中图分类号：　　单位代号：10280
密　　级：　　学　　号：21721395

上海大学

专业硕士学位论文

SHANGHAI UNIVERSITY PROFESSIONAL MASTER'S DISSERTATION

题目	空域及频域人脸图像通用对抗扰动技术研究

作　　者　金玺
学科专业　电子信息
导　　师　吴汉舟
完成日期　2024 年 8 月

姓　　名：金玺　　学号：21721395

论文题目：空域及频域人脸图像通用对抗扰动技术研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主任：

委员：王子驰 上海大学

吕东辉 上海大学

导　　师：吴汉舟

答辩日期：2024.8.15

姓　　名：金玺　　学号：21721395

论文题目：空域及频域人脸图像通用对抗扰动技术研究

原 创 性 声 明

本人声明：所呈交的论文是本人在导师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签　名：金玺　日 期：2020.8.25

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密的论文在解密后应遵守此规定）

签 名：金玺　导师签名：吴汉舟　日期：2020.8.25

上海大学工程硕士学位论文

空域及频域人脸图像通用对抗扰动技术研究

姓　　名：　金玺

导　　师：　吴汉舟

学科专业：电子信息

上海大学通信与信息工程学院

二〇二四年八月

A Dissertation Submitted to Shanghai University for the

Degree of Master in Engineering

Research on Universal Adversarial Perturbation Techniques for Facial Images in Spatial and Frequency Domains

Candidate： Xi Jin

Supervisor：Hanzhou Wu

Major：Electronic Information

School of Communication and Information Engineering

Shanghai University

August, 2024

摘 要

近年来，基于深度神经网络的人脸识别技术被广泛应用于不同领域。现有研究表明，基于深度神经网络的人脸识别模型容易受到精心设计的对抗性扰动的影响。攻击者通过在面部图像上施加微小的扰动，就能轻易诱导目标模型做出错误的判断，从而影响模型的可靠性和稳定性。在对抗攻击领域，大多数现有通用对抗扰动算法都是为图像分类任务设计的，当这些算法用于人脸识别模型时存在攻击成功率和攻击隐蔽性较低等问题。设计针对人脸识别技术的通用对抗攻击策略，对于增强模型的鲁棒性和发展有效的防御机制具有重要指导意义。在此背景下，本文从空域和频域两个角度出发，分别提出针对人脸识别模型的通用对抗扰动生成方案，具体内容如下：

（1）提出一种基于空域分析的人脸通用对抗扰动攻击方案。通过探索人脸关键区域对模型识别准确度的影响，提取针对数据集的语义关键区域位置，以此作为一种更符合人脸图像对抗攻击的先验区域。随后，使用可学习的流场来微调该区域的位置，确保生成的扰动能够集中于人脸的合理区域，通过控制不同语义区域扰动的强度，从而生成具有局部隐蔽性的通用对抗扰动。此外，本文设计针对人脸图像的对抗性损失和隐蔽性损失，实现了攻击性和隐蔽性的双重优化。实验结果表明，对比现有的通用对抗攻击手段，所提出的空域扰动方案在保持攻击性能的基础上，显著提升了扰动的隐蔽性。

（2）提出一种基于频域分析的人脸通用对抗扰动生成方案。通过分析自然图像和人脸图像在频域视角的差异，确定人脸图像可以采用频域维度的对抗攻击方案。通过学习高、中、低三个不同频段的滤波器来利用频域信息，为调整和优化人脸通用对抗扰动提供了额外的维度。此外，通过优化一幅特定的目标图像来将非目标攻击转化为目标攻击模式，并将该目标图像作为训练集分布之外的样本运用于对抗性损失中。实验结果表明，所提出的扰动生成方案与现有的通用对抗扰动生成方案相比，在公开人脸数据集上实现了更高的攻击成功率和更好的隐蔽性，验证了该方案引入频域信息的有效性。

关键词：深度学习；人脸识别；对抗攻击；通用对抗扰动

ABSTRACT

In recent years, face recognition technology based on deep neural networks has been widely applied in various fields. Existing research indicates that face recognition models based on deep neural networks are susceptible to the influence of carefully crafted adversarial perturbations. Attackers can easily induce target models to make incorrect judgments by applying subtle perturbations to facial images, thereby affecting the reliability and stability of the models. In the field of adversarial attacks, most existing universal adversarial perturbation algorithms are designed for image classification tasks. However, when these algorithms are applied to face recognition models, they suffer from issues such as low attack success rates and low attack stealthiness. Designing universal adversarial attack strategies specifically tailored for face recognition technology is of crucial importance for enhancing model robustness and developing effective defense mechanisms. Against this background, this thesis proposes universal adversarial perturbation generation schemes for face recognition models from both spatial and frequency domain perspectives. The specific contents are as follows:

(1) A universal adversarial perturbation attack scheme for face images based on spatial domain analysis is proposed. By exploring the influence of facial key regions on model recognition accuracy, semantic key region positions tailored to the dataset are extracted as a priori regions more suited for adversarial attacks on facial images. Subsequently, a learnable flow field is employed to fine-tune the position of these regions, ensuring that the generated perturbations are located within plausible facial areas. Controlling the intensity of perturbations in different semantic regions, universal adversarial perturbations with local stealthiness are generated. Additionally, adversarial loss and stealth loss are designed specifically for facial images, achieving dual optimization of attack and stealthiness. Experimental results demonstrate that compared to existing universal adversarial attack methods, the proposed spatial perturbation scheme significantly enhances the stealthiness of perturbations while maintaining

attack performance.

(2) A universal adversarial perturbation generation scheme for face images based on frequency domain analysis is proposed. Analyzing the differences between natural images and facial images from a frequency domain perspective, it is ascertained that facial images can adopt a frequency domain-based adversarial attack approach. Learning filters for high, medium, and low frequency bands provides an additional dimension for adjusting and optimizing facial universal adversarial perturbations by exploiting frequency domain information. Moreover, converting non-target attacks into target attack modes by optimizing a specific target image and utilizing this target image as an out-of-distribution sample in adversarial loss facilitates the transformation. Experimental results demonstrate that compared to existing universal adversarial perturbation generation schemes, the proposed scheme achieves higher attack success rates and better stealthiness on publicly available facial datasets, validating the effectiveness of incorporating frequency domain information.

Keywords: Deep learning, Face recognition, Adversarial attack, Universal adversarial perturbation

目 录

第一章　绪论

1.1　课题来源

本课题来源于 CCF-蚂蚁隐私计算专项科研基金项目“非同分布可迁移的人脸表征嵌入通用对抗扰动生成方法”，基金号：CCF-AFSG RF20220019。

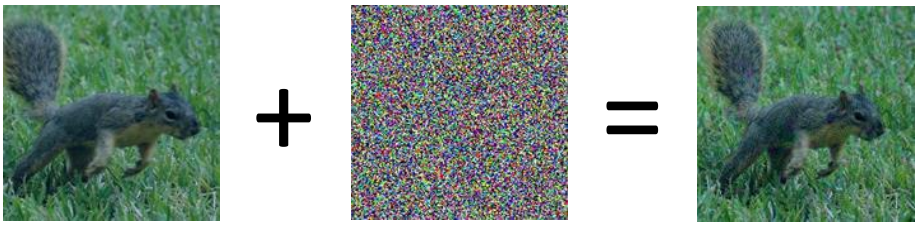
1.2　研究背景与意义

近年来，受益于大数据技术的飞速发展和计算硬件的持续进步，以深度学习[1]为代表的人工智能领域得到了研究者的高度重视。深度学习通过构建深层的神经网络，能够自动从原始数据中学习到有效的特征表示，从而大大提高了模型的表达能力和泛化能力。神经网络是一种启发自人类大脑神经元工作机制的计算模型，其通过模拟大脑神经元之间的连接和信息处理方式，以实现学习和模式识别的功能。神经网络中的神经元可类比为具有多个自变量和一个因变量的函数，用以处理输入信号。最早的卷积神经网络架构之一为 LeNet[2]，随后的研究不断对其宽度和深度进行改进，显著提升了神经网络的数据拟合能力。神经网络的研究和应用对社会产生深远影响，其成熟技术在图像识别[3]、自动驾驶[4]、形迹检测[5]等多个领域发挥关键作用，并应用于生活的各个方面。

作为图像识别领域的重要分支，人脸识别技术[6]专注于从图像或视频中识别和验证个体的身份，具有高效和准确的认证机制，在支付授权、社交网络、身份监控和公共安全等多个领域大放异彩。随着神经网络模型的性能越来越优异，通过神经网络对输入人脸数据进行特征提取和特征匹配等操作，现有的人脸识别技术在公开的人脸数据集上的识别准确率已接近百分之百。然而，当人脸识别这一技术发展到接近顶峰时，其安全问题也渐渐暴露在研究者视线中。数据投毒[7]、后门攻击[8]和对抗攻击[9]等针对神经网络模型的攻击手段，对其稳定性和可靠性构成了严重威胁。研究工作者发现神经网络的预测结果对输入数

据的微小变动非常敏感。通过在输入数据中叠加特定的干扰，可以轻易地引发模型输出的变化。这种显著的不稳定性如果被恶意利用，可能会对各行业的模型应用带来严重威胁。对抗扰动攻击技术[10]即是这种威胁之一，其通过对抗样本来实现针对现有模型的攻击。

对抗样本的产生是在原始数据样本上附加了一些细微且精确计算的扰动，这些干扰改变了样本的向量表达，导致模型输出错误的结果，从而引发模型进行错误的预测。Szegedy 等[11]在 2013 年提出了对抗攻击的概念，他们发现通过在输入图像上添加一些人眼无法感知的微小噪声后，图像分类模型能以很高的置信度将图像预测为错误标签。如图 1.1 所示，其中未被修改的原始输入图片被称为合法样本，基于不同攻击算法框架生成的可叠加噪声数据被称为对抗扰动，在合法样本上嵌入对抗扰动后产生的新图像被称为对抗样本，对抗攻击即是这整个生成对抗样本并愚弄目标模型的过程。



松鼠
95.6%置信度
(a)合法样本

(b)对抗扰动

蝴蝶
97.5%置信度
(c)对抗样本

图 1.1 生成对抗样本示意图

在人脸识别领域，对抗攻击对不同运用场景的安全性和准确性构成了巨大的威胁。目前，已有一些针对人脸识别系统的对抗攻击方案。Komkov 等[12]通过打印带有图案的贴纸，将该贴纸装饰在帽子上，戴帽子的攻击者虽然未遮挡五官，但依旧让人脸识别模型识别出错。2021 年 RealAI 团队利用了对抗样本攻击技术生成一副有特殊扰动的眼镜框架，使用 20 部手机作为人脸识别解锁试验机，该方式破解了 19 部手机的人脸解锁系统，此次攻击展示了人脸识别技术在商用领域的潜在隐患。这些实例揭示了人脸识别模型的脆弱性，证明了人脸识别系统在追求高准确率的同时，也需要考虑到模型的鲁棒性以应对潜在漏洞。

人脸识别技术是高效的生物识别方式之一，人脸作为独一无二的生物特征，与指纹、声纹和虹膜类似，可以充当个人独有的身份密码。但相较于其他

的生物信息，人脸数据没有很强的敏感性和隐私性，在现今这个信息化时代，人脸信息的获取非常方便。与数字密码不同，人脸的数据无法轻易改变，这种生物特征一旦泄露就无法恢复到泄露之前的状态，因此人脸数据隐私保护也至关重要。对抗样本因其自身属性可以作为一种有效的人脸隐私保护的方式，用户可以在自己照片上叠加扰动图片，即使该照片被不法手段获取，在人脸识别网络中也不会将该照片与用户真实身份关联，以此做到保护人脸隐私的目的。

人脸对抗攻击的研究为构建防御机制提供了重要的反面视角，使研究人员确定人脸识别模型的局限性和脆弱性，从而提高人脸识别的对抗鲁棒性。在对抗防御领域，不论是对抗样本检测[13]还是对抗扰动清洗[14]，都依赖于对抗攻击算法。此外，其还具备隐私保护功能，能够有效防止用户人脸数据泄露，从而避免由此引发的个人身份安全风险。人脸识别作为一种特殊的分类任务，目前针对其的对抗攻击方案多是倾向于攻击单个目标身份样本[15]，这种特定人脸的对抗攻击方式每次使用都需要再次产生相应的扰动。现有的针对人脸识别系统的通用对抗扰动攻击的方案还有待研究，本文受自然图像分类任务的通用对抗扰动方案启发，致力于研究面向人脸识别系统的通用对抗扰动生成方案，相关人脸通用对抗扰动作用效果如图 1.2 所示，在正常的人脸识别过程中，特征提取器获取干净的人脸嵌入向量特征，通过针对两者嵌入向量的相似性分析可以确定图像是否为同一个身份。然而在通用对抗攻击过程中，通过在一组人脸图像上叠加同一个通用对抗扰动生成相应人脸对抗样本，此时特征提取器只能获取修改后的人脸嵌入向量，导致人脸识别模型识别出错。通过分析人脸通用对抗扰动在不同数据集和不同模型下的有效性，有利于研究更为有效的防御机制，从而提升人脸识别模型的鲁棒性和安全性。

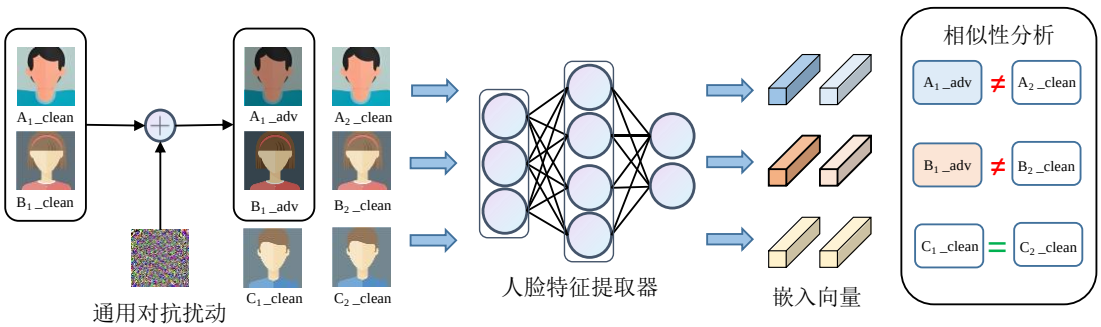


图 1.2 人脸通用对抗扰动作用效果示意图

1.3 国内外研究现状

1.3.1 对抗攻击技术研究现状

在 2014 年，Szegedy 等人[11]发现对抗样本的存在现象，他们通过在输入图像上添加精心设计的微小扰动，这些样本在视觉上与原始图像相似，但会导致模型做出错误的预测，他们揭示了深度神经网络的脆弱性，为后续对抗攻击的研究奠定了基础。快速梯度符号法（Fast Gradient Sign Method, FGSM）是由 Goodfellow 等人[16]提出的一种用于生成对抗样本的算法，该算法核心原理是利用深度神经网络的梯度信息来寻找图像上的微小扰动，这些扰动在添加到原图像后能够导致网络错误分类。其简单高效，具有较小的内存负担，后续研究者基于该算法的原理提出了多种改进方案。基本迭代方法（Basic Iterative Method, BIM）是 FGSM 的迭代版本[17]，通过多次迭代来逐步增加扰动，生成更有效对抗样本。后续改进方法[18]在基本迭代法的基础上引入了动量项，这有助于平滑扰动的更新方向，避免陷入局部最优，从而提高对抗样本的质量。DeepFool[19]算法的核心思想是寻找输入样本与其他类别决策边界的最短距离，并依此方向生成对抗样本。在同一时期，Papernot 等人[20]观察到对抗样本具有一定的黑盒攻击能力，攻击者可以构建一个替代模型来执行黑盒迁移攻击。进一步地，Papernot 等人还提出一种基于雅可比矩阵的显著性图攻击[21]，通过改变图像中极少数像素点进而创建出有效的对抗样本。2017 年，Madry 等人[22]引入投影梯度下降法（Projected Gradient Descent, PGD），其在初始迭代阶段采用一个按照均匀分布随机生成的扰动作为起始点，通过小步长和多次迭代来生成对抗样本，因其能生成高攻击性和强迁移性的对抗样本，成为最广泛使用的技术之一。Carlini 和 Wagner[23]开发了一种先进的扰动生成方法，核心在于针对 logits 的铰链损失函数，该方法基于优化的原理，通过在不同范数下限制扰动的大小，同时考虑了对分类器的攻击效果和与原图的差异，显著提升了攻击效力和视觉隐蔽性。Liu 等人[24]探究了对抗样本的迁移性，通过综合多个模型的信息来增强对抗样本的黑盒攻击能力。2018 年，Zhou 等人[25]基于在特征空间中距离较远的样本迁移到其他模型时能保持其对抗性这一观点，提出了 TAP 算法，旨在

生成具有高迁移性的对抗样本。2019 年，Su 等人[26]发现通过仅修改图像中的单个像素点，就能导致深度神经网络产生错误的预测结果，证明了尽管扰动十分微小，也足以让经过训练的模型做出错误判断。Huang 等人[27]通过在源模型的预指定层上增加扰动来增强现有对抗样本的黑盒迁移能力。在 2020 年，Wu 等人[28]提出一种利用残差攻击方法（Skip Gradient Method, SGM）的对抗样本生成技术，其核心思想是针对残差网络模型，通过增强在残差连接中传播的梯度的权重从而获得有更强迁移性的对抗样本。Duan 等人[29]提出一种对抗伪装攻击，通过为对抗样本设计一种伪装风格，使得生成的对抗样本在物理世界中不易被察觉，其揭示了深度学习模型在物理世界应用中的潜在脆弱性。Andriushchenko 等人[30]提出一种查询高效的黑盒对抗攻击方法，在每次迭代前随机采样扰动，并计算损失函数以指导扰动的更新方向。后续黄等人[31]提出基于进化策略和注意力机制的黑盒对抗攻击算法，旨在解决现有黑盒攻击方法在攻击效率和隐蔽性方面的不足，其通过协方差矩阵自适应进化策略和注意力机制的方式在提高攻击性同时增强了不可感知性。

上述方法都是属于针对特定图像的对抗攻击，每一张扰动图像都映射一张特定的原始图像。通用对抗攻击技术起步相对较晚，Moosavi-Dezfooli 等人[32]在 2017 年提出通用对抗扰动的概念，借助 DeepFool 方法计算出每个输入样本的最小扰动量，将扰动不断累加直至足以使大量样本跨过决策边界，这种扰动能够被添加到数据集中的多个样本上，使之都被错误分类。随后，Mopuri 等人[33]提出快速特征愚弄（Fast Feature Fool, FFF）的新方法，由于神经网络的特征层对输入数据具有敏感性，通过扰乱这些特征层可以影响网络的决策，该方法适用于不依赖于具体数据样本的场景，还有一种方法为 GD-UAP[34]，Mopuri 等人通过对 FFF 方法的优化损失函数进行改进，实现能够不依赖于具体数据情况下在多种任务上实现通用对抗扰动。Hayes 等人[35]利用生成模型来学习扰动的整体分布方法，为生成扰动提供了新的视角。Poursaeed 等人[36]提出创建通用对抗扰动的统一框架 GAP，其也是使用生成模型构建扰动的方式。类似的方式还有使用 NAG[37]来学习通用扰动的整体分布，其采用新型损失，运用生成器使得生成的扰动具有多样性。Liu 等人[38]为了增强通用对抗扰动的愚弄性能，使用蒙特卡洛采样方法激活神经元且利用纹理结构初始化扰动。Zhang 等人[39]将通用对抗

扰动视为包含关键决策信息的主要因素，而原始输入图像被视为背景噪声，通过设计一种特征主导的通用扰动算法，提高了扰动在目标攻击和非目标攻击场景下的迁移性。Dai 等人[40]采用方向优先的策略，在每次迭代中都选择与当前扰动方向最相似的扰动矢量，可以更快积累有效扰动，从而加快生成过程。Li 等人[41]提出一种针对目标检测场景下生成 UAP 的方法，使得在模型定位和识别图像时产生错误。Zhang 等人[42]探索了一种无需依赖于特定数据集即可生成通用对抗扰动的新方法，通过迭代最大化预训练模型的 logits 输出与干净图像及对抗样本之间的差异，从而生成有效扰动。

1.3.2 面向人脸识别的对抗攻击技术研究现状

针对人脸识别的对抗攻击可以根据其实施环境划分为物理世界攻击和数字域攻击两大类。物理世界攻击通常用于现实世界，攻击者通过将对抗性扰动以图案的形式呈现在实物上来进行攻击，如将扰动图案植入佩戴的眼镜和帽子中，或者在物理世界中布置特殊的标记或图像，可以诱导识别系统产生误判。数字域攻击通过在算法层面对数字图像进行微妙的修改，这些修改可能涉及到像素级别的小幅度调整或者更为复杂的空间变换等图像处理操作。

Sharif 等人[43]提出一种攻击方法，利用基于迁移的攻击方式生成对抗性眼镜的扰动，通过优化交叉熵损失并减少相邻像素间差异，生成平滑的对抗扰动。将该对抗扰动粘贴在眼镜框架上即可实现针对物理世界人脸识别模型的对抗攻击。类似的 AdvHat[12]通过打印一张带有图案的贴纸，将该贴纸装饰在帽子上即可攻击人脸识别模型。Ibsen 等人[44]将特制人脸图像打印在 T 恤上，穿着印有人脸图像的 T 恤会成为识别主体，以此混淆人脸识别系统。Adv-Makeup[45]算法着重于控制扰动的风格，其通过化妆的方式生成一种难以被察觉的对抗样本，利用化妆生成器来提高合成化妆的自然性，该方式不仅在数字图像中有效，也可以生成纹身贴的形式在物理世界中使用。Zolfi 等人[46]提出一种对抗性口罩的物理对抗攻击方式，通过在口罩上应用精心设计的图案作为扰动，以实现在现实世界中对人脸识别系统的攻击。

在数字域攻击中，Rozsa 等人[47]选择一个目标人脸作为人脸图像对抗样本的目标，通过最小化目标人脸和对抗样本间的欧式距离，来进行人脸识别模型

的目标攻击。Dabouei等人[48]提出一种针对人脸局部关键点的空间变换攻击，通过对人脸关键点进行一定程度的移位，使得人脸面部发生微小扭曲，以此实现快速的数字域攻击。Dong 等人[49]提出了一个在基于决策的黑盒场景下的演化攻击方法，可以对搜索方向的模型决策边界的几何结构建模，同时降低搜索空间的维度，提高了黑盒的攻击效率。Parmar 等人[50]通过在人脸上迭代应用一个透明的小补丁，针对不同人脸图像优化不一致的补丁以此欺骗人脸识别模型。

