Masked Face Recognition with Generative-to-Discriminative Representations

Shiming Ge [1 2] **Weijia Guo** [1 2] **Chenyu Li** [1 2 3] **Junzheng Zhang** [1 2] **Yong Li** [1 2] **Dan Zeng** [4]

Abstract

Masked face recognition is important for social good but challenged by diverse occlusions that cause insufficient or inaccurate representations. In this work, we propose a unified deep network to learn generative-to-discriminative representations for facilitating masked face recognition. To this end, we split the network into three modules and learn them on synthetic masked faces in a greedy module-wise pretraining manner. First, we leverage a generative encoder pretrained for face inpainting and finetune it to represent masked faces into category-aware descriptors. Attribute to the generative encoder's ability in recovering context information, the resulting descriptors can provide occlusion-robust representations for masked faces, mitigating the effect of diverse masks. Then, we incorporate a multi-layer convolutional network as a discriminative reformer and learn it to convert the category-aware descriptors into identity-aware vectors, where the learning is effectively supervised by distilling relation knowledge from off-the-shelf face recognition model. In this way, the discriminative reformer together with the generative encoder serves as the pretrained backbone, providing general and discriminative representations towards masked faces. Finally, we cascade one fully-connected layer following by one softmax layer into a feature classifier and finetune it to identify the reformed identity-aware vectors. Extensive experiments on synthetic and realistic datasets demonstrate the effectiveness of our approach in recognizing masked faces.

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100092, China. [2]School of Cyber Security at University of Chinese Academy of Sciences, Beijing 100049, China. [3]Cloud Music Inc., Hangzhou 311215, China. [4]Department of Communication Engineering, Shanghai University, Shanghai 200040, China. Correspondence to: Yong Li <liyong@iie.ac.cn>, Dan Zeng <dzeng@shu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

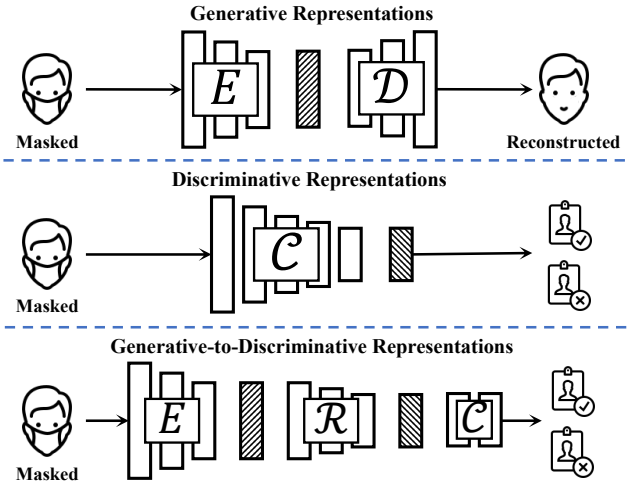


Figure 1. Our approach learns generative-to-discriminative representations for masked face recognition, which combines the advantages of generative representations and discriminative representations, providing general and robust solution to recover missing clues and capture identity-related characteristics.

1. Introduction

Deep face recognition models have delivered impressive performance on public benchmarks (Wen et al., 2016; Cao et al., 2018; Deng et al., 2019) and realistic scenarios (Anwar & Raychowdhury, 2020). In general, these models are designed for recognizing unmasked faces and often suffer from sharp accuracy drop in recognizing masked faces (Ngan et al., 2020a;b), which hinders real-world applications (Ge et al., 2017; Poux et al., 2022; Zhang et al., 2023; Wang et al., 2023; Al-Nabulsi et al., 2023). Unlike normal face recognition, masked face recognition is challenged by *insufficient or inaccurate representations*. Masks often occlude some important facial features, causing key information loss. With the occluded regions growing larger, the unmasked regions may become less sufficient for accurate identity prediction. Moreover, the ill-posed mapping between observation and possible groundtruth faces makes representations inaccurate. Therefore, an effective solution for masked face recognition should learn representations that could recover missing facial clues and calibrate inaccurate identity clues. Accordingly, many masked face recog-

nition approaches have been proposed, which are based on generative or discriminative idea (Deng et al., 2021).

"Generative" approaches aim to reconstruct the missing facial clues then perform recognition on completed faces. Deep generative models (Zheng et al., 2023; Choi et al., 2023) can provide general representations to noisy images (Chen et al., 2020; He et al., 2022a; Li et al., 2023). Generative face inpainting (Pathak et al., 2016; Li et al., 2017; Yu et al., 2018; Zhao et al., 2018; Yu et al., 2019; Wan et al., 2021; Dey & Boddeti, 2022; Yang et al., 2023) has successfully enabled recovery of high-quality visual contents. While it is very robust to generate consistent results to different masks, the benefit of face inpainting for masked face recognition is limited (Mathai et al., 2019) since they ignore the regularization on identity preservation, and the information about intra-identity and inter-identity relationships are lost during the process. Some attempts are made to enforce identity preservation via introducing face recognizer for regularization (Zhang et al., 2017; Ge et al., 2020), yet the help is limited, since the recognizers in these approaches cannot provide the upstream inpainting network with direct feedback and enough regularization on identity awareness.

In contrast, "discriminative" approaches aim to extract robust representations by reducing the effect of masked regions (Wen et al., 2016; He et al., 2022b; Song et al., 2019). They usually adopt part detection (Ding et al., 2020), compositional models (Kortylewski et al., 2021), knowledge transfer (Huber et al., 2021; Boutros et al., 2022; Zhao et al., 2022) or complementary attention (Cho et al., 2023), to localize or remove masked regions. However, the unmasked regions often hardly provide enough information for accurate recognition. Thus, some approaches propose to finetune existing face recognizers (Neto et al., 2021) or design powerful networks (Qiu et al., 2022; Boutros et al., 2022; Zhu et al., 2023; Zhao et al., 2024) to extract more information. Generally, directly finetuning general face recognizers may increase the accuracy on masked faces, while on the sacrifice of discriminative and generalization ability on the recognition of unmasked faces. Moreover, since the occlusions contain diverse mask types, these approaches usually show poor robustness on masked faces. In real-world masked face recognition scenarios, diverse masks could cause semantic divergences while the representations are expected to be consistent. Therefore, a key issue that needs to be carefully addressed in masked face recognition is the coordination between reconstructing general representations and enhancing their identity discrimination.

To facilitate masked face recognition, we propose learning generative-to-discriminative representations which combines the advantages from generative and dicriminative representations (Fig. 1). Specially, we cascade three modules and learn them in a greedy manner. First, generative encoder takes the encoder of a pretrained face inpainting model, and represents masked faces into category-aware descriptors with rich general information of masked faces to distinguish human faces from other objects. Then, discriminative reformer incorporates a 22-layer convolutional network and is learned to convert the category-aware descriptors into identity-aware vectors for enhancing recognition. Finally, feature classifier cascades a fully-connected layer and a softmax layer to identify the reformed vectors. In the approach, generative encoder and discriminative reformer are combined together, which serves as backbone for masked facial feature extraction and is progressively pretrained in a self-supervised manner, while feature classifier serves as recognition head. Finally, the backbone is frozen and feature classifier is finetuned on labeled masked faces.

Our main contributions can be summarized as: 1) we propose to learn generative-to-discriminative representations for masked face recognition, which combines the advantages of generative and discriminative representations to extract general and discriminative features for identifying masked faces; 2) we cascade generative encoder and discriminative reformer as the backbone and present a greedy module-wise pretraining strategy to improve representation learning via distillation in a self-supervised manner; and 3) we conduct extensive experiments and comparisons to demonstrate the effectiveness of our approach.

