

基于多尺度频率分解与元学习的人脸识别有目标攻击算法

蔡骏¹, 黄添强^{1,2}, 郑翱鲲¹, 叶锋^{1,2}, 徐超^{1,2}

(1. 福建师范大学计算机与网络空间安全学院, 福建 福州 350117; 2. 数字福建大数据安全技术研究所, 福建 福州 350117)

摘要: 随着人脸识别技术的日益普及, 人们对个人隐私泄露的担忧逐渐加剧。尽管近年来一些研究尝试通过生成对抗样本来保护照片隐私, 防止未经授权的人脸识别系统识别, 但这类方法常因攻击成功率低和可转移性弱而受限。针对这一问题, 提出了一种基于多尺度频率分解与元学习的人脸识别有目标攻击算法。首先设计了一个多尺度频率分解模块。该模块将目标人脸图像精细划分为不同频率成分的频带, 进而将频率信息与空间信息相融合, 从而提取目标人脸图像的完整信息。随后构建了以循环生成对抗网络 (cycle-consistent generative adversarial networks, CycleGAN) 作为主体、妆容转移为核心的对抗攻击算法。该算法使用源人脸图像生成高质量的妆容, 在这一过程中, 新增的元学习攻击模块负责计算损失函数并更新参数。元学习攻击模块可以对妆容进行精细化处理, 巧妙地将目标人脸的特征融入妆容中, 以生成具有强大对抗性的妆容图像, 从而实现对特定目标的有目标攻击。元学习攻击模块解决了以往针对白盒模型集成攻击时因过拟合而出现的泛化问题, 使得生成的对抗样本具有更强的攻击性和泛化能力。通过对不同攻击策略的实验效果进行分析, 该研究发现多尺度频率分解与元学习相结合的方法能大幅提高攻击人脸识别系统的成功率和鲁棒性。

关键词: 人脸识别有目标攻击算法; 多尺度频率分解; 循环生成对抗网络; 元学习

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2025001

Facial recognition targeted attack algorithm based on multiscale frequency decomposition and meta-learning

CAI Jun¹, HUANG Tianqiang^{1,2}, ZHENG Aokun¹, YE Feng^{1,2}, XU Chao^{1,2}

1. College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China

收稿日期: 2024-06-05; 修回日期: 2024-09-14

通信作者: 黄添强, fjhtq@fjnu.edu.cn

基金项目: 国家自然科学基金 (62072106); 福建省科技创新平台项目 (2023-P-003); 福建省自然科学基金 (2022J01188); 福建省教育厅中青年教师教育科研项目 (JAT210051); 福建省自然科学基金 (福建省科技厅校企合作项目) (2022J01190)

Foundation Items: The National Natural Science Foundation of China (62072106), Fujian Province Science and Technology Innovation Platform Project(2023-P-003), The Natural Science Foundation of Fujian Province (2022J01188), Fujian Provincial Department of Education Young and Middle-aged Teachers Education Research Project (JAT210051), The Natural Science Foundation of Fujian Province (Fujian Provincial Science and Technology Department School-Enterprise Cooperation Project) (2022J01190)

引用格式: 蔡骏, 黄添强, 郑翱鲲, 等. 基于多尺度频率分解与元学习的人脸识别有目标攻击算法[J]. 网络与信息安全学报, 2025, 11(1): 129-140.

Citation Format: CAI J, HUANG T Q, ZHENG A K, et al. Facial recognition targeted attack algorithm based on multiscale frequency decomposition and meta-learning[J]. Chinese Journal of Network and Information Security, 2025, 11(1): 129-140.

2. Digital Fujian Institute of Big Data Security Technology, Fuzhou 350117, China

Abstract: As the prevalence of facial recognition technology continued to grow, concerns about personal privacy breaches were also gradually intensifying. Despite recent studies attempting to safeguard photo privacy by generating adversarial examples to prevent unauthorized facial recognition systems from identifying individuals, these methods were often constrained by low attack success rates and weak transferability. To address this issue, a facial recognition targeted attack algorithm based on multiscale frequency decomposition and meta-learning was proposed. This algorithm initially devised a multiscale frequency decomposition module that meticulously partitioned the target facial image into frequency bands comprising distinct components. This module integrated frequency information with spatial information, thereby enabling the extraction of comprehensive information from the target facial image. Subsequently, a cycle-consistent generative adversarial networks (CycleGAN)-based adversarial attack algorithm with makeup transfer as its core was constructed. This algorithm employed a source facial image to generate high-quality makeup. During this process, a meta-learning attack module was introduced to calculate the loss function and update parameters. The meta-learning attack module enabled precise makeup processing, by which the target face's features were embedded into the makeup to create adversarial makeup images. This enabled the targeted attacks that were the focus of this study. The meta-learning attack module addressed the overfitting and generalization issues present in previous white-box model ensemble attacks, thereby enhancing the efficacy and generalization ability of the generated adversarial examples. The results of experimental analysis of different attack strategies demonstrate that the combination of multiscale frequency decomposition and meta-learning significantly enhances the success rate and robustness of attacks on facial recognition systems.

Keywords: facial recognition targeted attack algorithm, multiscale frequency decomposition, CycleGAN, meta-learning

0 引言

人脸识别 (facial recognition, FR) 技术作为现代科技的一大突破, 不仅极大提高了身份识别的准确性和效率, 在安全监控、公共服务以及城市智慧化建设等领域发挥着举足轻重的作用^[1]。然而, 随着人脸识别技术的广泛运用, 其安全性问题逐渐浮出水面, 引发了社会各界的关注。在实际应用中, 人脸识别系统给广大民众隐私保护带来的挑战日益严峻, 过度的人脸监控更是对个人隐私权的严重侵犯。因此, 迫切需要一种高效且切实可行的解决方案, 以保护个人面部图像不被未经授权的人脸识别系统滥用, 确保个人隐私得到充分保护。

