Transferable Facial Privacy Protection against Blind Face Restoration via Domain-Consistent Adversarial Obfuscation

Kui Zhang[1] Hang Zhou[2] Jie Zhang[3] Wenbo Zhou[1] Weiming Zhang[1] Nenghai Yu[1]

Abstract

With the rise of social media and the proliferation of facial recognition surveillance, concerns surrounding privacy have escalated significantly. While numerous studies have concentrated on safeguarding users against unauthorized face recognition, a new and often overlooked issue has emerged due to advances in facial restoration techniques: traditional methods of facial obfuscation may no longer provide a secure shield, as they can potentially expose anonymous information to human perception. Our empirical study shows that blind face restoration (BFR) models can restore obfuscated faces with high probability by simply retraining them on obfuscated (e.g., pixelated) faces. To address it, we propose a transferable adversarial obfuscation method for privacy protection against BFR models. Specifically, we observed a common characteristic among BFR models, namely, their capability to approximate an inverse mapping of a transformation from a high-quality image domain to a low-quality image domain. Leveraging this shared model attribute, we have developed a domain-consistent adversarial method for generating obfuscated images. In essence, our method is designed to minimize overfitting to surrogate models during the perturbation generation, thereby enhancing the generalization of adversarial obfuscated facial images. Extensive experiments on various BFR models demonstrate the effectiveness and transferability of the proposed method.

1. Introduction

Recently, the rapid development of deep learning has advanced computer vision comprehensively. To train vision models, a massive amount of data is being utilized, which may contain sensitive or unauthorized data collection, thus inevitably raising concerns about personal privacy invasion. Face privacy protection, one of the most significant issues among personal privacy has been studied promptly, and it can be categorized into two: reliable defense and reliable service. On the one hand, reliable defense (Cherepanova et al., 2020; Shan et al., 2020; Yang et al., 2021b; Huang et al., 2021; Tolosana et al., 2020; Sun et al., 2018)targets to protect face images from unauthorized face analysis models due to the widespread deployment of surveillance devices and the popularity of social media at the moment. Adversarial attacks are used to produce noise to defend against face recognition. On the other hand, reliable service (Hukkelås et al., 2019; Maximov et al., 2020; Gafni et al., 2019; Kuang et al., 2021; Li et al., 2021; Wu et al., 2019; Chen et al., 2021) takes advantage of generative models to produce massive realistic fake faces for computer vision tasks, which avoids the possibility of access to real face data, leading to reliably applying existing datasets for face-related tasks such as face detection and recognition.

Although these studies mitigate face privacy eavesdropping caused by unauthorized neural face recognition, the growing focus on privacy protection has overlooked another critical issue: privacy breach of obfuscated faces. In many public settings, law enforcement agencies, and the media anonymize unauthorized faces to protect privacy by operations like pixelation to influence perception purposely. However, recent work (Yang et al., 2020a; 2021a; Wang et al., 2021; 2022; Zhou et al., 2022) on blind face restoration (BFR) reveals that even low-quality face images that have undergone complex degradation can be recovered to faces with realistic details and rich identity information. Two examples of face privacy leakage are shown in Figure 1. Upon processing mosaic images through BFR models, obfuscated images are invalid and it is difficult for human observers to differentiate between restored and unpixelated faces.

Contemporary anonymization methods for face privacy protection are limited to a great extent. Existing obfuscation

[1]University of Science and Technology of China [2]Simon Fraser University [3]Nanyang Technological University. Correspondence to: Weiming Zhang <zhangwm@ustc.edu.cn>.

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

methods are mainly based on low-level image processing, such as blurring, pixelation, adding masks, and pixel replacement. Among them, a few effective obfuscation methods such as sticking black masks and markers are not pleasant to human eyes, while techniques like pixelation which are visually satisfying can be easily attacked by neural face restoration models. Although high-quality de-identified images can be obtained by facial manipulation, face generation models based on generative adversarial networks (GANs) tend to have pitfalls within themselves. On the one hand, a great number of real faces are required to train a preferable model, which cannot avoid face authorization. On the other hand, GANs are prone to be the inferred part of the training data through collisions (Hu et al., 2021), thus compromising face privacy in the training set.

To alleviate the above issues, we introduce a proactive defense against blind face restoration by adding adversarial perturbations to the obfuscated image. However, optimizing perturbations directly using existing attack algorithms often fails to generalize to other models. To overcome this limitation, we propose to design the defense using the shared characteristics of different BFR models. We observe that the nature of the BFR model is learning how to map degraded images back to corresponding high-quality ones. If the mapping from a high-quality image to a degraded-quality image can be obtained, the BFR model can be approximated as an inverse mapping. Despite the diverse structures and parameters of different BFR models, their underlying objective remains consistent: to be a subset of the numerous reasonable inverse mappings of the degradation transformation. Capitalizing on this insight, we propose a domain-consistent adversarial obfuscation approach (denoted as DomCo). This method ensures the perturbation remains aligned with the domain of the obfuscated face throughout the optimization phase. We demonstrate from the perspective of gradient propagation that the effectiveness of domain consistency is derived from its ability to transform partial gradients into those associated with the generalized characteristics of the BFR model, enhancing the similarity of the gradients with respect to the input across various BFR models. Expanding on this analysis, we also employ domain-consistent data augmentation to train the surrogate model to further stabilize the gradient transformation, enhancing the transferability of adversarially obfuscated faces.

We conduct extensive experiments on various BFR models and demonstrate that the proposed method can help users generate anonymized images with high transferability. Compared with baseline obfuscated images, the generated anonymous faces have much lower perceptual similarity and identity similarity after being restored, thus mitigating the risk of privacy breaches due to blind face restoration.

2. Related Work

We cover related research on face privacy protection with a focus on defenses against face analysis and reliable service methods. we also pay attention to blind face restoration methods that may inadvertently lead to privacy breaches.

2.1. Privacy Protection against Face Analysis

These researches focus on face analysis models and utilize anonymization or so-called obfuscation to avoid illegal identification or tampering misuse. Classic obfuscation methods aim to remove sensitive information from images, which can reduce the performance of face recognition to some extent (Wilber et al., 2016) Inspired by the observation that face recognition systems are also vulnerable to adversarial examples (Dong et al., 2019; Komkov & Petiushko, 2021; Yang et al., 2020b), adversarial attack-based methods (Cherepanova et al., 2020; Shan et al., 2020; Yang et al., 2021b) add imperceptible perturbations to face images to protect the user from unauthorized face recognition. Huang *et al.* (Huang et al., 2021) also add noise to the face image as an initiative defense, which deteriorates the quality of the face editing, leading to user protection from deepfake (Tolosana et al., 2020). In this paper, we borrow the idea of adversarial examples to generate anonymized faces.

2.2. Privacy Protection for Reliable Service

The ability to preserve privacy without compromising various downstream tasks of face analysis is the main focus of such research. With the emergence of generative adversarial networks (GANs) or adversarial autoencoder (Hukkelås et al., 2019; Maximov et al., 2020; Gafni et al., 2019; Kuang et al., 2021; Li et al., 2021; Wu et al., 2019; Chen et al., 2021), generative models are developed to generate anonymous faces. The technical means are mainly to modify face attributes and styles or manipulate faces in the semantic space to meet privacy requirements. For example, CIA-GAN (Wu et al., 2019) fuses the identity information used for guidance in the feature space to decode identity-specific face images. The goal is to generate varied identities while ensuring that the distribution of obfuscated faces aligns with that of the original dataset so that privacy-preserving faces can continue to be used for face detection and recognition. However, these researches require a large number of real faces to train the generation model to manipulate the real face, which may involve additional privacy risks.