2. Approach

2.1. Problem Formulation

Our objective is learning a deep model $\phi(\mathbf{x}, \mathbf{m}; \mathbf{w})$ for discriminative representations to identify a masked face $\mathbf{x}$. Here, the binary map $\mathbf{m}$ indicates whether a pixel p is occluded ($\mathbf{m}(p) = 1$) or not ($\mathbf{m}(p) = 0$), and $\mathbf{w}$ are model parameters. Let $\hat{\mathbf{x}}$ denote the groundtruth but generally unavailable unmasked face, having $\mathbf{x}(p) = \hat{\mathbf{x}}(p)$ when $\mathbf{m}(p) = 0$. Unlike the recognition of $\hat{\mathbf{x}}$, masked face recognition needs to learn representations from partly occluded faces where some informative clues are missing. Thus the key is to address the recovery of the missing clues from $\mathbf{x}$ to approximate the latent representations of $\hat{\mathbf{x}}$, ideally having:

$$\phi(\mathbf{x}, \mathbf{m}; \mathbf{w}) \doteq \Phi(\hat{\mathbf{x}}; \hat{\mathbf{w}}), \tag{1}$$

where $\Phi(\cdot)$ is a deep face recognizer well-trained on unmasked faces with parameters $\hat{\mathbf{w}}$. The symbol $\doteq$ means "equivalence" in some metric (*e.g.*, similarity of representations or consistency of predictions). To solve Eq. (1), there are three main challenges: 1) *greater complexity* due to the consideration of the joint distribution of $\mathbf{x}$ and $\mathbf{m}$, 2) *consistency requirement* that expects to extract consistent representations even when masked faces originated from the same $\hat{\mathbf{x}}$ have diverse $\mathbf{m}$, and 3) *insufficient data* due to difficulty of collecting real-world pairs $\{\hat{\mathbf{x}}, \mathbf{x}\}$. To sum up,

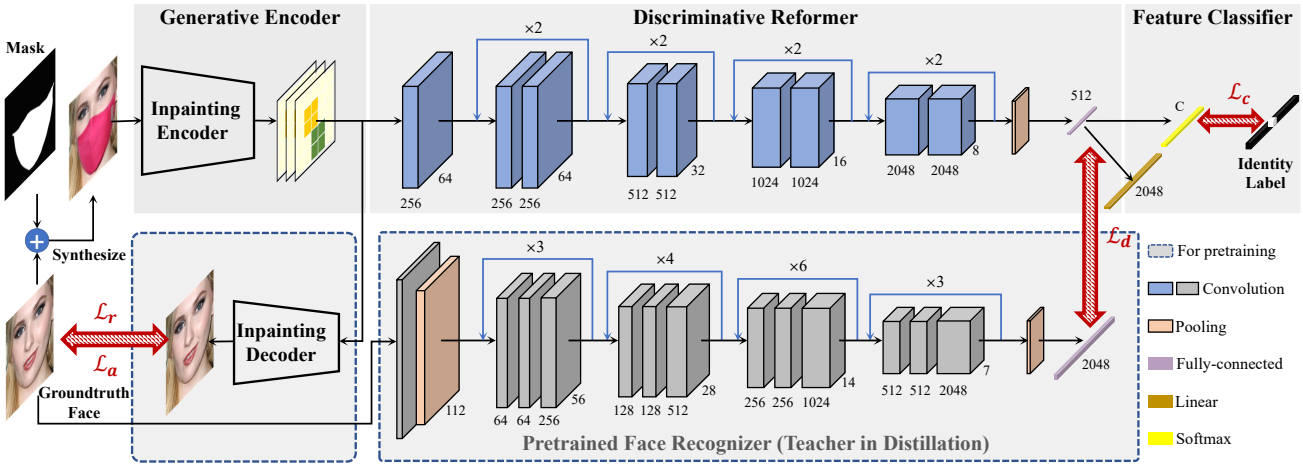


Figure 2. The framework of the proposed approach. It cascades three modules into a unified network and learns generative-to-discriminative representations on synthetic masked faces in a progressive manner. The approach first finetunes a generative encoder to represent a masked face into category-aware descriptors by initializing with a pretrained face inpainting model and finetuning via self-supervised pixel reconstruction. Then, it learns a CNN-based discriminative reformer to convert the category-aware descriptors into an identity-aware vector by distilling a general pretrained face recognizer via self-supervised relation-based feature approximation. Finally, it learns a feature classifier on identity-aware vectors by optimizing supervised classification task.

an effective solution for modeling masked faces is to learn representations through recovery, solving information reconstruction and representation clustering regularization in a unified and implicit way on synthetic data.

As shown in Fig. 2, we address masked face recognition by learning generative-to-discriminative representations. The unified network ϕ consists of generative encoder ϕ_e, discriminative reformer ϕ_r and feature classifier ϕ_c with parameters $\mathbf{w}_e$, $\mathbf{w}_r$ and $\mathbf{w}_c$, respectively. Given synthesized triplets $\{\hat{\mathbf{x}}_i, \mathbf{m}_i, \mathbf{x}_i\}_{i=1}^n$ with n samples, ϕ_e and ϕ_r are first learned by progressively reconstructing appearance and latent features, which are solved by minimizing the appearance reconstruction loss $\mathcal{L}_r$ and latent loss $\mathcal{L}_d$, separately:

$$\mathcal{L}_r(\mathbf{w}_e, \mathbf{w}_d) = \sum_{i=1}^{n} \ell(\psi(\phi(\mathbf{x}_i, \mathbf{m}_i; \mathbf{w}_e); \mathbf{w}_d), \hat{\mathbf{x}}_i), \quad (2)$$

$$\mathcal{L}_d(\mathbf{w}_r) = \sum_{i=1}^{n} \ell(\phi(\mathbf{x}_i, \mathbf{m}_i; \{\mathbf{w}_e, \mathbf{w}_r\}), \Phi(\hat{\mathbf{x}}; \hat{\mathbf{w}})), \quad (3)$$

where $\psi(\cdot)$ is inpainting decoder for training the encoder parameters $\mathbf{w}_e$ only. $\mathbf{w}_d$ are the decoder parameters. The trained $\mathbf{w}_e$ are then fixed and used in training the reformer parameters $\mathbf{w}_r$ by minimizing $\mathcal{L}_d$. $\Phi(\cdot)$ is a pretrained face recognizer used to guide the feature reconstruction of ϕ_r, with $\hat{\mathbf{w}}$ as its model parameters. $\ell(\cdot)$ denotes the distance function. Finally, $\mathbf{w}_e$ and $\mathbf{w}_r$ are frozen and all three modules are cascaded for finetuning classification loss $\mathcal{L}_c$ to learn the feature classifier in a supervised way:

$$\mathcal{L}_c(\mathbf{w}_c) = \sum_{i=1}^{n} \ell(\phi(\mathbf{x}_i, \mathbf{m}_i; \{\mathbf{w}_e, \mathbf{w}_r, \mathbf{w}_c\}), c_i), \quad (4)$$

where c_i denotes the groundtruth identity label for $\mathbf{x}_i$.

This architecture design can well address the three main challenges mentioned above in masked face recognition. First, the generative encoder and discriminative reformer are cascaded for the backbone, which decouples the burden of modeling *greater complexity* by jointly handling information reconstruction and representation clustering regularization in a progressive way. Second, the encoder aims to output a consistent reconstruction for the given masked faces originated from a same unmasked face regardless of diverse masks, that meets the *consistency requirement* of extracted representations. Third, it is easy to train the backbone on synthetic data in a self-supervised manner, alleviating the issue of *insufficient data* to avoid expensive and time-consuming annotation of training samples.

2.2. Generative Encoder

Generative encoder is responsible for extracting general face representations under mask occlusion. It is derived from ICT (Wan et al., 2021) pretrained on FFHQ (Karras et al., 2019), one of the state-of-the-art Transformer-based inpainting method. It consists of a Transformer network for face representations and a CNN for upsampling faces. We extract generative representations from the middle residual block of the upsample network. Given an input image and

a binary mask of size 256×256, the encoder computes a $64 \times 64 \times 256$ generative representation. To better adapt to the synthetic masked faces, we fix the Transformer and finetune the generative encoder on our training data with pixel reconstruction loss $\mathcal{L}_r$ as well as adversarial loss $\mathcal{L}_a$:

$$\mathcal{L}_r(\mathbf{w}_e, \mathbf{w}_r) = \sum_{i=1}^{n} \|\tilde{\mathbf{x}}_i - \hat{\mathbf{x}}_i\|^2, \tag{5}$$
$$\mathcal{L}_a = E_{\hat{\mathbf{x}}\,\mathbb{R}}[\zeta(\hat{\mathbf{x}}_i)] - E_{\hat{\mathbf{x}}\,\mathbb{G}}[\zeta(\tilde{\mathbf{x}}_i)] + E_{\check{\mathbf{x}}}(\|\nabla_{\check{\mathbf{x}}}\zeta(\check{\mathbf{x}})\|^2),$$

where adversarial loss is defined using modified WGAN-GP (Gulrajani et al., 2017), $\mathbb{R}$ is the real face distribution, $\mathbb{G}$ is the distribution implicitly defined by $\psi(\phi(*))$, $\tilde{\mathbf{x}} = \psi(\phi(\mathbf{x}_i, \mathbf{m}_i; \mathbf{w}_e); \mathbf{w}_d)$ is the inpainted face, ζ denotes the discriminators, $\check{\mathbf{x}}$ is sampled from the straight line between $\mathbb{G}$ and $\mathbb{R}$, having $\nabla_{\check{\mathbf{x}}}\zeta(\check{\mathbf{x}}) = (\tilde{x} - \check{\mathbf{x}})/\|\tilde{x} - \check{\mathbf{x}}\|$.

