南开大学

Nankai University

南　　开　　大　　学

计　算　机　学　院

数据安全

论文阅读心得

孟启轩 2212452

年级：2022 级

专业：计算机科学与技术

指导教师：刘哲理

2025 年 6 月 29 日

目录

一、 论文简介

论文题目

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

作者与会议

Kui Zhang, Hang Zhou, Jie Zhang, Wenbo Zhou, Weiming Zhang, Nenghai Yu
发表于 **ICML 2024（第 41 届国际机器学习大会）**

二、 研究动机

随着深度学习与计算机视觉技术的迅猛发展，人脸识别、图像增强与复原在社交媒体、安防监控、身份验证等领域获得广泛应用。但与此同时，个人隐私泄露问题也逐渐凸显，尤其是在公共场景下的人脸信息被非法识别和还原，构成了严重的安全隐患。

为了保护用户隐私，传统的人脸匿名方法如像素化（Pixelation）、模糊处理（Blurring）、遮挡（Masking）等被广泛应用。这些方法在人眼视觉层面看似有效，但近年来的研究表明，**盲人脸恢复（Blind Face Restoration, BFR）**模型对这些退化图像具有极强的还原能力。BFR 模型基于生成对抗网络（GAN）等深度结构，可在无原始图像参考的前提下，重建出结构、纹理与身份高度相似的人脸图像。

作者实验证明，即使是使用低质量的像素化图像进行训练，BFR 模型也能学习到退化与还原之间的强关联性，从而准确还原出接近真实的人脸图像。这意味着传统匿名手段已无法保障用户隐私安全。更为严峻的是，现实中攻击者通常使用黑盒模型，即用户无法得知其结构与参数，导致现有基于模型知识设计的防御手段难以应对。

因此，本文提出一种**具备迁移能力的主动对抗性防御方法**，通过添加领域一致的对抗扰动，**在不改变人眼可感知质量的前提下，有效防御多种未知 BFR 模型对匿名图像的还原攻击**，从根源上提升图像隐私保护的稳健性与实用性。

三、 主要贡献

本文围绕应对盲人脸恢复带来的隐私威胁，提出了具有高度迁移性的新型防御机制，主要贡献如下：

- **提出可迁移的对抗性人脸匿名方法：**本文设计了一种通用的人脸对抗扰动生成方法 **DomCo（Domain-Consistent Adversarial Obfuscation）**，可在**黑盒场景下实现跨模型防御**，有效防止不同结构的 BFR 模型还原匿名人脸。
- **引入领域一致性扰动理论：**作者观察到不同 BFR 模型本质上都在学习从退化图像域到高质量图像域的映射，即近似于某种退化函数的逆。基于此，DomCo 在扰动优化过程中保持**扰动在退化图像域的投影一致性**，从而提升在不同模型上的泛化能力。
- **提出结构感知损失函数：**传统感知损失侧重像素或特征差异，而本文进一步引入人脸解析模型，定义**结构混乱度（Structure Loss）**作为目标函数，有效破坏恢复后图像的面部结构一致性，达到更强的“去身份化”效果。

- **增强扰动迁移性的训练机制：**本文设计**领域一致性数据增强策略**，在扰动生成阶段增加退化函数一致性模拟，提高扰动对不同模型的通用性与稳定性。
- **丰富的实验验证：**作者在多个主流 BFR 模型（如 HiFaceGAN、GPEN、GFPGAN、CodeFormer、RestoreFormer）上验证方法，实验涵盖 ID 相似度（IDS）、感知距离（LPIPS）、图像质量（FID）、结构混乱度（LMD）等多维指标，显示出 DomCo 在保持图像质量的同时，显著提升了防御性能。
- **首次系统分析 BFR 模型带来的隐私风险：**本文是首批深入探讨 BFR 模型可能造成匿名失效的研究之一，为图像隐私保护技术的发展提供了新的理论视角与实践路径。

四、 方法概述

本文提出了一种面向盲人脸恢复攻击的可迁移人脸隐私保护方法——**DomCo**。该方法旨在解决传统像素化等人脸匿名手段无法有效抵御新兴 BFR 模型还原的问题，提出一种在不依赖目标模型结构与参数的前提下，能够**泛化于不同模型**的对抗性人脸扰动生成方法。

DomCo 的核心思想是：通过构造**域一致的对抗扰动**，将扰动限制在图像退化空间中，从而提升对抗扰动的通用性与迁移性。此外，为进一步提升跨模型鲁棒性，DomCo 在训练中引入了**域一致数据增强策略**，并设计了多种结合感知与结构信息的**损失函数**，以强化对扰动图像的身份破坏能力。

(一) 任务建模

设原始高清人脸图像为 z，退化函数 $T(\cdot)$ 表示像素化处理等匿名化操作，生成的模糊图像为 $x = T(z)$。攻击者使用 BFR 模型 $R(\cdot)$ 尝试从 x 恢复出高清图像 $\hat{z} = R(x)$。防御者的目标是在保持图像感知质量的前提下，通过对抗性扰动 δ 构造匿名图像 $x_a = x + \delta$，使得恢复图像 $R(x_a)$ 与原始图像 z 差异显著，降低身份信息泄露风险。

