generalization on variant datasets, no matter identity overlap exists or not.

Evaluating the generalization of cross-architecture. Another key advantage of our method is that the reconstructed faces can be used to train variant architectures, such as ResNet (R) He et al. (2016), MobileFaceNet (MBF) Chen et al. (2018) and Vision Transformer (ViT-S) Dosovitskiy et al. (2020). We train reconstructed faces using R18, R100, MBF, and ViT-S. The experimental results can be found in Tab. 5. The reconstructed faces are not sensitive to network architectures, which proves the good generalization of proposed FaceMAE. Using deeper network (R100) outperforms shallower network (R18) with 0.97%, 2.6%, and 1.38% on LFW, CFP-FP, and AgeDB. Therefore, it is feasible to utilize a light model to learn FaceMAE and apply to deeper model for boosting face recognition performance.

Exploring β. We define β as a trade-off hyper parameter between instance matching loss and relation matching loss. In order to evaluate the sensitiveness of β, we study the verification performances on evaluation datasets when setting β to be 0.1, 0.5, 1.0, and 10 in Tab. 6. One can find that the performance increases monotonically when β rises from 0.1 to 1. While continually increasing β to 10, the performance degrades significantly. It shows that over emphasizing the importance of relation matching may lead to less discriminative face reconstruction.

Evaluating several types of mask. We apply different mask strategies to evaluate the robustness of FaceMAE and show the results in Tab. 7. The comparable results of the first and second line show that changing random seed has very little influence on our proposed method. Masking eye or mouth consistently degrades the face recognition performances on three evaluation datasets, which indicates these areas are more important for the performance of face recognition. Meanwhile, masking these areas also cause difficulties for face alignment.

4.5 Analysis of Privacy Leakage Risk

As formulated in Sec. 3.4, we measure the risk of membership privacy leakage based on the face retrieval between original and reconstructed face images. The simulated retrieval experiments are explored on CASIA-WebFace. We here set K=2 and show the results in Fig. 3. One can find that the retrieval performances of 'Org. to MAE.' and 'Org. to FaceMAE' retrieval degrade largely with the number of identity increasing, while the performance of 'Org. to Org.' retrieval only

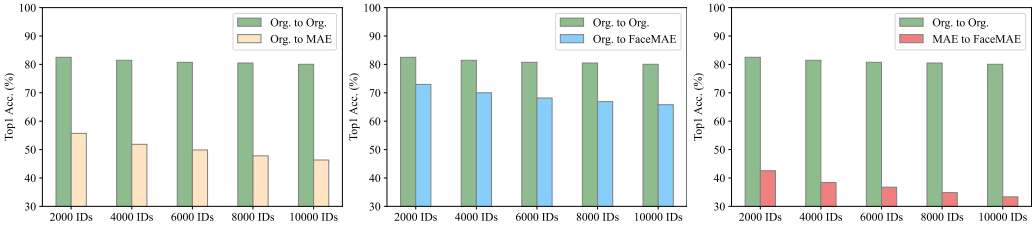

Figure 3: (a), (b), and (c) compare the retrieval performances on Original (Org.), MAE reconstructed and FaceMAE reconstructed datasets. The green bar ('Org. to Org.') represents using images from original dataset to retrieve each other. Other colors denote the retrieval experiments are completed among different settings.

degrades slightly. The comparison verifies that the difficulty of membership inference by retrieving the reconstructed faces in dataset with real faces is significantly increased by our method. Fig. 3(a) shows that 'Org. to MAE' retrieval is challenging compared to 'Org. to Org.' baseline. This result verifies that applying MAE to face dataset is the key to reduce the privacy leakage. 'Org. to FaceMAE' retrieval accuracy is also obviously lower than 'Org. to Org.' (illustrated in Fig. 3(b)), around 20% for 10,000 IDs, although it is higher than 'Org. to MAE'. The reason is that we trade-off the performance and privacy by introducing the instance relation matching module. Note that our FaceMAE achieves remarkable performance improvement in face recognition over MAE. Another possible privacy attack is that the adversary may imitate the masking process of MAE and retrieve the FaceMAE dataset with MAE processed images. Fig. 3(c) shows that this kind of attack is unpromising which is much more challenging than 'Org. to FaceMAE' retrieval.