We visualize the learned generative representations to check the consistency over diverse masks and clustering behaviors with t-SNE (Maaten & Hinton, 2008) in Fig. 3, finding high-overlapping among the same groundtrue faces and scatters among different groundtrue faces. It implies that the generative representations can eliminate mask effect and are robust towards diverse masks, but can not well describe inter- and intra-identity characteristics.

2.3. Discriminative Reformer

Discriminative reformer aims to turn the encoded generative representations into discriminative representations, so that the identity attributes can be better recovered and described. We cascade encoder and reformer as the backbone, which has several advantages. First, it reduces the accumulation of deviations. The reformer can shift the mapping from image space to latent space, avoiding the re-mapping loss during encoding of the completed faces. Second, latent space of higher level in neural network is proved to have flatter landscape (Bengio et al., 2013), so the reformation in latent manifold is more understandable for face representations. Third, it can make better use of the information that high-level representations contains, such as long-distance dependence. Finally, feature reformation can be seamlessly integrated with the recognition head, allowing more efficient end-to-end optimization. We apply a Resnet-like network due to its effectiveness in face representation (Cao et al., 2018; Deng et al., 2019) to construct the reformer, which consists of a convolutional layer, 4 residual blocks following by a pooling and a fully-connected layers, outputs $512d$ vectors, as shown in Fig. 2. We have experimentally found that shallower structures are poor in converting generative representations into discriminative ones, while deeper or Transformer-based networks are effective but greatly increase model complexity.

Inspired by previous success in integrating external knowledge to facilitate optimization of neural networks (Hinton et al., 2014; Park et al., 2019; Li et al., 2020), we take a pretrained general face recognizer as teacher to guide the generative-to-discriminative representation reforming via knowledge distillation, and leverage essential guidance from unmasked faces for reforming and represent the teacher knowledge with two and three order structural relations:

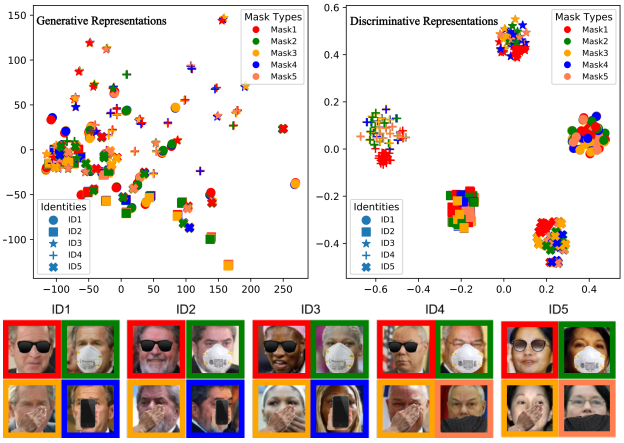


Figure 3. The t-SNE visualization of representations. We randomly sample five identities, use all sample images with these identities to synthesize masked faces with five random mask types, and extract generative and discriminative representations of masked faces. Generative representations are robust towards diverse mask occlusions but short in inter- and intra-identity discriminablility, while discriminative representations show good identity discriminablility. Bottom: some synthetic masked faces.

$$\mathcal{L}_2 = \sum_{(i,j)\in\mathbf{S}_2} \ell_H\left(\frac{1}{\mu_{\mathbf{t}}}\|\mathbf{t}_i - \mathbf{t}_j\|^2, \frac{1}{\mu_{\mathbf{s}}}\|\mathbf{s}_i - \mathbf{s}_j\|^2\right), \tag{6}$$

$$\mathcal{L}_3 = \sum_{(i,j,k)\in\mathbf{S}_3} \ell_H\Big(\Big\langle\frac{\mathbf{t}_i - \mathbf{t}_j}{\|\mathbf{t}_i - \mathbf{t}_j\|^2}, \frac{\mathbf{t}_k - \mathbf{t}_j}{\|\mathbf{t}_k - \mathbf{t}_j\|^2}\Big\rangle, \Big\langle\frac{\mathbf{s}_i - \mathbf{s}_j}{\|\mathbf{s}_i - \mathbf{s}_j\|^2}, \frac{\mathbf{s}_k - \mathbf{s}_j}{\|\mathbf{s}_k - \mathbf{s}_j\|^2}\Big\rangle\Big), \tag{7}$$

where ℓ_H denotes Huber loss, $\mathbf{t}_i = \Phi(\hat{\mathbf{x}}_i; \hat{\mathbf{w}})$ is the representation extracted by teacher recognizer, $\mathbf{s}_i = \phi(\mathbf{x}_i, \mathbf{m}_i; \{\mathbf{w}_e, \mathbf{w}_r\})$ is the discriminative representation output by the reformer. $\mu_{\mathbf{v}\in[\mathbf{t},\mathbf{s}]} = \frac{1}{|\mathbf{S}_2|}\sum_{(i,j)\in\mathbf{S}_2}\|\mathbf{v}_i - \mathbf{v}_j\|^2$ normalizes distances between teacher and student representations into the same scale, which enables relational structure transfer. $\mathbf{S}_2 = [(i,j)|1 \le i,j \le n, i \ne j]$ and $\mathbf{S}_3 = [(i,j,k)|1 \le i,j,k \le n, i \ne j \ne k]$ are pairwise set and triplet set, respectively. $\langle\rangle$ denotes cosine angle. The reformer training loss is re-formulated as:

$$\mathcal{L}_d(\mathbf{w}_r) = \mathcal{L}_1 + \alpha\mathcal{L}_2 + \beta\mathcal{L}_3, \tag{8}$$

where $\mathcal{L}_1 = \sum_{i=1}^{n}||\mathbf{t}_i - \ell_0(\mathbf{s}_i)||$ measures one order structural relation. $\ell_0(\cdot)$ is a linear mapping to convert the dimension of reformer output by adding a 2048-way linear

layer on its top, which can facilitate the pretraining. The two factors α and β are used for balancing the loss terms, and set as 0.01 and 0.02, respectively. As shown in Fig. 3, the reformed discriminative representations are effectively clustered according to identity and present clear separation between clustering of different identities, proving their identity discriminability. Thus, both encoder together with reformer plays an important role in representations for masked face recognition task, where the representations keep consistent with different masks and strengthen identity clues.

2.4. Feature Classifier

Feature classifier predicts a face identity from the reformed discriminative representation. It presents as a simple classification head, with a fully-connected layer and a softmax layer. The fully-connected layer uses 512-way to reduce the feature dimension and model parameters. We cascade feature classifier with the trained backbone and perform an end-to-end finetuning by minimizing the classification loss $\mathcal{L}_c$, which is defined as the cross-entropy loss between classifier output $p_i = \phi(\mathbf{x}_i, \mathbf{m}_i; \{\mathbf{w}_e, \mathbf{w}_r, \mathbf{w}_c\})$ and the groundtrue identity label c_i on training samples:

$$\mathcal{L}_c(\mathbf{w}_c) = -\frac{1}{n}\sum_{i=1}^{n} c_i \log(p_i). \tag{9}$$

2.5. Discussion

Relationship with other approaches Our approach can be seen as the fusion of generative approach with encoder-decoder architecture (Li et al., 2017; Kodirov et al., 2017) and discriminative approach focusing on knowledge transfer with two-stream framework (Park et al., 2019; Zhang et al., 2022), which transforms the representations from masked and groundtrue faces into a discriminative feature space. It learns general face knowledge with generative representations via inpainting like masked image modeling (He et al., 2022a; Xie et al., 2022) but focuses on more fine-grained inpainting where the input is masked face instead of complete one. Thus, generative representations can evaluate the relationship between masks and masked faces. Moreover, it converts generative representations into discriminative ones using a reformer and a pretrained face recognizer, where pairwise and triplet knowledge like (Schroff et al., 2015; Song et al., 2019; Li et al., 2020; Boutros et al., 2022) are transferred to facilitate identity recovery, rather than mean squared error in MaskInv (Huber et al., 2021) and cosine distance in CSN (Zhao et al., 2022). Specially, our approach is beyond learning two cascaded "vanilla" networks which is hard to ensure their roles, and our main novelty is the greedy module-wise pretraining that combines the advantages of generative and discriminative representations by: 1) generative encoder that is finetuned via reconstruction to ensure its role in mask-robust representations, and 2) discriminative reformer that is trained via distillation to ensure its role in identity-robust representations.