1.4 研究内容与结构安排

1.4.1 研究内容

本文主要研究了面向人脸识别系统的通用对抗扰动技术，通过对现有对抗攻击技术的分析，目前大部分针对人脸识别的对抗扰动生成方案都是依赖特定人脸图像的，为此本文通过空域和频域两个维度信息，分析自然图像与人脸图像的异同，探究人脸数据集和人脸损失函数的独有特征，针对人脸识别模型提出了两种有效的通用对抗扰动生成方案，主要内容如下：

（1）基于空域分析的人脸通用对抗扰动：本方案首先基于人脸识别可解释性的思想，通过获取不同人脸图像的语义关键区域，综合选定人脸数据集的整体关键区域位置，提取出对识别结果有显著影响的语义位置作为符合人脸图像通用攻击的先验区域，以此表征整个人脸数据集的总体特征。其次，通过训练一个可学习流场对表征数据集总体特征的先验区域进行微调，使其弥补区域中缺失的个体特征内容，确保生成的扰动定位于人脸的合理区域。此外，本方案通过设计针对人脸的隐蔽性损失和对抗性损失，有效实现攻击性和隐蔽性的双重优化。最终实验结果表明，对比现有攻击方案，本方案设计思路更符合人脸识别任务，所提出的语义关键区域控制的扰动生成方案在保持攻击性的同时能显著提升扰动的隐蔽性。

（2）基于频域分析的人脸通用对抗扰动：本方案引入频域信息作为额外维度，提出一种基于频带滤波器驱动的人脸通用对抗扰动生成方案。通过学习高频、中频和低频三个频段的扰动信息，分频段地利用频域信息生成对应的扰

动，有效规避了直接在空域全局添加扰动造成的视觉缺陷问题。通过对通用对抗扰动和频带滤波器进行调整和优化，自适应地获取合适信息来生成扰动，以此提高扰动的有效性。本方案还提前优化了一幅在数据集分布之外的定制目标图像，将其以目标攻击的模式纳入损失函数的设计中，并将对抗样本与干净图像之间的余弦相似度方差纳入计算，有效提高生成扰动的攻击性和隐蔽性。实验结果表明，与现有攻击方案相比，本方案引入频域信息驱动的扰动生成方法在常见的开源人脸数据集上有更高的攻击成功率和更好的客观隐蔽性。

1.4.2 结构安排

本文组织结构分为 5 个章节，其中各个章节的内容安排如下：

第一章为绪论，主要介绍了本课题的来源以及在人脸识别应用场景中存在的安全隐患，阐明了研究对抗攻击的重要意义。同时分析了针对该领域的国内外研究现状，总结并分析了现有研究的优点和不足，最后介绍了本论文的研究方向和结构安排。

第二章为相关技术基础部分，主要介绍了当前主流的卷积神经网络和人脸识别技术的基础，详细分析了对抗攻击的类别及算法细节，最后介绍了针对人脸识别任务的对抗攻击的应用场景及相关技术基础。

第三章为基于空域分析的语义关键区域控制的人脸通用对抗扰动，介绍了该扰动生成框架的整体流程，详细介绍了语义关键区域的作用及流场微调的原理，随后介绍了通用对抗扰动的生成过程及使用的损失函数，最终对该扰动生成方式的攻击性及隐蔽性进行评估。

第四章为基于频域分析的频带滤波器驱动的人脸通用对抗扰动，阐述了所提出算法的整体目标的分析，详细介绍自适应频带滤波器模块和定制目标函数模块，并分析了损失函数的设计思路，最终在人脸数据集上进行实验，对算法性能进行评估。

第五章为总结与展望，对本文的研究内容进行总结与概括，针对本文的不足之处进行分析，提出未来的改进及探索方向。

第二章　相关技术基础

2.1　深度神经网络模型

深度神经网络模型通过模拟人脑处理信息的方式，利用多层结构对数据进行高效抽象和特征提取。网络由多个层次的非线性处理单元构成，这些基本单元为神经网络的神经元，包括输入、求和运算和输出三个部分，如图 2.1 所示。

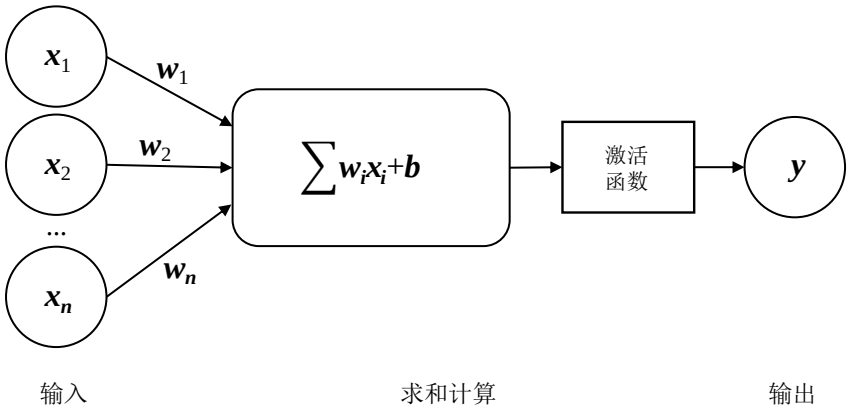


图 2.1 神经元结构示意图

网络中每一层都在对前一层的信息进行转换和编码，逐步提炼出数据中的复杂关系和模式。随着深度学习技术的不断进步，卷积神经网络作为深度神经网络的一种主要类型，因其在图像处理任务中的卓越表现而备受关注，其通过局部感受野和权重共享机制，优化了特征提取过程，显著提高了模型的计算效率和泛化能力。后续本文将介绍卷积神经网络模型的基础，揭示其如何通过卷积层、池化层和全连接层的协同工作，实现对图像数据的深度理解和分类。

作为最重要的深度学习模型之一，卷积神经网络以其出色的特征学习能力在计算机视觉领域占据着重要地位。卷积神经网络的架构设计巧妙，能够自动从数据中学习到层次化的特征表示，在图像分类任务中表现得尤为明显，其一般网络结构框架如图 2.2 所示[2]。卷积神经网络架构主要由三个部分组成：卷积层、子采样层和全连接层。卷积层是网络的核心模块，它由多个卷积核组成，这些卷积核在输入图像上滑动以捕捉局部特征。每个卷积核都是一个小型的权重矩阵，负责提取不同尺寸图像的输入特征。传统的全连接网络由各个单元相

互作用，由于卷积核比图像尺寸小很多，通过不同卷积核可以提取局部的细节特征，大大降低计算成本。除了局部感知外，卷积神经网络的另一大优势是全职共享，不同的通道使用相同权值的卷积核，该方式可以减少求解的参数，有效降低模型的复杂度，提高网络的运行效率。

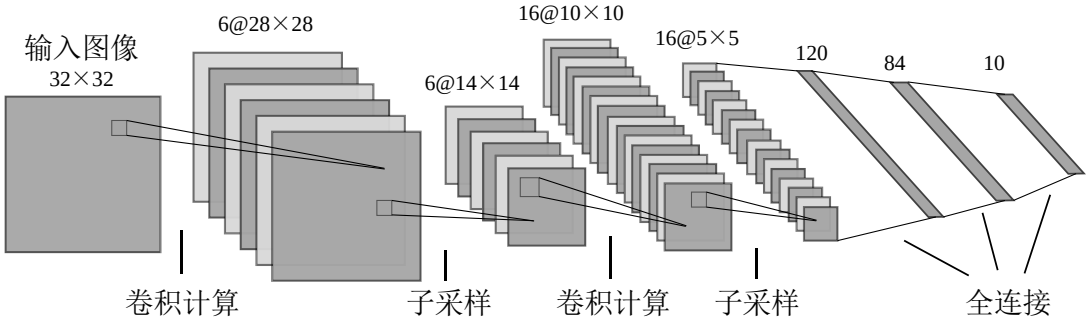


图 2.2 卷积神经网络示意图

卷积的计算过程为在输入图像上通过卷积核从左上角到右下角滑动来进行卷积运算以生成特征图。当输入数据为一个3×3的矩阵，使用2×2大小的卷积核，设置步长为 1，填充大小为 0，最后计算出的2×2矩阵，即为输出的特征图。激活函数通常应用于卷积层之后，帮助模型学习到更加丰富和抽象的特征表示，因其为网络引入了非线性关系，使得模型能够捕捉到输入数据中的复杂信息，从而提高了模型的性能。在卷积神经网络中常用的激活函数是 ReLU 函数[51]，其运算如公式(2.1)所示：

$$\mathrm{ReLU}(\boldsymbol{x})=\begin{cases}0, \boldsymbol{x}<0\\ \boldsymbol{x}, \boldsymbol{x}\geq 0\end{cases} \tag{2.1}$$

ReLU 函数的计算非常直观和简单，在实际训练中具有很快的收敛速度，这使得网络训练更加高效。相比于 Sigmoid 函数和 Tanh 函数，其在正区间上的导数恒为 1，有效解决了梯度消失问题。由于其在$\boldsymbol{x}<0$时输出为 0，这意味它会产生稀疏激活的特征图，网络的大部分神经元可能不会激活，有助于减少计算量以及过拟合，从而提高模型的泛化能力。卷积神经网络包含多个卷积层和激活函数，且卷积层和激活函数一般成对出现，帮助网络引入非线性，学习和执行更复杂的任务。

子采样层是卷积神经网络中的另一个关键组件，也可称其为池化层。池化层将输入图像相邻区域的信息考虑在内并输出该位置的统计特性运算结果，该

操作通过减少数据的空间大小来减少参数数量和计算量。平均池化操作和最大池化操作是两种常用的方式，通过分别提取特征图中的局部平均值和最大值，进一步增强了模型学习输入特征中的尺度不变信息的能力。

全连接层在卷积神经网络中扮演着整合和分类的角色，在网络的末端，全连接层将卷积层和池化层等提取到的局部特征进行整合，通过一维向量的加权求和以及非线性激活函数处理，综合局部特征的提取信息，最终输出分类决策的结果。卷积神经网络的优势是其端到端的学习模式，这意味着从原始图像到最终决策结果的整个处理过程都由网络自动完成，这种自动化的特征学习过程极大地提高了模型的泛化能力和效率。

卷积神经网络是深度学习技术在图像分析和计算机视觉任务中的应用典范。其巨大的成功很大程度上归功于两个关键因素，其一为计算机性能的显著提升，这为处理复杂的数据集合构建大型网络提供了可能；其二为大规模带标注的数据集的出现，这些数据集为训练高精度模型提供了丰富的信息。自 2012 年 AlexNet[52]在 ImageNet 大规模视觉识别挑战赛中取得突破性成绩以来，卷积神经网络模型的设计和优化成为了深度学习领域的焦点。VGGNet[53]通过加深网络层次展示了网络深度对提升性能的显著效果。GoogLeNet[54]引入的 Inception 模块通过并行处理不同尺度的特征图，进一步提高了网络的准确性。ResNet[55]通过引入残差学习解决了网络深度增加时的性能退化问题，实现了网络层次的大幅度扩展。DenseNet[56]的提出则是通过增强网络内部的连接密度，进一步提升了特征传递的效率。

2.2 人脸识别技术

人脸识别技术是一种生物特征识别方法，其通过分析个体脸部的视觉数据来识别或验证人的身份。本质上，人脸识别任务属于图像分类任务，不过人脸识别的任务数据集属于开集，其归属于开集识别任务，训练集和测试集的类别并不一致，这要求网络有更好的泛化能力学习到人脸图像的特征。人脸识别的流程可以分为如下四个步骤，首先为人脸图像的采集与检测，其次为人脸图像的预处理，之后为人脸图像的特征提取，最后是人脸图像的匹配与识别。目前

人脸图像的特征提取以及人脸图像的匹配与识别这两个步骤主要依靠基于深度学习的方法来实现。

2.2.1 人脸识别模型

人脸识别模型是实现上述流程的算法和计算框架。在深度学习时代，FaceNet[57]是经典的人脸识别模型之一，其依据 GoogleNet 基础设计网络结构，将人脸图像的识别、验证和聚类等问题统一映射到特征空间中处理。其本质是将对齐后的人脸图像输入到深度卷积神经网络中，学习人脸图像到欧几里得空间的映射。使用特征向量之间的距离来表示人脸图像之间的相关性，两幅人脸图像特征向量之间的欧式距离越小，表示两幅图像是同一身份的可能性越大。FaceNet 使用基于度量学习的三元组损失（Triplet Loss）作为训练过程中的损失函数，每次训练时随机选取一个锚点样本（Anchor）、一个与锚点样本同身份的正样本（Positive）和一个不同身份的负样本（Negative）构成三元组。该损失函数旨在使得来自同一身份的锚点样本和正样本在特征空间中彼此靠近，而来自不同身份的锚点样本和负样本彼此远离。通过这种方法，其能够学习到一种有效的人脸特征表示，使得即使在不同的光照、姿势和表情下，相同身份的人脸图像仍然具有高度相似的特征表示。三元组损失函数如公式(2.2)：

$$\mathcal{L}=\sum_{i}^{N}\left[\left\|\boldsymbol{f}\left(\boldsymbol{x}_i^a\right)-\boldsymbol{f}\left(\boldsymbol{x}_i^p\right)\right\|_2^2+\boldsymbol{m}-\left\|\boldsymbol{f}\left(\boldsymbol{x}_i^a\right)-\boldsymbol{f}\left(\boldsymbol{x}_i^n\right)\right\|_2^2\right] \tag{2.2}$$

其中，N 表示批量训练的样本数量，$\boldsymbol{f}(\cdot)$ 表示特征提取函数，$\boldsymbol{x}_i^a$ 表示选定的锚点样本，$\boldsymbol{x}_i^p$ 表示正样本，$\boldsymbol{x}_i^n$ 表示负样本，m 表示事先设定好的距离阈值，是一个超参数。FaceNet 的结构如图 2.3 所示，其由一个批量输入层和一个深度卷积神经网络构成，后续使用 $\boldsymbol{L}_2$ 进行规范化得到嵌入向量，最终使用定义好的三元损失函数进行计算来训练网络。

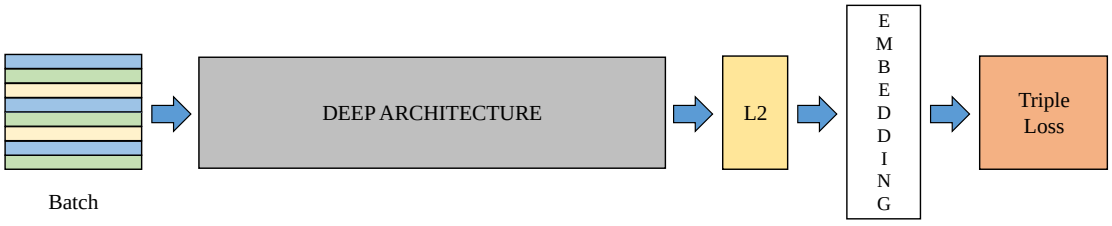


图 2.3 FaceNet 结构示意图

MobileFaceNet[58]是一种针对人脸识别的轻量级网络，其专为移动设备和实时应用设计，通过减少模型大小和计算量，同时保持高准确率，解决了在资源受限的设备上进行人脸识别的挑战。其通过使用深度可分离卷积和轻量级模块化设计，显著减少了模型的参数数量和计算复杂度。人脸图像的中心区域通常比边缘区域包含更多有用信息，因此该模型采用全局深度卷积（Global Deep Convolution, GDConv）替代传统的全局平均池化层，以获得不同位置的重要性权重系数，使网络学到特征图中不同点的权重，提高网络性能。GDConv 层是一个深度卷积层，其内核大小等于输入大小，其输出计算如公式(2.3)：

$$\boldsymbol{G}_{\boldsymbol{m}}=\sum_{\boldsymbol{i},\boldsymbol{j}}\boldsymbol{K}_{\boldsymbol{i},\boldsymbol{j},\boldsymbol{m}}\cdot\boldsymbol{F}_{\boldsymbol{i},\boldsymbol{j},\boldsymbol{m}} \tag{2.3}$$

其中，F 是大小为$\boldsymbol{W}\times\boldsymbol{H}\times\boldsymbol{M}$ 的输入特征图，K 是同样大小的深度卷积核，G 是 $1\times1\times\boldsymbol{M}$ 的输出，m 为通道的索引，$(\boldsymbol{i},\boldsymbol{j})$ 是 F 和 K 的空间坐标位置。该轻量级网络的开发为在移动设备上高效部署人脸识别系统提供强力的支持。

2.2.2 损失函数介绍

随着深度学习的不断发展，图像分类模型的深度也不断增加，分类模型的性能变得饱和，基于深度学习的人脸识别技术也常使用分类模型对人脸图像的特征进行提取，但分类模型所使用的损失函数并不能很好适配人脸识别任务的要求，为此研究者针对人脸识别任务开发了多种损失函数以提升网络的识别准确度。Softmax 损失函数是一种常见的用于图像分类网络的损失函数，其在早期人脸识别任务中也经常使用，Softmax 可由公式(2.4)描述：

$$\mathcal{L}_{\text{softmax}}=-\frac{1}{\boldsymbol{N}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{N}}\log\left(\frac{\boldsymbol{e}^{\boldsymbol{W}_{\boldsymbol{y}_i}^{\boldsymbol{T}}\boldsymbol{x}_i+\boldsymbol{b}_{\boldsymbol{y}_i}}}{\sum_{\boldsymbol{j}=1}^{\boldsymbol{n}}\boldsymbol{e}^{\boldsymbol{W}_{\boldsymbol{j}}^{\boldsymbol{T}}\boldsymbol{x}_i+\boldsymbol{b}_{\boldsymbol{j}}}}\right) \tag{2.4}$$

其中 W 表示权重，b 表示偏置，$\boldsymbol{x}_{\boldsymbol{i}}$ 表示输入样本特征，$\boldsymbol{y}_{\boldsymbol{i}}$ 表示真实标签，N 表示样本数量。Softmax 损失函数的所有类别的概率之和为 1。但是由于 Softmax 函数在特征空间中对样本进行分类，在决策边界附近的样本距离较少，其只考虑样本能否正确分类，对于样本之间的类间距离和类内距离没有涉及，在人脸识别中并没有理想效果。

SphereFace[59]通过将人脸识别中的特征空间映射到超球面角度特征空间之中，使得同一类别的特征向量角度接近，而不同类别的特征向量角度远离。其通过引入一个角度惩罚项 m 在特征向量之间创建一个明确的角度间隔，从而增加了决策边界的余量，最终样本分类结果只与参数θ有关，SphereFace损失函数的计算过程由公式(2.5)所示：

$$\mathcal{L}_{\text{sphereface}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{\|x_i\|\cos\left(m\theta_{y_i},i\right)}}{e^{\|x_i\|\cos\left(m\theta_{y_i},i\right)}+\sum_{j\neq y_i}e^{\|x_i\|\cos\left(\theta_j,i\right)}} \tag{2.5}$$

其中，x_i代表输入，y_i代表标签，N代表样本数，m代表惩罚项。由于角度惩罚项 m 是通过与角度相乘运算的，使每个类别间存在不同的间隔，这也导致不同类别间间隔会有区别，出现在决策空间中有些类别之间间隔比其他类之间间隔大的现象。且由于余弦函数不单调，在训练过程也存在难以优化的问题。

CosFace[60]损失函数弥补了这一点，其将角度惩罚项 m 移到了余弦函数之外，并且还对权重和特征都进行归一化操作，类与类之间也就有了相同的欧氏距离。此外 CosFace 还添加了参数 s，其可设置较大的数值，当超球面较少时可以帮助网络继续优化。相应流程如公式(2.6)：

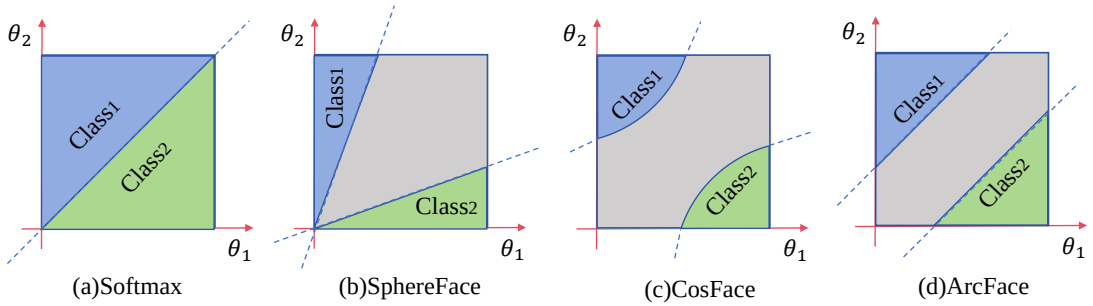
$$\mathcal{L}_{\text{cosface}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(\cos\left(\theta_{y_i},i\right)-m\right)}}{e^{s\left(\cos\left(\theta_{y_i},i\right)-m\right)}+\sum_{j\neq y_i}e^{s\cos\left(\theta_j,i\right)}} \tag{2.6}$$

CosFace 损失函数通过最大化类间余弦边距来训练模型，使人脸识别模型输出有较大区分度。ArcFace[61]损失函数也具有类似的思想。考虑到角度距离对角度的影响比余弦距离影响更大，ArcFace 在 CosFace 基础上将角度惩罚项移到了余弦函数内部，通过与角度相加来进行计算。在标准化后的超球面通过加性角度间隔，在加强不同类别间区分度的同时，保证了同一类别中样本的紧密度。其计算过程由公式(2.7)所示：

$$\mathcal{L}_{\text{arcface}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\left(\theta_{y_i}+m\right)}}{e^{s\cos\left(\theta_{y_i}+m\right)}+\sum_{j=1,j\neq y_i}^{N}e^{s\cos\left(\theta_j\right)}} \tag{2.7}$$

与其他损失函数相比，ArcFace 具有较好的稳定性，且实现较为简单，可以十分容易运用到训练数据集上，上述几种损失函数在进行二分类任务时其决策

边界可视化如图 2.4 所示，可以清晰看出几种函数之间的区别。



(a)Softmax (b)SphereFace (c)CosFace (d)ArcFace

图 2.4 不同损失函数决策边界示意图

2.3 对抗攻击技术

2.3.1 对抗攻击技术基本概念

对抗攻击即通过生成对抗样本对模型进行攻击的过程，通过对输入图像叠加人为精心设计的噪声以生成对抗样本，再将其输入模型中可诱导网络做出错误预测，且该样本几乎无法被人眼察觉。对抗样本的出现体现了神经网络的脆弱性，也是目前基于深度学习网络的一大隐患。本节将依据攻击时有无选定目标和对模型内部信息等的掌握程度来进行对抗攻击的分类。

根据对抗攻击是否有明确目标，可以分为目标攻击和非目标攻击两类。在非目标攻击中，只要所生成的对抗样本导致模型的预测结果与实际真实情况不一致，即可认为攻击成功，此方式不对模型预测的具体类别做要求。相对地，目标攻击则要求对抗样本必须使得模型将其错误地分类为一个预设定的目标，以此判定为一次成功的攻击。相较于非目标性攻击，目标攻击的威胁更大，但非目标攻击在实施时难度更小，且成功率方面表现也更优越。

