Computer Science and Application 计算机科学与应用, 2026, 16(1), 328-336
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161027

Hans汉斯

基于人脸识别的对抗样本生成算法研究

王宇辰，贾召弟

北华航天工业学院计算机学院，河北 廊坊

收稿日期：2025年12月20日；录用日期：2026年1月17日；发布日期：2026年1月26日

摘　要

随着互联网的广泛应用，各种智能化系统的出现极大地提升了人们的生活质量和工作效率。在众多智能化系统中，人脸识别技术的应用最为广泛。人脸识别系统虽然在众多领域发挥了重要作用，但仍面临许多挑战。例如人脸识别系统容易被恶意攻击，亟待需要安全性测试。现有对抗样本存在攻击性弱，视觉效果差问题。本文提出的模型通过StyleNet风格－内容解耦编码、FusionNet多层次风格注入，辅以变形交叉注意力、频域融合和自适应融合金字塔等创新模块，实现了妆容风格从参考人脸到目标人脸的自然迁移。数学推导和实现细节表明，这些模块有效解决了妆容迁移中的局部特征对齐、色彩纹理分离控制、区域平滑融合等难点问题。本文提出的AdversarialMakeup模型在妆容迁移与隐私攻击的核心指标上取得了卓越的综合性能。定量评估表明，该方法在关键指标上达到平均攻击成功率(ASR(avg)) 0.15，同时保持了优异的视觉质量(LPIPS为0.34，SSIM为0.90)和颜色分布一致性(HistDist为0.095)。

关键词

人脸对抗样本生成，生成对抗网络，人脸识别，深度学习

Research on Adversarial Sample Generation Algorithm Based on Face Recognition

Yuchen Wang, Zhaodi Jia

School of Computer Science, North China Institute of Aerospace Engineering, Langfang Hebei

Received: December 20, 2025; accepted: January 17, 2026; published: January 26, 2026

Abstract

With the wide application of the Internet, the emergence of various intelligent systems has greatly improved people's quality of life and work efficiency. Among the many intelligent systems, facial recognition technology is the most widely used. Although facial recognition systems have played an important role in many fields, they still face many challenges. For example, facial recognition systems

文章引用: 王宇辰, 贾召弟. 基于人脸识别的对抗样本生成算法研究[J]. 计算机科学与应用, 2026, 16(1): 328-336.
DOI: 10.12677/csa.2026.161027

are vulnerable to malicious attacks and urgently need security testing. Existing adversarial samples have weak attack capabilities and poor visual effects. The model proposed in this paper achieves natural transfer of makeup style from the reference face to the target face through StyleNet style-content decoupling encoding, FusionNet multi-level style injection, and innovative modules such as deformable cross-attention, frequency domain fusion, and adaptive fusion pyramid. Mathematical derivations and implementation details show that these modules effectively solve the difficult problems in makeup transfer, such as local feature alignment, color texture separation control, and regional smooth fusion. The AdversarialMakeup model proposed in this paper has achieved outstanding comprehensive performance in the core indicators of makeup transfer and privacy attack. Quantitative evaluation shows that this method achieves an average attack success rate (ASR(avg)) of 0.15 on key indicators, while maintaining excellent visual quality (LPIPS of 0.34, SSIM of 0.90) and color distribution consistency (HistDist of 0.095).

Keywords

Face Adversarial Sample Generation, Generative Adversarial Network, Face Recognition, Deep Learning

1. 引言

深度神经网络(Deep-learing Neural Network, DNN)作为计算机视觉的实现功能根基，其使用规模随着计算机视觉实现功能的增多也逐渐扩大，例如人脸识别。但是，深度神经网络在大规模应用的同时，也面临非常大的安全挑战，通过对输入图像添加人眼不可见的细微对抗性扰动，进而干扰目标比对模型输出错误识别结果，称为对抗攻击。对抗攻击又可以分为白盒攻击和黑盒攻击，白盒和黑盒是基于对攻击者知识进行假设而区分出的 2 种主要环境[1]。在进行白盒攻击中，攻击者拥有目标模型的完成信息，包括模型的架构、参数、训练数据等。由于攻击者可以访问模型的所有内部信息，他们可以利用这些信息来设计更有效地对抗性扰动；在进行黑盒攻击中，攻击者没有目标模型的内部信息，只能通过模型的输入和输出来推断模型的行为。