2.3. Blind Face Restoration

Blind face restoration (Li et al., 2018; Yang et al., 2020a; 2021a; Wang et al., 2021; Zhou et al., 2022; Wang et al., 2022; 2023) is an emerging technique for recovering high-quality faces from ambiguous counterparts, which aims to

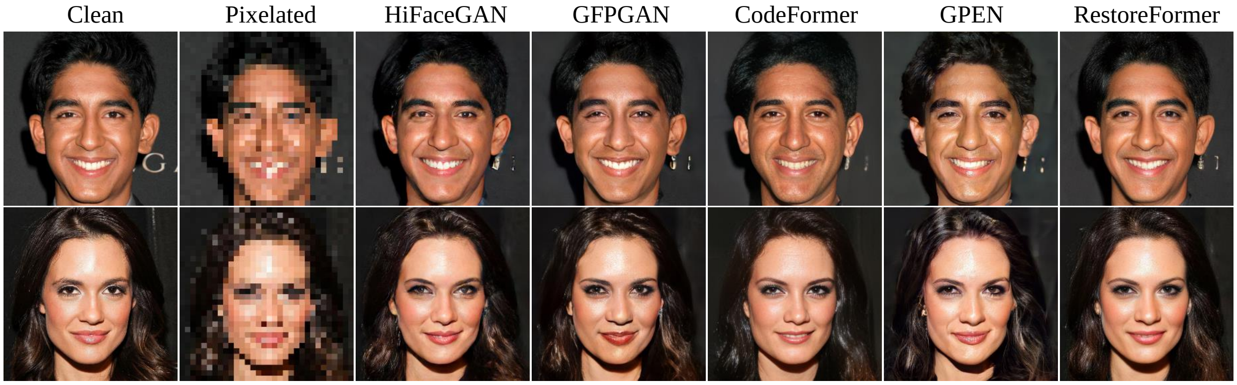


Figure 1. Some examples of personal face privacy breach. The pixelated image in the second column is transformed into a high-quality, detail-rich face through different BFR models. Although there are some artifacts in the restored images, they are very similar to the original face images. The privacy of the human face is completely exposed to the real world.

learn a restoration mapping from degraded faces to high-quality faces. Li *et al.* (Li et al., 2018) model image degradation as a combination of Gaussian blur, downsampling, Gaussian noise, and JPEG compression. Most of the subsequent studies use this as a basis to simulate degraded data from real-world scenarios and then design better network structures and training strategies. Yang *et al.* (Yang et al., 2020a) design a collaborative suppression and replenishment approach to end-to-end restoration. Menon *et al.* (Menon et al., 2020), Yang *et al.* (Yang et al., 2021a), Wang *et al.* (Wang et al., 2021), Zhou *et al.* (Zhou et al., 2022) and Wang *et al.* (Wang et al., 2022) exploit the richer generative prior, and the latter four yield surprising results. However, the powerful efficacy of face restoration raises new concerns that obfuscated face images are no longer secure. In this paper, we work to mitigate this issue by allowing the protected obfuscated faces to remain anonymous.

3. Methodology

In this section, we first formulate the problem, introduce the protection goal, and describe the threat model. Then we describe the proposed domain-consistent transferable adversarial obfuscation approach in detail.

3.1. Preliminaries

3.1.1. Problem Formulation

For an obfuscated face image $\boldsymbol{x} \in \mathcal{X}$ generated by a degenerate function $\mathcal{T}$, we aim to design an anonymization algorithm that generates a corresponding robust obfuscated face $\boldsymbol{x}_a = \boldsymbol{x} + \delta$ such that the BFR model is unable to restore $\boldsymbol{x}_a$ back to an identity-consistent, high-quality face while trying to make $\boldsymbol{x}_a$ visually consistent with $\boldsymbol{x}$. The generation of robust anonymized faces can be formalized as the following optimization problem:

$$\max \mathcal{L}(\boldsymbol{z}, \mathcal{R}(\boldsymbol{x}_a)), \text{ s.t. } \mathcal{S}(\boldsymbol{x}, \boldsymbol{x}_a) \leq \varepsilon \tag{1}$$

where $\boldsymbol{z} \in \mathcal{Z}$ is the ground-truth face, $\mathcal{L}$ is a function that measures the distance between the restored face and the real face and is used to hinder the recovery of confused images. $\mathcal{S}(\cdot)$ can be any function including the l_p norms used to characterize the similarity between the original obfuscated image and the deeply anonymized image.

3.1.2. Threat Model

We concentrate on the privacy breach risk due to no-reference BFR models, with a specific focus on the privacy vulnerabilities of commonly used pixelated obfuscations. We consider the privacy attack in the black-box scenario, where the BFR model used by the attacker and the restoration results are unknown to the user. This scenario is more challenging and more practical than the white-box scenario.

3.1.3. Goal of Protection

Proactive defense-hardened obfuscated face images have the following two necessary properties.

- *Unidentifiability*. Not only does the face need to be anonymous after deep obfuscation, but also the image recovered by the face restoration model needs to have the lowest possible resemblance to the real face and be unrecognizable by the AI model.
- *Transferability.* In the real world, on the one hand, we do not have access to the parameters of offline BFR models, and on the other hand, fine-tuning anonymous face images for a specific model may fail for other models. Therefore the anonymized faces should be able to transfer across different black-box models without the need for generating again.

3.2. Domain-Consistent Adversarial Obfuscation

Motivation. For the optimization problem shown in Equation 1, classical gradient descent-based methods such as PGD (Madry et al., 2018) or C&W-based joint optimization (Carlini & Wagner, 2017) can be used to satisfy the objective. However, directly applying these methods for optimization often yields a locally optimal solution that overfits the surrogate model and hinders the generalization of the adversarially obfuscated image to other restoration models. Creating transferable adversarial obfuscated faces that are effective against various BFR models poses a significant challenge. To tackle this issue, we first analyze the BFR model to discern characteristics shared by different models, so as to design a more general defense. Formally, for a collection of high-quality (HQ) face images denoted $\mathcal{Z}$, low-quality (LQ) images set $\mathcal{X}$ can be generated by a degradation function $\mathcal{T}(\cdot)$ that is used to simulate real-world degradation. $\mathcal{T}$ is usually defined as the set of a series of corrupted transformations:

$$\mathcal{T}(\boldsymbol{z}) = [(\boldsymbol{z} \otimes \boldsymbol{k}_\sigma) \downarrow_s + \boldsymbol{n}_\zeta]_{\texttt{JPEG}_q} \uparrow_s . \tag{2}$$

where $\boldsymbol{k}_\sigma$ denotes a Gaussian blur kernel with standard deviation σ, $\downarrow_s$ and $\uparrow_s$ represent upsampling and downsampling, respectively, where s is the scale factor, $\otimes$ denotes convolution, $\boldsymbol{n}_\zeta$ is the additive white Gaussian noise and $\texttt{JPEG}_q$ stands for JPEG compression with quality factor q. The BFR model is composed of multi-layer neural networks to remove the degraded information and generate a high-definition face image with a consistent identity. Most existing BFR models can be abstracted as a mapping $\mathcal{R}(\cdot) : \mathcal{X} \rightarrow \mathcal{Z}$ where $\mathcal{R}(\cdot)$ can be viewed as an inverse mapping of $\mathcal{T}$. Thus whatever structure and parameters BFR models have, they realize similar functions. If we consider the degradation and restoration as a black box and adapt this common function from the input-output perspective, we can naturally enhance the transferability of obfuscated face images.