该过程建模为如下约束优化问题：

$$\max_{\delta} \mathcal{L}(z, R(x+\delta)) \quad \text{s.t.} \quad S(x, x+\delta) \leq \epsilon \tag{1}$$

其中 $\mathcal{L}$ 为身份相似性损失函数，用于衡量恢复结果与真实人脸的相似度，$S(\cdot)$ 为扰动感知距离度量，ϵ 控制扰动强度，确保生成图像在视觉上与原图接近。

(二) 域一致扰动（DomCo）

DomCo 的关键思想在于避免传统扰动方法对 surrogate 模型（代理模型）的过拟合，提升扰动的跨模型迁移能力。为此，作者引入**域一致扰动**的概念：扰动 δ 应当来源于高清图像域 Z 中扰动后的图像 $z + \delta_z$ 经退化函数 $T(\cdot)$ 映射所致，即满足：

$$\exists \delta_z \text{ 使得} T(z+\delta_z) - T(z) = \delta \tag{2}$$

换言之，扰动应位于退化空间的自然图像流形上，确保 $x_a = T(z+\delta_z)$，从而避免扰动仅针对某一具体模型结构进行拟合。这种策略通过梯度传导路径的调整，有效将扰动对准于多个 BFR 模型间的**共性特征**，提高泛化能力。

作者进一步将扰动生成过程转化为对原图 z 的扰动优化，即：

$$\max_{\delta_z} \mathcal{L}(z, R(T(z+\delta_z))) \quad \text{s.t.} \quad S(T(z), T(z+\delta_z)) \leq \epsilon \tag{3}$$

该变换允许直接在高清图像域进行优化，再通过退化函数 T 映射至像素化空间，保证扰动的域一致性。

(三) 损失函数设计

为提升扰动的攻击效果与人脸结构破坏能力，DomCo 综合使用了以下三类损失函数：

- **内容损失** L_c：基于 L1 范数度量恢复图像与真实图像在像素空间的差异，强调整体图像重建质量差异。
- **感知损失** L_{percep}：基于 VGG 网络中间特征的距离，使用 LPIPS 或 VGG 感知差异指标，更贴近人眼感知，强调结构和纹理破坏。
- **结构损失** L_{struc}：利用预训练人脸分割网络（Face Parsing Network），输出每个像素的部位概率图，将扰动引导至破坏人脸结构的区域。定义如下：

$$L_{\text{struc}} = -\frac{1}{HW} \sum_{h,w,c} \hat{y}_{hwc} \cdot P_{hwc}(z+\delta_z) \cdot M_{hw} \tag{4}$$

其中 $\hat{y}$ 为原图结构标签的 one-hot 向量，$P(z+\delta_z)$ 为扰动图像的结构预测概率，M 为人脸区域掩码。

最终的总损失为多项加权组合，其中结构损失对身份相关结构扰动具有显著贡献。

(四) 数据增强与迁移性提升

考虑到扰动训练容易过拟合于 surrogate 模型，DomCo 引入了**域一致数据增强（Domain-Consistent Data Augmentation）**策略：在高质量图像 z 上施加轻微扰动（如高斯噪声、模糊等）形成 $A(z)$，再通过退化函数生成对应的训练样本 $x = T(A(z))$。训练目标变为：

$$\min \mathcal{L}_D \left(R(T(A(z))), z \right) \tag{5}$$

此策略使得 surrogate 模型在优化过程中暴露于更加广泛的输入扰动分布，提高其对迁移扰动的敏感性，从而间接提升 DomCo 的生成扰动在未知 BFR 模型上的有效性。

(五) 总结

总体而言，DomCo 的创新在于将“域一致性”理念引入对抗扰动构造，结合结构感知损失与域增强训练，大幅提高了在人脸隐私保护中的实用性和跨模型鲁棒性，为对抗盲人脸恢复提供了可行而通用的解决方案。

五、 实验结果

作者在多个主流的盲脸恢复模型上对所提出的对抗性人脸匿名技术进行实验验证，涵盖了 HiFaceGAN、GFPGAN、GPEN、CodeFormer 以及 RestoreFormer 等模型。图 1 展示了不同方法在上述五个模型下的 FID 指标对比结果，结果表明：所提出的 DomCo 方法在所有模型中

均获得最高的 FID 分数，显著优于传统像素化方法和经典对抗扰动方法，说明其在防御人脸还原任务中具备更强的有效性。

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

图 1: 不同方法（Pixelate、PPGD、DomCo）在五个 BFR 模型下的 FID 指标对比（FID 越高表示恢复越差，隐私保护效果越好）

在结构相似度指标 MS-SSIM 与感知图像距离指标 LPIPS 上，DomCo 同样表现出更强的扰乱能力。具体而言，DomCo 能有效降低恢复图像与原始图像的结构一致性，同时在视觉感知层面引入显著的差异，从而实现更高水平的匿名化保护。