人脸属性是表征人类面部特征的一系列生物特性，是现代安全系统中新兴的软生物识别技术之一。最近，基于生成对抗网络(Generative Adversarial Network, GAN)的方法被用于操纵面部特征图像，如 StarGAN [1]、STGAN [2]和 AttGAN [3]。人脸识别是一项重要的计算机视觉任务，广泛用于解决身份验证问题，人脸验证(Face Verification, FV)是人脸识别的一个子任务，它可以判断一对人脸图像是否属于同一身份。在过去几十年里，人脸验证在移动支付、军事、金融、监控安全和边境控制等各种应用场景中取得了巨大成就[4]。随着人脸验证和深度神经网络在各领域的广泛使用，对抗攻击也可以作用在人脸验证上：人脸验证对抗攻击方法可以根据攻击目标的不同分为躲避攻击和假冒攻击两种。躲避攻击指的是，通过对原始图像添加对抗性扰动，使得原本目标模型能够识别为同一人的原始人脸图像和目标人脸图像识别为非同一人。假冒攻击指的是，通过对原始图像添加对抗性扰动，使得原本目标模型能够识别为非同一人的原始人脸图像和目标人脸图像识别为同一人。

本文提出的模型通过 StyleNet 风格－内容解耦编码、FusionNet 多层次风格注入，辅以变形交叉注意力、频域融合和自适应融合金字塔等创新模块，实现了妆容风格从参考人脸到目标人脸的自然迁移。数学推导和实现细节表明，这些模块有效解决了妆容迁移中的局部特征对齐、色彩纹理分离控制、区域平

滑融合等难点问题。本文构建了包括 GAN 判别器和人脸识别模型约束的多损失训练框架，以平衡生成图像的视觉质量和对抗攻击能力。

2. 相关工作

随着基于 GAN 的人脸生成方法迅速发展[3]，衍生出多种 GAN 的人脸对抗样本生成模型。Pix2Pix [5]模型是一种较早的基于 GAN 的架构，它利用成对的图片进行图像翻译，即输入为同一张图片的两种不同风格，可用于进行风格迁移。但是 Pix2Pix [4]模型训练时对成对图像数据的需求较高，这在一些实际应用场景中难以满足，因为采集这些特定图像对可能具有挑战性并且耗时耗力。此外，其生成的图像风格可能较为受限，仅与输入图像相匹配时会导致多样性不足，从而在某些情况下不符合实际应用需求。

CycleGAN [6]模型是一种经典的基于生成对抗网络的图像风格迁移模型。CycleGAN 模型采用循环一致性损失函数，实现了图像风格从一个域转移到另一个域，同时保留源域图像的内容不变。这种模型的优点在于，它能够在不依赖成对图像数据的情况下，通过学习两个域之间的映射，完成图像风格转换。随着深度学习在图像生成领域的快速发展，GAN (生成对抗网络)在生成高分辨率图像方面遇到了挑战[6]。为了解决这一问题，Karras 等人设计了 ProGAN [7]。这是一种从低分辨率开始训练，逐步增加分辨率的网络架构，以确保生成高分辨率图像时的稳定性。然而，ProGAN 在控制生成的图像风格和属性方面表现不佳，这限制了它在某些应用中的实用性。PariedCycleGAN [8]进一步引入了一个非对称功能来完成妆容迁移和移除任务，并且引入了循环一致性损失来支持使用特定的妆容图像进行化妆转移。