Based on the above discussion we propose a domain-consistent adversarial obfuscated image generation method, capitalizing on the commonalities across these models. Definition 3.1 and 3.2 elucidate the idea of a domain and the criteria that should be required by a domain-consistent adversarial obfuscated image.

Definition 3.1. Given two image sets $\mathcal{X}$ and $\mathcal{Z}$, where $\mathcal{X}$ belongs to the domain $D_{\mathcal{T}}$ if there exists a transformation $\mathcal{T}$ satisfying $\mathcal{X} = \mathcal{T}(\mathcal{Z})$.

Definition 3.2. For an adversarial obfuscated image $x_a = \boldsymbol{x} + \delta$, where $\boldsymbol{x} = \mathcal{T}(\boldsymbol{z}) \in \mathcal{X}$ is a face image belongs to domain $D_{\mathcal{T}}$, δ is the adversarial perturbation. If there exists a δ_z satisfying $\mathcal{T}(\boldsymbol{z} + \delta_z) - \mathcal{T}(\boldsymbol{z}) = \delta$, then δ is a domain-consistent perturbation, x_a is a domain-consistent adversarial obfuscated image.

If the perturbations satisfy domain consistency, the optimization process can be extended to the original input level to develop attacks that adapt to the commonality. Since the transformation $\mathcal{T}$ is a process that loses information, it is not straightforward to determine the perturbations that satisfy the domain consistency by solving the inverse of $\mathcal{T}$. To incorporate domain consistency conditions into the optimization, we compute the perturbation in the $\mathcal{Z}$ domain without designing intricate constraints. Consequently, the optimization objective, as stated in Eq. 3, is reformulated as follows:

$$\max \mathcal{L}(\boldsymbol{z}, \mathcal{R}(\mathcal{T}(\boldsymbol{z}_a))), \text{ s.t. } \mathcal{S}(\boldsymbol{x}, \boldsymbol{x}_a) \leq \varepsilon, \tag{3}$$

where the final anonymized image $\boldsymbol{x}_a = \mathcal{T}(\boldsymbol{z}_a)$.

The objective can be solved using common gradient attack algorithms such as PGD. $\mathcal{S}$ can use L_p norm truncation or any distance constraints such as mean square error loss. In our approach, we constrain perturbations using perceptual distance and utilize Perceptual PGD (PPGD) (Laidlaw et al., 2020) for optimization. The integration of generating domain-consistent adversarial samples into PPGD is seamless, with further details provided in the appendix.

Domain consistency reduces overfitting. We illustrate how domain-consistent perturbation generation methods can mitigate overfitting to the surrogate model by analyzing gradient propagation. In one step of the optimization, the gradient of adversarial images in the high-quality domain is calculated as follows:

$$\nabla_{\boldsymbol{z}_a} \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \mathcal{R}(\mathcal{T}(\boldsymbol{z}_a))} \frac{\partial \mathcal{R}(\mathcal{T}(\boldsymbol{z}_a))}{\partial \mathcal{T}(\boldsymbol{z}_a)} \frac{\partial \mathcal{T}(\boldsymbol{z}_a)}{\partial \boldsymbol{z}_a}. \tag{4}$$

For simplicity, we use $\triangle$ to denote $\frac{\partial \mathcal{R}(\mathcal{T}(\boldsymbol{z}_a))}{\partial \mathcal{T}(\boldsymbol{z}_a)} \frac{\partial \mathcal{T}(\boldsymbol{z}_a)}{\partial \boldsymbol{z}_a}$. In the mosaic case, the gradient of the obfuscated image ends up as follows:

$$\delta = \mathcal{T}(\boldsymbol{z} + \nabla_{\boldsymbol{z}_a} \mathcal{L}) - \mathcal{T}(\boldsymbol{z}) = \mathcal{T}(\nabla_{\boldsymbol{z}_a} \mathcal{L}). \tag{5}$$

Eq. 4 and 5 show that domain consistency will additionally project the chained perturbation back into the domain $D_{\mathcal{T}}$. It is known that the product of the derivatives of a pair of inverse functions is equal to 1 at the same point. However, $\mathcal{T}$ is a non-invertible function in the task of blind face restoration. But $\mathcal{R}$ is a deterministic function and it can be considered that $\mathcal{R}$ has learned a reasonable approximation of one of the multiple inverse mappings of $\mathcal{T}$. Consequently, although the $\triangle$ does not equal to $\mathbf{I}$, as would be expected with inverse functions, the projection process neutralizes some of the non-robust gradients specifically associated with the BFR model, enhancing perturbations to the common characteristics of the BFR model.

Compared to domain-consistent adversarial obfuscation, vanilla domain-inconsistent perturbation gradient is

$\nabla_{x_a}\mathcal{L} = \frac{\partial \mathcal{L}}{\partial \mathcal{R}(x_a)}\frac{\partial \mathcal{R}(x_a)}{\partial (x_a)}$, the structure and parameters of the BFR model are deeply involved in the gradient propagation, which finally leads to the generated noise overfitting the current BFR model.

3.3. Transferability Enhancement

Despite the domain-invariant adversarial paradigm reducing overfitting to a surrogate BFR model, the transferability of adversarial obfuscated images is still affected by the surrogate model. Equation 4 describes the computation of the perturbation gradient in a single update, whereas the optimization is often repeated multiple times to enhance the adversarial performance of the obfuscated image. If the restoration model is not robust enough, a small change in $\boldsymbol{z}_a$ can degrade the image quality of $\mathcal{R}(\mathcal{T}(\boldsymbol{z}_a))$ heavily. This mismatch results in a significant weakening of the effect of $\triangle$, and fails to effectively enhance the focus on the commonalities of BFR models. Thus selecting a BFR model that exhibits higher robustness to the neighborhood of obfuscated data points is crucial for serving as a surrogate model. In this context, robustness pertains to the capacity to handle adversarial low-quality images. Notably, adversarial obfuscated faces generated using a robust BFR model demonstrate higher transferability. This insight also aligns with the findings in classification models (Springer et al., 2021). We propose a domain-consistent data augmentation technique to mitigate this problem.

Domain-consistent data augmentation. The idea of domain-consistent data augmentation is to align with the optimization process by placing the domain transformation $\mathcal{T}$ after the data augmentation to simulate the generation of domain-consistent adversarial perturbations. For mosaic transformation, we add random Gaussian noise to high-quality images for data augmentation, *i.e.*, $A(\boldsymbol{z}) = \boldsymbol{z} + \sigma$. There are two optional optimization objectives for augmented training:

$$\min \mathcal{L}_D \left(\mathcal{R}(\mathcal{T}(A(\boldsymbol{z}))), A(\boldsymbol{z})\right), \tag{6}$$

$$\min \mathcal{L}_D \left(\mathcal{R}(\mathcal{T}(A(\boldsymbol{z}))), \boldsymbol{z}\right). \tag{7}$$

Since the objective of the augmented training is to slow down the speed of restored image distortions when subjected to slight perturbations, the training objective described in Eq. 7 will exhibit stronger perturbation suppression than Eq. 6, so we use Eq. 7 as the final training objective.