根据攻击者对模型内部信息获取程度的多少，对抗攻击可以分为白盒攻击和黑盒攻击两类。在白盒攻击场景中，攻击者能够完全了解目标模型的内部架构、标注数据以及训练机制等信息，这使得攻击者能够依据详细信息直接生成对抗样本。然而在现实世界中，模型的详细信息较难获得，白盒攻击的使用范围受到限制。相比之下黑盒攻击则更为普遍，典型的黑盒攻击涉及与目标模型进行有限的交互，通过查询模型对特定输入的输出情况信息，来构建一个替代

模型，借助于对抗样本的迁移特性，可以利用该模型来对目标模型实施攻击。

当前对抗扰动的研究主要分为两个方向，一个是针对特定图像的对抗扰动，另一个是适用于整个数据集的通用对抗扰动。针对特定图像的对抗扰动仅对该图像产生攻击效果，其具有图像唯一性。而通用对抗扰动则追求在更大范围内的数据集上实现攻击效果，有更广泛的应用潜力。

2.3.2 特定图像对抗攻击

FGSM[16]是一种相对容易实现的对抗攻击算法，其核心思想是利用模型自身的梯度信息来生成对抗样本。具体来说，FGSM 通过计算损失函数相对于输入图像的梯度，随后利用梯度符号来构建对抗扰动，其运算过程如公式(2.8)：

$$\boldsymbol{x}' = \boldsymbol{x} + \varepsilon \cdot \operatorname{sign}\left(\nabla_{\boldsymbol{x}} \boldsymbol{J}\left(\boldsymbol{f}\left(\cdot\right), \boldsymbol{y}_{\text{true}}\right)\right) \tag{2.8}$$

其中，$\boldsymbol{x}'$ 表示对抗样本，ε 表示步长，用以控制扰动的强度，$\boldsymbol{J}\left(\cdot\right)$ 表示损失函数，将输入数据沿损失函数增大的方法移动以达到攻击的目的，$\operatorname{sign}\left(\cdot\right)$ 表示符号函数。但该方法生成的扰动不够隐蔽，且在一些情况下不够有效。

DeepFool[19]也是一种基于梯度的白盒攻击算法。其基本理念是探索高维空间中原始图像点邻近的决策边界，并调整图像使这些点越过边界，从而生成对抗扰动。针对分类器 $\boldsymbol{f}\left(\cdot\right)$ 和图像 x，通过找到数据点到决策超平面 $\mathcal{F} = \left\{\boldsymbol{x} : \omega^{T} + \boldsymbol{b} = 0\right\}$ 的最小位移，对抗扰动的最小距离如公式(2.9)所示，使用该方式能生成较小的扰动，这些扰动足以使模型识别出错，又保持了一定的隐蔽性，不过在处理复杂攻击场景时也有其局限性。

$$\delta\left(\boldsymbol{x}\right) = -\frac{\boldsymbol{f}\left(\boldsymbol{x}\right)}{\left\|\omega\right\|^{2}}\omega \tag{2.9}$$

PGD[22]算法是 FGSM 的扩展版本，其通过多步迭代优化来生成对抗样本，每一步都对输入图像进行微小的扰动，直到模型最终识别出现错误，整个扰动生成过程如公式(2.10)所示：

$$\boldsymbol{x}_{\boldsymbol{n}+1} = \prod\left(\boldsymbol{x}_{\boldsymbol{n}} + \alpha \cdot \operatorname{sign}\left(\nabla_{\boldsymbol{x}} \boldsymbol{J}\left(\boldsymbol{x}_{\boldsymbol{n}}, \boldsymbol{y}\right)\right)\right) \text{ s.t. } \boldsymbol{n} \in \left[0, \boldsymbol{K}\right) \tag{2.10}$$

其中 α 表示更新的步长，K 为迭代的次数。FGSM 每次进行单步操作，扰动一次

性获得较大更新，而 PGD 通过 K 轮迭代逐步调整扰动，每一轮的更新幅度都较小。对于非线性模型，PGD 通过逐步逼近，更有可能找到损失函数的局部或全局极值点，且其攻击性也与迭代轮次有关。

Carlini 和 Wagner 提出的迭代攻击算法称为 C&W[23]算法，是一种基于优化的对抗攻击方法，其核心思想通过精心涉及的损失函数和迭代优化过程来找到能够欺骗深度学习模型的对抗样本。C&W 攻击定义了一个损失函数，该函数最小化了目标类别和非目标类别之间的 logits 差异，通过调整损失函数中的参数，攻击者可以控制对抗样本的置信度和扰动的大小，这个过程中允许攻击者在不同的距离度量下进行攻击。其目标函数方程如公式(2.11)所示：

$$\mathcal{L}\left(\boldsymbol{x}+\delta\right)=\max\left(\max\left\{\boldsymbol{Z}\left(\boldsymbol{x}+\delta\right)_{\boldsymbol{i}}:\boldsymbol{i}\neq\boldsymbol{t}\right\}-\boldsymbol{Z}\left(\boldsymbol{x}+\delta\right)_{\boldsymbol{t}},-\boldsymbol{k}\right) \tag{2.11}$$

其中，$\boldsymbol{Z}\left(\boldsymbol{x}\right)_{\boldsymbol{i}}$表示模型 Softmax 层中第$\boldsymbol{i}$类的输出值，$\boldsymbol{t}$为指定的攻击目标标签，$\boldsymbol{k}$为超参数，其作用是控制误分类发生的置信度。通过梯度下降或其他优化算法，根据目标函数迭代更新扰动向量，直到找到满足条件的对抗样本。

2.3.3 通用对抗扰动攻击

Moosavi 等人最早发现 UAP[32]的存在，通过对 DeepFool 算法进行改进，首次实现生成的扰动添加到任意图像上都能导致模型分类错误。其通过不断迭代，将每一个数据点都往跨越决策边界的方向更新，最终获得独立于特定图像的扰动。其扰动的计算过程如公式(2.12)所示：

$$\nabla\delta_i\leftarrow\arg\min_{\delta'}\left\|\delta'\right\|_{\boldsymbol{p}}\quad\text{s.t.}\quad\boldsymbol{f}\left(\boldsymbol{x}_{\boldsymbol{i}}+\delta+\delta'\right)\neq\boldsymbol{f}\left(\boldsymbol{x}_{\boldsymbol{i}}\right) \tag{2.12}$$

针对当前的扰动δ，采用 DeepFool 方式获得一个使得数据点跨过决策边界的最小扰动$\Delta\delta_i$，将原有扰动δ更新为$\delta+\Delta\delta_i$，之后再叠加在下一个样本上重复寻找最小扰动，通过大量数据不断迭代，最终获得总扰动足以攻击整个数据集使得模型对大多数样本识别出错。

Mopuri 等人[37]受生成对抗网络的启发，提出利用生成器方式来构建通用对抗扰动，将符合正态分布的随机向量输入生成器，能够产生一系列通用对抗扰动，在生成器的输出端应用非线性激活函数 Tanh 可以控制扰动的幅度，满足视觉约束标准。该工作使用两部分损失，一方面确保生成扰动的愚弄性，另一方

面旨在提高生成的扰动的多样性。其损失函数如下公式(2.13)所示：

$$\mathcal{L}_{\mathrm{NAG}}=\sum_{n=1}^{B}-\log\left(1-\boldsymbol{q}_{n}\right)-\boldsymbol{d}\left(\boldsymbol{f}^{i}\left(\boldsymbol{x}_{n}+\delta_{n}\right),\boldsymbol{f}^{i}\left(\boldsymbol{x}_{n}+\delta_{n'}\right)\right)\ \text{s.t.}\ \left\|\delta\right\|_{\infty}\leq\xi \tag{2.13}$$

其中，愚弄性损失通过最小化输入数据在真实标签下的置信度 $\boldsymbol{q}_{n}$ 得到，多样性损失则通过最大化叠加在相同原始图像上的同组扰动在卷积层中特征距离来实现。其通过训练生成器的参数能产生具有多样性和攻击性的通用对抗扰动。

除了上述数据有关的通用对抗扰动生成方法，还有一些数据无关的通用对抗扰动生成方案。FFF[33]算法提出在不依靠任何训练数据的情况下训练通用对抗扰动，其优化目标如公式(2.14)：

$$\mathcal{L}_{\mathrm{FFF}}=-\log\left(\prod_{i=1}^{K}\overline{\boldsymbol{f}^{i}}\left(\delta\right)\right)\ \text{s.t.}\ \left\|\delta\right\|_{\infty}\leq\xi \tag{2.14}$$

其中，K表示模型特征总层数，由于不使用训练数据，只需通过最大化扰动δ在模型内部特征层的激活来引导模型对样本的错误分类，通过计算各卷积层的平均激活 $\overline{\boldsymbol{f}^{i}}\left(\delta\right)$，利用各卷积层的输出直接对通用扰动进行优化更新，以此达到不依赖训练数据集的目的。

2.4 面向人脸识别的对抗攻击技术

研究面向人脸识别的对抗攻击技术对于保护人脸隐私也有重要指导意义。通过将对抗扰动叠加在人脸图像上，淡化社交网络上的人脸图像与个人身份之间的关联，以此保护人脸信息的隐私安全。Fawkes[62]算法即是一种采用对抗攻击的方式保护个人面部隐私的手段，如公式(2.15)所示：

$$\max_{\delta}\mathrm{Dist}\left(\Phi\left(\boldsymbol{x}\right),\Phi\left(\boldsymbol{x}\oplus\delta\left(\boldsymbol{x},\boldsymbol{x}_{\mathrm{T}}\right)\right)\right)\ \text{s.t.}\ \delta\left(\boldsymbol{x},\boldsymbol{x}_{\mathrm{T}}\right)\leq\gamma \tag{2.15}$$

其中，$\Phi\left(\cdot\right)$表示特征提取器所提取的图像特征，$\mathrm{Dist}\left(\cdot\right)$为距离函数。首先对用户上传的人脸图像进行像素级修改，通过微小的扰动来改变图像中的特征，最大化对抗样本和原始干净样本之间的距离，同时保证扰动对人眼不可感知。使用这种方式引入对抗扰动，使得图像中的人脸难以被人脸识别算法准确地识别，干扰了人脸识别系统的特征提取和匹配过程，有效保护了隐私。

Zhong 等人也提出一种使用对抗攻击方式保护隐私信息的 OPOM[63]算法，该算法为每个用户生成一个定制的扰动，其核心思想是通过优化每个人脸图像在远离其原始特征空间的方向上生成针对单一身份的通用扰动，该扰动可以针对单个身份的进行人脸隐私保护，每个用户都有单独的针对性通用掩码，以此保护每一个用户在社交平台上的不同人脸图像。

2.5 本章小结

本章详细介绍了所提方案涉及的理论研究及技术基础。首先，对深度学习模型之一的卷积神经网络基础进行介绍，了解深度神经网络内部各层的作用及特征提取的流程。其次，介绍人脸识别模型的损失函数设计，理解针对人脸的对抗扰动的工作原理及机制。最后，将现有对抗攻击技术进行详细介绍，引出面向人脸识别系统的对抗攻击应用，这些工作为本文进一步探讨人脸识别的通用对抗攻击算法提供了重要的研究基础。

第三章　基于空域分析的人脸通用对抗扰动

3.1　研究动机

人脸识别技术被广泛应用于多个领域，随着技术应用的深入，其安全问题变得不容忽视。现有研究表明，人脸识别系统容易受到对抗扰动的影响，这种攻击方式给人脸识别系统的安全性和准确性带来了严峻挑战。目前大部分人脸对抗扰动都是针对特定人脸图像的攻击方式，通用对抗扰动通过在数据集中引入单一扰动，使模型识别结果出错，这种攻击方式在处理大规模数据集时，显著降低了计算成本。为此，本章依据现有的针对自然图像分类任务的通用对抗扰动攻击方案，设计出针对人脸图像的通用对抗扰动生成方案。

为了更有效地生成针对人脸识别任务的通用对抗扰动，应当充分考虑人脸识别系统的工作机制和人脸图像数据集的共性特征。本章考虑到在人脸对齐后，五官的位置在整个人脸区域具有一定的稳定性，且针对人脸识别的可解释性分析实验也证实人脸的有效区域集中在五官等语义区域之内，人脸识别提取信息主要关注这些语义关键区域内的特征，因此本章提出基于语义关键区域控制的人脸通用对抗扰动（Key Regions-Tuned via Flow Field for Facial Universal Adversarial Perturbation, KRT-FUAP），在不同区域叠加不同权重的扰动，让扰动叠加在更有利的区域内，以此区别现有的空域全局扰动，生成具有局部隐蔽性的人脸通用对抗扰动。

3.2　语义关键区域控制的通用对抗扰动框架

3.2.1　问题分析

计算机视觉领域的对抗攻击手段层出不穷，自对抗样本被发现以来，越来越多有效的对抗攻击算法被研究者提出。这些算法能为不同的输入图像产生不同的扰动，但针对新的图像其需要重新生成新的扰动，这极大增加了扰动的产生成本。通用对抗扰动是一类极具代表性的对抗攻击，解决了个体对抗攻击所

面临的问题，该方法从所选数据集中训练一张通用对抗扰动，使其能有效叠加在大量图像上生成相应的对抗样本，极大提高了对抗攻击的效率。

面向图像分类任务的通用对抗扰动技术的目的是使得叠加扰动后的自然图像输出错误的标签，而针对人脸图像的通用对抗扰动的目的为优化扰动使得对抗人脸样本和干净人脸样本的相似性跨过设定的阈值边界，即对抗样本与原始样本不再相似，相应流程如公式(3.1)所示：

$$\text{Similarity}_{x\in X}\left\{F(x+v),F(x)\right\}<t \tag{3.1}$$

其中 X 表示图像的分布，$F(\cdot)$ 定义为目标特征提取器，该提取器会对每一个输入图像 x 输出一个特征向量 $F(x)$，t 为设定的判别相似性阈值，v 表示生成的通用对抗扰动。通用对抗扰动的主要研究目的是在几乎所有采样于 X 分布数据中的 x 上寻找一个扰动 v 来愚弄神经网络，使得生成的对抗样本特征向量与原始图像特征向量相似性差距很大。如图 3.1 所示，人脸识别技术作为一项开集分类任务，其核心机制与传统图像分类任务有较大差异。在人脸识别领域，可解释性分析的侧重点在于深入探讨人脸嵌入向量之间的相似性关系，而非仅仅关注于预测特定的类别标签。由于不同个体的人脸在视觉特征上具有较高相似性，人脸识别技术面临的挑战在于如何精准识别并区分这些细微的特征差异。

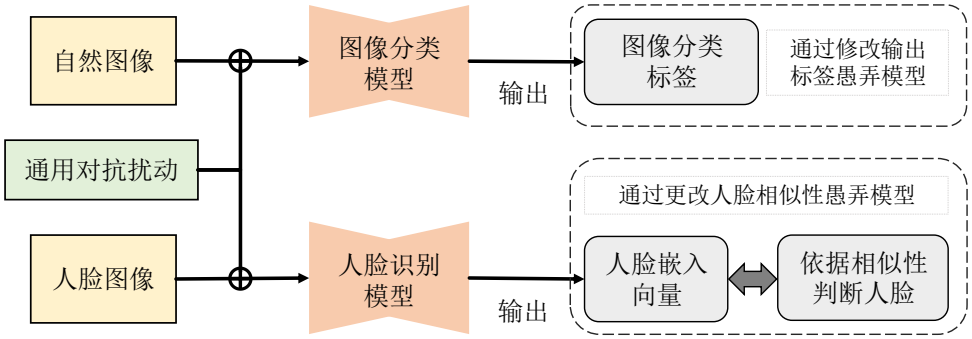


图 3.1 图像分类与人脸识别对抗攻击示意图

在进行可解释性分析的过程中，研究者会关注不同人脸区域对于嵌入向量提取过程的具体贡献和影响。鉴于此，针对人脸识别的通用对抗扰动研究，关键在于挖掘并利用数据集中不同身份人脸所共有的特征。本研究将焦点集中在人脸的关键区域，因为这些区域的特征对于最终嵌入向量的相似性判断起着决定性作用。为了实现这一目标，本章采用了一种可学习的流场方式来微调人脸

关键区域的掩码。通过对这些掩码进行精细修改，能够在空域中叠加不同权重的扰动，确保扰动主要集中在对人脸识别模型影响最大的区域。

本章节的整体方法框架如图 3.2 所示，第一部分为人脸关键区域提取，第二部分为利用流场微调关键区域来生成对抗样本。首先通过人脸关键点检测获取部分人脸图像的局部关键点，用凸包算法获取部分人脸图像的关键区域后取交集得到针对数据集的关键区域掩码。随后，初始化空间变换流场和噪声，用可学习流场控制关键区域掩码的空间变换获得一个更有效的关键区域，依据该区域的不同位置控制扰动的权重获得通用对抗扰动。最后将扰动叠加在干净图像上，分别通过目标模型和 VGG 模型获取对抗性损失和隐蔽性损失，不断迭代更新直到满足一定的标准。这种方法不仅提高了攻击的有效性，同时也增强了扰动的隐蔽性，从而在不引起察觉的情况下实现对人脸识别系统的攻击。

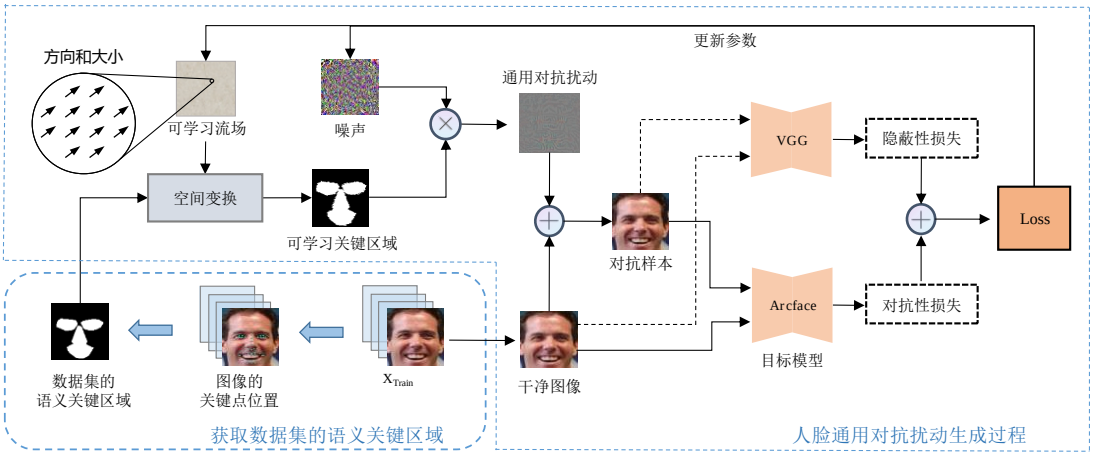


图 3.2 基于语义关键区域控制的扰动生成方案整体流程框图

3.2.2 流场微调关键区域设计

在人脸识别的可解释性任务中，Mery 等人[64]通过对人脸数据进行区域性遮挡，探究了不同人脸语义区域对于提取人脸嵌入向量的影响。不同关键区域在嵌入向量的不同维度上会表现出区域性特征，这意味着不同人脸的语义区域特征是影响人脸识别准确性的主要因素。为了符合面向人脸识别的通用对抗扰动任务的特点，需要寻找整个人脸数据集的普遍规律，本章将重点放在人脸的关键区域上。大多数人脸识别系统在预处理阶段会对人脸图像进行对齐操作，对齐完成后最终的人脸数据集的五官区域会具有固定的分布，这些五官关键区域

的语义分布位置将成为施加通用对抗扰动的理想位置。

人脸关键区域提取流程：为了获得人脸关键区域的掩码位置，本章收集对齐后的人脸数据集 $\boldsymbol{X}=\{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n\}$，取该数据集中的小批量数据提取人脸脸部的 68 个关键点，按照关键点序号的坐标将序号 26 到序号 36 区域划分为鼻子区域，同样的方式找出眼睛和鼻子的编号顺序获得相应区域，运用凸包算法计算得出各掩码区域，最终的语义区域即为这三类五官语义区域的叠加，相应的第 $\boldsymbol{i}$ 张人脸图像提取掩码 M_i 计算如公式(3.2)所示：

$$\mathrm{M}_i=\mathrm{H}_1\left(\mathrm{x}_i\right)+\mathrm{H}_2\left(\mathrm{x}_i\right)+\mathrm{H}_3\left(\mathrm{x}_i\right) \tag{3.2}$$

其中，$\mathrm{H}_1(\cdot)$，$\mathrm{H}_2(\cdot)$ 和 $\mathrm{H}_3(\cdot)$ 分别为不同凸包算法获得的眼睛、鼻子和嘴巴区域，因为不同人脸获得的掩码区域并不重合，本章通过计算该批量人脸图像关键区域的交集部分作为整个人脸数据集提取出的语义特征区域，关键区域 M_{o} 计算如公式(3.3)所示。

$$\mathrm{M}_{\mathrm{o}}=\boldsymbol{J}\left\{\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_n\right\} \tag{3.3}$$

其中 $\boldsymbol{J}\{\cdot\}$ 表示进行人脸关键区域的交集运算，该掩码的尺寸和人脸对齐后尺寸一致，本章针对人脸脸部关键区域和非关键区域叠加不同权重的扰动，以此提升通用对抗扰动攻击的有效性和隐蔽性。相应关键区域提取流程如图 3.3 所示

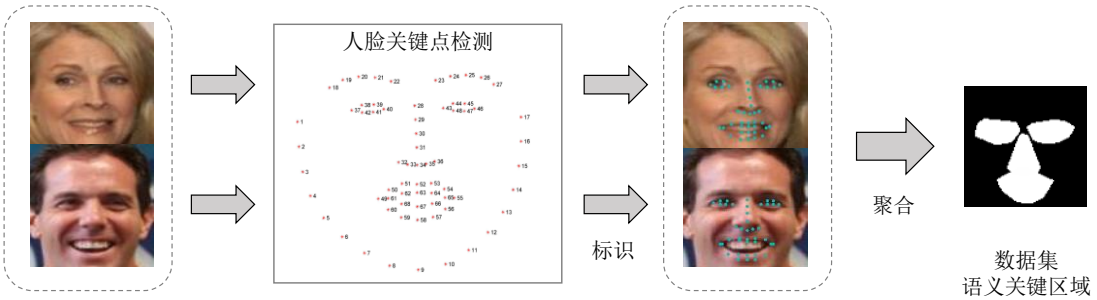


图 3.3 人脸关键点标识及语义关键区域聚合示意图

流场微调语义区域的空间变换：空间变换对抗样本由 Xiao 等[65]提出并运用在自然图像的对抗攻击中，通过对输入像素进行一定微小的位移，以此成功欺骗目标网络模型。但是上述空间变换对抗样本研究中涉及的变换方法是针对特定的原始干净样本采用的操作方式，在通用对抗扰动任务研究中，无法针对整个数据集产生一个通用的空间变换流场。因此，本章将目标转向人脸语义关键