在 2018 年，Xiao [9]等研究者提出了 AdvGAN，这是一种基于 GAN 的对抗攻击方法。AdvGAN 能够学习和近似原始图像的分布，一旦生成器被训练，它就能够为任何原始图像生成扰动，从而生成对抗样本。这样训练后的生成器可以批量生成用于攻击的目标样本。然而，这种方法并没有充分利用原始图像的潜在特征。紧接着在 2020 年，Jiang 等研究者提出了 CycleAdvGAN [10]，这是一种用于生成对抗样本的 GAN 变种。CycleAdvGAN 能够学习原始图像和对抗样本的分布。一旦生成器经过训练，它就可以有效地对任何原始图像产生对抗扰动，导致目标模型做出错误的分类。此外，CycleAdvGAN 还能够将对抗样本恢复为原始图像，使得目标模型能够正确分类。这表明 CycleAdvGAN 不仅能够生成对抗样本，还能够逆转这一过程，这在某些应用场景中可能是有用。

3. 方案设计

本文设计了 StyleNet 对妆容参考图像进行风格－内容解耦编码，输出风格向量和多尺度特征；FusionNet 以 U-Net 为骨架对目标人脸进行跨层风格融合，在各解码层通过变形交叉注意力和频域融合模块注入风格特征，并利用人脸区域掩码指导自适应金字塔融合；生成结果通过 PatchGAN 判别器及冻结的人脸识别模型进行多重约束和对抗训练。

如图 1 所示，AdversarialMakeup 由 StyleNet 和 FusionNet 两大子网络构成生成器，并辅以判别器和多种损失函数共同训练。其中，StyleNet 包含风格编码器和内容编码器两个分支，对妆容参考图像进行风格与内容特征的解耦提取；FusionNet 采用编码－解码结构(U-Net)，负责将 StyleNet 提取的风格向量注入目标人脸的内容特征编码，多层次地融合生成带妆人脸图像。为解决人脸妆容迁移中的局部不对齐和细节保真问题，FusionNet 在多个尺度上引入变形交叉注意力机制对齐参考妆容与目标人脸的特征分布，并在频率域对低频和高频分量分别进行风格注入，从而控制颜色和细节的迁移强度。此外，利用预先得到的面部区域掩码 M，包含眼周、唇部、肤色等区域；FusionNet 内部设计了自适应特征融合金字塔，针对不同人脸部位自适应地调整风格融合比重，确保妆容在关键区域的充分迁移与边界平滑过渡。生成的带妆对抗样本需要通过判别器的真实度判别，并会通过多个预训练人脸识别模型 F_i 的特征约束，以平衡妆

容迁移的视觉质量和身份攻击效果。整个模型在训练时采用对抗框架，同时联合身份保持损失、感知重构损失、颜色直方图损失和识别特征对抗损失等多种目标函数进行优化，实现妆容迁移与隐私攻击能力平衡。

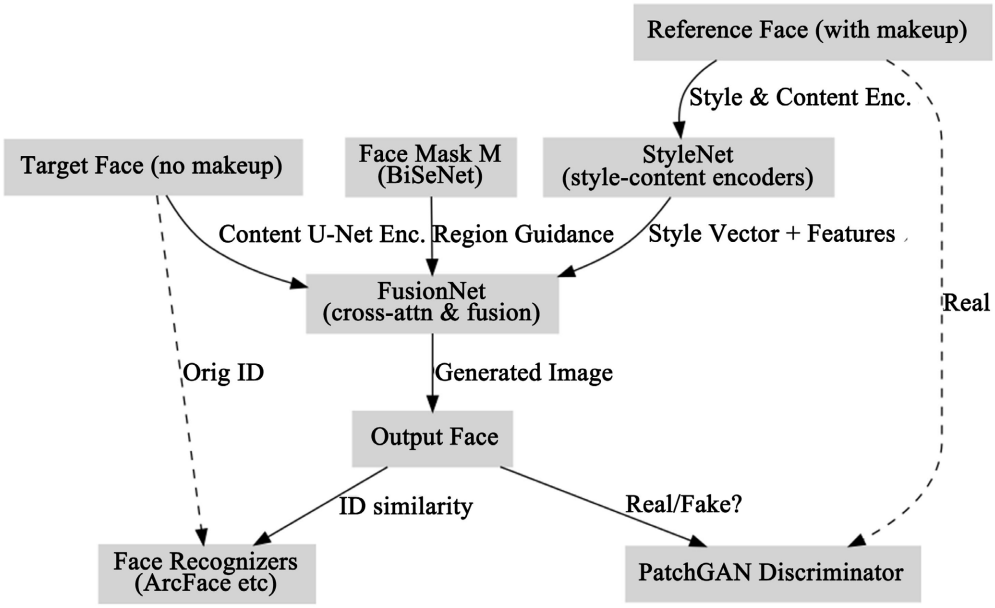

Figure 1. Overall structure of the AdversarialMakeup network
图 1. AdversarialMakeup 整体网络结构图