3.4. Loss Function

The key to the failure of the obfuscated image being recovered is the design of the loss function $\mathcal{L}$. We consider the following loss functions.

- Content loss. The content loss calculates the L_1 norm of the ground-truth image and the restored images to enlarge the distance of the face in the output space.
- Perceptual loss. Perceptual loss (Johnson et al., 2016) computes the correlation between features and is widely used to improve image quality in various image generation tasks. In turn, it can be used to reduce the texture and structural quality of the restored image.
- Task-guided face structure loss. The degree of disorganization of the face structure is difficult to measure with an explicit distance function. We therefore use an additional task-specific model to implicitly characterize the objective. Specifically, we design a face structure loss based on the face parsing model P. For an input face image $\boldsymbol{z}$, the output $P_z \in \mathbb{R}^{C\times H\times W}$ is a 19-channel tensor with the same size as the face image where every channel represent the prediction probability of different face attributes. The optimization goal is to make the recovered face structure as chaotic as possible, so the face structure loss is defined as the probability of minimizing the correct labels at each position of the face image:

$$L_{struc} = -\frac{1}{HW}\sum_{H,W}\sum_{C}\widehat{y}\cdot P(z_a)\cdot M, \tag{8}$$

where $\widehat{y}$ is a one-hot label vector of $P(z)$ and M is a binary mask where only the face region is 1 and the rest of the positions are 0.

Compared to the first two loss functions, which tend to indiscriminately enlarge the distance between the restored face and the original face, L_{struc} focuses more explicitly on the structure of the face. This can reduce meaningless perturbations, therefore we use L_{struc} as the final loss function.

4. Experiments

4.1. Experimental Setting

Datasets. Following previous work, we use the FFHQ (Karras et al., 2019) dataset and perform pixelation to train the model to help them have better performance in restoring the pixelated images. We resize the source image from 512 × 512 to 32 × 32 by nearest neighbor interpolation and resize it back to simulate the process of the pixelation. We select 2,000 images from the training set and test set of CelebA-HQ to form a test set of size 4,000, respectively, for evaluating the effectiveness of the proposed method.

Blind face restoration models. We select five models with excellent performance on face restoration, namely HIFaceGAN (Yang et al., 2020a), GPEN (Yang et al., 2021a), GFPGAN (Wang et al., 2021), RestoreFormer (Wang et al.,

Table 1. Comparison of the effectiveness of different obfuscation methods against BFR models. The surrogate model is the RestoreFormer which does not use domain-consistent data augmentation.

Attacker	Defense	FID	MS-SSIM	LPIPS	IDS	LMD	mPR	mIoU
HIFaceGAN (Yang et al., 2020a)	Pixelate	5.04	0.9224	0.1100	0.72	4.91	0.872	0.867
	PPGD	37.64	0.8600	0.1503	0.54	11.47	0.724	0.607
	DomCo	**165.18**	**0.7590**	**0.1962**	**0.25**	**83.41**	**0.323**	**0.247**
GFPGAN (Wang et al., 2021)	Pixelate	7.14	0.9298	0.1071	0.70	4.66	0.861	0.769
	PPGD	33.60	0.8787	0.1483	0.53	13.20	0.687	0.598
	DomCo	**115.37**	**0.8177**	**0.1859**	**0.33**	**40.74**	**0.374**	**0.304**
GPEN (Yang et al., 2021a)	Pixelate	6.51	0.9251	0.1141	0.76	4.86	0.856	0.766
	PPGD	73.67	0.8633	0.1623	0.55	15.84	0.702	0.602
	DomCo	**147.20**	**0.8061**	**0.1904**	**0.35**	**50.87**	**0.426**	**0.340**
CodeFormer (Zhou et al., 2022)	Pixelate	6.47	0.9181	0.1045	0.65	4.94	0.845	0.740
	PPGD	7.07	0.9126	0.1100	0.63	5.21	0.814	0.715
	DomCo	**74.35**	**0.8165**	**0.1596**	**0.32**	**25.22**	**0.512**	**0.410**
RestoreFormer (Wang et al., 2022)	Pixelate	8.43	0.9352	0.0984	0.74	4.54	0.868	0.772
	PPGD	**179.09**	**0.7670**	0.1935	**0.19**	**73.94**	**0.107**	**0.083**
	DomCo	143.13	0.7834	**0.1968**	0.23	58.67	0.160	0.122

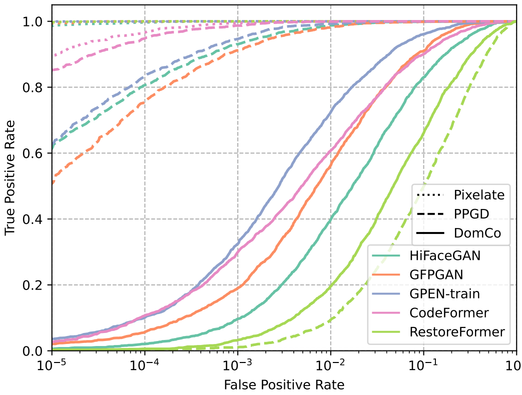


Figure 2. Comparison of ROC curves of different methods.

2022) and CodeFormer (Zhou et al., 2022). We use the suffix '$-$A' to indicate the augmented models for simplicity. Unless specifically stated, the models that appear are unaugmented.

Evaluation metrics. To evaluate the effectiveness of adversarially obfuscated faces, IDentity Similarity (IDS), MultiScale Structural Similarity (MS-SSIM) (Wang et al., 2003), Frechet Inception Distances (FID) (Heusel et al., 2017) and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) as quantitative metrics are used. We calculate the cosine similarity of the features drawn from the ArcFace (Deng et al., 2019) as the identity similarity. With the ground-truth image as the reference, FID, MS-SSIM, and LPIPS not only measure the quality of the restored image but also indicate the difference with the reference image in terms of similarity and statistics, aligning closely with human visual perception. In addition, we also use the face landmark distance (LMD), the mean intersection over union (mIoU) over classes of the face region, and the mean pixel recall (mPR) to measure the degree of disorder of restored faces. For metrics with the help of reference images, calculations are confined to the facial region. In our task, larger FID, LPIPS, and LMD indicate better defense performance, while for the remaining metrics, a smaller value is better.

Implementation details. We optimized each adversarial obfuscated image 20 times, using the bisection method to project the perturbations, the maximum ϵ is 1.5, and the step size defaults to $\frac{2\epsilon}{20}$. The face region mask M is specified by the labels of the clean face parsing map (Lee et al., 2020), excluding the background, hair, ear, and neck parts.

4.2. Main Results

4.2.1. Evaluation on Domain-Consistent Obfuscation

We first tested the effectiveness of adversarial obfuscated faces generated by the proposed method (denoted as DomCo) on the unaugmented model to defend BFR models. Table 1 shows the ability of anonymized faces generated by different obfuscation methods against different BFR models when RestoreFormer (Wang et al., 2022) is used as the surrogate model.

Table 2. Defense performance of different surrogate models trained with domain-consistent augmentation. The suffixes ”-GF” and ”-CF” denote the GFPGAN and CodeFormer as the surrogate model, respectively. The gray rows are the results of augmented surrogate models.