Network training Due to greater complexity of masked face recognition and different learning objectives between generative encoder and discriminative reformer, training all modules altogether is hard to converge. Thus our network training includes finetuning generative encoder, learning discriminative reformer via distillation and finetuning feature classifier in a progressive manner. The main training cost comes from the learning of discriminative reformer and is similar to the training of general face recognition models (Cao et al., 2018; Deng et al., 2019) even our entire network is larger.

3. Experiments

To verify the effectiveness of our generative-to-discriminative representation approach (**G2D**), we conduct experiments on both synthesized and realistic masked face datasets to provide comprehensive evaluations.

Datasets We use Celeb-A (Liu et al., 2015) for generating synthetic training data, LFW (Huang et al., 2007) for synthetic masked face evaluation, and RMFD (Huang et al., 2021) and MLFW (Wang et al., 2022) for real-world masked face evaluation. Celeb-A consists of 202,599 face images covering 10,177 celebrities. Each face image is cropped, aligned by similarity transformation and then scaled to 256×256. We randomly split it into training set and validation set with the ratio of $6:1$. RMFD consists of both synthetic and real-world masked faces with identity labels, covering various occlusion types and unconstrained scenarios. Our experiments only use the real-world masked face verification dataset, which contains 4,015 face images covering 426 subjects. The dataset is further organized to get 6,914 masked-unmasked pairs, including 3,457 positive and 3,457 negative pairs and serving as a valuable benchmark for cross-quality validation. MLFW is a relatively more difficult database to evaluate the performance of masked face verification. The dataset maintains the data size and the face verification protocol of LFW, considers that two faces with the same identity wear different masks and two faces with different identities wear the same mask, and emphasizes both the large intra-class variance and the tiny inter-class variance simultaneously.

For self-supervised backbone training, we synthesized massive masked faces via MaskTheFace (Anwar & Raychowdhury, 2020). For an input face, it detects the keypoints, applies affine transformation to a randomly selected mask, overlays the original image, and perform post-processing to obtain natural masked face. More details are given in Appendix A.

Baselines We consider four kinds of baselines: I) four gen-

eral face recognizers (CenterLoss (Wen et al., 2016) (CL), VGGFace (Parkhi et al., 2015) (VGG), VGGFace2 (Cao et al., 2018) (VGG2) and ArcFace (Deng et al., 2019) (AF)), II) generative approaches that equip the four general face recognizers with four face inpainting approaches (GFC (Li et al., 2017), DeepFill (Yu et al., 2018), IDGAN (Ge et al., 2020) and ICT (Wan et al., 2021)) and replace masked faces with inpainted faces as input, III) finetuning-based masked face recognizers, and IV) models trained on masked faces from scratch. Baselines in kind III and Kind IV are discriminative approaches. Baselines in kind IV adopt DoDGAN (Li et al., 2020) which first performs inpainting then learns a specialized recognizer with inpainted faces as input. To ensure fair comparisons, for each baseline, we use its published pretrained model to obtain the results and follow the same protocols for data preparation.

Evaluation We evaluate masked face verification under two settings: 1) MR-MP denoting masked reference against masked probe for evaluating over masked face pairs, and 2) UMR-MP standing for unmasked reference against masked probe, which is closer to real-world gallery-probe scenario. The evaluation is measured with 8 metrics, including verification accuracy (ACC), equal error rate (EER), Fisher discriminant ratio (FDR), false match rate (FMR), false non-match rate (FNMR), the lowest FNMR for a FMR $\leq 1.0\%$ (FMR100), the lowest FNMR for a FMR $\leq 0.1\%$ (FMR1000), and the average value calculated based on FMR100_Th and FMR1000_Th thresholds (AVG). The last 5 metrics are also used in (Huber et al., 2021).

Implementation details The experiments are implemented on Pytorch. To get facial masks, we perform simple segmentation based on Grabcut (Rother et al., 2004) automatically initialized the seeds with classical image features like colors and shapes. For generative encoder, we finetune ICT inpainting network with a batch size of 16 using Adam optimizer, where learning rate is 10^{-5} and $\beta_1 = 0.5, \beta_2 = 0.9$. For discriminative reformer, we employ pretrained VGGFace2 (Cao et al., 2018) as teacher since its input size is the same to generative encoder. All models are trained with a batch size of 64 and SGD optimizer. The initial learning rate is 0.1 and decreases to 0.5 times every 16 epochs. The momentum and weight decay are set as 0.9 and 5×10^{-4}, respectively.

3.1. Evaluation on Synthetic Masked Faces

We report the performance on synthetic masked faces. Similar to training data, we generate synthetic masked faces using images from LFW for a comprehensive evaluation, achieving 3,000 positive pairs with the same identities and 3,000 impostor pairs with different identities.

Comparison to baselines in kind I and kind II In this experiment, all recognizers and composite models extract features and then computes the cosine similarities for all the 6,000 face pairs. The accuracy is the percentage of correct predictions, where the threshold is decided as the one with the highest accuracy. The results are reported in Fig. 4. Three main conclusions can be drawn. First, diverse masks result in evident accuracy drop, which is in accord with previous research findings (Ngan et al., 2020a;b). Second, generative face inpainting sometimes are not always able to fill the gap. We notice that the combination of VGG2 and GFC achieves even lower accuracy than VGG2 alone, suggesting that the inpainting process may play a negative role if it cannot be regularized properly. We suspect it is due to the interference of similar mask patterns and poor robustness of inpainting model. Third, in the face inpainting plus recognition paradigm, adoption of the inpainting method do make a difference to the performance of the composite model. Moreover, on synthetic masked LFW, IDGAN delivers a 96.53% accuracy under 48×48 masks (Ge et al., 2020) when our G2D achieves 97.58% even under more complex masks. Finally, our G2D outperforms all combinations, proving the effectiveness of our approach.

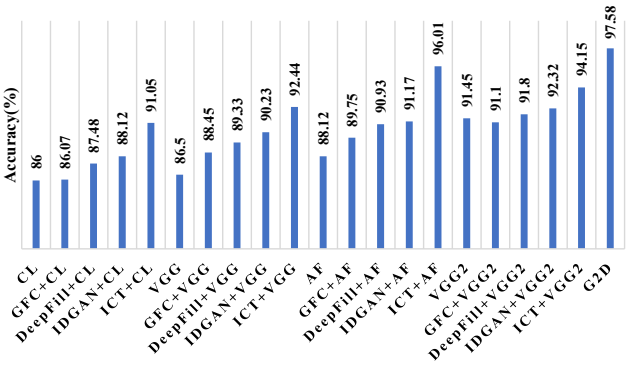


Figure 4. Evaluation on synthetic masked LFW. We report the accuracy of the proposed method (G2D), and make comparisons with combinations of general face recognizers (CenterLoss (Wen et al., 2016) or CL, VGGFace (Parkhi et al., 2015) or VGG, ArcFace (Deng et al., 2019) or AF, and VGGFace2 (Cao et al., 2018) or VGG2), and state-of-the-art generative face inpainting approaches (GFC (Li et al., 2017), DeepFill (Yu et al., 2018), IDGAN (Ge et al., 2020) and ICT (Wan et al., 2021)).

Comparison to baselines in kind III and kind IV Then, we employ the combinations of two inpainting approaches, DeepFill and ICT, with the four recognizers, together with two recently-proposed masked face recognition models, DoDGAN (Li et al., 2020) and Self-Restrained Loss (SRT) (Boutros et al., 2022), for more quantitative comparisons. Here, we do not adopt IDGAN, since it shares the same backbone with DeepFill and trained with full identity supervision. We intend to focus more on the efficacy of self-supervised representation learning. Tab. 1 presents the

Table 1. Verification Performance on LFW synthetic masked faces under MR-MP and UMR-MP settings.