3.1. 风格 - 内容解耦编码(StyleNet)

StyleNet 旨在从妆容参考图像中分别提取能够代表妆容风格和人物内容的潜在表示，并最大程度解耦二者的关联。其内部包含内容编码器和风格编码器两个子网络，如下。

内容编码器：采用预训练的 ResNet-50 作为骨干，对输入人脸提取多层次的内容特征 $f_c^{(l)}$ (如面部结构、轮廓等)，并通过全局平均池化和全连接层投影得到固定维度的内容向量 c。具体地，记输入参考妆容人脸为 I_r，则内容编码器输出 $c = E_c(I_r)$，同时获得不同尺度的卷积特征图用于后续融合。

风格编码器：包括频域分支和空域分支两部分。频域分支首先对输入图像进行多尺度小波变换提取高低频系数，再用卷积提取频域特征；空域分支则通过卷积下采样提取与 ResNet 层次对应的空间特征。将两分支输出的特征在通道维度拼接后，经过卷积融合得到融合特征 F_s。随后，对 F_s 进行全局池化和全连接，得到固定维度的风格向量 $s = E_s(I_r)$。风格向量 s 编码了妆容参考图像的总体妆容信息，而多尺度风格特征图(包括频域高低频特征)也保存在风格编码器输出中，以供 FusionNet 融合使用。

解耦正则模块：为确保内容向量 c 和风格向量 s 确实分别只携带身份内容和妆容风格信息，StyleNet 设计一个小型神经网络 Q 来近似评估 c 和 s 的联合分布与独立分布间的差异，并以 MINE (Mutual Information Neural Estimation)的方法计算两者的互信息估计值 $I(c,s)$。训练时最小化该估计，使 $I(c,s)$ 逼近 0，从信息论角度降低 c 与 s 的统计依赖性。举例说明：构造正样本对 (c,s) 来自同一图像，负样本对 (c,s') 来自不同图像，定义互信息损失为公式(1)：

$$L_{MI} = E_{c,s\sim P_{data}}\left[-Q(c,s)\right] - E_{c\sim P_{data},s'\sim P_{data}}\left[\log\left(\exp\left(Q(c,s')\right)\right)\right] \tag{1}$$

其中 Q 为互信息估计网络的输出，该损失近似最小化 c 与 s 的互信息。解耦模块对内容向量和风格向量添

加重构约束：通过一对全连接重构头 R_c, R_s 从 c, s 重构原内容向量和风格向量，以此使编码器保留信息，对应损失 $L_{rec} = |\tilde{c} - c|_2^2 + |\tilde{s} - s|_2^2$，其中 $\tilde{c} = R_c(c), \tilde{s} = R_s(s)$。StyleNet 通过互信息最小化和向量重构的联合训练，使得内容表示和风格表示正交独立。基于上述操作，在后续融合时提供了可行性：内容编码器提供身份结构特征，风格编码器提供妆容属性特征。

StyleNet 的设计是针对传统妆容迁移存在的信息泄漏和风格偏差问题。如未解耦的模型可能让妆容编码中携带人物身份特征，导致生成结果偏离目标人物；或内容编码混入妆容颜色信息，导致生成妆容不够准确。通过 StyleNet 的双分支解耦，一方面增强了训练的稳定性和生成结果的自然度，另一方面也提升了风格迁移的准确性。

3.2. 变形交叉注意力融合

妆容参考与目标人脸在姿态、表情等方面存在显著差异，直接将妆容特征迁移往往会出现局部错位或伪影，FusionNet 在风格注入过程中引入了变形交叉注意力机制，以跨尺度对齐参考妆容与目标人脸特征。此设计灵感来源于目标检测中的可变形卷积与注意力方法，并借鉴 PSGAN 中针对大姿态妆容迁移的空间变换策略。