5 Conclusion & Discussion

This paper deals with an urgent and challenging problem of privacy-preserving face recognition. FaceMAE is proposed to generate synthetic samples that reduce the privacy leakage and maintain recognition performance simultaneously. Instance relation matching module instead of MSE loss in vanilla MAE is designed to enable generated samples to train deep models effectively. The experiments verify that the proposed MAE surpasses the runner-up method by reducing at least 50% recognition error on popular face dataset. The privacy leakage risk decreases around 20% when our FaceMAE is applied.

Discussion. As a new framework for protecting the face membership privacy, our FaceMAE still has several limitations. First, our FaceMAE can reduce the membership privacy leakage of a face dataset rather than completely solve this problem. More future work is needed to advance large-scale privacy-preserving face recognition. Second, our method doesn't defend the inversion attack on model or training sample, which is out of the scope of this paper. Fortunately, those defence methods can be easily applied together with our method.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC*, pp. 308–318, 2016.

Divyansh Aggarwal, Jiayu Zhou, and Anil K Jain. Fedface: Collaborative learning of face recognition model. In *IJCB*, pp. 1–8. IEEE, 2021.

Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *ICCV Workshops*, pp. 1445–1449, 2021.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Fan Bai, Jiaxiang Wu, Pengcheng Shen, Shaoxin Li, and Shuigeng Zhou. Federated face recognition. *arXiv preprint arXiv:2105.02501*, 2021.

Blaž Bortolato, Marija Ivanovska, Peter Rot, Janez Križaj, Philipp Terhörst, Naser Damer, Peter Peer, and Vitomir Štruc. Learning privacy-enhancing face representations through feature disentanglement. In *FG*, pp. 495–502. IEEE, 2020.

Justin Brickell and Vitaly Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 70–78, 2008.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Tianshi Cao, Alex Bie, Arash Vahdat, Sanja Fidler, and Karsten Kreis. Don't generate me: Training differentially private generative models with sinkhorn divergence. In *NeurIPS*, volume 34, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pp. 428–438. Springer, 2018.

Emiliano De Cristofaro. An overview of privacy in machine learning. *arXiv preprint arXiv:2005.08679*, 2020.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *CVPR*, pp. 5154–5163, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Frédéric Dufaux and Touradj Ebrahimi. A framework for the validation of privacy protection solutions in video surveillance. In *ICME*, pp. 66–71. IEEE, 2010.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, volume 27, 2014.

Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *In ECCV*, 2016.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, volume 2019, pp. 133–152. De Gruyter, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Ran He, Wei-Shi Zheng, and Bao-Gang Hu. Maximum correntropy criterion for robust face recognition. *T-PAMI*, 33(8):1561–1576, 2010.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008.

Pavel Korshunov and Touradj Ebrahimi. Using warping for privacy protection in video surveillance. In *2013 18th International Conference on Digital Signal Processing (DSP)*, pp. 1–6. IEEE, 2013.

Ziyi Kou, Lanyu Shang, Yang Zhang, Siyu Duan, and Dong Wang. Can i only share my eyes? a web crowdsourcing based face partition approach towards privacy-aware face recognition. In *Proceedings of the ACM Web Conference 2022*, pp. 3611–3622, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *CVPR Workshops*, pp. 0–0, 2019.

Tiancheng Li, Ninghui Li, Jian Zhang, and Ian Molloy. Slicing: A new approach for privacy preserving data publishing. *TKDE*, 24(3):561–574, 2010.

Chengjun Liu and Harry Wechsler. A shape-and texture-based enhanced fisher classifier for face recognition. *IEEE TIP*, 10(4):598–608, 2001.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Qiang Meng, Feng Zhou, Hainan Ren, Tianshu Feng, Guochao Liu, and Yuanqing Lin. Improving federated learning face recognition via privacy-agnostic clusters. In *ICLR*, 2022.

Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Privacynet: semi-adversarial networks for multi-attribute face privacy. *TIP*, 29:9400–9412, 2020.

Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPR Workshops*, pp. 51–59, 2017.

Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *CVPR*, pp. 7044–7053, 2017.

Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*. British Machine Vision Association, 2015.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019.

Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *ICCV*, pp. 10880–10890, 2021.

Suman Ravuri and Oriol Vinyals. Seeing is not necessarily believing: Limitations of biggans for data augmentation. In *ICLR Workshops*, 2019.

Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *CVPR*, pp. 7892–7900, 2021.

Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646*, 2020.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *ICML*, pp. 5558–5567. PMLR, 2019.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pp. 815–823, 2015.

S. Sengupta, J. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *WACV*, pp. 1–9, 2016. doi: 10.1109/WACV.2016.7477558.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *NeurIPS*, volume 27, 2014.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *CVPR*, pp. 2746–2754, 2015.

Nikolaus F Troje and Heinrich H Bülthoff. Face recognition under varying poses: The role of texture and shape. *Vision research*, 36(12):1761–1771, 1996.

Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. In *ACMMM*, pp. 3779–3782, 2021.

Kai Wang, Shuo Wang, Zhipeng Zhou, Xiaobo Wang, Xiaojiang Peng, Baigui Sun, Hao Li, and Yang You. An efficient training approach for very large scale face recognition. In *CVPR*, 2022.

Xiaobo Wang, Shuo Wang, Shifeng Zhang, Tianyu Fu, Hailin Shi, and Tao Mei. Support vector guided softmax loss for face recognition. *TIP*, 2018.

Xiaobo Wang, Shuo Wang, Jun Wang, Hailin Shi, and Tao Mei. Co-mining: Deep face recognition with noisy labels. In *ICCV*, pp. 9358–9367, 2019.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pp. 499–515. Springer, 2016.

Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *TIFS*, 13(11):2884–2896, 2018.

Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34(1):47–60, 2019.

Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

Yang Yang, Yiyang Huang, Ming Shi, Kejiang Chen, Weiming Zhang, and Nenghai Yu. Invertible mask network for face privacy-preserving. *arXiv preprint arXiv:2204.08895*, 2022a.

Yu Yang, Xiaotian Cheng, Hakan Bilen, and Xiangyang Ji. Learning to annotate part segmentation with gradient matching. In *ICLR*, 2022b.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Zhengxin You, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Reversible privacy-preserving recognition. In *ICME*, pp. 1–6. IEEE, 2021.

Dan Zeng, Hailin Shi, Hang Du, Jun Wang, Zhen Lei, and Tao Mei. Npcface: Negative-positive collaborative training for large-scale face recognition. *arXiv preprint arXiv:2007.10172*, 2020.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *SPL*, 23(10):1499–1503, 2016.

Shifeng Zhang, Rui Zhu, Xiaobo Wang, Hailin Shi, Tianyu Fu, Shuo Wang, Tao Mei, and Stan Z Li. Improved selective refinement network for face detection. *arXiv preprint arXiv:1901.06651*, 2019.

Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *CVPR*, pp. 10145–10155, 2021.

Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. *arXiv preprint arXiv:2204.07513*, 2022.

Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *CVPR*, pp. 10492–10502, 2021.

A APPENDIX

In this appendix, we first introduce more details on the deployment of FaceMAE. Then, we show some visualization results of generated face images and more face retrieval results. After that, the generalization ability of FaceMAE is further tested on variant masked datasets.

A.1 MORE DETAILS ON THE DEPLOYMENT OF FACEMAE

By default, we train FcaeMAE on 10% data of WebFace260M and deploy it on CASIA-WebFace. Here, we use InsightFace[2], a well-known code base for face tasks, to train the reconstructed CASIA-WebFace dataset. Following the default settings in InsightFace, we use ArcFace as the loss function and ResNet50 He et al. (2016) as the backbone. The embedding dimension is 512, sampling rate is 1, momentum is 0.9 and the weight decay is 5e-4. The batch size is set to 128 per GPU. We use 4 GPUs to train ResNet50. The initialized learning rate is 0.1 and it decays at 20th, 28th, and 32th epoch, respectively. We warm up the learning rate for 1 epoch and end the training after 34th epoch.

Faces Reconstructed by **FaceMAE**

Faces Reconstructed by MAE

Figure 4: Visualization of faces reconstructed by FaceMAE and vanilla MAE. Please simply compare the areas in blue and orange boxes.

A.2 VISUALIZATION OF RECONSTRUCTED FACES

To better understand FaceMAE, we visualize the reconstructed faces that using FaceMAE and vanilla MAE. As shown in Fig. 4, FaceMAE can reconstruct more facial details than vanilla MAE. Especially, around the areas of eyes and mouth, the images generated by our method show more details that look more natural than the images generated by vanilla MAE. This explains why training models on FaceMAE generated faces can perform better than vanilla MAE generated faces. Previous works Troje & Bülthoff (1996); Liu & Wechsler (2001) have verified that such details are important for face recognition.