Chen 等^[2]提出了一种对恶意的人脸识别模型的训练数据集实施数据投毒攻击的面部隐私保护方案。然而, 这种方案有局限性, 因为它要求攻击者能够成功地将被投毒的人脸图像注入数据集中。一旦目标模型完成训练或在清洁的数据集上重新运行, 这些投毒图像可能变得无效, 无法持续干扰人脸识别系统的正常运行。

Ren 等^[3]提出了一种对抗性攻击保护方案, 其核心在于生成对抗性的人脸图像, 让人脸识别系统将源人脸图像 (用户的原始人脸图像) 错误地识别为其他个体或指定目标的人脸图像^[3], 从而有效保护用户隐私。与投毒攻击方案相比, 对抗性攻击方案无须向数据集中注入新的图像, 仅需对用户自己的图像进行微调即可实现隐私保护。这种方案不但更隐蔽, 而且能够在不改变系统训练数据集的情况下实施, 具有更高的实用性。然而, 该方案也存在不足之处。当对抗样本对人脸识别系统进行有目标的攻击时, 攻击者需要提取目标人脸特征, 以生成将源人脸图像识别为目标人脸的对抗样本。当前的目标人脸特征提取方法多局限于空域, 限制了攻击效果和隐蔽性^[4]。此外, 现有的人脸识别系统的对抗攻击算法在训练过程中常采用集成攻击算法来计算损失函数并更新参数。在这个过程中, 集成攻击算法会攻击多个白盒模型, 并将这些模型输出的损失进行集成, 进而生成对抗样本^[5]。但针对白盒模型集成攻击时会出现过拟合, 进而产生泛化问题,

导致生成的对抗扰动仅在相似模型上有效,对结构差异较大或结构未知的黑盒模型可能失效。

针对目标人脸图像特征提取的局限以及集成攻击算法导致的过拟合问题,本文深入研究并提出创新方案。本文设计了一种多尺度频率分解(multiscale frequency decomposition, MFD)模块,可以将目标人脸图像精准地分解为不同频率分量,针对性提取核心特征,以生成更具威胁性的对抗样本。同时,将元学习算法引入对抗性攻击算法中,通过攻击多个模型并综合其梯度信息更新生成器参数,学习通用攻击策略,以减少对特定模型的依赖,降低过拟合风险,显著提高模型的泛化能力。

本文的主要贡献如下。

1) 设计了一个提取目标人脸图像特征的多尺度频率分解模块,通过对目标人脸图像进行多尺度频率分解得到频率域特征,并与空间域特征结合,从而提取目标人脸图像的完整信息。

2) 构建了以循环生成对抗网络(CycleGAN)为主体框架、妆容转移为核心的对抗攻击算法。该算法可以使源人脸图像生成自然的妆容,将妆容作为扰动的媒介,实现对抗攻击。

3) 提出了利用元学习攻击模块来计算损失函数并更新参数的算法。该算法可以对妆容进行精细化的处理,将目标人脸的特征融入妆容中。元学习攻击算法解决了以往针对白盒模型集成攻击时因过拟合而出现的泛化问题,可以生成攻击性更强、更具有泛化能力的对抗性妆容。

1 相关工作

近年来,人脸识别对抗攻击领域的研究工作取得了丰硕的成果,研究者不断探索并提出了多种攻击方法。根据攻击方式的不同,人脸识别系统的对抗攻击可分为数字域攻击和物理域攻击两大类,接下来将分别介绍这两种攻击类型的研究进展。

1.1 数字域攻击

数字域攻击主要聚焦于利用人脸识别模型的算法缺陷,通常添加难以被人类察觉的微小扰动

来误导人脸识别模型的判断。

Goodfellow等^[6]提出的FGSM(fast gradient sign method)是基本的对抗攻击方法,利用模型的梯度信息来指导输入图像中每个像素点强度的调整,以生成能够误导模型分类的对抗性样本。该方法虽然简单直观,但仅基于单次梯度计算生成的对抗样本的攻击性很有限。为改进此不足,Kurakin等^[7]提出了迭代快速梯度符号法(iterative fast gradient sign method, IFGSM),IFGSM通过多次迭代应用FGSM策略对输入图像进行逐步微调,以找到更具攻击性的对抗样本。Madry等^[8]提出的投影梯度下降(projected gradient descent, PGD)法使用精心设计的扰动来构造对抗样本,同时确保扰动在一定范围内。PGD通过迭代更新并投影扰动,从而能够在复杂场景中保持高效攻击。

Carlini和Wagner提出的C&W(Carlini & Wagner)攻击^[9]利用不同范数精细约束扰动,转化为无约束凸优化,提高了攻击成功率,成为当前主流的攻击算法。在此基础上,Baluja等^[10]提出对抗性转换网络(adversarial transformation network, ATN),ATN可以同时优化对抗样本与原始图像的色彩相似度和模型输出置信度,使对抗样本更自然、隐蔽,从而进一步提高了攻击成功率和隐蔽性。

1.2 物理域攻击

物理域攻击更全面地考虑了整个人脸识别系统,要求对抗扰动能够自摄像头捕获的图像中传递至目标模型,因此必须兼顾扰动的可见性与有效性^[11]。

如图1(a)所示,Sharif等^[12]提出的Adv-glass攻击方法将对抗扰动限制在太阳镜形状区域,实现了可佩戴性,且物理上难以察觉,能轻松欺骗人脸识别系统。如图1(b)所示,Komkov等^[13]针对ArcFace算法的Adv-hat攻击,将对抗补丁置于帽子,通过弯曲和旋转增强物理世界中的鲁棒性,实现显著攻击效果。如图1(c)所示,Li等^[14]提出的SLAttack使用结构光技术,通过投射特定模式的光到人脸上来产生对抗性。