Setting	Model	ACC↑	EER↓	FDR↑	FMR100↓	FMR1000↓	FMR100_Th			FMR1000_Th		
							FMR↓	FNMR↓	AVG↓	FMR↓	FNMR↓	AVG↓
MR-MP	MFN (Chen et al., 2018)	81.53%	18.67%	1.53	61.23%	80.07%	2.63%	49.53%	26.08%	0.60%	69.13%	34.87%
	ResNet50 (He et al., 2016)	85.85%	14.73%	2.03	46.17%	64.00%	1.77%	39.73%	20.75%	0.07%	65.07%	32.57%
	ResNet100 (He et al., 2016)	92.27%	8.03%	3.42	21.53%	41.70%	2.53%	15.60%	9.07%	0.80%	24.27%	12.53%
	DeepFill (Yu et al., 2018)+CL (Wen et al., 2016)	87.48%	13.43%	2.61	46.43%	63.20%	0.70%	50.13%	25.42%	0.07%	70.40%	35.23%
	DeepFill (Yu et al., 2018)+VGG (Parkhi et al., 2015)	89.33%	11.00%	2.37	36.57%	57.43%	1.03%	36.57%	18.80%	0.10%	58.40%	29.25%
	DeepFill (Yu et al., 2018)+AF (Deng et al., 2019)	90.93%	9.27%	3.44	27.73%	54.67%	1.33%	25.13%	13.23%	0.10%	54.90%	27.50%
	DeepFill (Yu et al., 2018)+VGG2 (Cao et al., 2018)	91.80%	8.37%	4.42	27.40%	52.97%	0.50%	36.87%	18.68%	**0.00%**	65.43%	32.72%
	ICT (Wan et al., 2021)+CL (Wen et al., 2016)	91.05%	8.98%	3.91	34.50%	73.13%	0.68%	37.08%	18.88%	0.14%	57.06%	28.60%
	ICT+VGG (Parkhi et al., 2015)	92.44%	7.59%	4.19	23.95%	45.86%	1.32%	21.37%	11.35%	0.17%	40.10%	20.13%
	ICT+AF (Deng et al., 2019)	96.01%	3.97%	6.66	**7.53%**	**15.03%**	**0.24%**	12.79%	**6.51%**	**0.00%**	43.49%	21.74%
	ICT+VGG2 (Cao et al., 2018)	94.15%	6.00%	5.65	20.90%	38.36%	1.32%	18.89%	10.11%	**0.00%**	49.39%	24.69%
	MFN (SRT) (Boutros et al., 2022)	78.23%	22.30%	1.23	68.40%	85.10%	4.60%	46.07%	25.33%	1.03%	67.57%	34.30%
	ResNet50 (SRT) (Boutros et al., 2022)	78.87%	21.70%	1.22	66.97%	79.17%	5.60%	44.27%	24.93%	0.90%	68.43%	34.67%
	ResNet100 (SRT) (Boutros et al., 2022)	92.80%	7.63%	3.54	20.97%	35.37%	2.03%	14.77%	8.40%	0.67%	23.23%	11.95%
	DoDGAN (Li et al., 2020)	95.44%	6.12%	5.60	22.45%	58.97%	34.93%	**0.46%**	17.70%	10.20%	**3.52%**	6.86%
	Our G2D	**97.58%**	**3.27%**	**7.01**	10.74%	33.44%	20.94%	5.83%	13.39%	6.40%	3.65%	**5.02%**
UMR-MP	MFN (Chen et al., 2018)	90.28%	9.87%	3.17	33.40%	49.23%	0.73%	37.90%	19.32%	0.07%	62.00%	31.03%
	ResNet50 (He et al., 2016)	88.83%	11.70%	2.79	27.37%	51.70%	0.40%	33.67%	17.03%	0.03%	57.90%	28.97%
	DeepFill (Yu et al., 2018)+CL (Wen et al., 2016)	90.22%	7.53%	4.69	23.87%	48.23%	0.40%	31.60%	16.00%	0.10%	52.90%	26.50%
	DeepFill (Yu et al., 2018)+VGG (Parkhi et al., 2015)	86.90%	6.63%	3.53	21.13%	43.30%	0.87%	22.47%	11.67%	0.13%	42.27%	21.20%
	DeepFill (Yu et al., 2018)+AF (Deng et al., 2019)	93.28%	10.63%	3.05	30.67%	50.90%	0.43%	39.80%	20.12%	**0.00%**	73.47%	36.73%
	DeepFill (Yu et al., 2018)+VGG2 (Cao et al., 2018)	92.65%	5.70%	5.96	18.67%	37.67%	0.30%	29.57%	14.93%	**0.00%**	62.70%	31.35%
	ICT (Wan et al., 2021)+CL (Wen et al., 2016)	91.73%	8.33%	4.50	30.15%	59.50%	0.54%	34.72%	17.63%	0.13%	54.18%	27.16%
	ICT+VGG (Parkhi et al., 2015)	92.81%	7.26%	4.55	22.66%	43.60%	0.40%	29.55%	14.97%	0.07%	53.75%	26.91%
	ICT+AF (Deng et al., 2019)	93.28%	7.36%	4.32	17.48%	26.99%	0.03%	34.55%	17.29%	**0.00%**	75.09%	37.55%
	ICT+VGG2 (Cao et al., 2018)	94.99%	5.21%	6.41	17.04%	48.13%	0.91%	18.25%	9.58%	0.07%	50.99%	25.53%
	MFN (SRT) (Boutros et al., 2022)	87.97%	12.30%	2.65	40.53%	59.47%	0.23%	55.13%	27.68%	**0.00%**	82.50%	41.25%
	ResNet50 (SRT) (Boutros et al., 2022)	82.90%	17.70%	1.73	48.23%	65.27%	**0.00%**	94.77%	47.38%	**0.00%**	99.97%	49.98%
	DoDGAN (Li et al., 2020)	94.32%	5.02%	5.46	19.41%	73.52%	4.28%	8.92%	6.55%	0.42%	51.50%	25.96%
	Our G2D	**97.75%**	**3.05%**	**8.02**	**8.93%**	**22.55%**	2.14%	**2.67%**	**2.41%**	0.17%	**13.65%**	**6.96%**

results under UMR-MP and MR-MP settings. For SRT, the performance of both baselines (ResNet50 (He et al., 2016) and MobileFaceNet (Chen et al., 2018)) and those along with an extra module trained with SRT loss are reported.

As shown in Tab. 1, for SRT which finetunes existing deep recognizers with an extra module on top, the original baselines, instead of the refined ones, show better performance. This suggest that, although these solutions can recover some performance on masked samples, the generalization ability of deep models can be easily suffered. Similarly, the recent work DoDGAN (Li et al., 2020) experienced an evident drop on cross-quality evaluation. In essence, these approaches do not appropriately handle the distribution divergence between masked and non-masked samples in the latent space. Our G2D achieves the highest accuracy on both MR-MP and UMR-MP settings. Tab. 1 also reports the fisher discriminant ratio (FDR), which measures the distinguish ability of positive and negative pairs. Our approach shows better capacity to deal with the identity obscuring of masked faces.

Analysis on FMR and FNMR results It is worth to note that, the approaches based on off-the-shelf face recognizers show lower false match rate (FMR). It suggests that they tend to predict more positive pairs (which share the same identity) as negative, while prediction over negative pairs is less affected. To the contrary, our G2D shows evident superiority in false non-match rate (FNMR). This reveals a basic difference in our motivation. When occlusions occurs, for general face recognizers, the main challenge is the invalidation of pre-existing intra-class characteristics. Our approach, differently, teaches the model to doubt, and re-calibrate. It is also worth to note that, our model presents lower average values of FMR and FNMR, especially under UMR-MP setting. It suggests our proposed G2D achieves a better balance between FMR and FNMR, in another word, a better generalization over unmasked and masked faces.

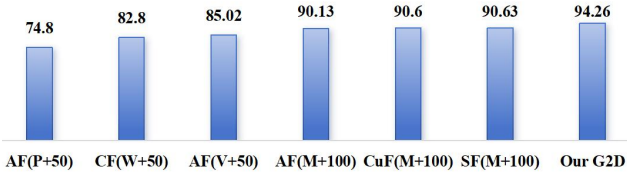


Figure 5. Verificaiton accuracy (%) on MLFW (Wang et al., 2022). AF: ArcFace (Deng et al., 2019), CF: CosFace (Wang et al., 2018), CuF: CurricularFace (Huang et al., 2020), SF: SFace (Zhong et al., 2021). P: Private-Asia, W: WebFace, V: VGGFace2, M: MS1MV2. 50 means ResNet50 and 100 means ResNet100.