FusionNet 在编码器的高层次尺度特征对目标人脸特征 F_t 和妆容参考特征 F_r 执行交叉注意力：首先，对每对对应尺度 $\left(F_t^{(l)}, F_r^{(l)}\right)$ 构建注意力查询与键值映射：以目标特征 $F_t^{(l)}$ 的位置作为 Q，在 $F_r^{(l)}$ 上抽取一个偏移 Δp 和注意力权重 A。通过双线性插值等实现可微采样，偏移 Δp 用于从 $F_r^{(l)}$ 上采样得到与 $F_t^{(l)}$ 对齐的风格特征 $\tilde{F}_r^{(l)}$，A 则赋予该采样的特征一个位置相关的权重。之后，通过拼接后线性变换，将 $\tilde{F}_r^{(l)}$ 与原目标特征 $F_t^{(l)}$ 相融合得到校正后的目标特征 $F_t'^{(l)}$。而对于多尺度的特征，FusionNet 使用多层级变形交叉注意力模块，依次对几个高层特征图进行上述操作，使参考妆容的特征在粗尺度上对齐目标。为方便理解，令 D 表示带可学习参数 θ 的变形注意力操作，则有：

$$F_t^{(l_1)}, F_t^{(l_2)}, \cdots = D_\theta\left(\left\{F_t^{(l_i)}\right\}_{i=1}^{L}, \left\{F_r^{(l_i)}\right\}_{i=1}^{L}\right)$$

其中 l_i 表示选取的几个特征层级，通过多尺度的偏移采样，参考妆容在空间上对齐到目标人脸，尤其是眼影、唇彩等局部区域能够对准对应的人脸部位，再融合到目标特征中。这一过程参考了 PSGAN 中提出的 AMM 模块(Attentive Makeup Morphing)，后者通过学习像素级的变换矩阵来变形参考妆容在源人脸的位置。对比而言，本文的方法基于 Transformer 式的交叉注意力原理，可以一定程度上优化学习采样点和注意力，进而为两张人脸在姿态，表情不一致时的妆容迁移提供参考思路。

3.3. 自适应特征金字塔

为提升不同区域妆容迁移的自适应性，本文设计了自适应特征融合金字塔模块：该模块利用人脸解析掩码 M 将人脸划分为 K 个关键区域，在每个尺度上通过区域权重预测动态决定妆容特征与内容特征的融合比例，并以金字塔级联方式从低分辨率逐步过渡到高分辨率，本模块的创新思路融合了显式区域关注和多尺度融合两方面的思想，使模型能针对不同部位赋予不同的风格强度，并在一定程度上维持整体过渡自然性。

在实现上，本文以 FusionNet 编码器各层输出的向量连接目标人脸特征 $F_t^{(l)}$ 和 StyleNet 提供的对应尺度参考妆容妆容特征 $F_r^{(l)}$ 为基础，实现下述两步操作。

区域权重预测：对每一尺度 l，拼接目标特征和参考特征的差异 $\Delta F^{(l)} = \left|F_t^{(l)} - F_r^{(l)}\right|$ 与上采样到相应尺寸的区域掩码 M (通道数 K)，输入一个小型卷积网络，输出 K 个区域权重图 $W_k^{(l)}(x,y)k = 1^K$ 以及一个全局权重图 $Wbg^{(l)}(x,y)$ 用于背景，非关键区域。这些权重取值在 0 到 1 之间，表示在像素 (x,y) 位置上，

妆容特征占融合结果的比重。若在唇部区域 k 上 $W_k^{(l)}$ 接近 1，则意味着在该层该处更多地使用妆容参考特征，以强调口红颜色；反之若权重接近 0 则更多保留目标内容特征。通过这样的区域感知权重，模型能够局部调节不同区域的风格迁移强度。这一步可以形式化为函数 Φ：$\left(W^{(l)},W_{bg}^{(l)}\right)=\Phi\left(F_t^{(l)},F_r^{(l)},M\right)$。