近年来,人脸识别系统通过识别社交媒体

上的公开照片来追踪用户关系,对个人隐私构成严重威胁。为应对这一挑战,研究者开始探索在网上发布带有对抗性妆容的照片来保护个人隐私。Yin 等^[15]提出 Adv-Makeup 攻击方法,首次用妆容来生成对抗性人脸图像,通过生成具有眼影风格的对抗扰动,成功欺骗人脸识别模型。如图 1(d)所示, Hu 等^[16]提出的 AMT-GAN (adversarial makeup transfer GAN) 攻击,通过 GAN 打造自然妆容并藏匿对抗扰动,结合判别器提升视觉效果,在黑盒攻击中展现出强大的攻击能力。本文所提攻击算法也正是通过妆容的方式自然、隐蔽、有效地攻击人脸识别系统。

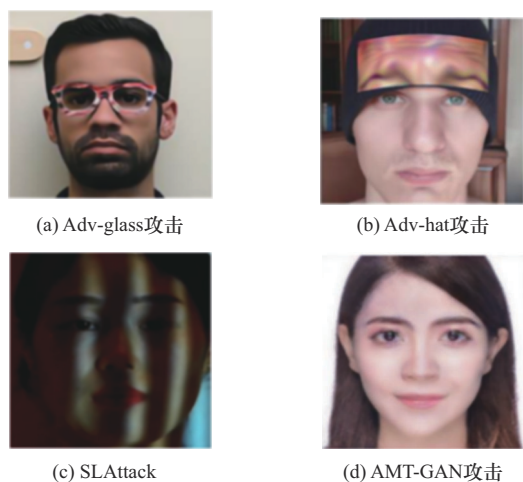


图1 物理域攻击人脸识别算法

Figure 1 Physical domain attack on facial recognition algorithm

2 攻击算法

为了生成FR模型中可攻击转移的对抗性样本,有必要深入研究基于多尺度频率分解与元学习的人脸识别有目标攻击算法。该算法通过更加全面、深入地提取目标人脸图像的特征信息,并以妆容作为媒介,实施元学习攻击。其目标是生成带妆容的源人脸图像,这种图像能够在FR模型中被错误地识别为指定的目标人脸图像。基于多尺度频率分解与元学习的人脸识别有目标攻击算法的框架如图2所示。

2.1 多尺度频率分解

在人脸识别对抗攻击领域,尽管在数据集内的人脸图像攻击上取得显著进展,但在跨数据集测试时,攻击成功率往往大幅下滑。这主要是因为提取的目标人脸特征不够全面。为此,本文设计了多尺度频率分解模块,该模块可将目标人脸图像分解为多个频率分量,从而提取更全面的特征。

频域分析作为图像信号处理的经典方法,被广泛应用于图像与纹理分类^[17]。现有的人脸攻击方法^[18]应用小波变换在频域内实现人脸老化生成。在本文所提方法中,使用了一组频率滤波器来对目标人脸的频率域特征进行提取。首先,使用 N 个手动设计的二进制基滤波器 f_b^i 将频域划分为低频带、中频带和高频带。这些二进制基滤波器的目标是确保从

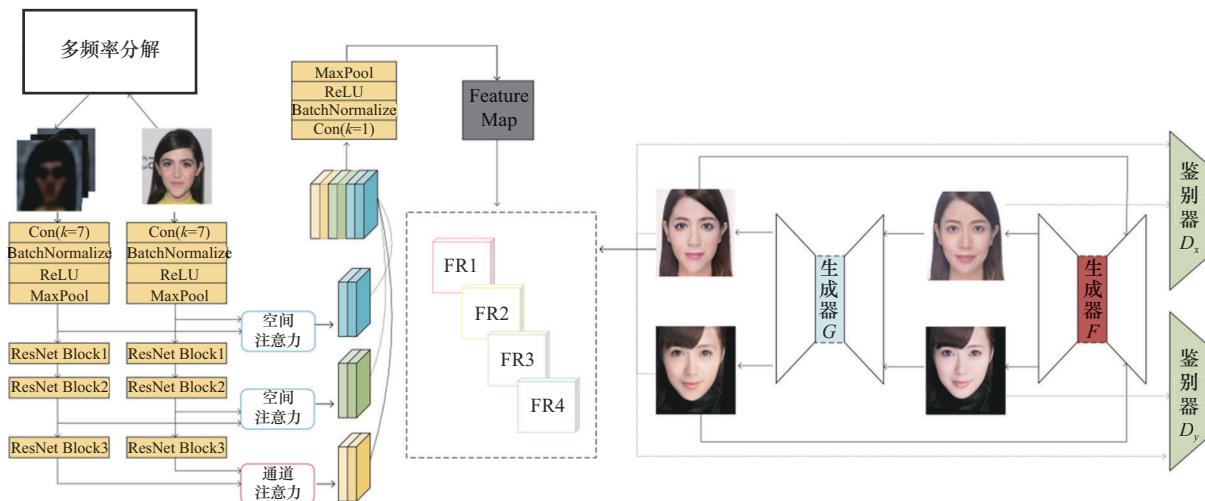


图2 基于多尺度频率分解与元学习的人脸识别有目标攻击算法的框架

Figure 2 Framework of facial recognition targeted attack algorithm based on multiscale frequency decomposition and meta-learning

低频到高频的 N 个频带内的频谱大致相等划分^[19]，二进制基滤波器得到的频域分量 C 的计算式如下。

$$C_i = D^{-1} \{ D(T) \odot f_b^i \}, i = \{ 1 \leq i \leq N \} \quad (1)$$

D 是离散余弦变换； D^{-1} 是离散余弦逆变换； $\odot$ 是点积，代表两个矩阵对应元素相乘。在本文的具体应用中，将 N 设置为 4，从而得到了低、中、高 3 个频带以及一个互补的全频带。具体来说，滤波器设置选择了 3 个波段：① 低频带基底是整个频谱的前 1/16；② 中频带基底在整个频谱的 1/16 和 1/8 之间；③ 高频带基底在整个频谱的 1/8 和 7/8 之间。然而，这种频率的划分可能不足以分析出目标人脸的所有细微线索。为此，添加了一个额外的滤波器，其中频带涵盖了整个频谱。通过上述的多尺度频率分解，目标人脸图像被分解为 4 个分量，即 C_1 、 C_2 、 C_3 、 C_4 ，并沿着通道轴进行堆叠。如图 3 所示，从左往右，第 1 张为目标人脸图像，第 2 张到第 5 张为多尺度频率分解的频域分量。