区域掩码上。关键区域掩码是依据整个人脸数据集提取出的信息，该信息可以间接表征整个人脸数据集的语义关键区域分布，作为一种先验的全局信息进行利用。但该语义区域缺少了数据集中每一个人脸身份个体的独有特征，为此本章通过设计一个可学习的空间变换流场来对该关键区域掩码进行微小的位置调整，该可变流场的学习过程将训练集中的每一个人脸身份特征都纳入运算过程，弥补了最初提取出的语义关键区域缺失的个体特征，有利于更好地达到通用对抗扰动针对整个人脸数据集的目的。相应的空间变换如公式(3.4)：

$$\mathrm{M_f} = \mathrm{Flow}\left\{ \hat{\boldsymbol{f}}_{\mathrm{flow}}, \mathrm{M_o} \right\} \tag{3.4}$$

其中，$\mathrm{Flow}\{\cdot\}$ 为流场空间变换函数，$\hat{\boldsymbol{f}}_{\mathrm{flow}}$ 为可学习的流场，$\mathrm{M_f}$ 为经过空间变换后的掩码区域。流场表示一种空间变换的规律，其规定了图像中不同像素点的空间改变量大小。定义空间坐标转换流场为 $\boldsymbol{f}_{\mathrm{flow}} \in [-1,1]^{2\times h\times w}$，其中 $\boldsymbol{f}_{\mathrm{flow}}^{(i)}$ 是图像中第 $\boldsymbol{i}$ 个像素点的位移规则，表示该像素坐标改变的方向和大小。具体来说，本章中需要进行空间变换的图像为数据集中提取出的人脸语义区域掩码，该初始掩码图像的第 $\boldsymbol{i}$ 个像素点的二维坐标可以表示为 $\left(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)}\right)$，针对该像素点的流场空间变换量为 $\boldsymbol{f}_{\mathrm{flow}}^{(i)} = \left(\Delta\boldsymbol{a}^{(i)}, \Delta\boldsymbol{b}^{(i)}\right)$。$\left(\hat{\boldsymbol{a}}^{(i)}, \hat{\boldsymbol{b}}^{(i)}\right)$ 表示空间转换后掩码的第 $\boldsymbol{i}$ 个像素点对应的坐标。其中本章定义流场向量是由转换后掩码图像导向初始掩码图像，其坐标与原图像中相同位置像素点坐标的关系如公式(3.5)所示：

$$\left(\boldsymbol{a}^{(i)}, \boldsymbol{b}^{(i)}\right) = \left(\hat{\boldsymbol{a}}^{(i)} + \Delta\boldsymbol{a}^{(i)}, \hat{\boldsymbol{b}}^{(i)} + \Delta\boldsymbol{b}^{(i)}\right) \tag{3.5}$$

鉴于可学习的空间转换流场内的参数 $\left(\Delta\boldsymbol{a}^{(i)}, \Delta\boldsymbol{b}^{(i)}\right)$ 并不一定都是整数，而网格图像中各点的坐标只接受整数，这意味着无法依据变换后的坐标直接匹配对应位置的像素值。本章使用双线性插值的方式计算非整数坐标的像素值，同时这种方式也可以保证在训练过程中流场参数可微。

给定流场运算后图像的第 $\boldsymbol{i}$ 个像素点 $\mathrm{M_f^{(i)}}$，其坐标为 $\left(\hat{\boldsymbol{a}}^{(i)}, \hat{\boldsymbol{b}}^{(i)}\right)$，考虑到其坐标可能不是整数，通过取整操作获得与之相邻的四个像素点的坐标，分别为 $\left(\left\lfloor \boldsymbol{a}^{(i)} \right\rfloor, \left\lfloor \boldsymbol{b}^{(i)} \right\rfloor\right)$，$\left(\left\lfloor \boldsymbol{a}^{(i)} \right\rfloor + 1, \left\lfloor \boldsymbol{b}^{(i)} \right\rfloor\right)$，$\left(\left\lfloor \boldsymbol{a}^{(i)} \right\rfloor, \left\lfloor \boldsymbol{b}^{(i)} \right\rfloor + 1\right)$ 和 $\left(\left\lfloor \boldsymbol{a}^{(i)} \right\rfloor + 1, \left\lfloor \boldsymbol{b}^{(i)} \right\rfloor + 1\right)$。并用

$\boldsymbol{N}\left(\boldsymbol{a}^{(i)},\boldsymbol{b}^{(i)}\right)$ 表示上述坐标对应的像素点的集合。依据邻域集合中各点的像素值，采用双线性插值的方式，更新图像中全部的像素值，获得改变过后的关键区域掩码。插值公式如公式(3.6)所示：

$$\mathrm{M}_{\mathrm{f}}^{(i)}=\sum_{\boldsymbol{j}\in\boldsymbol{N}\left(\boldsymbol{a}^{(i)},\boldsymbol{b}^{(i)}\right)}\mathrm{M}_{\mathrm{o}}^{(j)}\left(1-\left|\boldsymbol{a}^{(i)}-\boldsymbol{a}^{(j)}\right|\right)\left(1-\left|\boldsymbol{b}^{(i)}-\boldsymbol{b}^{(j)}\right|\right) \tag{3.6}$$

同时为了防止经过空间变换后的掩码与从数据集中提取出的掩码位置差距过大，本章在每次迭代更新时约束流场中各点的位移方向和大小，控制其不超过设定的 $[-1,1]$ 阈值，以免关键区域掩码变化过大丢失过多数据集的全局特征。最终计算得出的掩码相比于初始提取出的掩码位置，考虑了更多数据集中的个体特征，同时纳入了损失函数中提升隐蔽性的因素，所述方式对于后续生成的通用对抗扰动有更佳的表现效果。

3.2.3 通用对抗扰动的生成

通用对抗扰动 v 由获得的关键区域掩码 M_{f} 和可学习噪声 n 生成。为了提升生成的通用对抗扰动的隐蔽性，相比于以前常用方法在全局直接叠加扰动，本方案将非关键区域叠加扰动的强度设置为关键区域扰动强度的一半，以此确保有效噪声集中在关键区域，非关键区域叠加更少扰动可以更好提高扰动的隐蔽性。最终生成的对抗样本 $\mathrm{x}_{\mathrm{adv}}$ 如公式(3.7)所示：

$$\mathrm{x}_{\mathrm{adv}}=\mathrm{x}+\mathrm{v}=\mathrm{x}+\mathrm{Mask}\left\{\hat{\mathrm{n}},\mathrm{M}_{\mathrm{f}}\right\} \tag{3.7}$$

其中 v 表示生成的通用对抗扰动，n 表示学习到的噪声，$\mathrm{Mask}\{\cdot\}$ 表示使用掩码和可学习噪声产生扰动的函数，为确保非关键区域的扰动强度为关键区域扰动强度的一半，将掩码区域的值归一化到区间 $[1/2,1]$ 内。该方案的整个优化过程可用公式(3.8)表示：

$$\left(\hat{\mathrm{n}},\hat{\boldsymbol{f}}_{\mathrm{flow}}\right)=\arg\min_{\mathrm{n},\boldsymbol{f}_{\mathrm{flow}}}\left[\mathcal{L}_{\mathrm{adv}}\left(\mathrm{x}+\mathrm{v},\mathrm{x}\right)+\lambda\mathcal{L}_{\mathrm{ste}}\left(\mathrm{x}+\mathrm{v},\mathrm{x}\right)\right]\ \text{s.t.}\ \left\|\mathrm{v}\right\|_{\mathrm{p}}\le\xi \tag{3.8}$$

其中 n 和 $\boldsymbol{f}_{\mathrm{flow}}$ 表示可学习的噪声和空间变换流场，$\mathcal{L}_{\mathrm{adv}}$ 表示对抗性损失，$\mathcal{L}_{\mathrm{ste}}$ 表示隐蔽性损失，λ 控制两者之间的平衡，ξ 是控制扰动强度大小的参数，通用对抗扰动生成的具体流程如算法 1 所示。

算法 1：语义关键区域控制的扰动生成算法

输入：预处理训练集 X，随机噪声 n，人脸识别目标网络 $\mathrm{F}(\cdot)$，愚弄率 δ，扰动的无穷范数 ξ，决策阈值 t。

输出：通用对抗扰动 v，可学习流场 $\boldsymbol{f}_{\mathrm{flow}}$。

1：随机初始化 $(\mathrm{n}, \boldsymbol{f}_{\mathrm{flow}})$
2：从 X 中获得初始掩码 M_{o}
3：**While** 愚弄率 $< \delta$ **do**
4： **for** x_i in X **do**
5： 空域变换：$\mathrm{M}_{\mathrm{f}} \leftarrow \mathrm{Flow}\left\{\boldsymbol{f}_{\mathrm{flow}}, \mathrm{M}_{\mathrm{o}}\right\}$
6： 获得 $\mathrm{v} \leftarrow \mathrm{Mask}\left\{\mathrm{n}, \mathrm{M}_{\mathrm{f}}\right\}$
7： **if** $\mathrm{Similarity}\left\{\mathrm{F}\left(\mathrm{x}_i+\mathrm{v}\right), \mathrm{F}\left(\mathrm{x}_i\right)\right\} > \mathrm{t}$ **then**
8： $\left(\Delta \mathrm{n}, \Delta \boldsymbol{f}_{\mathrm{flow}}\right) \leftarrow \arg \min_{\left(\mathrm{n}, \boldsymbol{f}_{\mathrm{flow}}\right)}\left\|\left(\mathrm{n}, \boldsymbol{f}_{\mathrm{flow}}\right)\right\|_2$
9： s.t. $\mathrm{Similarity}\left\{\mathrm{F}\left(\mathrm{x}_i+\mathrm{v}\right), \mathrm{F}\left(\mathrm{x}_i\right)\right\} \leq \mathrm{t}$
10： 更新噪声：$\mathrm{n} \leftarrow \mathrm{n}+\Delta \mathrm{n}$
11： 更新流场：$\boldsymbol{f}_{\mathrm{flow}} \leftarrow \boldsymbol{f}_{\mathrm{flow}}+\Delta \boldsymbol{f}_{\mathrm{flow}}$
12： **end if**
13： 裁剪 v 以满足无穷范数
14： **end for**
15：**end while**
16：**return** v

3.2.4 损失函数的设计

本方案的损失函数设计包括两个部分，分别通过针对对抗性和隐蔽性两个角度来不断优化生成的扰动。对抗性损失的作用是控制对抗性扰动的攻击性能。在图像分类任务中，通过给定预测标签的方式来使用交叉熵损失进行对抗攻击，控制最终计算出的各类别的概率来攻击分类模型。而在人脸识别任务中，为了衡量干净样本和对抗样本之间的差异性，研究者们会采用余弦相似度或者欧氏距离等度量方法。这些度量方法能够量化样本间的相似性，从而为对抗样本的生成提供指导。

本方案使用的对抗性损失方法是基于 ODFA[66]方式，其对传统对抗性损失函数进行了一定的改进，不依赖于中间的度量值，而是直接在特征空间内对原始的干净样本和生成的对抗样本两者的特征向量进行操作。具体而言，该对抗攻击的目标是通过优化扰动来控制特征向量的方向，使得对抗样本的特征向量

方向与原始样本的特征向量方向形成显著的偏差。这种方式的核心在于通过将对抗样本的特征向量推向与原始样本特征向量相反的方向，以此有效增加两者之间的差异性，从而绕开中间度量值的计算使得对抗样本在特征空间中远离原始样本。对抗性损失如公式(3.9)所示：

$$\mathcal{L}_{\text{adv}} = \sum_{\mathrm{x}^{(i)} \in \boldsymbol{D}} \left(\frac{\mathrm{F}\left(\mathrm{x}^{(i)}\right)}{\left\|\mathrm{F}\left(\mathrm{x}^{(i)}\right)\right\|_2} + \frac{\mathrm{F}\left(\mathrm{x}_{\text{adv}}^{(i)}\right)}{\left\|\mathrm{F}\left(\mathrm{x}_{\text{adv}}^{(i)}\right)\right\|_2} \right)^2 \tag{3.9}$$

其中 D 为训练样本集，i 表示图像在小批量数据集中的编号，$\mathrm{F}\left(\mathrm{x}^{(i)}\right)$ 和 $\mathrm{F}\left(\mathrm{x}_{\text{adv}}^{(i)}\right)$ 表示原始干净样本和生成的对抗样本从目标模型中提取出的特征向量，为了使得损失函数不断变小，对抗样本的特征向量将被限制在与干净样本特征向量相反的方向。该方式为对抗样本的特征向量选择了优化方向，这相当于该方向与原方向之间的余弦相似度计算值为-1，代表余弦相似度度量最低得分的边界，通过约束该方向有效地规避了余弦相似度地计算，直接将对抗样本的特征向量往反方向推移，从而也能达到实现对抗攻击的目的。

隐蔽性损失的设计宗旨在于确保生成的对抗扰动对于观察者而言难以察觉。这种损失函数的引入是为了在不显著改变输入数据外观的同时，使人脸识别模型的最终输出结果产生错误。VGG[53]网络全称为 Visual Geometry Group 网络，在图像识别和其他计算机视觉任务中发挥着重要作用，能够学习到丰富的图像特征表示。本任务采用 VGG 网络的浅层输出来捕捉正常图像和对抗样本之间的低级视觉特征，如边缘和纹理，进而计算隐蔽性损失以调控通用对抗扰动的隐蔽性。相应流程如图 3.4 所示。

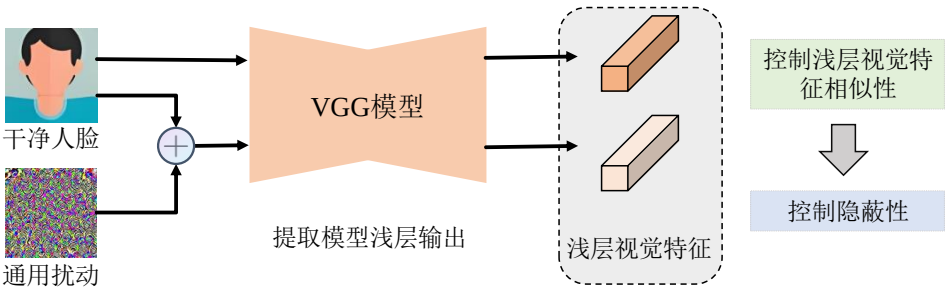


图 3.4 通用对抗扰动的隐蔽性损失调控示意图

首先将对抗样本和干净图片输入至 VGG 网络，并利用网络的浅层所提供的输出信息。这些浅层网络主要负责捕捉图像的纹理和边缘等低级特征，通过比

较两者的低级特征，量化两者之间的差异，这一差异的度量构成了本节的隐蔽性损失，其直接关联到生成的对抗扰动的隐蔽程度。隐蔽性损失的计算旨在最小化对抗样本与干净图片在低级特征上的分布差异，从而推动生成的对抗扰动在视觉上的不可感知性。其隐蔽性损失计算表达式如公式(3.10)所示：

$$\mathcal{L}_{\text{ste}} = \sum_{\mathrm{x}^{(i)} \in \boldsymbol{D}} \left(\left\| \varphi_j\left(\mathrm{x}^{(i)}\right), \varphi_j\left(\mathrm{x}_{\text{adv}}^{(i)}\right) \right\|_2 \right) \tag{3.10}$$

其中，$\varphi_j(\cdot)$为VGG网络中第 $\boldsymbol{j}$ 层的特征映射输出。通过控制对抗样本和干净图片在纹理边缘等低维特征的相似性，提高最终生成的通用对抗扰动的不可感知性。本方案最终的损失函数设置综合了对抗性损失和隐蔽性损失，最终损失函数 $\mathcal{L}_{\text{all}}$ 表达式如公式(3.11)所示：

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{adv}} + \lambda \mathcal{L}_{\text{ste}} \tag{3.11}$$

3.3 实验与分析

在本小节设置了全面的实验来验证所提出方法的有效性。首先提供了实验的整体参数设置，后续验证了语义关键区域对人脸识别系统准确率的影响，并且将本章提出的扰动生成框架与现有的分类任务的通用对抗扰动方法进行比较，结果表明本方案在具备一定攻击性的同时有较好的隐蔽性。之后进行的黑盒测试表明本方案对黑盒方式的对抗攻击有一定的攻击有效性。最后针对方案中不同模块的有效性进行消融实验，验证多种因素对所提方法有效性的影响。

3.3.1 实验设置

本方案所有实验使用单张 GTX TITAN XP 显卡进行加速运算，显存容量为12G，采用 Pytorch 深度学习框架进行编程，Pytorch 版本为 1.12.0，Python 版本为 3.8.16，通过上述软硬件配置，充分利用设备性能，提高实验效率和测试精度。本方案选用常见的开源人脸图像数据集 LFW[67]和 CASIA-WebFace[68]生成面向人脸识别的通用对抗扰动。其中 LFW 收录了 13000 多张从互联网上搜集来的面部图像，共有用户的身份 5749 人，每张图像都标注了对应的人物姓名。CASIA-WebFace 数据集包含 494414 张人脸图像，涵盖了 10575 个不同身份的人

物，每个人脸身份有着多张个人图片。本方案将所有数据进行检测和对齐后，输出的图像固定尺寸为112×112×3。使用的人脸识别模型为 ArcFace 模型，通过比较两张人脸图像提取出的特征相似性来判别是否为同一身份，为此本文使用的两个数据集都由相同身份人脸图像对的形式构成。针对两个不同数据集，训练集采用 6000 对人脸图像构成，测试集采用 3000 对人脸图像来进行通用对抗扰动的生成。在以 Arcface 模型预训练的 IResNet50[69]、MobileFaceNet[58]和 MobileNetV1[70]三个主干提取网络上，使用上述两个人脸数据集的初始参数如下表 3.1 和表 3.2 所示：

表 3.1 基于 LFW 数据集的不同预训练模型的基本参数

数据集	LFW		
模型结构	IResNet50	MobileFaceNet	MobileNetV1
准确率(%)	99.25	99.11	99.31

表 3.2 基于 CASIA-WebFace 数据集的不同预训练模型的基本参数

数据集	CASIA-WebFace		
模型结构	IResNet50	MobileFaceNet	MobileNetV1
准确率(%)	98.90	99.01	99.10

在通用对抗扰动的训练过程中，输入的图像数据的像素值范围都归一化在$[-1,1]$区间内，选择无穷范数作为扰动的强度限制方式，其扰动强度ξ设为 0.08，训练的批处理大小设置为 10，实验选择 Adam[71]优化器优化扰动，扰动的学习率设定为 0.01，损失函数的权重参数λ设置为 0.05。对于不同的数据集和主干提取网络获得的扰动，使用结构相似性指数[72]（Structural Similarity Index Measure, SSIM）和峰值信噪比[73]（Peak Signal-to-Noise Ratio, PSNR）两个客观隐蔽性指标来评估扰动的隐蔽性性能。

3.3.2 语义关键区域有效性验证

本方案的核心思想认为将扰动叠加在语义的关键区域会对人脸识别模型的识别性能产生更大的影响，为了验证这一想法，本章设计了一个展示关键区域和非关键区域对人脸识别准确率影响的实验，使用不同的局域噪声叠加在数据集上，并按照图 3.5 所示方式测试目标人脸识别模型的准确率。

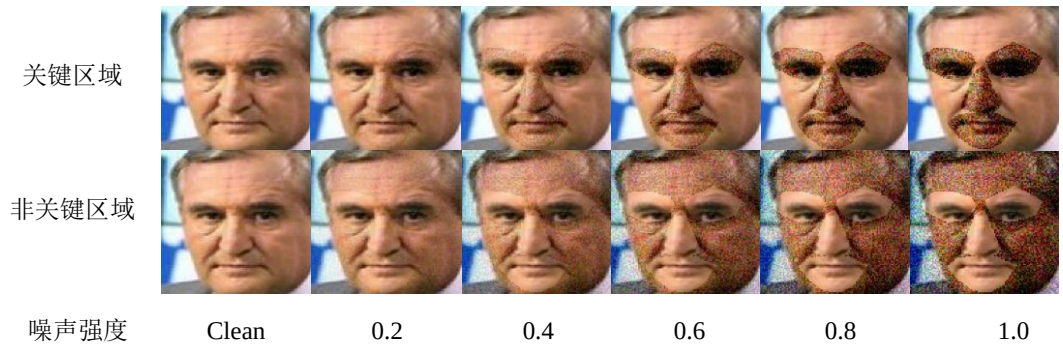
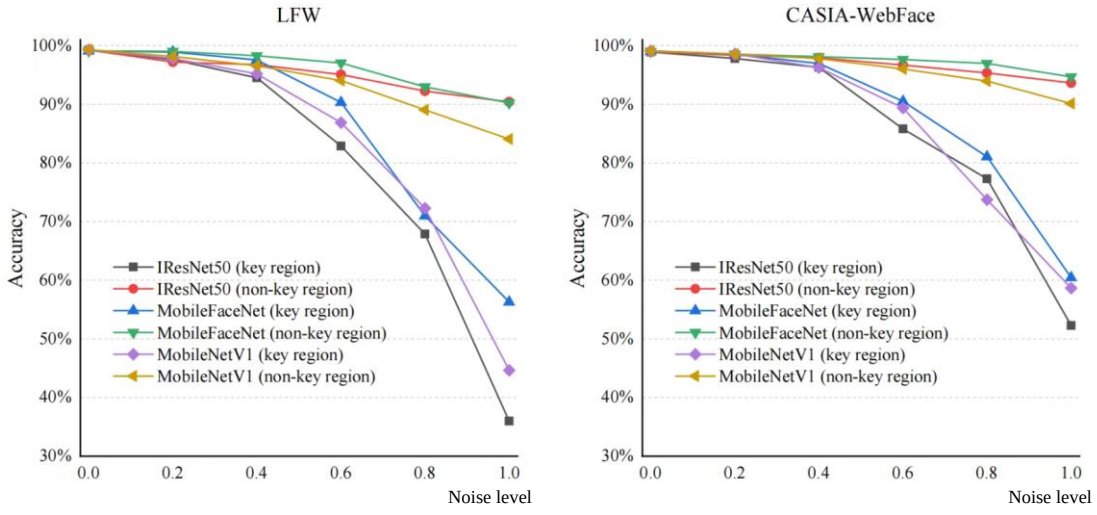