Attacker	Defense	FID	MS-SSIM	LPIPS	IDS	LMD	mPR	mIoU
HIFaceGAN (Yang et al., 2020a)	DomCo-GF	**155.03**	0.7821	0.1869	0.32	59.49	0.496	0.392
	+Aug	142.66	**0.7435**	**0.1931**	**0.24**	**70.09**	**0.306**	**0.230**
	DomCo-CF	46.95	0.8442	0.1570	0.49	13.29	0.716	0.580
	+Aug	**74.43**	**0.7910**	**0.1775**	**0.36**	**27.10**	**0.559**	**0.430**
GFPGAN (Wang et al., 2021)	DomCo-GF	**176.77**	**0.7600**	**0.1911**	**0.24**	**73.55**	**0.143**	**0.114**
	+Aug	96.80	0.7794	0.1744	0.28	34.13	0.232	0.182
	DomCo-CF	24.73	0.8810	0.1448	0.55	9.54	0.740	0.634
	+Aug	**38.51**	**0.8392**	**0.1618**	**0.43**	**15.14**	**0.614**	**0.498**
GPEN (Yang et al., 2021a)	DomCo-GF	**136.92**	0.8279	0.1810	0.45	31.54	0.616	0.509
	+Aug	136.02	**0.7750**	**0.1897**	**0.29**	**48.49**	**0.322**	**0.250**
	DomCo-CF	54.04	0.8723	0.1536	0.59	10.04	0.764	0.646
	+Aug	**72.14**	**0.8227**	**0.1726**	**0.43**	**18.92**	**0.602**	**0.481**
CodeFormer (Zhou et al., 2022)	DomCo-GF	42.06	0.8401	0.1451	0.39	13.93	0.655	0.532
	+Aug	**75.35**	**0.7985**	**0.1633**	**0.30**	**26.42**	**0.465**	**0.370**
	DomCo-CF	**181.64**	**0.7331**	**0.1963**	**0.18**	**90.42**	**0.140**	**0.106**
	+Aug	48.36	0.8171	0.1569	0.35	20.82	0.544	0.430
RestoreFormer (Wang et al., 2022)	DomCo-GF	37.80	0.8776	0.1522	0.56	9.18	0.739	0.632
	+Aug	**68.70**	**0.8161**	**0.1702**	**0.35**	**20.99**	**0.418**	**0.333**
	DomCo-CF	13.66	0.9040	0.1265	0.64	6.12	0.810	0.698
	+Aug	**29.79**	**0.8567**	**0.1514**	**0.47**	**10.95**	**0.670**	**0.545**

It can be intuitively seen that both DomCo and PPGD are well defended against the white-box model RestoreFormer and the restored face images show a significant decrease in structural and perceptual similarity as well as in image quality, indicating that the obfuscated information in the images undermines the ability of the model to maintain structural consistency. PPGD shows a significant performance degradation when oriented to other models, while the proposed method can maintain a similar defense performance against other BFR models, proving the superior generalization ability of DomCo.

In addition, we sampled 2,000 identity-differentiated face images from the test set to further evaluate the effectiveness of DomCo on face identity protection. Figure 2 shows the ROC curves on the restored anonymized face images. For most BFR models, anonymized faces generated by the proposed method have lower face verification TAR at the same FAR, indicating that DomCo effectively prevents the restoration model from reconstructing the identity information. Furthermore, other adversarial obfuscation methods exhibit significant gaps in performance between black-box models and white-box models. DomCo, by contrast, consistently exhibits superior performance in defending a range of BFR models, illustrating better transferability.

4.2.2. Effectiveness of Domain-Consistent Augmentation

Table 2 illustrates the effect of domain-consistent data-augmented training on the transferability of adversarially obfuscated face images. The transferability of adversarially obfuscated images generated by the surrogate model augmented with domain-consistent data significantly improves, leading to a greater reduction in image-level and feature-level similarities, as well as a further increase in the degree of disorder of the faces. This performance gain is even more evident when GFPGAN is used as a surrogate model to defend against RestoreFormer. The defensive performance of the augmented model decreases in the white-box scenario because the parameters of the augmented model have changed compared to the original model. It is not strictly a white-box defense, and the augmented model also helps alleviate overfitting to the original model.

4.2.3. Qualitative Results

Figure 3 presents the restoration results of the obfuscated face images generated by different approaches. The restored images show varying degrees of degradation, and the distortion of the adversarially obfuscated face image is acceptable. The restored face exhibits a marked increase in confusion and entanglement, leading to a significant displacement of

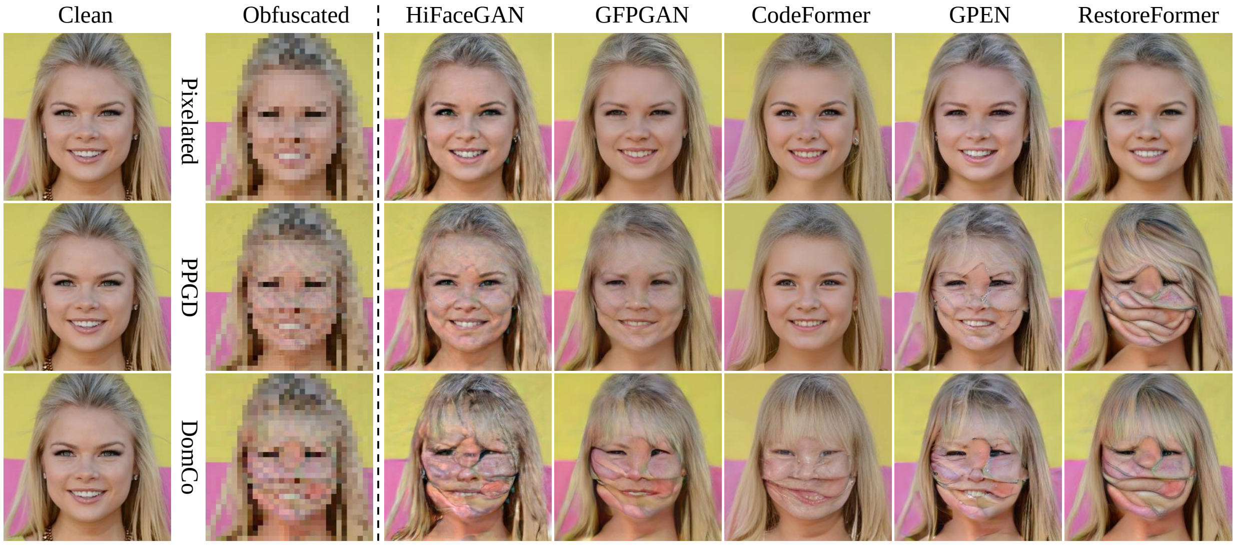
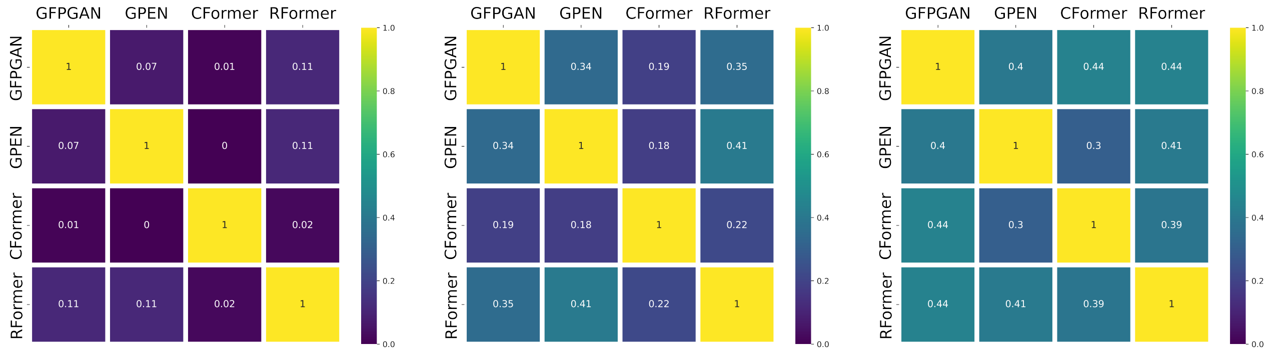


Figure 3. Qualitative comparison of the restored results of obfuscated face images using different BFR models.