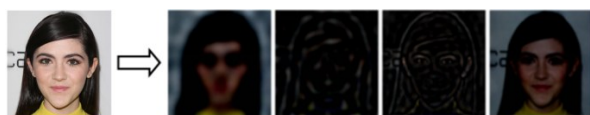


图3 基于多尺度频率分解目标人脸图像

Figure 3 Decomposition of target face image based on multi scale frequency

在频域中，较高频分量强调边缘和纹理，较低频分量与色域空间分布有关。本文采用了卷积块注意力机制（convolutional block attention module, CBAM）^[20]进行特征融合。结合图 2 具体来说，在第 1 个卷积块和第 2 个 residual network（ResNet）块后引入空间注意力模块，以关注边缘和纹理细节；第 3 个 ResNet 块与通道注意力模块共同学习高级语义特征。空间注意力模块与通道注意力模块如图 4 所示，其中卷积核大小分别为 7×7（第 1 空间注意力模块）和 5×5（第 2 空间注意力模块），因为较小的卷积核更适合定位小尺度纹理线索。

最后，特征被融合，以更全面地捕获目标人脸的重要信息，并得到目标人脸图像的特征图 I_t 。

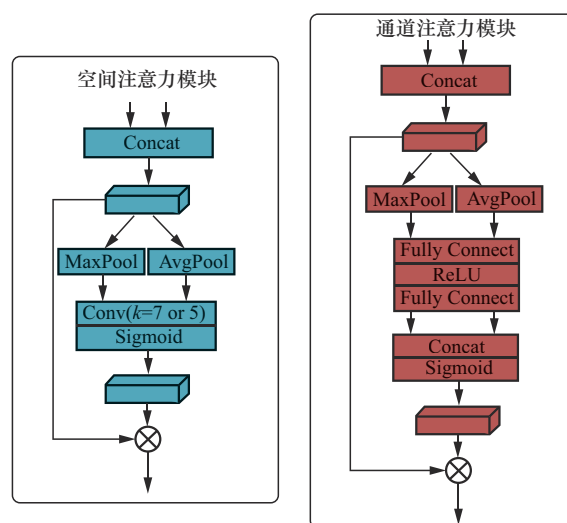


图4 空间注意力模块与通道注意力模块

Figure 4 Spatial attention module and channel attention module

2.2 CycleGAN 化妆转移

提取完目标人脸信息后，接下来需要进行源人脸图像的妆容生成。此过程旨在精准地将已化妆人脸的妆容转移到未化妆的源人脸图像，同时保持其面部特征。近年来，GAN 技术发展迅速，其在化妆迁移方面取得显著进步。其中，PSGAN^[21]专注于处理表情差异时的妆容转移问题，确保特定部位妆容精准迁移。CycleGAN^[22]则专为解决不同图像域间的妆容转换而设计，能够实现高效全脸妆容迁移，更贴合本文对源人脸图像整张脸妆容转移的需求。

本文研究的核心正是基于 CycleGAN 来完成源人脸图像的妆容生成。首先，设定域 X 和域 Y ，分别表示源人脸图像和参考妆容人脸图像的化妆风格域。具体来说， $x \in X$ 表示源人脸图像， $y \in Y$ 表示参考妆容人脸图像。本文的目标是训练一个生成器 G ，它可以将源人脸图像 x 和参考妆容图像 y 作为输入，输出一个与 y 具有相同化妆风格、与 x 具有相同视觉特征的对抗性图像 $G(x,y)$ 。与此同时，另外设计一个生成器 F ， F 负责将参考人脸图像去妆，通过对比 F 生成的图像和源人脸图像，可以更有效地更新 G 的参数。鉴别器 D_X 和 D_Y 分别负责区分源域图像 x 和生成的伪参考域图像 $F(y,x)$ ，以及参考域图像 y 和生成的伪源域图像 $G(x,y)$ 。这两个鉴别器的任务是将由生成器生成的伪图像分布与真实图像分布

区分开来。利用鉴别器 D_X 和 D_Y ，可以评估生成的图像与真实图像之间的差异。生成器和鉴别器的损失函数公式如下。

$$\begin{aligned} L_G &= -\log(D_Y(G(x,y))) - \log(D_X(F(y,x))) \quad (2) \\ L_D &= -\log D_Y(y) - \log(1 - D_X(G(y,x))) \\ &\quad -\log D_X(X) - \log(1 - D_Y(F(y,x))) \quad (3) \end{aligned}$$

在训练过程中，CycleGAN 使用无监督学习方式不断优化生成器性能，以最小化生成的伪造图像与真实图像之间的差异。为了确保图像翻译过程中的循环操作一致性，本文引入了循环一致性损失函数。该损失函数确保图像在完成从 X 到 Y 的映射后再进行从 Y 到 X 的映射，以尽可能地还原为原始人脸图像。这是通过比较原始人脸图像与循环后的图像之间的差异来实现的。循环一致性损失函数表示为

$$L_{\text{cycle}} = \|F(G(x,y),x) - x\|_1 \quad (4)$$

其中， $\|\cdot\|_1$ 表示 $L1$ 范数。循环一致性损失函数的引入，确保了 CycleGAN 在训练过程中能够学习到更准确、更一致的映射关系。

最后，还需考虑当参考图像与源人脸图像为同一图像时，生成器 G 必须保留原始内容及风格等重要信息，这就需要引入身份损失。

$$L_{\text{idt}} = \text{LPIPS}(G(x,x),x) \quad (5)$$

其中，LPIPS 函数^[23]可以衡量两个图像之间的感知相似性，这一损失在维护资源图像的结构信息以及防止人脸属性失真方面起到了至关重要的作用。

利用上面的 CycleGAN 可以实现将参考妆容人脸图像的妆容转移到源人脸图像上，获得一个带有妆容的源人脸图像 $G(x,y)$ 。

2.3 元学习攻击

为了生成具有强大对抗性的妆容图像，本文提出了元学习攻击模块。该模块用于攻击预训练好的 FR 模型，通过计算输出损失并更新参数，以精准地调整源人脸上的妆容细节，从而生成高对抗性的妆容图像。