图 3.5 语义关键区域和非关键区域叠加不同噪声的可视化示意图

针对遮挡语义关键区域和遮挡非关键区域两种方式，使用 LFW 和 CASIA-WebFace 两个数据集在 IResNet50、MobileFaceNet 和 MobileNetV1 三个主干提取网络上测试人脸识别模型识别准确率的变化趋势，通过控制用于遮挡的随机噪声的强度，得出不同强度噪声对不同区域的影响程度。从两个数据集中随机选择 2000 对相同身份的人脸图像，通过在这一对图像的其中一张图片的不同区域叠加一定强度的噪声，噪声设置为随机高斯噪声，噪声强度设置从 0 到 1 增强，随着设定叠加的噪声强度变大，使用两个数据集在三个主干网络上的测试结果如图 3.6 所示。



(a)在LFW上准确率变化　　(b)在CASIA-WebFace上准确率变化

图 3.6 人脸识别准确率随着叠加噪声强度变化趋势

由图 3.6 测得的人脸识别准确率趋势变化曲线可知，不论是针对 LFW 数据集还是针对 CASIA-WebFace 数据集，在噪声强度为 0，即不添加任何噪声在人脸图像上时，在 IResNet50、MobileFaceNet 和 MobileNetV1 这三个主干网络上测得的人脸识别准确率都接近 100%。随着噪声强度逐渐上升，不论是遮挡关键

语义区域的实验准确率还是遮挡非关键语义区域的实验准确率都有了不同程度的下降。当噪声强度达到 1 时，使用 LFW 数据集在 MobileNetV1 网络上遮挡非关键语义区域测试得到的准确率下降幅度达到了 15%，而在其他两个主干网络中的准确率也有 10%的下降，说明在人脸识别任务中，非关键语义区域的特征信息对最终的识别结果有一定的影响。但相比于遮挡关键语义区域测试得到的准确率结果，当噪声强度达到 0.6 时，人脸识别模型的准确率就开始有了明显的下降，随着后续噪声强度的增加，识别准确率下降的幅度越来越大。在 CASIA-WebFace 数据集中同样可以找到相似的规律，这说明人脸的关键语义区域相比于非关键区域含有的特征信息对人脸识别最终判别结果有更大的权重，也进一步验证了本章依据人脸关键区域叠加扰动这一方式的合理性。

3.3.3 攻击性评估

在面向人脸识别的通用对抗扰动任务中，攻击性用于评估通用对抗扰动生成策略能否有效欺骗正常的人脸识别模型，是衡量对抗扰动有效性的重要指标。本实验在 IResNet50 和 MobileFaceNet 网络上测试了使用 LFW 和 CASIA-WebFace 两个数据集训练人脸通用对抗扰动的性能，实验结果在各个测试集上实现了大约 80%的攻击成功率。

为了强调本方案精心生成的扰动的有效性，本实验使用一组随机生成的噪声进行扰动的叠加操作，并测试使用随机噪声后模型的攻击成功率情况。由于目前针对人脸识别任务的通用对抗扰动生成方案尚缺乏充足研究，本实验将提出的方案与已有的其他领域的对抗扰动生成方案进行对比，其中 UAP[32]和 FG-UAP[74]是针对自然图片分类任务的方案，前者为经典的通用对抗扰动生成方案，通过控制对抗样本跨过决策边界实现攻击，后者通过控制最后阶段的特征改变实现攻击。FTGAP[75]是针对纹理图片的方案，通过在频域上叠加扰动并在频域控制扰动强度实现攻击。虽然上述方法最初是为其他任务设计的，本实验通过一些改动使之契合人脸识别任务。相关的攻击性结果见表 3.3 所示，随机生成的噪声与其他精心设计的扰动相比，攻击性相差巨大，这一点可以证明精心设计的扰动对人脸识别模型危害性更大。UAP 是在空域全局产生扰动的方式，FG-UAP 是针对损失函数进行改进过的空域扰动生成方式，两者都是仅考虑空

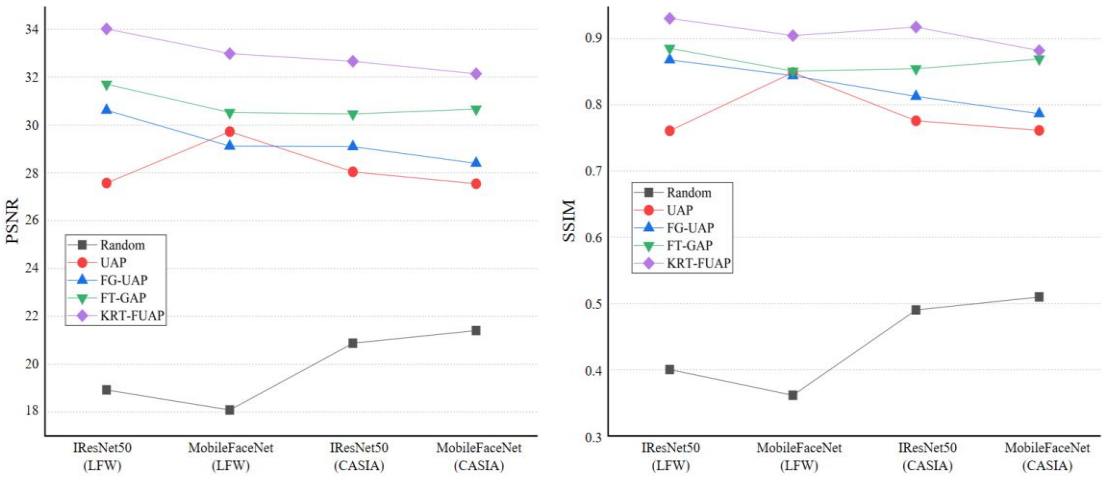
域的简单攻击模式，而 FTGAP 扰动生成方式考虑了频域维度方面的信息，因此其攻击成功率指标略优于两种空域扰动。本方案依据人脸识别可解释性方式，通过在语义关键区域产生扰动，实验结果证明由此方式产生扰动攻击性更佳。

表 3.3 不同扰动生成方案的攻击性对比结果（单位：%）

数据集	网络结构	Random	UAP	FG-UAP	FTGAP	KRT-FUAP
LFW	IResnet50	20.8	76.6	80.4	82.1	81.9
	MobileFaceNet	23.2	79.1	78.8	79.4	80.1
CASIA-WebFace	IResnet50	37.1	71.2	74.7	76.3	78.4
	MobileFaceNet	30.3	76.3	77.8	78.4	80.2
均值		27.9	75.8	77.9	79.1	80.1

3.3.4 隐蔽性评估

通用对抗扰动任务的隐蔽性用于评估该扰动生成策略能否有效逃避过人眼视觉的检验，是衡量对抗样本是否可感知的重要标准，一般会使用一些客观的隐蔽性指标来度量。为了度量本方案生成的对抗扰动的隐蔽性，除了使用正常的人眼主观观测外，还使用了 PSNR 和 SSIM 两种客观的图像质量指标来衡量，使用这些指标有助于量化生成的通用对抗扰动的隐蔽性性能。最终的指标也表明当把通用对抗扰动叠加在人脸图像上时，能够表现出良好的隐蔽性性能。如图 3.7 所示为在不同情况下测得 PSNR 和 SSIM 的对比情况。



(a)不同情况下PSNR比较　　(b)不同情况下SSIM比较

图 3.7 不同情况下对抗样本的隐蔽性对比示意图

使用随机生成的噪声图像测得的隐蔽性指标与其他精心设计的扰动隐蔽性指标相差巨大，UAP 和 FG-UAP 两者都是空域扰动，隐蔽性指标都仅依托于范数的约束，而 FTGAP 扰动隐蔽性达成的手段是直接在频域限制扰动的强度，因

此相比之下，频域信息因具有一定的不可察觉性，使 FTGAP 的隐蔽性指标略优于两种空域扰动。

本方案提出的 KRT-FUAP 不仅精心设计了关键语义区域的扰动叠加方式，还采用了有效的隐蔽性损失和对抗性损失来控制最终生成的通用扰动，这使得 KRT-FUAP 在与其他方法进行比较时，能在攻击性相差不大的情况下取得更大的隐蔽性的提升。上述方式相应的可视化结果如图 3.8 所示，其中（a,f,k）表示三个干净的人脸身份图像，（b,g,l）是使用 UAP 方式生成的人脸对抗样本的可视化图像，（c,h,m）是使用 FG-UAP 生成的人脸对抗样本的可视化图像，（d,i,n）是使用 FTGAP 生成的人脸对抗样本的可视化图像，最后的（e,j,o）则是使用 KRT-FUAP 方案生成的人脸对抗样本可视化图像。

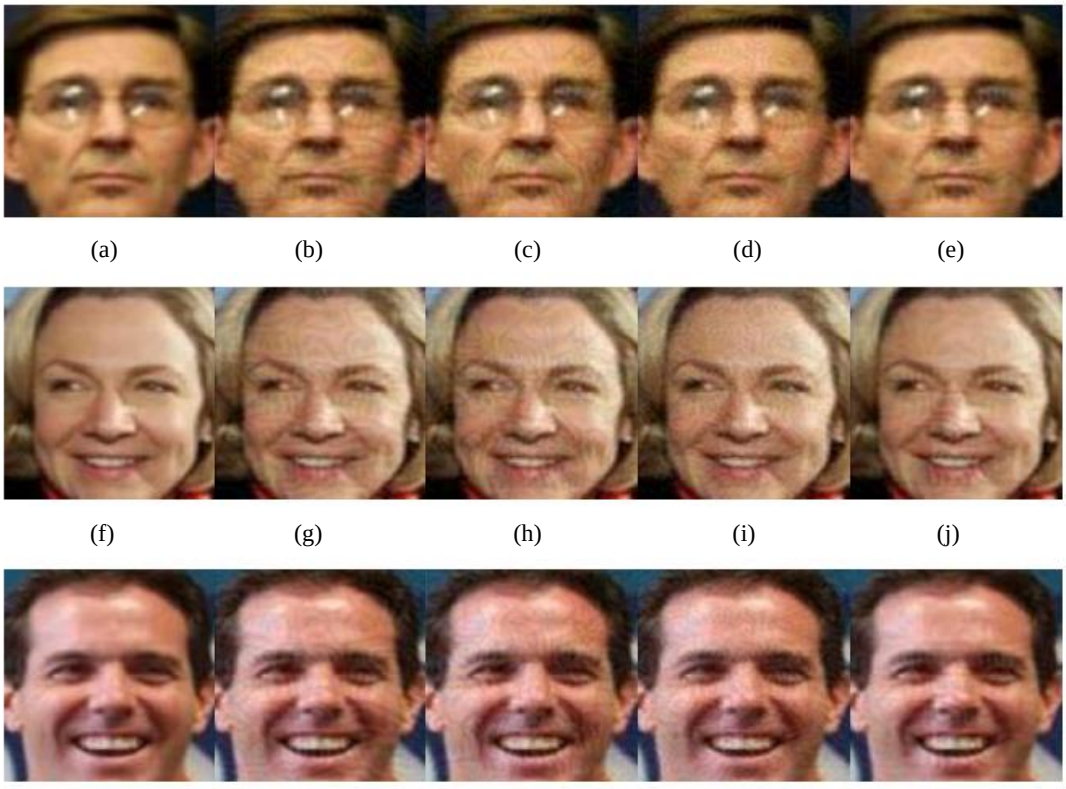


图 3.8 不同方法生成的对抗样本可视化示意图

虽然在所有方案中都有对视觉隐蔽性限制的手段，但将扰动加入到原始图片后仍会对图片视觉效果造成一定程度的破坏，不过总体来衡量，上述各方案生成的对抗样本依然具有较高的视觉质量。其中 UAP 和 FG-UAP 两种方式在可视化结果图中也能较明显看出叠加的扰动情况，尤其是在额头及脸颊等光滑区

域，FTGAP 方式生成的扰动考虑了频域信息，其可视化图在光滑低频区域的扰动相对较少，隐蔽性视觉效果也更优异，KRT-FUAP 通过控制扰动在语义关键区域及非关键区域叠加扰动的权重大小，使得更多的扰动叠加在了更有效的区域，以此平衡了攻击性和隐蔽性之间的关系，与其他集中方式相比，本方案生成的对抗样本有着优越的隐蔽性，图片看起来也更正常，最终不同方案展示的可视化结果也符合隐蔽性实验测出的客观性指标结果。

3.3.5 黑盒攻击性能评估

在面向人脸识别的通用对抗扰动任务中，黑盒攻击性能用于评估在没有访问目标模型内部结构和参数的情况下，通用对抗扰动对人脸识别模型准确率的影响程度。大多数实际的攻击模式都是针对黑盒的场景，攻击者无法获得目标人脸识别模型的内部详细参数和具体实现细节，只能通过观察模型对输入数据的响应来推断行为。

为此本章也针对 KRT-FUAP 方案进行了黑盒测试，实验使用 IResNet50、MobileFaceNet 和 MobileNetV1 三个主干提取网络进行测试，将其中一个网络作为训练人脸通用对抗扰动的可访问参数的白盒模型，剩下的两个网络作为无法访问内部详细参数的黑盒模型，在 LFW 数据集上进行相应的黑盒测试，测试结果如下表 3.4 所示，该实验评估了在数据集使用三个不同主干网络的攻击性性能，其中第一列的网络表示可访问内部参数的白盒方式，后续列的网络则表示不可直接访问内部参数的黑盒方式，因此对角线上的数据表示白盒方式下的攻击成功率，其他位置的数据表示在不同主干网络情况下的黑盒攻击成功率。由表格中数据可知，在黑盒攻击的方式下，通用对抗扰动的攻击成功率都有不同程度的下降，其表明该方案在攻击未知模型的情况下还是会受到很大影响。不过在 MobileFaceNet 上生成的对抗扰动迁移到 IResNet50 上测试能有 54.6%的攻击成功率，这也一定程度说明该扰动生成策略在黑盒模式下有一定的有效性。因本方案将扰动主要叠加在语义关键区域，不同模型对于语义关键区域的响应不一致，但语义关键区域对最终提取出的特征向量影响的权重相对还是比较大，所以扰乱关键语义区域对黑盒方式具有一定有效性。不过相比于真正有效的黑盒攻击，本方案在该方面的提升还面临巨大的挑战。

表 3.4 KRT-FUAP 方案的黑盒攻击性能

愚弄率(%)	IResNet50	MobileFaceNet	MobileNetV1
IResNet50	81.9	26.8	25.4
MobileFaceNet	54.6	80.1	41.3
MobileNetV1	45.2	33.5	79.7

3.3.6 消融实验

人脸关键区域掩码的影响：针对特定个体身份的对抗扰动主要集中于识别单一人脸图像的特征，并通过对这些特征的修改来误导识别模型。相反，通用对抗扰动则需要探究整个数据集中的共性规律，并基于这些规律设计出能够普遍应用于整个数据集的单一扰动，以此影响人脸识别模型的判断，显著提高在整个数据集上的欺骗成功率。

在生成通过对抗扰动的过程中，直接在全局生成扰动与区分局部语义关键区域和非关键区域扰动权重对最终生成的扰动性能有显著影响。为了探究两种方式产生扰动的性能，依据前文分析的人脸语义关键区域的有效性，将语义关键区域掩码从小批量数据集中提取出来，具体操作为从数据集中随机选定 1000 张人脸图片，提取每张人脸图片的关键点坐标，运用凸包算法计算出不同人脸的关键区域位置，最后取交集得到针对整个数据集的关键区域掩码位置。为了显示选定关键区域后使用 KRT-FUAP 方式的有效性，本实验通过对比不使用关键区域掩码直接优化全局的通用对抗扰动方式，并测试在 LFW 数据集下，使用 IResNet50 和 MobileFaceNet 主干提取网络所得的攻击成功率和隐蔽性客观评价指标，相应结果如表 3.5 所示，在两种网络上，使用全局方式生成的扰动在攻击性和隐蔽性性能方面都有一定的降低，且因为控制扰动在关键语义区域的权重大小后，可以更好地分配扰动的强度，使得噪声能着重于攻击关键区域，并且针对非关键语义区域的噪声强度相对变小也更有利于扰动隐蔽性的提升。

表 3.5 KRT-FUAP 中人脸全局区域与关键区域的影响

网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑
IResNet50	KRT-FUAP	81.9	0.9304	34.0157
	全局区域	77.8	0.8926	31.6674
MobileFaceNet	KRT-FUAP	80.1	0.9044	32.9812
	全局区域	77.7	0.8639	30.5034

可学习流场对扰动有效性的影响：在数据集中识别并提取的固定人脸语义

掩码区域可视为整个数据集的共同特征，其概括了人脸关键区域的整体分布趋势。然而，这种方法可能会忽略或舍弃掉一些个体独有的特征。为了弥补这一缺陷，在固定人脸语义掩码区域的基础上引入了一个可调节的流场机制。通过持续的训练过程，该空间变换流场能够对固定区域进行微调，以逐步适应并整合个体特征。这样设计不仅原有的统一特征得到了保留，而且通过学习过程融入了训练数据中的个体差异，从而使得产生扰动的过程中在最终的信息处理上更为全面和精准。

本实验采用了一种动态学习的方法来定义人脸关键区域掩码，这一掩码是可训练的，其通过一个可变流场进行空间变换以确定最终的关键区域。通过预设的损失函数不断优化流场内的参数，进而精细化从数据集中初步提取的固定语义掩码。该方式优势在于其不受限于初始选定的人脸图像子集，而是通过持续的迭代学习过程，增强了所生成扰动的泛化能力。为了验证可学习流场的有效性，本节首先使用固定的区域掩码作为对照进行实验，评估了使用 LFW 数据集使用不同主干网络条件下测试的攻击成功率和隐蔽性指标。相应的对比结果如表 3.6 所示，不使用流场微调关键区域在不同网络上攻击成功率有小幅度下降，且其隐蔽性也略有不足。该结果表明，可学习流场的方式能够更全面地捕捉信息，其不仅综合了数据集的共性特征，还融合了不同个体的独特特征，这种方法生成的通用对抗扰动在攻击性和隐蔽性方面都表现出了更显著效果。

表 3.6 KRT-FUAP 中可学习流场的影响

网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑
IResNet50	KRT-FUAP	81.9	0.9304	34.0157
	固定区域	80.2	0.9147	33.1845
MobileFaceNet	KRT-FUAP	80.1	0.9044	32.9812
	固定区域	78.9	0.8916	32.1108

平衡攻击性和隐蔽性损失的影响：本实验精心设计了一个包含对抗性损失和隐蔽性损失的复合损失函数。对抗性损失的优化目标是直接确定对抗样本的生成方向，本方案绕过计算中间值的度量方式，直接选择与原始图像特征方向相反的方向进行优化以增强攻击效果。同时为了提升生成对抗样本的不可感知性，本方案的隐蔽性手段通过引入 VGG 网络来精细调控隐蔽性损失，从而在维持一定攻击性的情况下提高图像的隐蔽性。

为了验证该方式设计损失函数的有效性，本节选取了人脸识别任务中广泛使用的余弦相似度损失函数和欧式距离损失函数作为对比。在保持其他实验条件一致的情况下，余弦相似度损失函数通过减少特征间的相似度来实现攻击性，而欧式距离损失函数通过增加特征间的欧氏距离来达到相同的目的。不过这些传统方式仅关注于提高攻击性，而未考虑隐蔽性的指标。

本节通过比较使用 KRT-FUAP 框架下的复合损失函数与上述两种常用的人脸识别损失函数在相同条件下的测试情况，在使用 LFW 数据集在两个主干提取网络上的攻击成功率和隐蔽性指标测试结果如

表 3.7 所示，三者的攻击成功率的指标相差并不是很大，说明本方案的相反方向损失方式对攻击性有更好的作用，而在隐蔽性指标方面，通过在损失函数中同时兼顾攻击性和隐蔽性，KRT-FUAP 的隐蔽性度量结果展现出更优越的性能。这表明，本方案的损失函数设计在实现攻击效果的同时，也成功提升了对抗样本的隐蔽性，为构建更为高效和隐蔽的对抗攻击提供了一种新的策略。

表 3.7 KRT-FUAP 中平衡隐蔽性和攻击性损失函数的影响

网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑
IResNet50	KRT-FUAP	81.9	0.9304	34.0157
	欧式距离	80.2	0.8875	31.7418
	余弦相似度	80.4	0.8963	31.9427
MobileFaceNet	KRT-FUAP	80.1	0.9044	32.9812
	欧氏距离	78.9	0.8622	30.8143
	余弦相似度	79.1	0.8657	30.8807

3.4 本章小结

本章提出了一种语义关键区域控制的人脸通用对抗扰动。通过分析人脸识别可解释性领域的观点，阐述本章聚焦人脸图像语义关键区域的研究动机。后续详细介绍了本章所提方法依据流场微调关键区域的操作步骤以及产生通用对抗扰动的流程，最后通过全面的实验评估本章所提方法在攻击及隐蔽方面的性能。实验结果证明，相较于现有的通用对抗扰动生成方案，本方案生成的扰动在保持攻击性能的同时显著提升了隐蔽性。

第四章　基于频域分析的人脸通用对抗扰动

4.1　研究动机

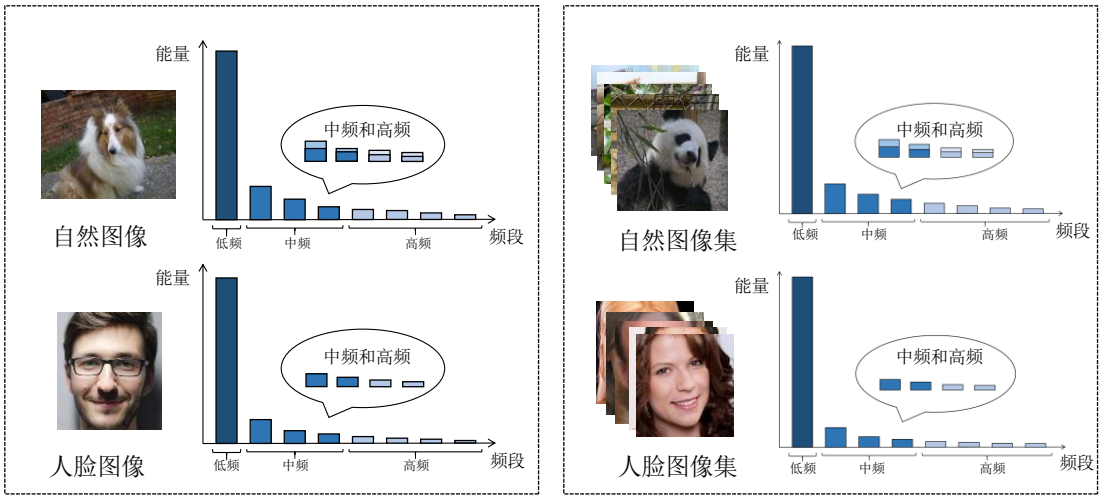
为了在人脸识别任务中生成更加高效的通用对抗扰动，需要精心权衡生成扰动的攻击性能和隐蔽性能之间的平衡。尽管上一章节通过在空域分析人脸语义关键区域对人脸识别系统识别准确率的影响，利用人脸数据集的共性来生成通用对抗扰动，这种方法在一定程度上提升了扰动对人脸图像的适应性，但仍未能精确捕捉到面向人脸识别通用对抗扰动的全面信息。为了进一步提高对抗扰动的有效性，需要更深入地理解人脸图像数据集的共性特征，并针对人脸识别系统的关键识别环节进行扰动设计。