(a) Domain-Inconsistent

(b) Domain-Consistent

(c) Domain-Consistent Augmentation

Figure 4. The cosine similarity between perturbation gradients of blind face restoration models when using different defense paradigms. CFormer and RFormer represent CodeFormer and RestoreFormer, respectively. In (c), GFPGAN and CodeFormer are trained with domain-consistent data augmentation.

identity information. Furthermore, the proposed method demonstrates superior defense capabilities against different BFR models with better generalization performance compared to the domain-inconsistent baseline method.

4.3. Ablation Study

4.3.1. Domain-Consistent Gradient Analysis

To further understand the improvement of domain consistency in the generalization of obfuscated images, we analyzed the gradients of the input images at the initial time of optimization. Specifically, we utilized a subset of 100 images from the test set to calculate the gradients of the loss on the input under various strategies using different surrogate models and compared their cosine similarity. Figure 4 illustrates the gradient similarity under different defense paradigms. The results demonstrate that domain consistency enhances the similarity of perturbation gradients between surrogate models and strengthens the adaptation of perturbations to model commonalities, while domain-consistent data augmentation further stabilizes and amplifies this characteristic. Note that even if the gradient attention remains similar, the images generated by a more robust surrogate model will have stronger obfuscation capability, as its outputs are less susceptible to perturbations.

4.3.2. Analysis of loss functions

We conducted experiments using the content loss function (denoted as $\mathcal{L}_c$) and the perceptual loss function (denoted as $\mathcal{L}_{percep}$) which uses the pre-trained VGG19 network (Si-

Table 3. Defense performance when using different loss functions. The upper bound of the perturbation ϵ is uniformly set to 1. Values in bold represent the best performance.

Attacker	Loss	FID	MS-SSIM	LPIPS	IDS	LMD	mPR	mIoU
HiFaceGAN (Yang et al., 2020a)	$\mathcal{L}_c$	76.81	0.7845	0.1785	0.34	18.61	0.640	0.476
	$\mathcal{L}_{percep}$	92.06	0.7867	0.1854	0.33	23.88	0.625	0.454
	$\mathcal{L}_{struc}$	**157.89**	**0.7666**	**0.1927**	**0.27**	**72.48**	**0.352**	**0.271**
GFPGAN (Wang et al., 2021)	$\mathcal{L}_c$	45.70	**0.8224**	0.1674	0.41	12.91	0.670	0.523
	$\mathcal{L}_{percep}$	65.56	0.8292	0.1804	0.40	15.07	0.675	0.512
	$\mathcal{L}_{struc}$	**107.34**	0.8253	**0.1813**	**0.34**	**37.20**	**0.407**	**0.333**
GPEN (Yang et al., 2021a)	$\mathcal{L}_c$	86.64	**0.8090**	0.1788	0.42	15.90	0.684	0.527
	$\mathcal{L}_{percep}$	104.15	0.8179	**0.1877**	0.41	17.93	0.698	0.519
	$\mathcal{L}_{struc}$	**138.52**	0.8148	0.1860	**0.37**	**43.64**	**0.460**	**0.370**
CodeFormer (Zhou et al., 2022)	$\mathcal{L}_c$	28.28	0.8273	0.1463	0.36	10.04	0.687	0.532
	$\mathcal{L}_{percep}$	40.40	0.8287	0.1535	0.34	11.27	0.684	0.516
	$\mathcal{L}_{struc}$	**68.46**	**0.8213**	**0.1565**	**0.33**	**23.62**	**0.528**	**0.423**
RestoreFormer (Wang et al., 2022)	$\mathcal{L}_c$	51.32	**0.7582**	0.1910	0.37	13.11	0.623	0.445
	$\mathcal{L}_{percep}$	93.62	0.8034	**0.2017**	0.38	16.58	0.647	0.458
	$\mathcal{L}_{struc}$	**136.82**	0.7912	0.1929	**0.25**	**53.69**	**0.180**	**0.140**

monyan & Zisserman, 2015) as the feature extractor and computes the Frobenius norm between the original image features and the recovered image features. The step size was fixed at $3/20$ to ensure that ϵ could reach the bound of 1. Table 3 compares the defensive performance when using different loss functions. The results indicate that both content-related losses are effective in reducing the quality of the recovered images. The content loss is more effective in reducing pixel-level similarity between restored and original faces, as evidenced by lower MS-SSIM values, while the perceptual loss achieves lower perceptual similarity by widening the differences at the feature level. The proposed face structure loss, on the other hand, has a more balanced performance and is particularly effective in preventing identity and structure recovery. The results suggest that simply reducing pixel-level or perceptual similarity is insufficient to disrupt face structure, while disrupting face structure can naturally reduce perceptibility.

5. Conclusion

This paper studied the risk of face privacy breaches from blind face restoration. Our findings indicate that while certain obfuscation techniques may thwart human observers, BFR models can decipher them with high accuracy. To alleviate this issue, we leveraged insights into the BFR model's inherent characteristics to develop a domain-consistent adversarial obfuscation approach. By constraining the perturbation to be within the same domain as the obfuscated image throughout the optimization, we illustrate that the proposed method can reduce the likelihood of overfitting from the perspective of gradient similarity. Extensive experiments across various BFR models demonstrate the effectiveness of the proposed method in safeguarding face privacy. We believe this research offers valuable insights for the future development of protection for such privacy concerns.

Acknowledgements

This work is supported in part by the Natural Science Foundation of China under Grants 62372423, 62121002, U20B2047, 62072421, 62206009, supported by the National Research Foundation, Singapore, and the Cyber Security Agency under its National Cybersecurity R&D Programme (NCRP25-P04-TAICeN) and by Fundamental Research Funds for the Central Universities under Grant WK2100000041. It is also supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTC-RGC-04).

Impact Statement

To the best of our knowledge, this is the first study to investigate the privacy breach risks associated with blind face restoration. We believe our method will inspire future research on the robustness and privacy protection of image-to-image-based models. A potential side effect is that this privacy-preserving technique might be employed to thwart normal restoration in some rare scenarios.

References

Bar Tal, O., Haviv, A., and Bermano, A. H. Omg-attack: Self-supervised on-manifold generation of transferable evasion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3696–3706, 2023.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. *2017 Ieee Symposium on Security and Privacy (Sp)*, pp. 39–57, 2017. ISSN 1081-6011. doi: 10.1109/Sp.2017.49. URL <GotoISI>://WOS:000413081300003.

Chen, J.-W., Chen, L.-J., Yu, C.-M., and Lu, C.-S. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J. P., Taylor, G., and Goldstein, T. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*, 2020.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., and Zhu, J. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Gafni, O., Wolf, L., and Taigman, Y. Live face de-identification in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hu, A., Xie, R., Lu, Z., Hu, A., and Xue, M. Tablegan-mca: Evaluating membership collisions of gan-synthesized tabular data releasing. In *Conference on Computer and Communications Security in press*, 2021.