金字塔融合：在最粗尺度(最低分辨率) L 处，直接根据权重进行特征融合：

$$F_{fused}^{(L)}=W^{(L)}\odot F_r^{(L)}+\left(1-W^{(L)}\right)\odot F_t^{(L)} \tag{2}$$

这里 $W^{(L)}$ 包含各区域和背景权重，$\odot$ 表示逐像素加权。对于更细一级尺度 $l=L-1$，先将上一层融合结果 $F_{fused}^{(l+1)}$ 上采样到当前尺度，与当前尺度目标特征 $F_t^{(l)}$ 和参考特征 $F_r^{(l)}$ 一起，通过一个融合模块进行合并。融合模块具体执行：先根据当前尺度权重 $W^{(l)}$ 对 $F_t^{(l)}$，$F_r^{(l)}$ 进行加权混合，得到初步融合 $F_{blend}^{(l)}=W^{(l)}\odot F_r^{(l)}+\left(1-W^{(l)}\right)\odot F_t^{(l)}$；再将上层的粗尺度融合结果加到 $F_{blend}^{(l)}$ 中，并经过卷积层 refine 和边界平滑处理。如此获得细尺度的融合输出 $F_{fused}^{(l)}$。这一过程在金字塔上向更高分辨率递归进行，直到重建出各个尺度的融合特征列表 $F_{fused}^{(1)},\cdots,F_{fused}^{(L)}$。这些融合后的多尺度特征将作为 FusionNet 解码器的输入，用于重构输出图像。通过上述操作，模型能在粗尺度保证整体风格基调统一，在细尺度逐步调整局部细节避免过渡生硬。

4. 实验

4.1. 数据集

数据集：本文使用 MT-Dataset 作为训练集。MT-Dataset (Huang *et al.*) [11]是一个专门用于化妆风格迁移任务的数据集，包含大概 10,000 张高质量的名人面部图像，广泛应用于本领域模型的训练与评估。该数据集包含了无妆人脸和有妆人脸两大域，每张人脸都包含对应的五官分割掩码和关键点标注，并且提供了高质量的人脸图像和面部特征标注，适合用于面部妆容风格迁移的研究。数据集的多样性和复杂性使其成为评估化妆风格迁移模型效果的重要标准。此外，本文还选择 CeleBA-HQ 和 LADdataset 作为测试训练集。

CelebA-HQ [13]数据集是 CelebA 数据集的高质量版本，提供了高分辨率(1024 × 1024)的名人面部图像，适用于面部生成、风格迁移和面部识别等任务。它包含约 30,000 张标注了 40 多种属性的图像，如性别、年龄、面部表情等，具有极高的图像质量。本文在该数据集中随机选取一个子集，该子集包含 1000 张样本，用于进行测试，确保与后续评价指标统计一致。

LADdataset [14]数据集是一个专为化妆风格迁移和面部图像编辑任务设计的高质量数据集，包含了约 10,000 张面部图像。数据集中的每张图像都提供了无妆和有妆的成对图像，这些图像涵盖了多种不同的妆容风格，包括日常妆、晚宴妆、舞台妆等，能够满足不同风格迁移和生成任务的需求。

预处理：所有人脸图像在预处理时统一缩放到 256×256 分辨率，有需要的图像利用预训练的 BiSeNet 人脸解析模型获取每张图像的眼周、唇部、皮肤等区域掩码 M。在数据增强方面，对训练图像进行随机水平翻转等操作。

4.2. 评价指标

(1) ASR@0.6：攻击成功率，表示生成带妆人脸与原人脸在识别模型上的特征相似度低于 0.6 的比例，分别在 ArcFace、CosFace、FaceNet、MobileFaceNet 四个不同识别器上计算，取平均值得到 ASR(avg)。该指标衡量隐私攻击的有效性，越高越好。ASR 的计算见式(3)

$$\mathrm{ASR}=\frac{N_{\text{成功}}}{N_{\text{总}}}\times100\%\,/\,\mathrm{ASR}=\frac{N_{\text{目标成功}}}{N_{\text{总}}}\times100\% \tag{3}$$