3.2. Evaluation on Realistic Masked Faces

We then evaluate on RMFD (Huang et al., 2021), where realistic masked faces have various mask types and complicated photographic conditions. We use 6,992 sample pairs to examine model performance and present comparison results in. Tab. 2. We can find that general models trained on

Table 2. Performance on RMFD realistic masked faces under UMR-MP setting.

Model	ACC↑	EER↓	FDR↑	FMR100↓	FMR1000↓	FMR100_Th			FMR1000_Th		
						FMR↓	FNMR↓	AVG↓	FMR↓	FNMR↓	AVG↓
MFN (Chen et al., 2018)	69.90%	30.16%	0.49	88.69%	95.70%	1.01%	88.69%	44.85%	0.09%	95.70%	47.89%
ResNet50 (He et al., 2016)	71.75%	28.44%	0.65	81.65%	94.32%	1.01%	81.65%	41.33%	0.09%	94.32%	47.20%
MFN (SRT) (Boutros et al., 2022)	69.25%	31.28%	0.45	88.83%	97.09%	0.09%	97.14%	48.62%	0.03%	99.22%	49.63%
ResNet50(SRT) (Boutros et al., 2022)	65.92%	34.76%	0.35	87.80%	96.88%	**0.00%**	100.00%	50.00%	**0.00%**	100.00%	50.00%
ArcFace (Deng et al., 2019)	72.35%	27.71%	0.68	81.59%	93.48%	0.99%	81.59%	41.29%	0.09%	93.48%	46.78%
VGGFace2 (Cao et al., 2018)	72.22%	27.91%	0.60	88.08%	98.67%	0.99%	88.08%	44.53%	0.09%	98.67%	49.38%
DoDGAN (Li et al., 2020)	72.55%	28.26%	0.54	83.12%	95.24%	1.01%	83.12%	42.07%	0.09%	95.24%	47.66%
Our G2D	**79.18%**	**21.64%**	**1.31**	**72.18%**	**86.89%**	0.99%	**72.18%**	**36.58%**	0.09%	**86.89%**	**43.49%**

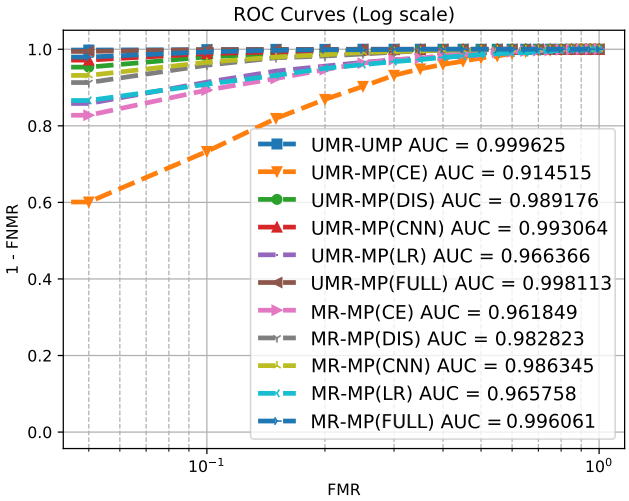


Figure 6. The achieved log-scale ROC curves by G2D models trained with different losses. In each plot, the curves of UMR-MP cases are marked with a dashed line. The curves of MR-MP cases are marked with a dotted line. For each ROC curve, the area under the curve (AUC) is listed inside the plot.

normal faces all exhibit more violent drop on accuracy. For example, VGGFace2 achieves a 91.45% accuracy on synthesized masked faces, while only gets a 72.22% accuracy on realistic masked faces. The results prove the difficulty of the dataset. Our G2D achieves the highest accuracy of 79.18%, which proves it capable of adapting to masked face recognition in the wild. The models with SRT show extreme imbalance between FMR and FNMR, which suggests they get overfitting to the masked face recognition scenarios while almost completely sacrificing the discriminant over unmasked faces. Instead, our G2D show better capacity to connect unmasked and masked faces, which is rather valuable in realistic applications. The evaluation on MLFW (Fig. 5) also shows that our G2D delivers the best accuracy.

3.3. Ablation Studies

Generative encoder To evaluate our generative encoder design, we simulate the case when available information is reduced so that the model training cannot use the higher-level features, and compare with two variants: 1) G2D(CNN) that uses a CNN-based inpainting network DeepFill as encoder and extracts generative representations from the layer before the decoding part of its second fine-grained network, and 2) G2D(LR) that removes convolutional layers from ICT upsampling network and takes appearance prior output as generative representations. From Tab. 3, it is obvious that G2D outperforms G2D(CNN) due to better encoder, while the reduced input information leads G2D(LR) to overfitting and the learned representations poor generalization to different data domains. Fig. 6 provides their log-scale ROC curves, which also shows the similar conclusion, implying that Transformer-based generative encoder is more suitable for masked face recognition. We suppose that the representation space constructed by pretrained Transformer allows it to simulate and explore the distribution correlation among masked face, the corresponding mask and original face.

Table 3. Ablation study of G2D variants with different encoders and reformers under UMR-MP and MR-MP settings.

Synthetic masked LFW

Setting	Model	ACC↑	EER↓	FDR↑	FMR100↓	FMR1000↓
UMR-MP	G2D(CNN)	96.42%	3.60%	7.20	14.20%	46.10%
	G2D(LR)	93.99%	6.54%	4.87	17.98%	38.77%
	G2D[CE]	83.50%	16.60%	1.94	69.83%	90.67%
	G2D[DIS]	95.25%	4.80%	6.04	18.30%	70.90%
	G2D	**97.75%**	**3.05%**	**8.02**	**8.93%**	**22.55%**
MR-MP	G2D(CNN)	96.14%	5.77%	5.93	18.40%	55.63%
	G2D(LR)	91.48%	9.96%	3.82	23.83%	39.12%
	G2D[CE]	82.72%	10.40%	3.36	36.03%	69.37%
	G2D[DIS]	93.53%	6.53%	5.49	24.37%	60.07%
	Full	**97.58%**	**3.27%**	**7.01**	**11.74%**	**38.44%**

Realistic masked RMFD

Setting	Model	ACC↑	EER↓	FDR↑	FMR100↓	FMR1000↓
UMR-MP	G2D(CNN)	73.26%	27.77%	0.61	83.69%	93.27%
	G2D(LR)	73.45%	27.02%	0.70	86.11%	95.29%
	G2D[CE]	64.87%	35.37%	0.28	94.20%	99.02%
	G2D[DIS]	70.80%	30.59%	0.48	87.21%	97.14%
	G2D	**79.18%**	**21.64%**	**1.31**	**72.18%**	**86.89%**

Discriminative reformer First, we argue that discriminative reformer is very necessary due to poor identity discriminability of generative representations, *e.g.*, the model achieves only a low accuracy of 57.10% on synthetic masked LFW

if discriminative reformer is discarded from the whole network. Then, we further check the learning process of discriminative reformer by comparing two models trained with different losses: 1) G2D[CE] trained with $\mathcal{L}_c$ only, and 2) G2D[DIS] trained with $\mathcal{L}_1$ only. The models trained with $\mathcal{L}_2$ and $\mathcal{L}_3$ can hardly converge, therefore the results are not presented. We report the results in Tab. 3 and Fig. 6, which suggest that directly enforcing the model to approximate the hard identity label is less efficient. Thus, it is necessary to perform student learning supervised by a pretrained teacher whose features contain rich identity relationship (Li et al., 2020). A better teacher may lead to improved performance, *e.g.*, we replace VGGFace2 with ArcFace as teacher where the inputs are resized into 112×112, achieving a higher verification accuracy of 98.02% on synthetic masked LFW than 97.58% with VGGFace2 as teacher (Tab. 1).

3.4. Further Analysis

Table 4. Test accuracy (%) on CPLFW (Zheng & Deng, 2018)

CL	SphereFace	VGG2	AF	MaskInv	G2D
77.48	81.40	84.00	92.08	**92.86**	92.23

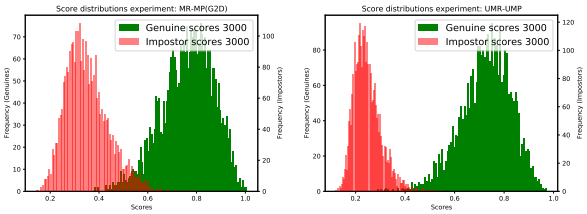


Figure 7. Similarity score distributions of our G2D under MR-MP setting (left) and the ideal case under UMR-UMP setting (right). The positive and negative pairs are marked in green and red, respectively. Our G2D delivers small overlapping, which is close to the ideal case. More results are shown in Appendix B.