Huang, Q., Zhang, J., Zhou, W., Zhang, W., and Yu, N. Initiative defense against facial manipulation. In *AAAI*, 2021. ISBN 2374-3468.

Hukkelås, H., Mester, R., and Lindseth, F. Deepprivacy: A generative adversarial network for face anonymization. In *International Symposium on Visual Computing*. Springer, 2019.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Komkov, S. A. and Petiushko, A. Advhat: Real-world adversarial attack on arcface face id system. In *International Conference on Pattern Recognition*, 2021.

Kuang, Z., Liu, H., Yu, J., Tian, A., Wang, L., Fan, J., and Babaguchi, N. Effective de-identification generative adversarial network for face anonymization. In *ACM MM*, 2021.

Laidlaw, C., Singla, S., and Feizi, S. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2020.

Lee, C.-H., Liu, Z., Wu, L., and Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., and Cao, X. Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *ACM MM*, 2021.

Li, X., Liu, M., Ye, Y., Zuo, W., Lin, L., and Yang, R. Learning warped guidance for blind face restoration. In *ECCV*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Maximov, M., Elezi, I., and Leal-Taixé, L. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., and Zhao, B. Y. Fawkes: Protecting privacy against unauthorized deep learning models. In *USENIX Security Symposium (USENIX Security)*, 2020. ISBN 1939133173.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., and Zhao, B. Y. Glaze: protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y. (eds.), *International Conference on Learning Representations*, 2015.

Springer, J., Mitchell, M., and Kenyon, G. A little robustness goes a long way: Leveraging robust features for targeted transfer attacks. *Advances in Neural Information Processing Systems*, 34:9759–9773, 2021.

Sun, Q., Ma, L., Oh, S. J., Van Gool, L., Schiele, B., and Fritz, M. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion*, 64, 2020.

Wang, X., Li, Y., Zhang, H., and Shan, Y. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, volume 2. Ieee, 2003. ISBN 0780381041.

Wang, Z., Zhang, J., Chen, R., Wang, W., and Luo, P. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17512–17521, 2022.

Wang, Z., Zhang, Z., Zhang, X., Zheng, H., Zhou, M., Zhang, Y., and Wang, Y. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1704–1713, 2023.

Wilber, M. J., Shmatikov, V., and Belongie, S. J. Can we still avoid automatic face detection? In *Winter Conference on Applications of Computer Vision*, 2016.

Wu, Y., Yang, F., Xu, Y., and Ling, H. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34(1), 2019. ISSN 1000-9000.

Yang, L., Liu, C., Wang, P., Wang, S., Ren, P., Ma, S., and Gao, W. Hifacegan: Face renovation via collaborative suppression and replenishment. In *ACM MM*, 2020a.

Yang, T., Ren, P., Xie, X., and Zhang, L. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021a.

Yang, X., Wei, F., Zhang, H., and Zhu, J. Design and interpretation of universal adversarial patches in face detection. In *ECCV*, Computer Vision – ECCV 2020. Springer International Publishing, 2020b. ISBN 978-3-030-58520-4.

Yang, X., Dong, Y., Pang, T., Zhu, J., and Su, H. Towards privacy protection by generating adversarial identity masks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhou, S., Chan, K., Li, C., and Loy, C. C. Towards robust blind face restoration with codebook lookup transformer. *Advances in Neural Information Processing Systems*, 35: 30599–30611, 2022.

A. Domain-Consistent Optimization

According to PPGD (Laidlaw et al., 2020), we rewrite the optimization objective as follows:

$$\begin{aligned} &\max_{\delta} \quad \mathcal{L}(\mathcal{R}(\mathcal{T}(\boldsymbol{z}+\delta))) \\ &s.t. \quad d(\mathcal{T}(\boldsymbol{z}), \mathcal{T}(\boldsymbol{z}_a)) = \|\phi(\mathcal{T}(\boldsymbol{z})) - \phi(\mathcal{T}(\boldsymbol{z}_a))\|_2 \leq \epsilon \end{aligned} \tag{9}$$

where ϕ is the flattened activations of AlexNet and the constant high-quality image z is omitted for simplicity. Let $\hat{f}(\boldsymbol{z}) := \mathcal{L}(\mathcal{R}(\mathcal{T}(\boldsymbol{z})))$ for an input $\boldsymbol{z} \in \mathcal{Z}$ and J be the Jacobian of $\phi(\mathcal{T}(\cdot))$ at $\boldsymbol{z}$ and $\nabla\hat{f}$ be the gradient of $\hat{f}(\cdot)$ at $\boldsymbol{z}$. The first-order approximation of (9) is

$$\max_{\delta} \quad \hat{f}(\boldsymbol{z}) + (\nabla\hat{f})^\top\delta \qquad \text{subject to} \quad \|J\delta\|_2 \leq \eta, \tag{10}$$

and can be solved in closed form.

It can be deduced that by adding an image transformation layer $\mathcal{T}$ before the BFR model and recognition model, we can apply the algorithm in (Laidlaw et al., 2020) directly.

B. Experiments

B.1. Different Distance Constraints

We have tried different distance constraints, including the PGD approach with an l_p norm constraint. Table 4 demonstrates the effectiveness of the adversarial perturbed images generated under the l_∞ norm. The obfuscation ability is weakened in both white-box and transferability scenarios, which may be due to the optimization method based on perceptual distance constraining the overall perceptibility, while l_∞ norm restricts the magnitude of pixel modification.

Table 4. Defense performance when using the l_p norm as the distance constraint. $\epsilon_\infty = 8$ and $\epsilon_\infty = 16$ indicate that the perturbation bounds are $L_\infty = \frac{8}{255}$ and $L_\infty = \frac{16}{255}$, respectively.

Attacker	Budget	FID	MS-SSIM	LPIPS	IDS	LMD	mPR	mIoU
HiFaceGAN (Yang et al., 2020a)	$\epsilon_\infty = 8$	29.56	0.8570	0.1484	0.31	12.77	0.689	0.577
	$\epsilon_\infty = 16$	120.11	0.7838	0.1906	0.30	46.02	0.451	0.360
GFPGAN (Wang et al., 2021)	$\epsilon_\infty = 8$	22.07	0.8843	0.1425	0.53	11.34	0.687	0.595
	$\epsilon_\infty = 16$	78.28	0.8220	0.1855	0.35	27.76	0.486	0.408
GPEN (Yang et al., 2021a)	$\epsilon_\infty = 8$	38.84	0.8801	0.1472	0.58	10.62	0.719	0.620
	$\epsilon_\infty = 16$	116.40	0.8114	0.1911	0.38	33.77	0.533	0.436
CodeFormer (Zhou et al., 2022)	$\epsilon_\infty = 8$	13.71	0.8813	0.1230	0.50	8.18	0.724	0.612
	$\epsilon_\infty = 16$	48.40	0.8366	0.1506	0.37	15.62	0.606	0.495
RestoreFormer (Wang et al., 2022)	$\epsilon_\infty = 8$	29.20	0.8800	0.1412	0.49	12.08	0.630	0.531
	$\epsilon_\infty = 16$	100.74	0.8006	0.1936	0.29	34.09	0.319	0.252

B.2. More Qualitative Results

In Fig. 5 we show more examples of generated adversarial obfuscated face images.