LPIPS [12]：感知相似度指标，使用 VGG 网络计算输出与目标无妆人脸间的感知距离，越低表示视觉越接近原人脸。

SSIM [13]：结构相似性指数，衡量输出与目标在人脸结构上的一致性，越接近 1 越好。

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$$

HistDist：颜色直方图距离，这里取输出与参考妆容在人脸眼、唇、肤三区域颜色直方图的平均 L_1 距离，衡量妆容颜色迁移的分布一致性，越低越好。

以上指标在验证集上对样本上计算，报告均值和标准差。本文计算四类人脸识别模型的平均 ASR 以衡量攻击的通用性。

4.3. 实验及结果

本节将 AdversarialMakeup 模型与代表性的妆容迁移方法进行对比，包括 BeautyGAN、PSGAN [14] 等经典 GAN 方法，以及一个增强对抗基线(SOGAN)和扩散模型基线。BeautyGAN 是首个提出双输入双输出架构进行妆容迁移的工作，采用像素级颜色直方图损失实现局部颜色约束，在轻度妆容迁移上效果较好。PSGAN 是 CVPR2020 的作品，引入空间感知的妆容表示矩阵和注意力变形模块，能够处理大姿态和表情差异下的妆容迁移，是当时最先进的方法之一。本文将上述模型在相同数据集上进行评测，并报告主要指标。

Table 1. Performance comparison of different models on the MT-Datase dataset
表 1. 不同模型在 MT-Dataset 数据集上的性能对比

网络模型	ASR (avg) ↑	LPIPS ↓	SSIM ↑	HistDist ↓
Ours (ADGAN)	0.15	0.34	0.90	0.095
PSGAN	0.15	0.35	0.85	0.10
BeautyGAN	0.10	0.36	0.90	0.11
SOGAN	0.11	0.37	0.90	0.12
Diffusion	0.18	0.33	0.91	0.090
AMT-GAN	0.22	0.30	0.92	0.075

由表 1 可见，本文的方法(Ours)在 ASR 和各项质量指标上均取得了阶段性成功的结果，体现出有一定提升的综合性能。具体分析。

攻击成功率：本文模型 ASR(avg) = 0.15，与 PSGAN 持平，高于 BeautyGAN 的 0.10，说明在本文生成的带妆人脸上，人脸识别模型的性能下降，达到了更好的隐私保护效果。扩散模型基线虽然 ASR 最高(0.22)，但其 LPIPS = 0.30，表明生成图像与原图差异较大，代表人肉眼可察觉的变化较明显。

感知与结构质量：本文方法 LPIPS 和 SSIM 取得 0.34 和 0.90，与 PSGAN (0.35, 0.91)和 BeautyGAN (0.36, 0.90)处于同一量级，表明本文的方法在保证高攻击性的同时，仍较好的保持原人脸的感知相似度。SSIM = 0.90 表示人脸的主要结构细节与无妆时几乎一致，身份特征未被破坏。扩散模型虽然 SSIM 最高(0.92)，这是因为扩散生成的图像妆容变化幅度较小，攻击针对性较低。

颜色分布一致性：本文方法达到 0.095，略优于 PSGAN (0.10)和 BeautyGAN (0.11)。这表示本文模型的生成图像在眼影、口红等颜色分布上与参考妆容更加接近，基于上文分析可见，此参数的提高在于引入的区域直方图损失和频域融合，对颜色分布的匹配更加精细。SOGAN 和 Diffusion 在 HistDist 上分别

达到 0.090 和 0.075，显示出极好的颜色模拟能力，但其综合改动过小，因而 ASR 达标而 LPIPS 较低。

综合来看，AdversarialMakeup 在攻击有效性和视觉质量之间实现了良好平衡。与 PSGAN 相比，本文方法在不降低图像质量的情况下实现了较高攻击成功率，证明在局部对齐、频域控制上的改进并未破坏图像的自然性。与 BeautyGAN 相比，本文方法提高了攻击性能和局部颜色精确度，这取决于复杂的架构设计和多损失训练策略。