现有的人脸识别系统的对抗攻击算法在训练时基本使用集成攻击算法计算损失，通过攻击多个已知的白盒模型并集成其输出的损失，生成对抗样本^[24-25]。针对白盒模型集成攻击时因过拟合而容易出现泛化问题，即生成的对抗扰动过于依赖

特定模型，这些扰动可能在结构相似模型上有效，但在差异较大或完全未知的黑盒模型上则可能失效。为了解决这一问题，本文设计了元学习攻击模块，用来优化损失函数的计算和参数更新的过程。

近期，元学习在对抗性攻击与防御领域的应用取得了显著进展，特别是在文献[26]和文献[27]等研究中，元学习被成功应用于对抗性攻击的实践，给予本文研究很大的启发。元学习攻击模块是通过攻击 L 个人脸识别模型，再由得到的损失计算损失函数关于生成器参数的梯度来获得每个模型的梯度信息，对梯度信息进行加权平均，以得到一个综合的梯度信号，然后使用该信号来更新生成器的参数。本文预先训练了 L 个受害者 FR 模型，记作 $F1, F2, \dots, FL$ ，优化后的损失函数公式及参数更新过程公式如下。

$$\begin{aligned} L_{\text{adv}}^i &= 1 - \cos[F_i(I_t), F_i(G(x,y))] \quad (6) \\ \theta_G^{\text{new}} &= \sum_{i=1}^L \omega_i \nabla_{\theta_G^{\text{old}}} \{1 - \cos[F_i(I_t), F_i(G(x,y))]\} \quad (7) \end{aligned}$$

其中， I_t 为目标人脸图像的特征图， $G(x,y)$ 为带扰动源人脸图像， G 为生成带扰动源人脸图像的生成器， ω 为加权平均数。由上可知，元学习攻击中梯度信息是被求和然后用于参数更新，这个加权平均梯度反映了在所有模型上攻击策略的平均改进方向，生成的对抗样本对模型结构的依赖性也较低。通过这个过程训练一个通用的生成器，生成在不同模型上均有效的对抗性扰动，从而有效地解决了针对白盒模型集成攻击时因过拟合而出现的泛化问题。

最后，为了收集妆容生成和攻击阶段所有的信息，本文提出了一种联合优化策略，如式(8)所示。

$$\begin{aligned} \theta_G^{\text{new}} &\leftarrow \theta_G^{\text{old}} \leftarrow \alpha_1 \sum_{i=1}^L \omega_i \nabla_{\theta_G^{\text{old}}} L_{\text{adv}}^i - \beta_1 \nabla_{\theta_G^{\text{old}}} L_G \\ &\quad - \beta_2 \nabla_{\theta_G^{\text{old}}} L_{\text{cycle}} - \beta_3 \nabla_{\theta_G^{\text{old}}} L_{\text{idt}} \quad (8) \end{aligned}$$

其中， α_1 是攻击的平衡权重， β_1 、 β_2 、 β_3 是生成的平衡权重。该策略通过调整攻击的平衡权重和妆容生成的平衡权重，确保两个阶段的优化目标能够同时被考虑并达到平衡。通过调整这些权重，能够在保持妆容生成质量的同时提升攻击的强度和效率。

3 实验

3.1 实验设置

(1) 参数设置

在攻击模型中, 本文精心打造了 G 、 F 、 D_x 和 D_y 的架构, 并运用了一系列超参数来精细调控训练过程。本文将 α_1 、 β_1 、 β_2 、 β_3 的值分别设置为1、0.5、0.5、0.5, 以确保模型性能的最优化。使用ADAM优化器^[28]来进行训练, 学习率为0.000 2, 同时将指数衰减率参数设为(0.9, 0.999)。

(2) 数据集

实验采用化妆转移(makeup transfer, MT)数据集^[29]进行训练, 其中包含1 115张未化妆的人脸图像和2 719张化妆后的人脸图像。为了验证模型的性能, 选用了两个数据集进行测试。其中一个为CelebA-HQ数据集^[30]的子集, 包含10 000张具有不同人脸特征的高质量人脸图像; 另一个为LADN数据集^[31], 它是一个聚焦于化妆的数据集, 含有333张未化妆的图像和302张化妆后的图像。

(3) 目标模型

本文进行了大量实验, 针对4种FR模型(包括IR152^[32]、IRSE50^[33]、FaceNet^[34]和MobilFace^[35])及2种商业人脸识别API(包括Face++和Aliyun)分别进行攻击。

(4) 对比算法的选择

本文挑选了一系列具有代表性的对抗性攻击算法来对比: 经典的PGD^[9]、MI-FGSM (momentum iterative fast gradient sign method)^[23]; 利用对抗性身份掩码有效保护面部隐私的TIP-IM (targeted identity-protection iterative method)^[36]; 与本文研究类似的通过妆容巧妙生成对抗性人脸图像的Adv-Makeup^[15]和AMT-GAN算法^[16]。

(5) 评估指标

本文采用攻击成功率(attacks success rate, ASR)这一核心指标来评估各种攻击方法的性能, 并且设定人脸识别系统在0.01错误接受率的条件下, 攻击方法成功欺骗系统的比例。在针对商业API进行评估时, 直接获取其返回的平均置信度得分作为评估依据。评估生成图像的质量采用的是FID^[37]、PSNR和SSIM^[38]指标。FID反映生成数据与真实数据间的分布差异, FID值越小, 表示生成的数据与真实数据在特征空间中的分布