进一步地，在身份相似度指标 IDS 上，DomCo 显著降低了恢复图像与真实身份之间的特征相似度，有效抑制了人脸识别模型对恢复图像的判别能力。此外，LMD（人脸关键点平均距离）、mPR（像素级召回率）以及 mIoU（平均交并比）等评估指标也显示，DomCo 能够极大地破坏恢复图像的面部结构一致性，使恢复图像在语义层面失真。

从跨模型迁移的角度看，DomCo 在黑盒场景下对其他未知 BFR 模型同样展现出优秀的通用性。相比于仅对特定模型有效的扰动，DomCo 能在多个恢复模型上保持稳定且较强的隐私防御能力，具有良好的模型不可知适应性。

在定性实验中，恢复图像展示出显著的“去身份化”特征，表现为五官混乱、结构错位等现象。尤其在人眼无法察觉明显扰动的前提下，DomCo 能成功干扰恢复模型，使输出图像的身份信息与原始图像完全脱钩，验证了其在实际应用中兼具鲁棒性与实用性。

最后，作者还通过消融实验探讨了不同损失函数、扰动约束和数据增强方式对匿名效果的影响，进一步证明了域一致性设计在提升扰动迁移性和恢复破坏能力方面的重要作用。

六、 个人思考与未来设想

(一) 方法局限与改进方向

尽管 DomCo 方法在提升对抗性扰动的迁移能力方面展现了显著优势，但在实际应用中仍存在若干值得关注的局限性：

- **扰动感知性较强**：DomCo 依赖较大的扰动幅度（ϵ）以增强跨模型的防御效果，虽然在黑盒攻击场景下取得了较好结果，但其产生的图像可能引起人眼感知上的不适或视觉异样，降低了匿名图像在现实社交网络、媒体发布等场景中的可接受性。

- **对退化模型 T 依赖较强**：域一致扰动的核心在于模拟高质量图像退化至模糊图像的过程，但实际中攻击者可能采用与设计不一致的退化流程，甚至复合多种未知噪声源，此时 DomCo 的鲁棒性可能受到削弱，对退化映射 T 的建模假设仍存在理想化风险。
- **扰动生成计算成本高**：当前 DomCo 使用 PPGD 等优化方法生成扰动图像，每张图像需迭代计算梯度，对处理大规模图像或实时匿名化场景（如视频会议、直播）存在计算开销高、不易部署的问题。
- **缺乏攻击对抗演化机制**：尽管 DomCo 针对现有 BFR 模型展示了优越防御效果，但其自身扰动模式可能被未来攻击者识别并规避，尚未设计动态演化或自适应防御策略，存在对抗“过拟合”的风险。

(二) 未来工作思路

DomCo 提出的“领域一致性扰动”机制，为图像隐私保护开辟了新路径。在此基础上，可探索以下方向以拓展其理论深度与实用价值：

- **多模态隐私保护拓展**：当前方法聚焦于静态图像的人脸恢复场景，未来可进一步扩展至视频匿名化（跨帧扰动一致性）、语音掩码（面向语音识别模型的对抗性掩蔽）、文本去身份化（文本生成模型的身份抹除）等多模态输入空间，实现更全面的个人隐私保护。
- **轻量化扰动生成器设计**：引入生成式扰动网络（如轻量级 U-Net 或 Transformer 结构），替代迭代优化过程，实现扰动的一次性前向生成，进而支持边缘设备部署，适应实时处理场景。
- **视觉质量与隐私保护协同优化**：结合生成模型（如 StyleGAN、Diffusion Model）提升防御图像的结构合理性与自然度，同时引入感知对抗损失，在保留扰动效果的同时提升图像的主观视觉质量。
- **用户可控匿名化机制**：开发允许用户自主设定匿名区域（如仅遮挡眼部、嘴部等）的可交互匿名工具，并结合人脸部件分割技术，实现灵活、定制化的匿名策略，提升算法在人机交互场景中的适应能力。
- **动态对抗博弈机制**：引入生成对抗网络（GAN）思路构建扰动与恢复模型之间的长期博弈机制，模拟攻击与防御之间的迭代演化，提升方法在未来更强大攻击者面前的可持续性与适应性。

七、 总结

本文围绕日益严重的盲脸恢复隐私泄露问题，从攻击者可通过图像还原模型重构被像素化人脸入手，系统性地分析了现有匿名化技术的脆弱性，并提出了具备跨模型迁移性的主动防御方法 DomCo。该方法基于“领域一致性”理论构建对抗性扰动，不仅有效扰乱主流 BFR 模型的重建效果，还在不依赖攻击模型参数的黑盒场景下展现了优良的迁移性能。作者从理论动机、优化方法、实验评估等多个维度系统论证了 DomCo 的有效性，为图像隐私保护领域提供了全新视角与技术路线。从研究范式来看，DomCo 将对抗学习与图像退化建模深度融合，展示出“主动匿名”防御思想的强大潜力。这一工作不仅具有较高的工程实用性，也为未来多模态安全、对抗泛化、用户可控隐私等关键议题奠定了理论基础，值得在后续研究中持续跟进与拓展。