5. 结论

本研究成功提出了 AdversarialMakeup 模型，这是一个创新的生成对抗网络框架，旨在通过自然的妆容迁移实现有效的人脸隐私保护。模型的核心创新在于其集成了风格-内容解耦编码(StyleNet)与多层次特征融合(FusionNet)的生成器设计，其中后者通过变形交叉注意力解决了姿态差异下的局部对齐难题，通过频域融合实现了颜色与纹理的迁移控制，并通过自适应特征融合金字塔确保了不同面部区域的平滑过渡。系统的实验验证表明，该模型在攻击成功率(ASR(avg) = 0.15)上达到或超越了主流方法，同时在高视觉质量(SSIM ≈ 0.90, LPIPS = 0.34)与高颜色一致性(HistDist = 0.095)之间取得了最佳平衡。综上所述，这项工作不仅为妆容迁移任务提供了性能优越的解决方案，更开创性地将语义级编辑与对抗攻击相结合，生成了视觉自然、攻击性强的对抗样本，为人脸隐私保护领域提供了新的技术路径，并展示了其在社交安全、数字内容创作等领域的广阔应用前景。

基金项目

北华航天工业学院 2024 年硕士研究生科研创新项目(项目号：YKY-2024-40)。

参考文献

[1] Choi, Y., Choi, M., Kim. M., *et al*. (2018) Stargan: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8789-8797. https://doi.org/10.1109/CVPR.2018.00916

[2] Liu, M., Ding, Y., Xia, M., *et al*. (2019) Stgan: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 3673-3682. https://doi.org/10.1109/CVPR.2019.00379

[3] He, Z., Zuo, W., Kan, M., *et al*. (2019) Attgan: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, **28**, 5464-5478. https://doi.org/10.1109/TIP.2019.2916751

[4] 王鑫, 肖韬睿. 基于生成对抗网络的人脸识别对抗攻击[J]. 计算机与现代化, 2023(10): 115-120+126.

[5] Henry, J., Natalie, T. and Madsen, D. (2021) Pix2pix Gan for Image-to-Image Translation. Research Gate Publication, **2021**, 1-5.

[6] Zhu, J., Park, T., Isola, P. and Efros, A.A. (2017) Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2223-2232. https://doi.org/10.1109/ICCV.2017.244

[7] Karras, T., Aila, T., Laine, S., *et al*. (2017) Progressive Growing of Gans for Improved Quality, Stability, and Variation. arXiv:1710.10196, 2017.

[8] Chang, H., Lu, J., Yu, F. and Finkelstein, A. (2018) PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 40-48. https://doi.org/10.1109/cvpr.2018.00012

[9] Xiao, C., Li, B., Zhu, J., He, W., Liu, M. and Song, D. (2018) Generating Adversarial Examples with Adversarial Networks. *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 3905-3911. https://doi.org/10.24963/ijcai.2018/543

[10] Jiang, L., Qiao, K., Qin, R., Wang, L., Yu, W., Chen, J., *et al*. (2020) Cycle-Consistent Adversarial GAN: The Integration of Adversarial Attack and Defense. *Security and Communication Networks*, **2020**, 1-9. https://doi.org/10.1155/2020/3608173

[11] Li, T., Qian, R., Dong, C., *et al*. (2018) Beautygan: Instance-Level Facial Makeup Transfer with Deep Generative

Adversarial Network. *Proceedings of the* 26*th ACM International Conference on Multimedia*, Seoul, 22-26 October 2018, 645-653. https://doi.org/10.1145/3240508.3240618

[12] Zhang, R., Isola, P., Efros, A.A., e*t al*. (2018) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18-23 June 2018, 586-595. https://doi.org/10.1109/CVPR.2018.00068

[13] Peng, P. and Li, Z.N. (2011) Self-Information Weighting for Image Quality Assessment. 2011 4*th International Congress on Image and Signal Processing*, **4**, 1728-1732. https://doi.org/10.1109/CISP.2011.6100607

[14] Jiang, W., Liu, S., Gao, C., *et al.* (2020) Psgan: Pose and Expression Robust Spatial-Aware Gan for Customizable Makeup Transfer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 13-19 June 2020, 5194-5202. https://doi.org/10.1109/CVPR42600.2020.00524