本章通过分析自然图像和人脸图像在频域视角的差异性，自然图像数据集包含多种图像的语义类别，在不同的类别意象中，物体在形状、样式和大小等方面都有很大的变化。然而，在人脸图像数据集中，因人脸识别一般都需要经过人脸对齐的预处理操作，最终获得的人脸图像的大致轮廓会有很大程度相似，其整个图像的主体区域趋于平滑。上述自然图像和人脸图像的特征在频域中也表现出显著的差异性，自然图像不同种类之间差异较大，因此单个自然图像的频谱表现出高度的不一致性，不同自然图像的中高频区域能量分布相对更丰富。而人脸图像不同个体之间相似区域的特征更集中，在频谱中展现出较强的一致性，其低频区域能量分布更多，而中高频区域能量分布相对较少。因此，若按照一般的自然图像通用对抗扰动生成步骤直接在空域叠加扰动，人脸图像的低频区域可能会产生明显的视觉缺陷，从而致使在人脸图像上生成的通用对抗扰动的隐蔽性效果不佳。本章将频域作为额外维度来调整和优化人脸图像的通用对抗扰动，通过引入空频转换的操作，获得频域内的有效信息。将干净人脸图像转换到频域后进行分频段处理，通过训练三个不同的频带滤波器来提取频域信息，利用在频域中优化通用对抗扰动的方式，获得基于频带滤波器驱动的通用对抗扰动（Facial Universal Adversarial Perturbations Driven by Frequency Band Filters, FaUAP-FBF）。

4.2 基于频带滤波器驱动的通用对抗扰动

4.2.1 问题分析

本章考虑到人脸图像的频率信息相比于自然图像的频率信息有较大的差异，尤其针对图像分类任务的数据集和人脸识别任务的数据集，两者差异更为明显。因为自然图像涉及到的种类繁多，不同类别图像之间的差异较大，而人脸图像虽然不同身份之间也存在着较大差异，但相较于自然图像各类之间的区别，人脸图像不同身份之间的人脸特征具有一定相似性。为了展示两者频域所含信息分布的区别，通过随机选择200幅自然图像和200幅人脸图像进行验证。详细结果如图4.1所示。



(a)单张图像能量比较情况　　(b)图像集能量比较情况

图4.1 自然图像与人脸图像频域能量分布比较示意图

对于每个单独的图像，计算其频率的能量分布情况，通过汇总两种类别图像的单独分布状况，分别计算出自然图像和人脸图像在不同频段的平均能量分布。其中左侧的图像表示单张自然图像和人脸图像在频域能量分布的比较情况，由图4.1可知，在中频段和高频段区域，单张自然图像所含信息更丰富。右侧图像则表示该小批量自然图像和人脸图像在高频、中频和低频的平均能量分布比较情况，同样可以分析得出自然图像在中高频段区域含有的信息比人脸图像更丰富，而人脸图像所含的低频信息占比更高。该结果证实了本章针对频域信息的推测，而人类的视觉感知和频域信息有一定关联，能够更敏锐地捕捉到

低频区域的变化。为此本章通过在频域视角对频率变化进行精确控制，通过额外考量频域的有效信息，以此达到提高通用对抗扰动攻击性和隐蔽性的目的。

4.2.2 框架设计概述

在数字图像处理领域，离散余弦变换（Discrete Cosine Transform, DCT）是一种重要的数学工具，用于将图像从空域转换到频域，以提取图像的频率特征。DCT 通过将图像分解为一系列频率成分，将图像信息表示为一组余弦函数的线性组合。其在图像压缩、图像编辑、图像增强和噪声滤除等方面发挥重要作用。通过对图像进行 DCT 变换，可以在频域中操作和分析图像，以实现各种图像处理任务。本章假设有一个矩阵 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$，每一个元素 $\boldsymbol{a}_{i,j}$ 对应 $\boldsymbol{A}$ 中的一个位置 $(\boldsymbol{i},\boldsymbol{j})$，使用 DCT 变换后可以得到变换后的矩阵 $\boldsymbol{B}\in\boldsymbol{R}^{n\times n}$，$\boldsymbol{B}$ 中位于位置 $(\boldsymbol{i},\boldsymbol{j})$ 的元素 $\boldsymbol{b}_{i,j}$ 可以表示为公式(4.1)：

$$\boldsymbol{b}_{i,j}=\boldsymbol{c}_i\boldsymbol{c}_j\sum_{\boldsymbol{p}=0}^{\boldsymbol{n}-1}\sum_{\boldsymbol{q}=0}^{\boldsymbol{n}-1}\boldsymbol{a}_{\boldsymbol{p},\boldsymbol{q}}\cos\left(\frac{(2\boldsymbol{p}+1)\boldsymbol{i}\pi}{2\boldsymbol{n}}\right)\cos\left(\frac{(2\boldsymbol{q}+1)\boldsymbol{j}\pi}{2\boldsymbol{n}}\right)\tag{4.1}$$

其中 $\boldsymbol{i}=\boldsymbol{j}=0$ 时，$\boldsymbol{c}_i=\boldsymbol{c}_j=1/\sqrt{4\boldsymbol{n}}$，否则，$\boldsymbol{c}_i=\boldsymbol{c}_j=1/\sqrt{2\boldsymbol{n}}$。$\boldsymbol{B}$ 也可以通过逆变换重新转为 $\boldsymbol{A}$，如公式(4.2)所示：

$$\boldsymbol{a}_{i,j}=\sum_{\boldsymbol{p}=0}^{\boldsymbol{n}-1}\sum_{\boldsymbol{q}=0}^{\boldsymbol{n}-1}\boldsymbol{c}_{\boldsymbol{p}}\boldsymbol{c}_{\boldsymbol{q}}\boldsymbol{b}_{\boldsymbol{p},\boldsymbol{q}}\cos\left(\frac{(2\boldsymbol{i}+1)\boldsymbol{p}\pi}{2\boldsymbol{n}}\right)\cos\left(\frac{(2\boldsymbol{j}+1)\boldsymbol{q}\pi}{2\boldsymbol{n}}\right)\tag{4.2}$$

对于8×8 的空间块，DCT 变换包含了 64 个不同频率的分量，将图像从空域变换为一系列频域的块。本章采用常见的第二类离散余弦变换来完成图像从空域到频域的相互转换。

与特定于单幅图像的对抗性扰动不同，通用对抗扰动是一种与图像无关的攻击方式，其通过训练集与目标函数不断迭代学习并受到视觉隐蔽性准则的约束。随后，使用学习到的通用对抗扰动来修改从与训练集相同的分布中提取的任何未见过的人脸图像，以此欺骗人脸识别的目标模型。本章提出了一种基于频带滤波器驱动的人脸通用对抗扰动生成框架，该框架对现有的面向自然图像分类任务的通用对抗扰动生成框架进行了改进，本框架如图 4.2 所示，通过将人脸图像和对抗扰动转换到频域，在频域进行扰动的产生操作，最终再转换回空

域输入人脸识别特征提取器中进行损失函数的运算。

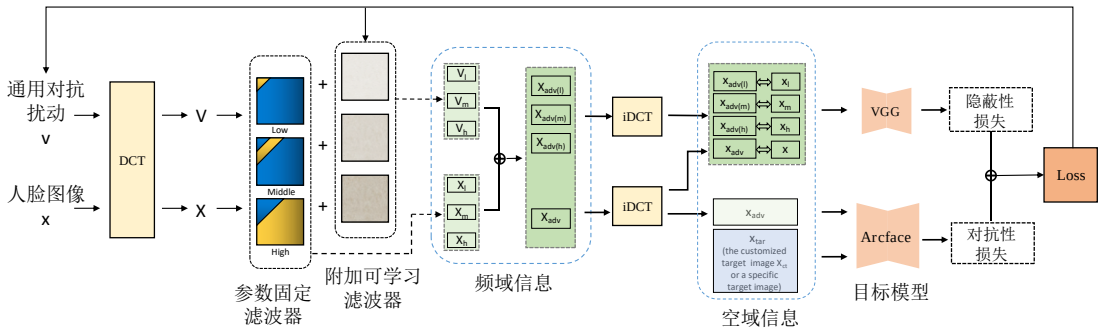


图 4.2 频带滤波器驱动的通用对抗扰动产生方案框架示意图

首先通过 DCT 模块将训练集中的人脸图像转换到频域，用高斯噪声初始化通用对抗扰动，并用三个固定的高频、中频和低频滤波器初始化频带可学习滤波器。随后依据本框架设定的对抗性损失和隐蔽性损失的加权组合方式对通用对抗扰动和三频段可学习滤波器进行迭代更新，其中通用对抗扰动的强度还受到无穷范数的约束。隐蔽性损失借助 VGG 网络模型进行计算，对抗性损失通过预先训练的一个定制目标图像在选定的 Arcface 人脸识别目标模型中进行计算。在迭代过程中，分别使用频带滤波器对输入人脸图像和通用对抗扰动进行分频段信息的提取操作，获得对应的三频段信息。在频域中将对应频段的分量进行叠加，获得不同频段的对抗样本和全频段的对抗样本的频域信息，利用离散余弦变换的逆变换方式将频域信息转回到空域，用于后续设计的对抗损失和隐蔽损失的计算。至此完成整个通用对抗扰动框架的搭建，通过后续不断循环更新参数。图 4.2 框架示意图中涉及的符号及相应的定义如表 4.1 所示。

表 4.1 框图符号及相应定义

符号	定义
v / V	空域对抗扰动 / 频域对抗扰动
x / X	空域人脸图像 / 频域人脸图像
$V_l / V_m / V_h$	低频 / 中频 / 高频扰动分量
$X_l / X_m / X_h$	低频 / 中频 / 高频图像分量
$X_{adv(l)} / X_{adv(m)} / X_{adv(h)}$	低频 / 中频 / 高频对抗样本分量
x_{adv} / x_{ct}	空域对抗扰动 / 定制的目标图像

当达到设定的标准阈值时，停止迭代更新，输出最终获得的通用对抗扰动。其目标函数的表述如公式(4.3)和公式(4.4)所示：

$$\left(\hat{\mathrm{v}},\hat{\mathrm{f}}_{\mathrm{l}},\hat{\mathrm{f}}_{\mathrm{m}},\hat{\mathrm{f}}_{\mathrm{h}}\right)=\arg\min_{\mathrm{v},\mathrm{f}_{\mathrm{l}},\mathrm{f}_{\mathrm{m}},\mathrm{f}_{\mathrm{h}}}\left[\mathcal{L}_{\mathrm{adv}}\left(\mathrm{x}_{\mathrm{adv(tr)}},\mathrm{x}_{\mathrm{tar}}\right)+\lambda\mathcal{L}_{\mathrm{ste}}\left(\mathrm{x}_{\mathrm{adv(tr)}},\mathrm{x}_{\mathrm{tr}}\right)\right]\mathrm{s.t.}\left\|\mathrm{v}\right\|_{\infty}\le\xi \quad (4.3)$$

$$\mathrm{x}_{\mathrm{adv(tr)}}=\mathrm{IDCT}\left[\mathrm{DCT}\left(\mathrm{x}_{\mathrm{tr}}\right)+\mathrm{DCT}\left(\mathrm{v}\right)\cdot\sum\nolimits_{\mathrm{s=l,m,h}}\mathrm{f}_{\mathrm{s}}\right] \quad (4.4)$$

其中，$\mathcal{L}_{\mathrm{adv}}$ 和 $\mathcal{L}_{\mathrm{ste}}$ 表示对抗性损失和隐蔽性损失，λ 控制对抗性损失和隐蔽性损失之间的平衡，保证生成的扰动能够兼顾攻击成功率和隐蔽性指标两个维度。v 和 $\left\{\mathrm{f}_{\mathrm{l}},\mathrm{f}_{\mathrm{m}},\mathrm{f}_{\mathrm{h}}\right\}$ 分别表示通用对抗扰动和三个可学习频带滤波器的组合，$\hat{\mathrm{v}}$ 和 $\left\{\hat{\mathrm{f}}_{\mathrm{l}},\hat{\mathrm{f}}_{\mathrm{m}},\hat{\mathrm{f}}_{\mathrm{h}}\right\}$ 分别表示学习到的通用对抗扰动和频带滤波器。x_{tr} 和 $\mathrm{x}_{\mathrm{adv(tr)}}$ 表示干净的训练集样本数据和与之对应的对抗样本数据，$\mathrm{x}_{\mathrm{tar}}$ 表示目标图像，在本章节框架中，其可以表示预先定制的目标图像也可以是空间域中已有的特定人脸目标图像。由上述公式运算流程可知，人脸识别模型最终还是要在空域中对人脸图像进行特征提取，对抗性损失和隐蔽性损失也都是在空域中进行计算，这意味着在学习过程中必须使用离散余弦逆变换方式将滤波后的扰动反向转换到空域中。空频转换流程如图 4.3 所示，针对对抗性损失，使用 ArcFace 人脸识别模型来评估最终的攻击性指标。针对隐蔽性损失，使用 VGG 网络来计算对抗性样本的浅层视觉特征，以此控制其隐蔽性指标。

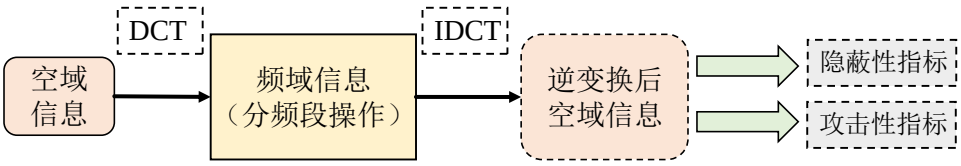


图 4.3 空频转换流程示意图

一旦训练完成获得上述的 $\hat{\mathrm{v}}$ 和 $\left\{\hat{\mathrm{f}}_{\mathrm{l}},\hat{\mathrm{f}}_{\mathrm{m}},\hat{\mathrm{f}}_{\mathrm{h}}\right\}$，就可以利用训练好的参数为一个干净的人脸图像测试样本生成一个对抗样本，相应的表述如公式(4.5)：

$$\mathrm{x}_{\mathrm{adv(te)}}=\mathrm{IDCT}\left[\mathrm{DCT}\left(\mathrm{x}_{\mathrm{te}}\right)+\mathrm{DCT}\left(\hat{\mathrm{v}}\right)\cdot\sum\nolimits_{\mathrm{s=l,m,h}}\hat{\mathrm{f}}_{\mathrm{s}}\right] \quad (4.5)$$

其中，x_{te} 和 $\mathrm{x}_{\mathrm{adv(te)}}$ 表示干净的人脸测试集数据和对应生成的对抗样本数据。在后续小节中将分别详细阐述通用对抗扰动、频带滤波器和对抗样本之间的交互关系，以及定制目标图像设计模块的原理和损失函数设计。本方案的算法流程如算法 2 所示。

算法 2：基于频带滤波器驱动的扰动生成算法

输入：预处理数据集 D，定制目标图像 x_{ct}，人脸识别网络 $F(\cdot)$，愚弄率 δ，扰动的无穷范数 ξ，三个固定滤波器 $\{\tilde{f}_s\}$，决策阈值 t。

输出：通用对抗扰动 v，可学习频带滤波器 $\{f_s\}$。

1：随机初始化 v
2：**While** 愚弄率 < δ **do**
3： **for** x_i in D **do**
4： 进行 DCT 变换获得 V 和 X_i
5： 使用 $\{f_s\}$ 和 $\{\tilde{f}_s\}$ 获得 V_s 和 $X_s^{(i)}$
6： 进行 IDCT 变换：
7： $x_{adv}^{(i)} \leftarrow \mathrm{IDCT}\left[\sum\left(X_s^{(i)} + V_S\right)\right]$
8： **if** $\mathrm{Similarity}\left\{F\left(x_{adv}^{(i)}\right), F\left(x_i\right)\right\} > t$ **then**
9： $\left(\Delta v_i, \Delta f_s\right) \leftarrow \arg\min_{(v_i, f_s)} \left\|\left(v_i, f_s\right)\right\|_2$
10： s.t. $\mathrm{Similarity}\left\{F\left(x_i + v\right), F\left(x_i\right)\right\} \geq t$
11： 更新可学习滤波器：$f_s \leftarrow f_s + \Delta f_s$
12： 更新扰动：$v \leftarrow v + \Delta v$
13： 裁剪 v 以满足无穷范数
14： **end if**
15： **end for**
16：**end while**
17：**return** v

4.2.3 自适应频带滤波器模块

通常对抗样本都是由原始数据图像和对抗扰动在空域直接线性相加产生的，这为生成对抗样本提供了一种简单有效的方法。然而，直接在空域中不经限制地添加扰动可能会导致视觉缺陷，其隐蔽性指标会相对较差。相比之下，频率的改变与人类的视觉感知没有直接的关系。因此，为了解决这一问题生成具有更强隐蔽性的对抗样本，本章节采用 DCT 变换的方式将图像转换到频域，从而在频域引入扰动，最终再逆变换回空域，以此实现对抗样本的生成。

通过对自然图像和人脸图像两者在频域视角的差异性分析，本模块采用了一种更精细的方式来定制针对人脸识别任务的通用对抗扰动，相应框架如图 4.4 所示，其中彩色块表示的是参数固定的滤波器，三个灰色的块表示参数可变的

附加滤波器，通过将高频、中频和低频的固定参数滤波器与对应频带附加滤波器叠加获得参数可变的三个频带滤波器，使用不同滤波器提取不同人脸图像和对抗扰动的频域信息。使用 DCT 模块将空域人脸图像转换到频域之后，区别于直接在其中添加扰动的方式，本方案采用分频带的策略，将变换后的人脸图像分成不同的频率段。在每个片段中设计并使用三个频带自适应滤波器来提取特定于该频率范围的信息。本框架中使用的三个频带滤波器由高频、中频和低频的固定滤波器和可学习附加滤波器组合而成。其中本模块设置固定滤波器的低频频带占整个频谱的前 1/16 区域，固定滤波器的中频区域占整个频谱的 1/16 至 1/8 区域，剩下的其他区域都定义为滤波器的高频区域。所有参数固定滤波器都是8×8 空间块，其使用 0 和 1 两个参数组成的掩码来选择不同的频域信息。

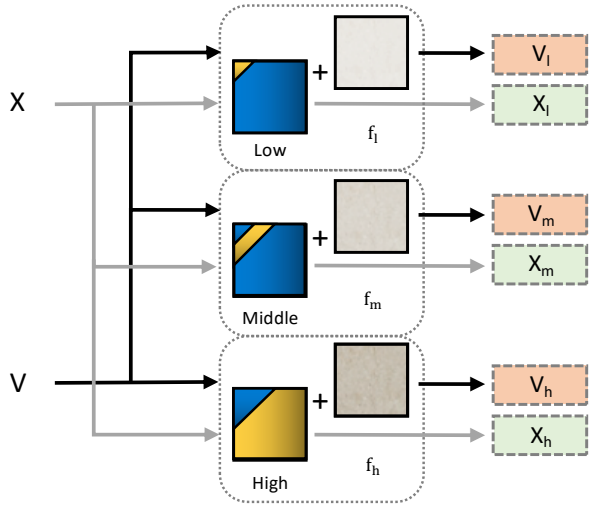


图 4.4 自适应频带滤波器框架

设空间域的人脸图像为 x，空间域扰动为 v，则高频、中频和低频情况下的提取过程如公式(4.6)所示：

$$\begin{cases} X_s = \mathrm{DCT}(x)\cdot \tilde{f}_s \\ V_s = \mathrm{DCT}(v)\cdot f_s \end{cases} \quad s = l, m, h \tag{4.6}$$

其中 $\tilde{f}_s$ 表示参数固定的滤波器，f_s 表示参数可学习的频带滤波器。参数固定的滤波器作用为提取干净人脸图像的不同频段信息，参数可变的频带滤波器用于提取通用对抗扰动的不同频段信息。$\mathrm{DCT}(x)$ 表示 DCT 变换函数，将 x 从空间域变换到频域。随后，将图像的低频信息 X_l 和扰动的低频信息 V_l 叠加，获得对抗样本的低频信息 $X_{adv(l)}$。同样，应用相同的运算可以得到对抗样本的中频信息

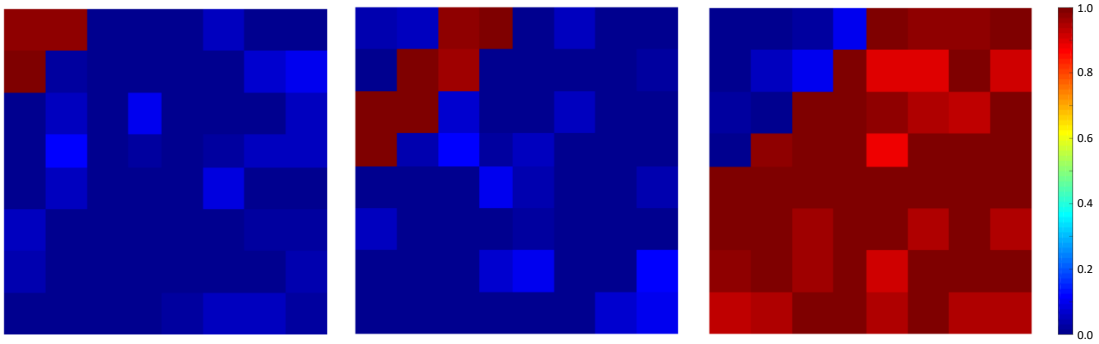
$X_{adv(m)}$ 和对抗样本的高频信息 $X_{adv(h)}$。最终使用离散余弦逆变换将频域信息转回空域后得到对应频段的空域对抗样本 $x_{adv(s)}$，相应的运算方式如公式(4.7)：

$$x_{adv(s)} = iDCT\left\{DCT(x)\cdot \tilde{f}_s + DCT(v)\cdot f_s\right\},\ s = l,m,h \tag{4.7}$$