越接近, 表明生成的图像质量越高。而PSNR通过计算原始图像和生成图像之间的均方误差(MSE)来评估图像的相似度。PSNR值越大, 表示图像间的像素差异越小, 即失真程度越低, 说明图像质量越好。SSIM是一种结构相似度评估指标, 它综合考虑了图像的亮度、对比度和结构信息。SSIM值越接近1, 表示两个图像在结构上的相似度越高, 说明图像质量越好。

3.2 实验内容

(1) Loss曲线

图5所示为本文算法在训练阶段的Loss曲线, 横坐标代表训练的轮次, 纵坐标代表Loss值。其中 G_total_loss 表示对生成器的所有损失, 包括 L_G 、 L_{cycle} 、 L_{idt} 、 L_{adv} 加权平均的总和; D_total_loss 表示对判别器的所有损失。在训练初期, G_total_loss 迅速收敛, 至大约50轮次时, 下降趋势放缓, 基本维持着上下浮动, 到120轮次时, 又开始下降, 大约在165轮次后趋于平稳, 至300轮次仍保持这一稳定状态。而 D_total_loss 一直稳步下降, 在170轮次趋于平稳, 这一状态直至300轮次也保持稳定。这一训练过程标志着模型已实现良好的收敛。

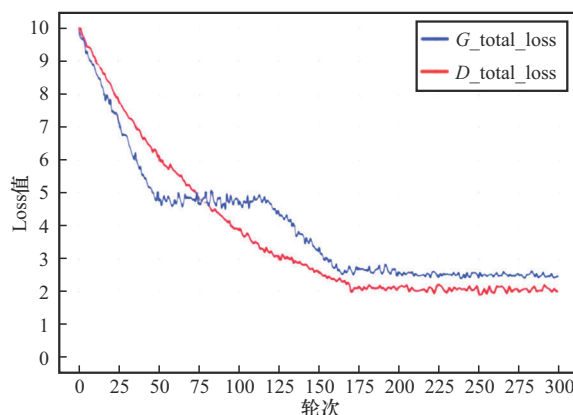


图5 Loss曲线
Figure 5 Loss curve

(2) 时间/空间复杂度

本文主要关注算法的主干网络模型的时间复杂度和空间复杂度, 如表1所示, Params为模型的参数量, 用来评估模型的空间复杂度, FLOP (floating point operation) 为浮点运算量, 用来评估模型的时间复杂度。对于空间复杂度, 本文算

法的 Params 达到了 17.34 MB；而在时间复杂度上，通过输入尺寸 $I \in \mathbf{R}^{3 \times 256 \times 256}$ 计算时间复杂度，本文算法的 FLOP 为 10.02 G（G 表示十亿次操作，即 10^9 次）。相较于传统且易于实现的算法 PGD 和 MI-FGSM，本文算法的空间复杂度和时间复杂度更高，这是因为本文算法运行的计算量和参数量更大。与近 5 年的 3 种算法相比，本文算法在平衡计算效率与性能方面展现出了显著的优势。这两个指标直观地展示了本文算法在数据处理上的高效性，并兼顾了存储空间优化。

表 1 不同算法的空间复杂度和时间复杂度对比
Table 1 Comparison of space complexity and time complexity of different algorithms

攻击算法	Params	FLOP
PGD	4.21 MB	3.96 G
MI-FGSM	8.65 MB	5.88 G
TIP-IM	26.09 MB	19.03 G
Adv-Makeup	25.49 MB	20.73 G
AMT-GAN	20.08 MB	14.46 G
本文算法	17.34 MB	10.02 G

（3）黑盒攻击

将测试图像分为 4 组，以攻击 4 种预先训练好的 FR 模型，每个目标模型都由其他 3 个模型作为训练模型进行攻击。表 2 展示了针对 4 种预训练模型的黑盒攻击的成功率，实验结果表明，AMT-GAN 在 LADN 数据集上针对 MobileFace 模型的攻击表现尤为出色，其攻击成功率超过了本文算法。然而，在其他几种测试场景中，AMT-GAN 攻击成功率相对较低，不如本文算法有效。TIP-IM 针对 FaceNet 模型的攻击成功率非常高，表明该算法对 FaceNet 模型的结构和特性有较深入的理解。这种针对特定模型的优化策略使得 TIP-IM 在攻击 FaceNet 模型时具有显著的优势。尽管 AMT-GAN 在特定场景下表现优异，TIP-IM 针对 FaceNet 模型的攻击也表现出色，但本文算法在大部分测试场景下能保持较高的攻击成功率，证明本文算法具有良好的稳定性和泛化能力。图 6 展示了本文算法针对不同目标人脸图像的黑盒攻击效果。

（4）图像质量

在图像质量评估方面，本文考量的是整张图

表 2 黑盒攻击的成功率
Table 2 Success rate of black box attack

攻击算法	CelebA-HQ				LADN			
	IRE50	IRE152	FaceNet	MobileFace	IRE50	IRE152	FaceNet	MobileFace
Clean	7.29%	3.80%	1.08%	12.68%	2.71%	3.61%	0.60%	5.11%
PGD	36.87%	20.68%	1.85%	43.99%	40.09%	19.59%	3.82%	41.09%
MI-FGSM	45.79%	25.03%	2.58%	45.85%	48.9%	25.57%	6.31%	45.01%
TIP-IM	54.4%	37.23%	40.74%	48.72%	65.89%	43.57%	63.50%	46.48%
Adv-Makeup	21.95%	9.48%	1.37%	22.00%	29.64%	10.03%	0.97%	22.38%
AMT-GAN	76.96%	35.13%	16.62%	50.71%	89.64%	49.12%	32.13%	72.43%
本文算法	79.65%	38.11%	37.46%	52.65%	90.55%	55.12%	34.49%	50.39%