A.3 MORE FACE RETRIEVAL RESULTS

In the main paper, we evaluated the risk of membership privacy leakage based on face retrieval and set the retrieval hyper-parameter K = 2. Here, we present more results with K = 3 and K = 4. As shown in Fig. 5, the proposed FaceMAE can consistently reduce the retrieval performance of 'Org. to FaceMAE' when compared to 'Org. to Org.'. Although MAE has lower retrieval performance, *i.e.* better privacy preserving, the experiments in main paper have proved that our FaceMAE significantly improves the recognition performance, *i.e.* better utility.

[2]https://insightface.ai/

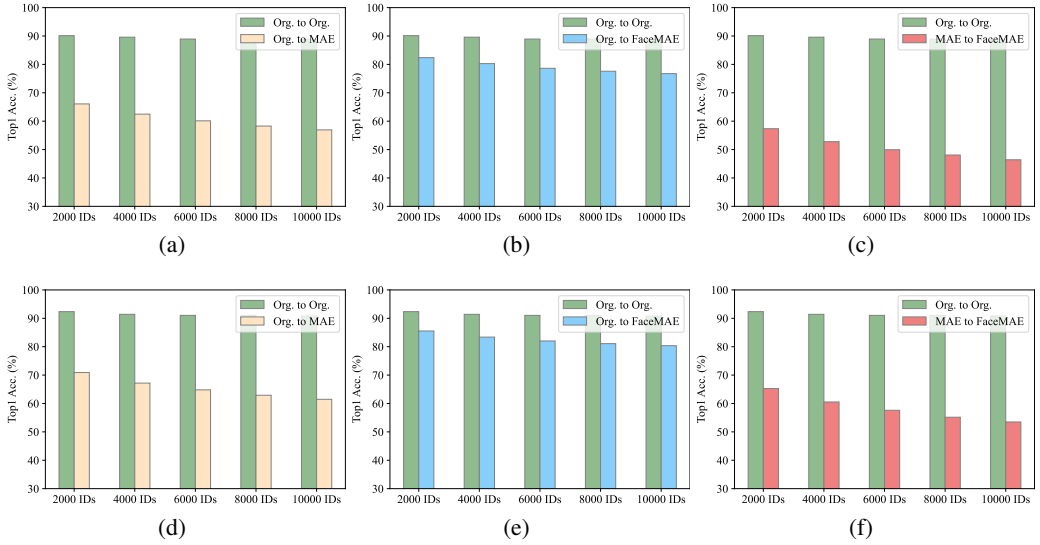

Figure 5: We report the face retrieval performance with different settings. For example, 'Org. to MAE' means that we retrieve the MAE generated face dataset with original images and report the averaged Top-K retrieval accuracy. K is set to be 3 and 4 in the upper and below rows respectively.

A.4 Generalization of FaceMAE on Variant Masked Datasets

Training Dataset	Deployment Dataset	Evaluation Datasets		
		LFW	CFP-FP	AgeDB
75% Masked WebFace260M	30% Masked CASIA-WebFace	99.28%	93.92%	92.88%
	50% Masked CASIA-WebFace	99.25%	92.78%	91.83%
	75% Masked CASIA-WebFace	99.23%	90.80%	90.25%
	85% Masked CASIA-WebFace	98.66%	83.38%	84.55%
	90% Masked CASIA-WebFace	88.66%	69.20%	66.50%

Table 8: Evaluation of FaceMAE on variant ratios masked datasets.

In order to explore the generalization ability of FaceMAE on variant masked datasets, we deploy the trained FaceMAE on 30%, 50%, 75%, 85%, and 90% masked CASIA-WebFace datasets. As shown in Fig. 8, FaceMAE can perform better on 30% and 50% masked CASIA-WebFace datasets. Although too large masking ratio, such as 85%, decreases the performance significantly, FaceMAE can still perform better than GAN-based and mask-based methods with a large margin. Another interesting finding is that training on 75% masked WebFace260M and deploying it on 90% masked CASIA-WebFace improves around 10% performance than training FaceMAE on 90% masked WebFace260M (in Tab. 3 of Experiments section). These results indicate the good generalization of proposed FaceMAE when deploying it on variant masked datasets.