因本方案中提取原始干净人脸图像的滤波器为参数固定的滤波器，其进行离散余弦正变换和逆变换后，干净人脸图像的信息不会发生损失，所以最终生成的空域全频段对抗样本 x_{adv} 运算方式如公式(4.8)：

$$x_{adv} = iDCT\left\{DCT(x) + DCT(x)\cdot(f_l + f_m + f_h)\right\} \tag{4.8}$$

为了产生针对人脸图像的通用对抗扰动，本章利用了大量人脸图像进行训练，通过在频域内不断迭代的过程，将通用对抗性扰动的频域信息应用于新的人脸图像数据，同时允许设定的优化器不断更新通用扰动以实现欺骗人脸识别网络模型的任务。其中，可学习的频带滤波器的使用确保提取的频率段信息不局限于预定义的高频、中频和低频范围。结合这些可调参数滤波器的微小改变有助于捕获每个频率段内更有价值的信息，以此生成针对各个频段的更高效的对抗扰动。本模块最终迭代生成的可学习频带滤波器的可视化滤波器核如图 4.5 所示，滤波器大小都为8×8，由图 4.5 可知，经过不断训练迭代参数之后，三个可学习的频带滤波器的内部参数由固定的 0 和 1 两个数值转变成了在 0 至 1 区间内的参数，其仍然保留了初始的高频、中频和低频的大致范围，只不过内部的参数进行了微小的调整，使之更符合分频段叠加扰动的任务。



(a) low　　(b) middle　　(c) high

图 4.5 可学习的频带滤波器核的可视化示意图

4.2.4 定制目标图片设计模块

通常对抗攻击可以分为目标攻击和非目标攻击两种类别，目标攻击需要选定一个确定的攻击目标，使对抗样本都能被错误识别为该目标，而非目标攻击没有一个明确的目标，只需要模型能识别错误就达到基本要求。本模块将两种攻击方式进行合理统一，提出一种新型的有目标攻击方式，不过该目标并非传统意义的某个确定的人脸身份，而是按照特定的方式生成的一张优化图像，该图像的特征远离用于训练通用对抗扰动的人脸图像数据集内的数据特征分布。常规的针对人脸识别模型的非目标攻击通过控制人脸对抗样本与原始干净人脸图像之间相似性，迫使对抗样本特征远离该原始人脸身份的特征，以此达到攻击的目的。然而，本方案通过计算人脸对抗样本与该自定义的优化图像之间相似性，控制对抗样本特征不断靠近该优化图像特征，通过该方式间接达到控制对抗样本远离原始人脸图像数据集分布的目的，以此达到使用目标攻击的模式完成非目标对抗攻击的要求。本方案中定制目标图像的流程如图 4.6 所示。

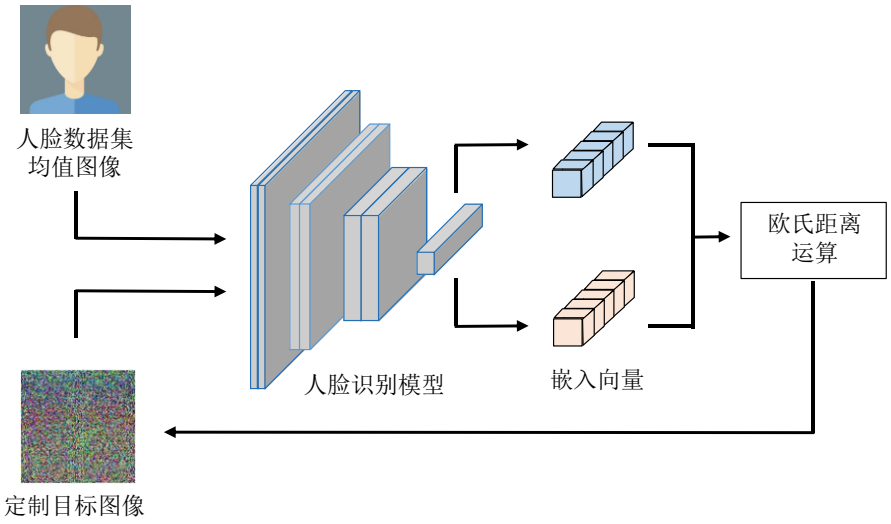


图 4.6 定制目标图像的生成流程

通过自定义一个图像作为目标，将针对不同受攻击对象的非目标攻击转换为针对单一图像的目标攻击。具体而言，通过最大化该目标优化图像与干净人脸数据集的总体分布距离来寻找该目标优化图像 x_{ct}。从人脸数据集中随机选定一小批量的人脸图像数据，计算该批量数据的平均像素值，以此定义为该数据集人脸图像的数据分布情况。随机初始化一个定制的目标图像，将人脸数据集

的均值图像和该定制的目标图像共同输入模型中，通过比较两者提取出的嵌入向量的欧氏距离，损失函数控制该欧氏距离不断变大以此不断优化自定义的目标图像。最终获得的该定制目标图像特征会远离原始人脸数据集的特征分布。

首先，从干净人脸数据集 X 中随机选择包含 6000 张人脸的图像子集 $\{x_i, \boldsymbol{i}=1,2,\ldots,\boldsymbol{n}\}$，该数据集包含每个身份的一幅人脸面部图像，计算该子集的平均图像 x_o 如公式(4.9)所示：

$$x_o = \frac{1}{\boldsymbol{n}}\sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} x_i,\ x_i \sim X \tag{4.9}$$

随后用高斯噪声初始化本方案定制的目标图像 x_{ct}，通过逐步增大目标图像 x_{ct} 与数据集的平均图像 x_o 之间嵌入向量特征的欧氏距离进行迭代更新，该过程导致自定义的图像 x_{ct} 明显偏离原始干净人脸图像数据集的分布，以此作为本方案的定制目标图像。其生成过程表示为公式(4.10)：

$$\hat{x}_{ct} = \arg\max_{x_{ct}} \left\| F(x_{ct}), F(x_o) \right\|_2 \tag{4.10}$$

其中 F 表示目标替代模型（ArcFace 模型），$F(x_{ct})$ 和 $F(x_o)$ 表示从目标替代模型中提取出的人脸图像嵌入向量。由于 $\hat{x}_{ct}$ 与小批量干净的人脸图像数据集的平均值有显著差异，即表明该定制目标图像 $\hat{x}_{ct}$ 偏离了整个干净人脸图像集的特征分布。因此，减小定制目标图像 $\hat{x}_{ct}$ 和对抗样本两者特征之间的距离与增大对抗样本和原始干净人脸图像两者特征之间距离是一致的，从而达到使用目标攻击的模式完成针对人脸识别模型的非目标对抗攻击的要求。在真正的目标攻击任务中，本方案流程中需要做的改变就是将定制的目标图像替换为确定的人脸受攻击者图像，从而将非目标攻击和目标攻击统一为目标攻击框架的整体。

4.2.5 损失函数的设计

本方案的损失函数设计包括两个部分，通过设计有针对性的对抗性损失和隐蔽性损失，实现了通用对抗扰动的攻击性和隐蔽性的双重优化。通过限制扰动达到预定的视觉隐蔽性标准的情况下，提升扰动的攻击成功率。

隐蔽性损失函数通过使用 VGG 网络的浅层给出的特征映射来评估扰动的视

觉隐蔽性。因为本框架在训练扰动的过程中，采用自适应频带滤波器驱动的方式将人脸图像和通用扰动划分为了三个不同的频段，获得了各个频率范围的对抗样本和干净图像的信息。随后再将上述频域信息转换回空域，即可获得高频、中频和低频的对抗样本信息和全频段的对抗样本信息。将干净人脸图像的全频段信息和对抗样本的全频段信息输入 VGG 网络，仅使用其浅层的输出来提取低级的纹理特征，同时将干净人脸图像的分频段信息和对应对抗样本的分频段信息输入 VGG 网络，最终将全频段和分频段的两者提取出特征向量的差异性作为隐蔽性损失，控制该差异性不断变小以此来控制对抗扰动的隐蔽性。隐蔽性损失的计算如公式(4.11)所示：

$$\mathcal{L}_{\text{ste}} = \sum_{\mathrm{x}^{(i)}\in \mathrm{D}} \left(\left\| \varphi_j\left(\mathrm{x}^{(i)}\right), \varphi_j\left(\mathrm{x}_{\text{adv}}^{(i)}\right) \right\|_2 + \sum_{\mathrm{s=l,m,h}} \left\| \varphi_j\left(\mathrm{x}_{\mathrm{s}}^{(i)}\right), \varphi_j\left(\mathrm{x}_{\text{adv(s)}}^{(i)}\right) \right\|_2 \right) \tag{4.11}$$

其中，$\mathcal{L}_{\text{ste}}$ 包含两个部分的内容，第一部分计算整个频段的干净人脸图像和对抗样本之间的差异性，第二部分则分别计算干净人脸图像和对抗样本在高频、中频和低频段的特征之间差异性。D 表示用于训练的人脸数据集，$\boldsymbol{i}$ 表示图像在数据集中的序号，$\varphi_j(\cdot)$ 表示 VGG 网络中第 $\boldsymbol{j}$ 层的特征映射输出。

对抗性损失的作用是控制对抗性扰动的攻击性能。在人脸识别中，两个嵌入向量之间的余弦相似度是最常用的度量两个样本的差异的方式。区别于一般针对人脸识别的非目标对抗攻击通过度量对抗样本和原始图像之间相似性作为对抗性损失的方式，本方案非目标攻击采用度量生成的对抗样本和定制的目标图像之间的相似性的方式，通过对抗样本的特征不断靠近定制的目标图像的特征，以此达到对抗样本的特征远离原始人脸图像数据集分布的目的。对抗性损失函数设计为 1 减去对抗样本和定制目标图像之间的余弦相似度的样式，如公式(4.12)所示：

$$\mathcal{L}_{\text{ste}} = \sum_{\mathrm{x}^{(i)}\in \mathrm{D}} 1 - \cos\left(\mathrm{F}\left(\mathrm{x}_{\text{adv}}^{(i)}\right), \mathrm{F}\left(\mathrm{x}_{\text{tar}}\right) \right) \tag{4.12}$$

其中，x_{tar} 表示目标图像，其可以是本方案定制的目标图像，也可以是用于真正目标攻击所选定的目标人脸图像。为了进一步提高对抗损失的有效性，本章还将一批量数据的余弦相似度的方差纳入了对抗性损失的运算中。通过计算一组人脸对抗样本与设定的定制目标图像的余弦相似度，随后计算该组余弦相似度

的方差，其表示余弦相似度分布的散度。通过控制迭代过程中不断减小方差，可以有效地缩小该批次对抗样本与定制的目标图像之间的距离，从而提高攻击的成功率。对抗性损失中的方差项如公式(4.13)和(4.14)所示：

$$\mathcal{L}_{\text{var}} = \sum_{x^{(i)} \in \boldsymbol{D}} \left[\cos\left(F\left(x_{\text{adv}}^{(i)}\right), F\left(x_{\text{tar}}\right)\right) - \mu \right]^2 \tag{4.13}$$

$$\mu = \frac{1}{\boldsymbol{N}} \sum_{x^{(i)} \in D} \cos\left(F\left(x_{\text{adv}}^{(i)}\right), F\left(x_{\text{tar}}\right)\right) \tag{4.14}$$

其中，μ 表示一组数据计算得出的余弦相似度的均值，$\mathcal{L}_{\text{var}}$ 表示这一组数据计算得出的余弦相似度的方差。为了可视化方差计算对于最终对抗样本和定制目标图像之间离散度变化，本节将学习过程中产生的部分数据做了记录，对于选中的 20 组余弦相似度数值的直方图进行计数，分别使用 T_1、T_2、T_3 和 T_4 标记四个训练过程的瞬间，记录该组数据余弦相似度数值随着训练的进行发生的变化，相应变化如图 4.7 所示。

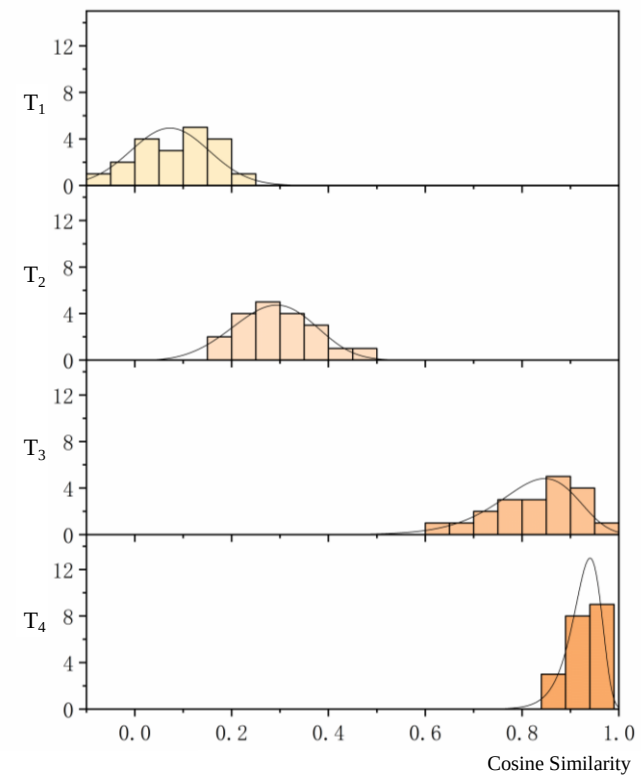


图 4.7 训练不同时刻的余弦相似度数值直方图变化趋势示意图

其中横坐标表示对抗样本与定制的目标图像之间的余弦相似度，纵坐标表示该范围内数值出现的个数，从上到下四个部分表示训练随着时间的变化过

程，即 T_1 的结果为训练开始时刻，T_2 和 T_3 表示为两个中间的时刻，T_4 表示最终训练结束的时刻。由图可知，初始 T_1 时余弦相似度数值还比较分散，没有集中的分布，随着训练不断进行，在余弦相似度数值不断向 1 靠近的过程中，其散度也越来越聚集，最终达到一个比较高的聚集度。通过该方式可以使得生成的对抗样本与设定的目标图像距离更近，以此证实在对抗性损失中引入余弦相似度方差有利于提高攻击性能。

4.3 实验与分析

本小节设置了一系列实验来验证本章所提出方法的有效性。首先给出了总体的实验设置。后续将本章提出的方法与其他有关通用对抗扰动方法进行了比较，结果表明本方案纳入频域信息这一维度后在扰动的攻击性和隐蔽性上都有了显著的提升。随后，验证了本框架针对真正的人脸识别目标攻击的性能，通过更改目标图片为确切的人脸身份，在目标攻击方式上同样取得一定的效果。之后进行的黑盒测试表明本方案在黑盒攻击模式下具有一定的攻击成功率，但还有很大的改进空间。最后，针对整个方案中不同模块的有效性设置了一系列消融实验，验证多种因素对本方案最终性能的影响。

4.3.1 实验设置

本方案所有实验同样使用单张 GTX TITAN XP 显卡进行加速运算，其显存容量为 12G，整体实验代码采用 Pytorch 深度学习框架进行编程。为验证本节所提方案在不同数据集上生成人脸通用对抗扰动的可行性，本实验同样选择 LFW 和 CASIA-WebFace 两个数据集生成人脸通用对抗扰动，其人脸图像尺寸经过预处理后变为$112\times112\times3$。针对两个不同数据集，训练集采用 6000 对人脸图像构成，测试集采用 2000 对人脸图像来进行通用对抗扰动的生成。在以 Arcface 模型预训练的 IResNet50、MobileFaceNet 和 MobileNetV1 三个主干提取网络上完成实验。在实验过程中，输入的图像数据的像素值范围都归一化在$[-1,1]$区间内，选择无穷范数作为扰动的强度限制方式，其扰动强度ξ设置为 0.12，训练

的批处理大小设置为 10，实验选择 Adam 优化器优化扰动，其扰动的学习率设定为 0.01，表 4.2 列出了针对不同数据集合主干提取网络的损失函数权重参数 λ_1、λ_2 和 λ_3 的选取数值。对于不同的数据集和主干提取网络获得的通用对抗扰动，本章节额外使用学习感知图像块相似度[76]（Learned Perceptual Image Patch Similarity, LPIPS）来衡量生成的对抗样本的隐蔽性。

表 4.2 不同权重的损失函数超参数

数据集	网络结构	λ_1	λ_2	λ_3
LFW	IResNet50	1	0.5	0.09
	MobileFaceNet	1.2	0.5	0.05
CASIA-WebFace	IResNet50	2	0.4	0.09
	MobileFaceNet	2.5	0.5	0.01

4.3.2 攻击性和隐蔽性评估

在对抗攻击领域中，为衡量生成的通用对抗扰动的有效性，通常都使用攻击性和隐蔽性两个指标来度量。攻击性用于评估生成的对抗样本是否具有愚弄人脸识别模型的能力，模型的愚弄率越高表明叠加通用对抗扰动后，这个数据集样本被模型识别出错的概率越高，即对抗样本攻击性越强。隐蔽性用于评估该扰动生成的对抗样本能否有效逃避过人眼视觉的检验，隐蔽性越强表明叠加通用对抗扰动后的对抗样本和原始干净样本之间视觉差异越小。本实验使用 LFW 数据集和 CASIA-WebFace 数据集在 IResNet50 和 MobileFaceNet 两个预训练的面部特征提取骨干网络上学习通用对抗扰动。本章提出的方案在各自的人脸图像测试集上实现了 80% 左右的攻击成功率。值得注意的是，在 IResNet50 骨干网络上的攻击成功率达到了 85% 。为了度量隐蔽性，除了主观的人类视角观察外，本实验还使用了三个客观的图像质量评价指标来衡量。其中 SSIM 和 PSNR 都是计算的评分数值越高表示图像质量越好，而 LPIPS 计算得出的评分数值越小表示图像的质量越好。对抗样本的隐蔽性与图像质量指标成正相关，当应用于生成的面部图像时，这些图像质量的评分有助于量化生成的对抗样本的隐蔽性。本章最终的实验结果如表 4.3 所示，证明本方案产生的扰动在添加到面部图像时表现出良好的攻击性和隐蔽性。

表 4.3 不同扰动生成方案的攻击性和隐蔽性对比结果

数据集	网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑	LPIPS↓
LFW	IResnet50	Random	20.8	0.4003	18.9271	0.6155
		UAP	76.6	0.7607	27.5761	0.2538
		FTGAP	82.1	0.8852	31.7033	0.2361
		FaUAP-FBF	85.0	0.9219	33.4856	0.1365
	Mobile-FaceNet	Random	23.2	0.3614	18.0833	0.6331
		UAP	79.1	0.8487	29.7213	0.2281
		FTGAP	79.4	0.8507	30.5232	0.1942
		FaUAP-FBF	81.0	0.8952	32.1317	0.1831
CASIA-WebFace	IResnet50	Random	37.1	0.4901	20.8767	0.4595
		UAP	71.2	0.7759	28.0403	0.2947
		FTGAP	76.3	0.8544	30.4638	0.2649
		FaUAP-FBF	78.6	0.9014	32.1339	0.1155
	Mobile-FaceNet	Random	30.3	0.5099	21.4036	0.4418
		UAP	76.3	0.7612	27.5446	0.3006
		FTGAP	78.4	0.8691	30.6668	0.2156
		FaUAP-FBF	80.4	0.8767	31.3463	0.2269

由于目前还缺乏针对人脸识别领域的通用对抗扰动技术方案的研究，在对本方案性能的比较分析中，本章将已有的针对其他领域的通用对抗扰动产生方案与本章提出的方案进行对比，其中 UAP[32]是一种针对自然图像的图像分类任务设计的通用对抗扰动生成方案，FTGAP[75]是对纹理图像进行识别任务所提出的通用对抗扰动生成方式。虽然上述两种方案都是为其他任务设计的，本实验通过进行微小修改使之适配人脸识别的任务。同时为了验证通过精心设计产生扰动的方案是有效的，本实验还采用随机生成的高斯噪声作为扰动进行对抗样本生成实验，通过测试使用该方式获得的对抗样本在同样模型上的攻击成功率和隐蔽性指标来完善论证。所有方案在选定的数据集和测试的主干提取网络上的实验指标结果如表 4.3 所示。从表中可以看出，为自然图像设计的 UAP 确实表现出了一定程度的愚弄效果。然而，由于其只考虑了空域的全局信息，与其他两种方案相比，其隐蔽性指标的最终表现效果相对较弱。另一种针对纹理图像的 FTGAP 方法引入了频域信息的概念，但相比本章提出的方法，其缺乏对频域信息更细节层次的探索，不能更好地挖掘频域内的有效信息进而提升通用对抗扰动的性能。

本方案采用分频段的方式对人脸数据集和通用对抗扰动进行分析，通过结合不同频段的有效信息，将频域作为一个额外的维度来控制扰动的生成，以此

生成更具攻击性和更佳隐蔽性的扰动，同时本方案的损失函数也采用目标攻击的模式，通过产生一张在训练数据集分布之外的定制目标图像，并使用方差将数据离散度的度量标准纳入衡量指标，以此控制扰动的攻击性和隐蔽性指标。上述不同方案的视觉可视化图像如图 4.8 所示，其中（a,e,i,m）表示干净的人脸面部图像，（b,f,j,n）表示使用 UAP 生成的对抗样本可视化图像，（c,g,k,o）表示使用 FTGAP 生成的对抗样本可视化图像，最后（d,h,l,p）表示使用本方案生成的人脸对抗样本可视化图像。其中 UAP 方式在全局添加扰动，在人脸光滑区域会有明显的视觉缺陷，FTGAP 生成扰动的方式考虑了频域信息，其隐蔽性指标有一定提升，本方案采用更细致的频域信息，生成的扰动在低频区域视觉缺陷不明显，综上本方案的视觉可视化效果更显著，生成的图像看起来更正常。

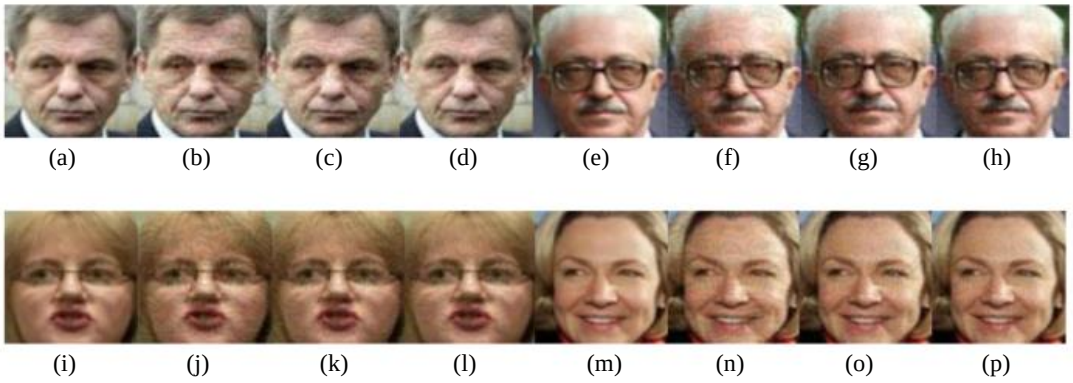


图 4.8 不同扰动生成方案产生对抗样本可视化图像

4.3.3 目标攻击性能评估

本研究方案将针对人脸识别模型的非目标对抗攻击和目标对抗攻击整合到一个统一的目标攻击框架中，之前的实验结果证实了定制的目标图像可以实现非目标攻击的任务，且取得了一定的攻击有效性。针对目标攻击的性能，本小节在 IResNet50 和 MobileFaceNet 两个主干网络上使用 LFW 和 CASIA-WebFace 两个数据集进行了目标攻击的实验。通过将原框架中的定制目标图像替换为确切的某身份的人脸图像，那么非目标的对抗攻击任务就变成了针对该确切人脸身份的目标攻击任务，在其他方面不变的情况下，测试了攻击的成功率和客观的隐蔽性指标如表 4.4 所示，同样的攻击成功率指标都能达到 80%以上，且其隐蔽性指标也达到人眼不可感知要求。这一系列实验的结果充分验证了本方案提