Evaluation on normal face recognition To evaluate the effect of our G2D on normal face recognition, we conduct an experiment on the normal face benchmark CPLFW (Zheng & Deng, 2018) and report the results in Tab. 4. We can find that our model achieves competitive performance, *e.g.*, just lower than MaskInv (Huber et al., 2021). We suppose the main reasons include: 1) generative encoder that provides general and robust representations towards normal and masked faces, and 2) discriminative reformer that remains performance on normal faces by distilling on pretrained high-accuracy face recognizer.

Representation discriminability Fig. 3 has showed that the reformed discriminative representations cluster the masked faces with the same identity together and present strong discriminability between different identities. We further conduct evaluation by using similarity score distributions on synthetic masked LFW and report the results in Fig. 7. We can find that our G2D delivers a small overlapping between positive and negative samples, which is close to the ideal case, demonstrating strong discriminability of our generative-to-dicriminative representations.

Inference efficiency Due to greater complexity of masked face recognition, our model has 178.5M parameters, larger than normal face recognition models (*e.g.*, VGGFace2 and Arcface) who use Resnet50 as backbone and have 25.6M parameters. However, it is still efficient. We conduct efficiency analysis on a NVIDIA GeForce RTX 3090 GPU by performing inference on 100 masked faces with size of 256 × 256. The average inference time cost of a face image is 0.0428 seconds, leading to an inference speed of 23.35 FPS, implying the deployment feasibility in practical scenarios like urban governance.

4. Conclusion

Masked face recognition has been gathering intensive attention over the past few years due to its real-world applications (*e.g.*, fighting the COVID-19 pandemic). In this work, we propose to address masked face recognition by learning generative-to-discriminative representations. Our approach splits a unified network into three modules and learn them in a greedy module-wise pretraining way. Generative encoder and discriminative reformer are cascaded as the backbone to provide occlusion-robust and discriminative representations towards masked faces. The experiments are conducted on synthetic and realistic datasets to verify the effectiveness of our approach. In the future, we will design simultaneous training with synthetic and realistic datasets, and extend the framework to more vision tasks like occluded person ReID.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning for Social Good, particularly addressing the challenge of masked face recognition. The proposed method would contribute positively to society by identifying masked faces and facilitating the development of Safety AI, *e.g.*, improving urban governance and fighting the COVID-19 pandemic. There are many other potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements

This work was supported by grants from the Pioneer R&D Program of Zhejiang Province (2024C01024), and Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China (SKLMCC2022KF004).

References

Al-Nabulsi, J. et al. Iot solutions and ai-based frameworks for masked-face and face recognition to fight the COVID-19 pandemic. *Sensors*, 23(16):7193, 2023.

Anwar, A. and Raychowdhury, A. Masked face recognition for secure authentication. *arXiv preprint*, 2020.

Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. Better mixing via deep representations. In *ICML*, pp. 552–560, 2013.

Boutros, F., Damer, N., Kirchbuchner, F., and Kuijper, A. Self-restrained triplet loss for accurate masked face recognition. *PR*, 124:108473, 2022.

Cao, Q., Shen, L., Xie, W., et al. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pp. 67–74, 2018.

Chen, M., Radford, A., Child, R., et al. Generative pretraining from pixels. In *ICML*, pp. 1691–1703, 2020.

Chen, S., Liu, Y., Gao, X., and Han, Z. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR*, pp. 428–438, 2018.

Cho, Y., Cho, H., Hong, H. G., et al. Localization using multi-focal spatial attention for masked face recognition. In *FG*, pp. 1–6, 2023.

Choi, J., Park, Y., and Kang, M. Restoration based generative models. In *ICML*, pp. 5787–5816, 2023.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pp. 4690–4699, 2019.

Deng, J., Guo, J., An, X., et al. Masked face recognition challenge: The insightface track report. In *CVPRW*, pp. 1437–1444, 2021.

Dey, R. and Boddeti, V. N. Generating diverse 3d reconstructions from a single occluded face image. In *CVPR*, pp. 1537–1547, 2022.

Ding, F., Peng, P., Huang, Y., et al. Masked face recognition with latent part detection. In *MM*, pp. 2281–2289, 2020.

Ge, S., Li, J., Ye, Q., and Luo, Z. Detecting masked faces in the wild with LLE-CNNs. In *CVPR*, pp. 2682–2690, 2017.

Ge, S., Li, C., Zhao, S., and Zeng, D. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE TCSVT*, 30(10):3387–3397, 2020.

Gulrajani, I., Ahmed, F., Arjovsky, M., et al. Improved training of wasserstein gans. In *NeurIPS*, pp. 5769–5779, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

He, K., Chen, X., Xie, S., et al. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022a.

He, M., Zhang, J., Shan, S., et al. Locality-aware channel-wise dropout for occluded face recognition. *IEEE TIP*, 31:788–798, 2022b.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2014.

Huang, B., Wang, Z., Wang, G., et al. Masked face recognition dataset and validation. In *ICCVW*, 2021.

Huang, G., Ramesh, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, 2007.

Huang, Y., Wang, Y., Tai, Y., et al. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *CVPR*, pp. 5900–5909, 2020.

Huber, M., Boutros, F., Kirchbuchner, F., and Damer, N. Mask-invariant face recognition through template-level knowledge distillation. In *FG*, pp. 1–8, 2021.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *CVPR*, pp. 4396–4405, 2019.

Kodirov, E., Xiang, T., and Gong, S. Semantic autoencoder for zero-shot learning. In *CVPR*, pp. 4447–4456, 2017.

Kortylewski, A., Liu, Q., Wang, A., et al. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *IJCV*, 129 (3):736–760, 2021.

Li, C., Ge, S., Zhang, D., and Li, J. Look through masks: Towards masked face recognition with de-occlusion distillation. In *MM*, pp. 3016–3024, 2020.

Li, T., Chang, H., Mishra, S. K., et al. MAGE: masked generative encoder to unify representation learning and image synthesis. In *CVPR*, pp. 2142–2152, 2023.

Li, Y., Liu, S., Yang, J., and Yang, M.-H. Generative face completion. In *CVPR*, pp. 3911–3919, 2017.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *ICCV*, pp. 3730–3738, 2015.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-SNE. *JMLR*, 9:2579–2605, 2008.

Mathai, J., Masi, I., and AbdAlmageed, W. Does generative face completion help face recognition? In *ICB*, pp. 1–8, 2019.

Neto, P., Boutros, F., Pinto, J. R., et al. Focusface: Multi-task contrastive learning for masked face recognition. In *FG*, pp. 1–8, 2021.

Ngan, M., Grother, P., and Hanaoka, K. Ongoing face recognition vendor test (frvt) part 6a: Face recognition accuracy with masks using pre- covid-19 algorithms, 2020a.

Ngan, M., Grother, P., and Hanaoka, K. Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms, 2020b.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *CVPR*, pp. 3967–3976, 2019.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. Deep face recognition. In *BMVC*, pp. 41.1–41.12, 2015.

Pathak, D., Krahenbuhl, P., Donahue, J., et al. Context encoders: Feature learning by inpainting. In *CVPR*, pp. 2536–2544, 2016.

Poux, D., Allaert, B., Ihaddadene, N., et al. Dynamic facial expression recognition under partial occlusion with optical flow reconstruction. *IEEE TIP*, 31:446–457, 2022.

Qiu, H., Gong, D., Li, Z., et al. End2end occluded face recognition by masking corrupted features. *IEEE TPAMI*, 44(10):6939–6952, 2022.

Rother, C., Kolmogorov, V., and Blake, A. 'GrabCut': interactive foreground extraction using iterated graph cuts. *ACM TOG*, 23:309–314, 2004.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pp. 815–823, 2015.

Song, L., Gong, D., Li, Z., et al. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *ICCV*, pp. 773–782, 2019.

Wan, Z., Zhang, J., Chen, D., and Liao, J. High-fidelity pluralistic image completion with transformers. In *ICCV*, pp. 4692–4701, 2021.

Wang, C., Fang, H., Zhong, Y., and Deng, W. MLFW: A database for face recognition on masked faces. In *CCBR*, pp. 180–188, 2022.

Wang, H., Wang, Y., Zhou, Z., et al. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pp. 5265–5274, 2018.