C. Discussion

C.1. Relationship with Classification Task

We demonstrate that domain consistency improves the generalizability of defenses (attacks as defenses) against image-to-image-based blind face restoration models. This property may be able to explain adversarial transferability in other tasks. In a wide range of tasks, the transformation of the domain $D_{\mathcal{T}}$ may be unknown, so we consider the general form of domain consistency, i.e., for $\boldsymbol{x} \in \mathcal{X}$, there exists a $\boldsymbol{z} \in \mathcal{Z}$ aligned with the output that satisfies $\boldsymbol{x} = \mathcal{T}(z)$ by a transformation $\mathcal{T}$. In

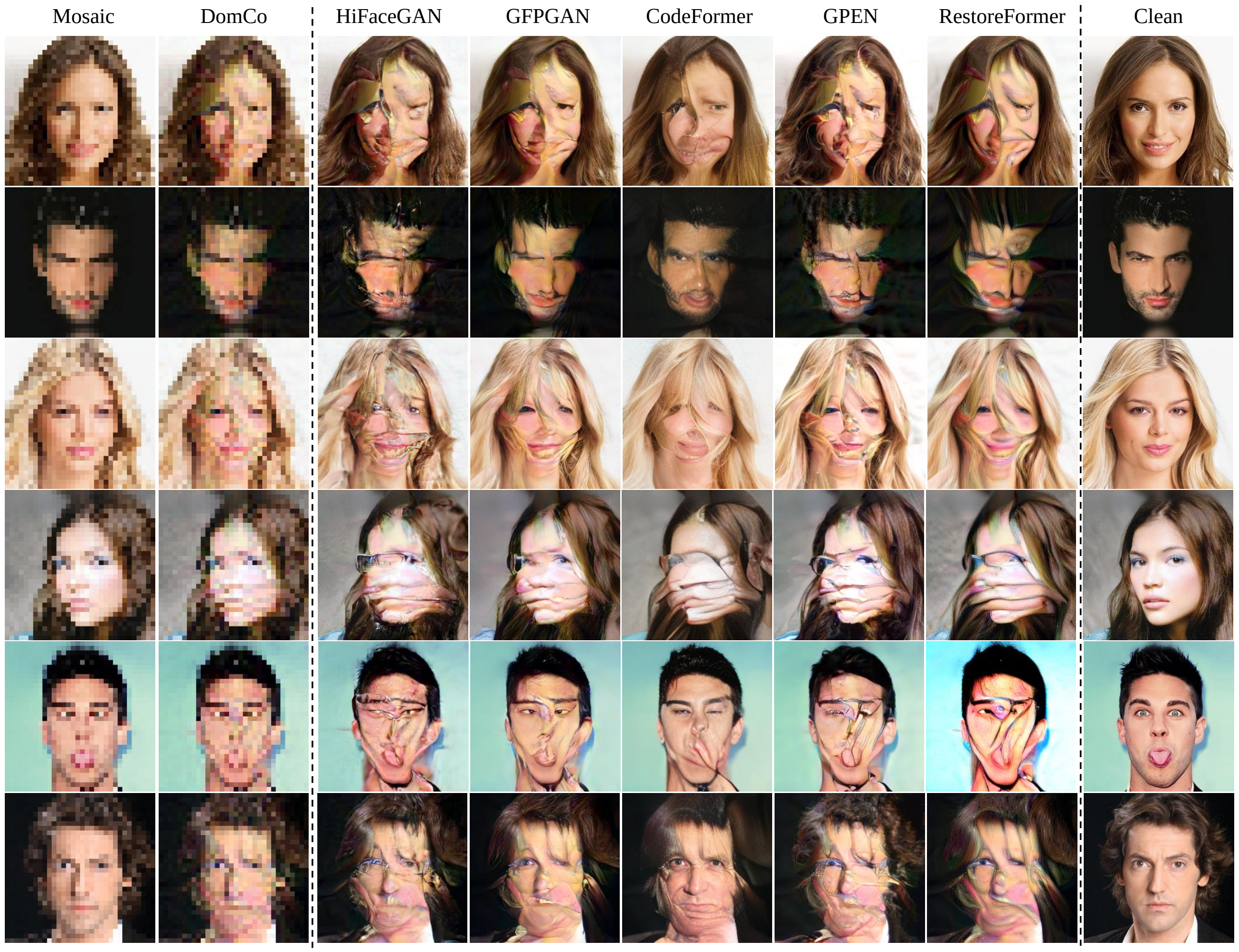


Figure 5. Qualitative results of obfuscated face images restored by different BFR models.

the classification task, since the domain transformation $\mathcal{T}$ is unknown, we can fit the $\mathcal{T}$ by an additional neural network G. The input of G is the discriminative output of the model, and $\mathcal{Z}$ can be an output probability vector, a logits vector, or an intermediate representation. It is easy to see that G can be considered as a generator, at which point perturbations are projected onto low-dimensional manifolds, i.e., GAN-based attacks. This may explain why GAN-based adversarial samples are more transferable than conventional methods (Bar Tal et al., 2023). We hope to provide insights to the future design of generalized attack and defense methods for diverse tasks.

C.2. Limitation

To improve the transferability of adversarial obfuscated images, the DomCo approach compromises the imperceptibility of perturbations. The perceptibility of the perturbation is influenced by many factors especially the choice of optimization method and the loss function used. Domain-consistent adversarial obfuscation methods modulate the intermediate perturbations during iterations, requiring a larger perturbation budget. As mentioned in GLAZE (Shan et al., 2023), the image-to-image generative model needs to be reconstructed with features extracted from the input image, which will retain more information from the original image, requiring a larger perturbation budget. For image-to-image-based models, the perceptibility of the perturbation can be reduced by utilizing features that are salient to the attack target, or features of a clean image, for instance, by setting the loss function to target the attack with a clean image target.

For methods of proactive defense, various countermeasures may emerge in the future. Theoretically, the generated adversarial obfuscated face image patterns can be defeated to some extent by pairwise or adversarial training. However, attack and defense is a continual game, where stronger attacks motivate the development of more robust defenses. Studying obfuscated face privacy protection in complex scenarios is also part of our future work. Additionally, there are some limitations inherent in existing countermeasures. Adversarial training can decrease performance on clean images and incurs high training costs. Moreover, diffusion-based methods require a careful equilibrium between fidelity and the realism of the outputs. Notably, maintaining high fidelity may inadvertently conceal identity information, challenging the objective of identity preservation.

C.3. Discussion with Stronger Degenerate Operators

A stronger degenerate function can somewhat prevent restoration, however it will result in poorer visual quality. Table 5 compares the image quality of obfuscated face images generated by DomCo with those generated by stronger degenerate operators. The reference image is the image processed by a 16x16 scale operator. It can be observed that the images generated by the DomCo operator outperform the others in all metrics, demonstrating superior visual quality. Moreover, degenerate operators of different strengths can be combined with DomCo to further enhance privacy.

Table 5. Comparison of obfuscated face image quality.

Obfuscation	FID	PSNR	MS-SSIM	LPIPS
20×	37.00	23.56	0.8108	0.1187
24×	45.80	22.57	0.7770	0.1396
28×	85.94	22.23	0.7527	0.1481
32×	116.66	22.53	0.7567	0.1282
DomCo	**24.42**	**30.41**	**0.9287**	**0.0873**