图 6 对不同目标人脸图像的黑盒攻击效果

Figure 6 Black box attack effect on different target face images

像的质量评估。PGD和MI-FGSM在技术上较传统，并不直接关注生成的对抗样本的图像质量；Adv-Makeup仅限于修改眼睛区域的妆容，会导致不公平的比较结果。因此，在实验中未将这几种算法纳入图像质量的对比范围。表3展示了其他几种算法攻击同一目标图像的整张图像的质量评估结果，与AMT-GAN相比，本文算法在定量评估上明显高于AMT-GAN。与TIP-IM相比，本文算法虽然在SSIM指标上稍逊一筹，但在FID和PSNR指标上的表现更加优秀。这要归功于本文的妆容转移方法，它让图像的整体视觉效果更加自然。图7展示了几种算法的视觉质量，可以看出本文算法生成的对抗样本与源人脸图像相似度最高。

表3 图像质量评估
Table 3 Image quality evaluation

算法	FID (↓)	PSNR (↑)	SSIM (↑)
TIP-IM	38.7357	33.2089	0.9214
AMT-GAN	34.4405	19.5045	0.7873
本文算法	28.3213	35.2550	0.8041



图7 视觉质量比较

Figure 7 Visual quality comparison

(5) MFD模块性能测试实验

为了评估MFD模块在提取目标人脸图像的完整信息时的准确率，本文设计了MFD模块性能测试实验。表4为在CelebA-HQ和LADN数据集上4种人脸识别模型的识别准确率，可以看出，增加MFD模块后的人脸识别准确率都有了一定的提高，说明MFD模块提取目标人脸图像

的完整信息的功能是比较强大的，可以更全面地提取人脸图像的关键特征。

(6) 消融实验

表5展示了在CelebA-HQ和LADN数据集上黑盒攻击的成功率，可以看出，在没有MFD模块的情况下，黑盒攻击成功率明显下降。这是因为通过多尺度频率分解，可以从不同的频率域中提取目标人脸图像的特征，可以更好地捕捉图像的局部细节和全局结构。MFD模块对提高黑盒攻击成功率的重要性是显而易见的。

(7) 对比实验

为了评估元学习在人脸攻击中的效果，本文进行了元学习攻击与集成攻击对比实验。实验结果如表5所示，可以看出，元学习的人脸攻击算法优于集成攻击，能够有效解决针对白盒模型集成攻击时因过拟合而出现的泛化问题。这是因为元学习攻击是在多个模型上学习到一个通用的攻击策略，而不是仅仅针对几个模型进行优化。

表5 消融和对比实验结果
Table 5 Ablation and comparative experimental evaluation

算法	攻击成功率	
	FaceNet	MobileFace
本文算法 (w/o Frequency)	32.52%/32.91%	45.19%/34.33%
本文算法 (w/o Meta)	28.58%/29.76%	37.81%/33.07%
本文算法	37.46%/34.49%	52.65%/50.39%

这种通用性有助于减少对模型的过度依赖，从而降低过拟合的风险，大幅提高模型的泛化能力。

(8) 商业API的攻击性能

图8为对Aliyun和Face++两种商业API攻击性能测试结果对比。这些结果是通过收集大量对抗性样例并取其平均置信度得分（百分比）而得出的。数据表明，对于这两种商业API，本文算法在能力上普遍优于其他算法。

表4 MFD模块性能测试结果

Table 4 The result of MFD module performance test

模块	CelebA-HQ				LADN			
	IRE50	IRE152	FaceNet	MobileFace	IRE50	IRE152	FaceNet	MobileFace
Clean	92.71%	96.20%	98.92%	87.32%	97.29%	96.39%	99.40%	94.89%
MFD	95.14%	97.23%	99.10%	90.82%	97.79%	96.85%	97.95%	95.90%

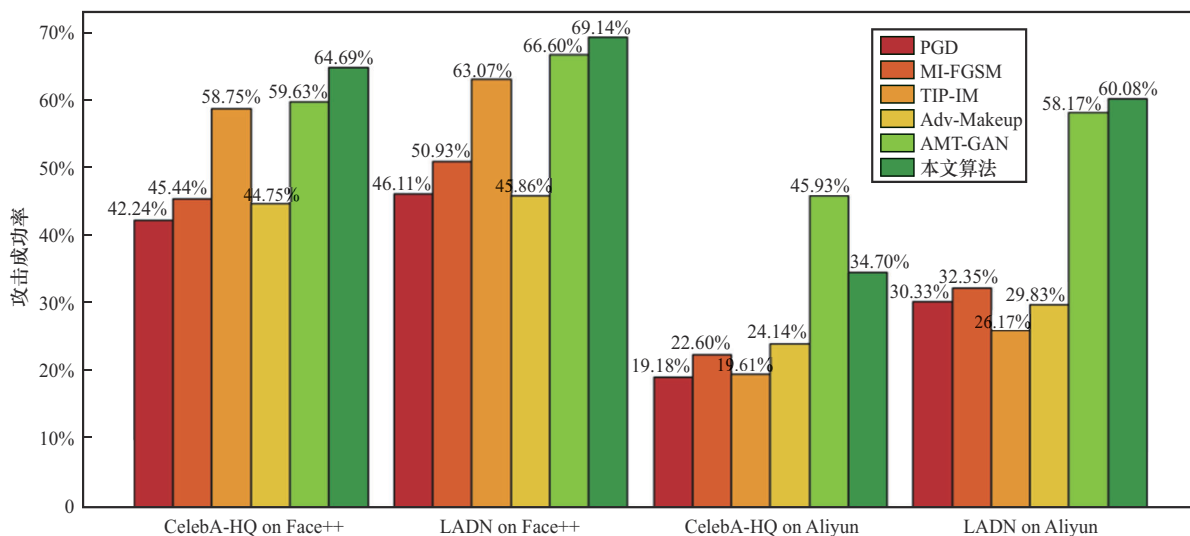


图 8 商用 API 攻击性能测试结果对比

Figure 8 Comparison of commercial API attack performance test results

4 结束语

本文提出了一种基于多尺度频率分解与元学习的人脸识别有目标攻击算法，旨在生成高质量的对抗性妆容。从对比实验结果来看，本文算法能有效解决对目标人脸图像特征提取不足以及针对白盒模型集成攻击时因过拟合而出现的泛化问题。与现有算法相比，本文提出的算法在攻击成功率和鲁棒性方面均表现出优越的性能。本文的研究成果对于增强人脸识别技术的隐私保护具有重要意义。然而，尽管本文在人脸攻击算法方面取得了一定的进展，但仍然需要进一步的研究来提高攻击方法的可转移性，以更好地保护个人隐私和安全。