Wang, Z., Huang, B., Wang, G., et al. Masked face recognition dataset and application. *IEEE TBBIS*, 5(2):298–304, 2023.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. A discriminative feature learning approach for deep face recognition. In *ECCV*, pp. 499–515, 2016.

Xie, Z., Zhang, Z., Cao, Y., et al. Simmim: A simple framework for masked image modeling. In *CVPR*, pp. 9643–9653, 2022.

Yang, X., Han, M., Luo, Y., et al. Two-stream prototype learning network for few-shot face recognition under occlusions. *IEEE TMM*, 25:1555–1563, 2023.

Yu, J., Lin, Z., Yang, J., et al. Generative image inpainting with contextual attention. In *CVPR*, pp. 5505–5514, 2018.

Yu, J., Lin, Z., Yang, J., et al. Free-form image inpainting with gated convolution. In *ICCV*, pp. 4471–4480, 2019.

Zhang, K., Zhang, C., Li, S., et al. Student network learning via evolutionary knowledge distillation. *IEEE TCSVT*, 32 (4):2251–2263, 2022.

Zhang, P., Huang, F., Wu, D., et al. Device-edge-cloud collaborative acceleration method towards occluded face recognition in high-traffic areas. *IEEE TMM*, 25:1513–1520, 2023.

Zhang, S., He, R., Sun, Z., and Tan, T. Demeshnet: Blind face inpainting for deep meshface verification. *IEEE TIFS*, 13(3):637–647, 2017.

Zhao, F., Feng, J., Zhao, J., et al. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE TIP*, 27(2):778–790, 2018.

Zhao, W., Zhu, X., Shi, H., et al. Consistent sub-decision network for low-quality masked face recognition. *IEEE SPL*, 29:1147–1151, 2022.

Zhao, W., Zhu, X., Guo, K., et al. Masked face transformer. *IEEE TIFS*, 19:265–279, 2024.

Zheng, C., Wu, G., Bao, F., et al. Revisiting discriminative vs. generative classifiers: Theory and implications. In *ICML*, pp. 42420–42477, 2023.

Zheng, T. and Deng, W. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical report, BUPT, 2018.

Zhong, Y., Deng, W., Hu, J., et al. Sface: Sigmoid-constrained hypersphere loss for robust face recognition. *IEEE TIP*, 30:2587–2598, 2021.

Zhu, Y., Ren, M., Jing, H., et al. Joint holistic and masked face recognition. *IEEE TIFS*, 18:3388–3400, 2023.

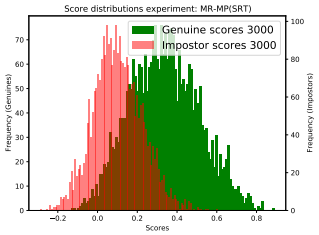
A. The Synthesis of Masked Faces

Our approach uses synthetic masked faces to train the models in a self-supervised manner. To this end, the masked faces are generated by synthesizing from normal faces. We take 202,599 normal facial images from Celeb-A dataset and synthesize massive masked facial images via pasting diverse mask patterns onto the images. To achieve that, we collected 45 transparent mask images (some examples are shown in Fig. 8) online and resized them to cover an average of about 1/5 of the face. For a normal facial image, a random mask pattern is selected and simple alignment based on the facial landmarks is conducted to better simulate the realistic masked faces. Fig. 8 also presents some examples of the synthesized masked faces. To improve model generalizability, we further perform data augmentation by flipping and translation.

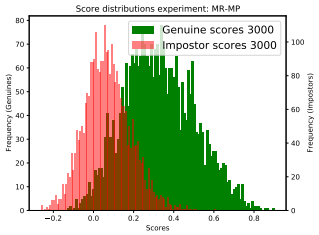
Figure 8. Some examples of mask images (top) used for generating masked faces (bottom).

B. Representation Discriminability

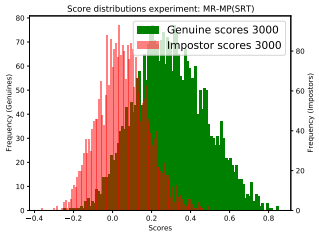
We can use similarity score distributions to evaluate the representation discriminability. Fig. 9 reports the results achieved by different models on synthetic masked LFW under MR-MP setting. The scores of the genuine pairs are in green color, while the scores of the impostor pairs (negative pairs) are presented in red color. Smaller overlapping areas suggest a more distinct separation between pairs with same and different identities. It illuminates that our G2D delivers smaller overlapping region than other models and is close to the ideal case (Fig. 9 (n)), indicating that G2D can extract robust representations and provide stronger discriminative ability for masked faces.



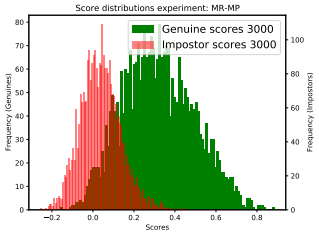
(a) MFN (SRT)



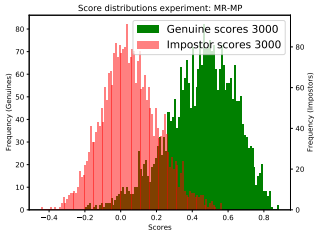
(b) MFN



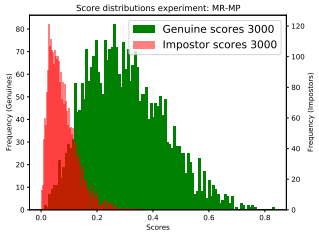
(c) ResNet50 (SRT)



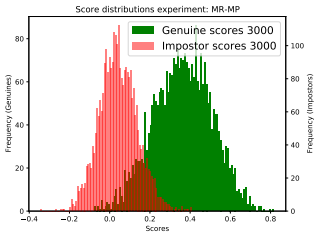
(d) ResNet50



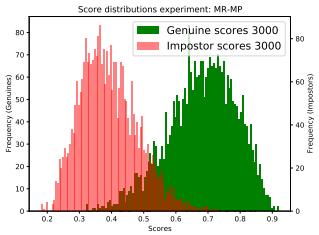
(e) DeepFill+CL



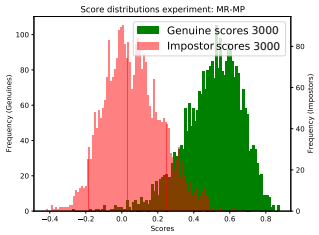
(f) DeepFill+VGG



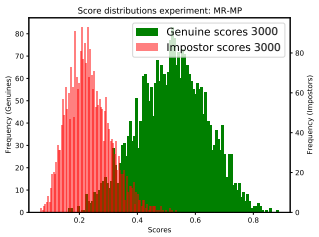
(g) DeepFill+AF



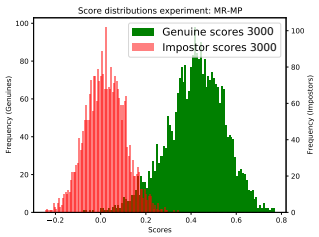
(h) DeepFill+VGG2



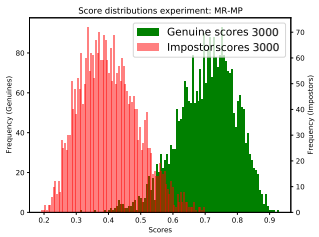
(i) ICT+CL



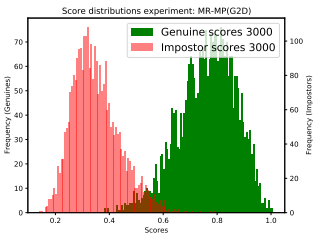
(j) ICT+VGG



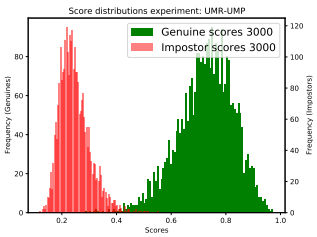
(k) ICT+AF



(l) ICT+VGG2



(m) Our G2D



(n) Ideal Case with UMR-UMP

Figure 9. The similarity score distributions achieved by different models under MR-MP setting. MR-MP briefs for masked/unmasked reference and masked/unmasked probes. The similarity score of the genuine pairs are in green color, and impostor pairs in red. Smaller overlapping suggest better discriminative ability. UMR-UMP refers to the circumstance where both probe and reference are unmasked, indicating the normal ideal case.