参考文献：

- [1] 张博林, 朱春陶, 殷琪林, 等. 基于噪声注意力的伪造人脸检测方法[J]. 网络与信息安全学报, 2023,9(4):155-165.
ZHANG B L, ZHU C T, YIN Q L, et al. Noise-attention-based forgery face detection method[J]. Chinese Journal of Network and Information Security, 2023,9(4):155-165.
- [2] CHEN X Y, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv: 1712.05526, 2017.
- [3] REN K, ZHENG T, QIN Z, et al. Adversarial attacks and defenses in deep learning[J]. Engineering, 2020, 6(3): 346-360.
- [4] GUO H, LI J, LIU J, et al. A survey on space-air-ground-sea integrated network security in 6G[J]. IEEE Communications Surveys &

Tutorials, 2021, 24(1): 53-87.

- [5] BHAMBRI S, MUKU S, TULASI A, et al. A survey of black-box adversarial attacks on computer vision models[J]. arXiv preprint arXiv: 1912.01667, 2019.
- [6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412.6572, 2014.
- [7] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world [M]. London: Chapman and Hall/CRC, 2018: 99-112.
- [8] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv: 1706.06083, 2017.
- [9] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). Piscataway:IEEE Press, 2017: 39-57.
- [10] BALUJA S, FISCHER I. Adversarial transformation networks: learning to generate adversarial examples[J]. arXiv preprint arXiv: 1703.09387, 2017.
- [11] 陈晋音, 赵晓明, 郑海斌, 等. 基于光学的物理域对抗攻防综述[J]. 网络与信息安全学报, 2024, 10(02): 1-21.
CHEN J Y, ZHAO X M, ZHENG H B, et al. Survey of optical-based physical domain adversarial attacks and defense[J]. Chinese Journal of Network and Information Security, 2024, 10(02): 1-21.
- [12] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM Sigsac Conference on Computer and Communications Security. New York: ACM Press, 2016: 1528-1540.
- [13] KOMKOV S, PETIUSHKO A. Adv-hat: real-world adversarial attack on ArcFace face ID system[C]// Proceedings of 2020 25th International Conference on Pattern Recognition (ICPR). Piscataway:

- IEEE Press, 2021: 819-826.
- [14] LI Y J, LI Y Q, DAI X L, et al. Physical-world optical adversarial attacks on 3D face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2023: 24699-24708.
- [15] YIN B, WANG W, YAO T, et al. Adv-makeup: a new imperceptible and transferable attack on face recognition[J]. arXiv preprint arXiv:2105.03162, 2021.
- [16] HU S, LIU X, ZHANG Y, et al. Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2022: 15014-15023.
- [17] JI H, YANG X, LING H B, et al. Wavelet domain multifractal analysis for static and dynamic texture classification[J]. IEEE Transactions on Image Processing, 2012, 22(1): 286-299.
- [18] RUSIA M K, SINGH D K. A comprehensive survey on techniques to handle face identity threats: challenges and opportunities[J]. Multimedia Tools and Applications, 2023, 82(2): 1669-1748.
- [19] FANG M, DAMER N, KIRCHBUCHNER F, et al. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway:IEEE Press, 2022: 3722-3731.
- [20] WOO S, PARK J, LEE J Y, et al. Cbam: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.
- [21] JIANG W, LIU S, GAO C, et al. PSGAN: pose and expression robust spatial-aware gan for customizable makeup transfer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2020: 5194-5202.
- [22] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press, 2017: 2223-2232.
- [23] ZHANG R, ISOLA P, EFROS A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2018: 586-595.
- [24] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9185-9193.
- [25] ZHOU L, CUI P, ZHANG X, et al. Adversarial eigen attack on black-box models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2022: 15254-15262.
- [26] SHAO R, LAN X, YUEN P C. Regularized fine-grained meta face anti-spoofing[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11974-11981.
- [27] SHEN M, LI C, YU H, et al. Decision-based query efficient adversarial attack via adaptive boundary learning[J]. IEEE Transactions on Dependable and Secure Computing, 2023.
- [28] KINGMA D P. ADAM: a method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
- [29] LI T, QIAN R, DONG C, et al. BeautyGAN: instance-level facial makeup transfer with deep generative adversarial network[C]//Proceedings of the 26th ACM International Conference on Multimedia. 2018: 645-653.
- [30] KARRAS T, AILA T, LAINE S, et al. Progressive growing of GANs for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.
- [31] GU Q, WANG G, CHIU M T, et al. LADN: local adversarial disentangling network for facial makeup and de-makeup[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE Press, 2019: 10481-10490.
- [32] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [33] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2018: 7132-7141.
- [34] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and cluster-ing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2015: 815-823.
- [35] DENG J, GUO J, XUE N, et al. ArcFace: additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press, 2019: 4690-4699.
- [36] YANG X, DONG Y, PANG T, et al. Towards face encryption by generating adversarial identity masks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2021: 3897-3907.
- [37] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium[J]. arXiv preprint arXiv:1706.08500, 2017.
- [38] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[作者简介]



蔡骏 (2000-), 男, 福建莆田人, 福建师范大学硕士生, 主要研究方向为多媒体内容安全。



黄添强（1971-），男，福建莆田人，博士，福建师范大学教授、博士生导师，主要研究方向为机器学习安全和数字媒体取证。



叶锋（1978-），男，福建福州人，博士，福建师范大学副教授，主要研究方向为计算机视觉和视频图像编码。



郑翱鲲（1999-），男，河南郑州人，福建师范大学硕士生，主要研究方向为多媒体内容安全。



徐超（1981-），男，湖北天门人，福建师范大学讲师，主要研究方向为多媒体内容安全。