出的统一框架的有效性，无论针对目标攻击任务还是针对非目标攻击任务都有良好的性能指标，产生有效的对抗样本。

表 4.4 基于频带滤波器驱动的人脸通用对抗扰动方案的目标攻击性能

数据集	网络结构	愚弄率(%)↑	SSIM↑	PSNR↑	LPIPS↓
LFW	IResnet50	84.9	0.9199	33.2516	0.1031
	MobileFaceNet	82.9	0.9056	32.3951	0.1122
CASIA-WebFace	IResnet50	81.5	0.9184	32.9812	0.1171
	MobileFaceNet	80.0	0.8896	31.3293	0.1391

4.3.4 黑盒攻击性能评估

本文也针对所提出方案进行了黑盒测试，与前文设置一样，实验使用 IResNet50、MobileFaceNet 和 MobileNetV1 三个主干提取网络进行测试，将其中一个网络作为训练人脸通用对抗扰动的可访问参数的白盒模型，剩下的两个网络作为无法访问内部详细参数的黑盒模型，在 LFW 数据集上进行相应的黑盒测试，测试结果如下表 4.5 所示，该实验评估了使用三个不同主干网络的攻击性性能，其中第一列的网络表示可访问内部参数的白盒方式，后续列的网络则表示不可直接访问内部参数的黑盒方式，使用对角线上的数据表示白盒方式下的攻击成功率，其他位置的数据表示在不同主干网络情况下的黑盒攻击成功率。

表 4.5 黑盒攻击性能

愚弄率(%)	IResNet50	MobileFaceNet	MobileNetV1
IResNet50	85.0	11.4	15.9
MobileFaceNet	42.7	80.9	35.1
MobileNetV1	48.3	23.1	79.2

通过表格中数据可知，在黑盒攻击的模式下，因不可访问模型内部参数及运算过程，通用对抗扰动的攻击成功率都有很大程度的下降，说明该方案在攻击未知模型时会受到一定影响。不过在 MobileNetV1 上生成的扰动迁移到 IResNet50 上测试能有 48.3%的攻击成功率，这也一定程度说明该方案在黑盒模式下也有一定的有效性。由此实验结果证实，在人脸识别领域实现强大的黑盒通用对抗攻击性能仍然具有挑战性。

4.3.5 消融实验

分频段操作对实验结果的影响：在本实验中通过在频域中叠加三个不同频

段的扰动来产生最终的通用对抗扰动。通过学习三个频带滤波器的参数，本方案将对抗性扰动和干净人脸图像分离到三个不同的频带信息中进行操作。通过对应频段的扰动信息和干净图像信息进行叠加获得对应频段的对抗样本信息。使用这种频率分割的方式可以更详细和有效的考虑频域中的信息，以此在不同频段产生有效的扰动。为了证明该分频段操作的有效性，本实验通过在整个频谱上直接产生全频段的通用扰动来生成最终的人脸对抗样本。在本框架其他内容不变的情况下，通过分频段和全频段两种生成扰动的方式，测试了在 IResNet50 和 MobileFaceNet 两个主干网络上的攻击成功率和隐蔽性指标，其结果如表 4.6 所示，由表格内容可知，全频段方式产生的扰动在两个网络上测试得到的数据在攻击性方面有 3% 到 4% 的攻击成功率差距，且其隐蔽性性能也有一定程度的下降，说明本方案采用分频段操作可以为生成的对抗扰动考虑到更详细的频域信息，以此生成更佳的扰动。

表 4.6 分频段对结果的影响

网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑	LPIPS↓
IResNet50	FaUAP-FBF	85.0	0.9219	33.4856	0.1365
	全频段	81.6	0.8929	32.0886	0.2129
MobileFaceNet	FaUAP-FBF	80.9	0.8952	32.1317	0.1831
	全频段	77.2	0.8586	30.6407	0.2149

可学习频带滤波器对实验结果的影响：在本实验中通用对抗扰动的生成主要利用频域提取阶段的三个可学习的频带滤波器。如前文所述，本方案首先选择三个固定参数的高频、中频和低频滤波器，随后引入了三个参数可变的附加滤波器，其参数通过指定的优化器依据损失函数的计算进行迭代调整。通过将这三个参数可变的附加滤波器对应叠加到三个固定参数的滤波器中，得到了本方案使用的三个可学习的频带滤波器。通过这种方式可以使生成扰动的过程中从不同的高频、中频和低频分段中提取更有效的信息，从而提高扰动的攻击性和隐蔽性。为了证明使用可学习频带滤波器的有效性，本实验通过只使用固定参数的高频、中频和低频滤波器来替代三个可学习的频带滤波器，在其他框架内容不变的情况下，以此来训练通用对抗扰动。在上述方式下测试了在 IResNet50 和 MobileFaceNet 两个主干网络上的攻击成功率和隐蔽性指标，其结果如表 4.7 所示，由表格可得，只是用固定参数的滤波器最终提取出的信息具有

一定的局限性，通过使用可学习的滤波器，可以获得针对攻击性指标和隐蔽性指标的更有效的参数信息，从而在生成的对抗样本上，有更优异的攻击性和隐蔽性指标。

表 4.7 可学习频带滤波器对结果的影响

网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑	LPIPS↓
IResNet50	FaUAP-FBF	85.0	0.9219	33.4856	0.1365
	固定滤波器	84.5	0.8816	31.5579	0.1957
MobileFaceNet	FaUAP-FBF	80.9	0.8952	32.1317	0.1831
	固定滤波器	80.2	0.8607	30.7332	0.2511

设置定制的目标图像对实验结果的影响：在本实验中，该定制的目标图像通过了一定的训练过程获取，如前文所述通过从数据集中随机选择 6000 张人脸图像，进行均值计算以此获得代表数据集分布的均值图像，通过不断训练对设定的目标图像进行调整，使其特征与均值图像的相似性最小化，以此生成远离原始数据集分布的定制目标图像。随后，使用该定制目标图像进行通用扰动的产生，以降低人脸对抗样本与原始干净人脸图像之间的相似性，从而实现针对人脸识别的对抗攻击。为了验证使用该定制目标图像的有效性，本节采用一张随机生成的高斯噪声模拟目标图像的作用，在 LFW 数据集上测试两个主干提取网络的攻击成功率和隐蔽性指标，相应的结果如表 4.8 所示。因随机高斯噪声图像也可以认为是数据集分布之外的数据，因此使用该方式也会有一定的攻击效果。但本方案通过预先训练的方式将定制的目标图像有目的地远离原始图像分布，所以使用定制的目标图像能准确产生一个有效的目标，最终生成的扰动在攻击性和隐蔽性方面能产生更优异的效果。

表 4.8 定制目标图像对结果的影响

网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑	LPIPS↓
IResNet50	FaUAP-FBF	85.0	0.9219	33.4856	0.1365
	随机噪声	82.7	0.9149	33.0586	0.1529
MobileFaceNet	FaUAP-FBF	80.9	0.8952	32.1317	0.1831
	随机噪声	78.1	0.8809	31.5233	0.1946

损失中使用方差对实验结果的影响：在本实验的通用对抗扰动生成框架的损失函数中，对抗性损失部分纳入了分布的概念。具体来说，通过计算对抗性样本和定制的目标图像之间的余弦相似度的方差，获取一批量数据集的余弦相

似度的聚集度情况。通过不断减小这个方差的值，可以确保最终一个批处理大小内的对抗样本特征更接近定制目标图像的特征，即生成的对抗样本特征不断远离原始人脸数据集的分布。该方式旨在提高生成的通用对抗扰动的攻击成功率性能。为了验证使用该方式的有效性，在其他条件不变的情况下，通过在学习通用对抗扰动的过程中删除基于分布的损失函数分量。使用 LFW 数据集在两个主干提取网络上测试攻击成功率和隐蔽性指标的结果如表 4.9 所示。从表中可以得出，在损失函数中不使用方差这一指标最终生成的对抗样本在攻击成功率上有大幅的下降，而在隐蔽性指标的比较方面两者差异不大。说明在损失函数中使用方差的计算方式可以有效提高攻击的性能。

表 4.9 计算余弦相似度方差对结果的影响

网络结构	方法	愚弄率(%)↑	SSIM↑	PSNR↑	LPIPS↓
IResNet50	FaUAP-FBF	85.0	0.9219	33.4856	0.1365
	无方差	80.2	0.9199	32.8674	0.1392
MobileFaceNet	FaUAP-FBF	80.9	0.8952	32.1317	0.1831
	无方差	76.0	0.8857	31.7936	0.1893

4.4 本章小结

本章提出了一种基于频带滤波器驱动的人脸通用对抗扰动。通过分析现有通用对抗扰动攻击在人脸识别领域的不足，阐述本章引入频域信息的研究动机。后续详细介绍了本章方案的框架设计情况，并分两个模块解析本方案生成通用对抗扰动的原理，最后通过全面的实验评估本章所提方法在各方面的性能。实验结果证明，相较于现有的通用对抗扰动生成方案，本方案在考虑频域信息的基础上，生成的扰动在攻击性指标和隐蔽性指标方面都有显著提升。

第五章　总结与展望

5.1　总结

本文主要研究了适用于人脸识别系统的通用对抗扰动技术，现有的通用对抗扰动生成算法大都是针对图像分类任务设计的，这种通用对抗攻击对人脸识别系统的影响尚未得到充分研究。本文基于自然图像通用对抗扰动的研究思路，通过分析自然图像与人脸图像的差异性，深入挖掘了人脸识别数据集及损失函数的独特特性，据此在空域和频域两个角度上，分别提出了面向人脸识别任务的通用对抗扰动生成方案，本文主要工作总结如下：

（1）提出了基于空域分析的人脸通用对抗扰动生成技术，与现有的工作直接在人脸图像全局优化扰动的方法不同，本方案从人脸识别的可解释性分析出发，探究了人脸图像中不同区域的语义特征对最终识别准确率的影响，基于此前提，将人脸图像的语义关键区域作为生成扰动的核心依据，以此作为制定扰动策略的重要准则。首先利用关键点检测技术识别出数据集中具有显著语义特征的区域，作为表征整个人脸数据集的总体特征。随后通过迭代优化过程训练可学习的流场，利用该流场微调表征数据集总体特征的先验区域，使该区域弥补缺失的个体特征信息，确保扰动生成在人脸更有效的区域。此外，本方案目标函数包含了适用于人脸图像的隐蔽性损失和对抗性损失，确保有效实现攻击性和隐蔽性的双重优化。最终实验结果表明，相较于现有通用对抗扰动生成算法，本方案在维持攻击性同时能显著提升隐蔽性。

（2）提出了基于频域分析的人脸通用对抗扰动生成技术，相较于仅依赖于空间域信息的扰动生成方式，本方案通过引入频域信息为生成扰动提供了一个额外的考量维度。通过深入分析自然图像和人脸图像在频域视角上的差异性，在频域生成扰动能更适应人脸图像的特点。为了对频域信息做更细致地分析，本方案通过分别学习高频、中频和低频段的扰动信息，采用分频段的方式生成对应的扰动，以此利用了不同频段的信息且有效规避了直接在空域全局叠加扰动造成的视觉缺陷。本方案以目标攻击的模式完成非目标攻击任务，提前优化

了一幅在数据集分布之外的定制目标图像，并且将对抗样本和干净图像之间的余弦相似度方差纳入计算，有效提升了生成扰动的攻击性和隐蔽性。实验结果表明，与现有攻击方式相比，所提出的综合考虑频域信息生成扰动的方案在攻击性能上有更显著的提升。

5.2 展望

本文研究了面向人脸识别系统的通用对抗扰动技术，并从空域和频域两个角度分别提出了语义关键区域控制的扰动生成算法和基于频带滤波器驱动的扰动生成算法，揭示了人脸识别系统普及的环境下存在的安全问题，促进针对人脸识别系统防御机制的完善。在未来的研究中，为进一步提升通用对抗扰动的性能，可以在下面两个方向展开研究：

本文在空域角度提出的语义关键区域控制的扰动生成算法属于加性扰动，通过在空间域中选定人脸图像的有效位置叠加通用扰动生成对抗样本。因此，扰动的有效性还取决于叠加这一操作，该方式针对侧脸图像会有一定局限性，后续研究可以探讨一种更高层语义的加性的扰动生成方式，通过结合人脸表情、姿势等高层语义特征进行更细致的空间扰动叠加。

本文在频域角度提出的频带滤波器驱动的扰动生成算法能够有效利用频域信息进行扰动的生成，频域信息相对空域信息在不可感知性方面有显著提升。后续研究可以通过纳入更多维度信息综合考量生成扰动的攻击性和隐蔽性，比如在其他感知色彩空间或者隐空间中生成有效对抗扰动，以此提升扰动性能。

参考文献

[1] Samek W, Montavon G, Lapuschkin S, et al. Explaining deep neural networks and beyond: A review of methods and applications[J]. Proceedings of the IEEE, 2021, 109(3): 247-278.

[2] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[3] Maurício J, Domingues I, Bernardino J. Comparing vision transformers and convolutional neural networks for image classification: A literature review[J]. Applied Sciences, 2023, 13(9): 5521.

[4] 卢笑, 竺一薇, 阳牡花, 等. 联合图像与单目深度特征的强化学习端到端自动驾驶决策方法[J]. 武汉大学学报, 2021, 46(12): 1862-1871.

[5] 李琳辉, 周彬, 连静, 等. 基于社会注意力机制的行人轨迹预测方法研究[J]. 通信学报, 2020, 41(06): 175-183.

[6] 彭之军. 人脸识别技术进展综述[J]. 信息与电脑(理论版), 2023, 35(15): 168-171.

[7] Yerlikaya F A, Bahtiyar Ş. Data poisoning attacks against machine learning algorithms[J]. Expert Systems with Applications, 2022, 208: 118101.

[8] Li Y, Jiang Y, Li Z, et al. Backdoor learning: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[9] 曹灿, 司强, 游雪松, 等. 针对人脸识别的物理对抗攻击研究综述[J]. 数据与计算发展前沿, 2023, 5(03): 49-65.

[10] Baniecki H, Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: A survey[J]. Information Fusion, 2024: 102303.

[11] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]. In: Proceedings of International Conference on Learning Representations, 2014: 1-10.

[12] Komkov S, Petiushko A. Advhat: Real-world adversarial attack on arcface face id system[C]. 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 819-826.

[13] Deng Z, Yang X, Xu S, et al. Libre: a practical bayesian approach to adversarial detection[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 972-982.

[14] Zhang S, Gao H, Rao Q. Defense against adversarial attacks by reconstructing images[J]. IEEE Transactions on Image Processing, 2021, 30(01): 6117-6129.

[15] 崔廷玉, 张武, 贺正芸, 等. 针对眼部掩模的人脸识别对抗贴片研究[J]. 计算机技术与发展, 2023, 33(06): 139-146.

[16] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]. In: Proceedings of International Conference on Learning Representations, 2015: 1-11.

[17] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale[C]. In: Proceedings of the 5th International Conference on Learning Representations, 2017:1-12.

[18] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9185-9193.

[19] Moosavi-Dezfooli S-M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2574-2582.

[20] Papernot N, Mcdaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arxiv preprint arxiv: 160507277, 2016.

[21] Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]. In: Proceedings of the 2016 IEEE European Symposium on Security and Privacy, 2016: 372-387.

[22] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]. In: Proceedings of International Conference on Learning Representations, 2018: 1-28.

[23] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy, 2017: 39-57.

[24] Liu Y, Chen X, Liu C, et al. Delving into transferable adversarial examples and black-box attacks[C]. In: Proceedings of the 5th International Conference on Learning Representations, 2017:1-24.

[25] Zhou W, Hou X, Chen Y, et al. Transferable adversarial perturbations[C]. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018: 452-467.

[26] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(05): 828-841.

[27] Huang Q, Katsman I, He H, et al. Enhancing adversarial example transferability with an intermediate level attack[C]. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 4733-4742.

[28] Wu D, Wang Y, Xia S-T, et al. Skip connections matter: On the transferability of adversarial examples generated with resnets[C]. In: Proceedings of the 8th International Conference on Learning Representations, 2020:1-15.

[29] Duan R, Ma X, Wang Y, et al. Adversarial camouflage: hiding physical-world attacks with natural styles[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1000-1008.

[30] Andriushchenko M, Croce F, Flammarion N, et al. Square attack: a query-efficient black-box adversarial attack via random search[C]. Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII, 2020: 484-501.

[31] 黄立峰, 庄文梓, 廖泳贤, 等. 一种基于进化策略和注意力机制的黑盒对抗攻击算法[J]. 软

件学报, 2021, 32(11): 3512-3529.

[32] Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1765-1773.

[33] Mopuri K R, Garg U, Babu R V. Fast feature fool: a data independent approach to universal adversarial perturbations[J]. arXiv preprint arXiv: 170705572, 2017.

[34] Mopuri K R, Ganeshan A, Babu R V. Generalizable data-free objective for crafting universal adversarial perturbations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(10): 2452-2465.

[35] Hayes J, Danezis G. Learning universal adversarial perturbations with generative models[C]. 2018 IEEE Security and Privacy Workshops (SPW), 2018: 43-49.

[36] Poursaeed O, Katsman I, Gao B, et al. Generative adversarial perturbations[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4422-4431.

[37] Mopuri K R, Ojha U, Garg U, et al. Nag: network for adversary generation[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 742-751.

[38] Liu H, Ji R, Li J, et al. Universal adversarial perturbation via prior driven uncertainty approximation[C]. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 2941-2949.

[39] Zhang C, Benz P, Imtiaz T, et al. Understanding adversarial examples from the mutual influence of images and perturbations[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 14521-14530.

[40] Dai J, Shu L. Fast-UAP: An algorithm for expediting universal adversarial perturbation generation using the orientations of perturbation vectors[J]. Neurocomputing, 2021, 422(01): 109-117.

[41] Li D, Zhang J, Huang K. Universal adversarial perturbations against object detection[J]. Pattern Recognition, 2021, 110(01): 1-12.

[42] Zhang C, Benz P, Karjauv A, et al. Data-free universal adversarial perturbation and black-box attack[C]. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 7868-7877.

[43] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]. In: Proceedings of the 2016 Acm Sigsac Conference on Computer and Communications Security, 2016: 1528-1540.

[44] Ibsen M, Rathgeb C, Brechtel F, et al. Attacking Face Recognition with T-shirts: Database, Vulnerability Assessment and Detection[J]. IEEE Access, 2023.

[45] 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.

[46] Zolfi A, Avidan S, Elovici Y, et al. Adversarial mask: Real-world adversarial attack against

face recognition models[J]. arxiv preprint arxiv:2111.10759, 2021, 2(3).

[47] Rozsa A, Günther M, Boult T E. LOTS about attacking deep features[C]. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017: 168-176.

[48] Dabouei A, Soleymani S, Dawson J, et al. Fast geometrically-perturbed adversarial faces[C]. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), 2019: 1979-1988.

[49] Dong Y, Su H, Wu B, et al. Efficient decision-based black-box adversarial attacks on face recognition[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7714-7722.

[50] Parmar R, Kuribayashi M, Takiwaki H, et al. On fooling facial recognition systems using adversarial patches[C]. In: 2022 International Joint Conference on Neural Networks (IJCNN). IEEE, 2022: 1-8.

[51] Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks[J]. Journal of Machine Learning Research, 2011, 15:315-323.

[52] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012, 1097-1105.

[53] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]. In: Proceedings of International Conference on Learning Representations, 2015: 1-14.

[54] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[55] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[56] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[57] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 815-823.

[58] Chen S, Liu Y, Gao X, et al. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices[C] Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13. Springer International Publishing, 2018: 428-438.

[59] Liu W, Wen Y, Yu Z, et al. Sphereface: Deep hypersphere embedding for face recognition[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 212-220.

[60] Wang H, Wang Y, Zhou Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 5265-5274.

[61] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4690-4699.

[62] Shan S, Wenger E, Zhang J, et al. Fawkes: Protecting privacy against unauthorized deep learning models[C]. In: Proceedings of the 29th USENIX Security Symposium, 2020

[63] Zhong Y, Deng W. OPOM: Customized Invisible Cloak towards Face Privacy Protection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.

[64] Mery D. True black-box explanation in facial analysis[C]. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1596-1605.

[65] Xiao C, Zhu J Y, Li B, et al. Spatially transformed adversarial examples[J]. arxiv preprint arxiv:1801.02612, 2018.

[66] Zheng Z, Zheng L, Hu Z, et al. Open set adversarial examples[J]. arxiv preprint arxiv:1809.02681, 2018, 3.

[67] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]. 2008.

[68] Yi D, Lei Z, Liao S, et al. Learning face representation from scratch[J]. arxiv preprint arxiv:1411.7923, 2014.

[69] Duta I C, Liu L, Zhu F, et al. Improved residual networks for image and video recognition[C]. In: 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 9415-9422.

[70] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arxiv preprint arxiv:1704.04861, 2017.

[71] Kingma D P, Ba J. Adam: a method for stochastic optimization[C]. In: Proceedings of International Conference on Learning Representations, 2015: 1-13.

[72] 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.

[73] Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM[C]. In: 2010 20th International Conference on Pattern Recognition. IEEE, 2010: 2366-2369.

[74] Ye Z, Cheng X, Huang X. Fg-uap: Feature-gathering universal adversarial perturbation[C] International Joint Conference on Neural Networks (IJCNN). IEEE, 2023: 1-8.

[75] Deng Y, Karam L J. Frequency-tuned universal adversarial perturbations[C]. Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer International Publishing, 2020: 494-510.

[76] Korhonen J, You J. Peak signal-to-noise ratio revisited: Is simple beautiful?[C] Fourth International Workshop on Quality of Multimedia Experience. IEEE, 2012: 37-38.

作者在攻读硕士学位期间的研究成果

一、论文/专利

[1] **Jin X**, Liu Y, Sun G, et al. KRT-FUAP: Key Regions Tuned via Flow Field for Facial Universal Adversarial Perturbation[J]. Applied Sciences, 2024, 14(12): 4973. (本人第一作者，SCI 三区)

[2] **Jin X**, Wu H, Malik A, et al. Leveraging Universal Adversarial Perturbations and Frequency Band Filters against Face Recognition[J]. Information Sciences(Under Review) (本人第一作者，SCI 一区，CCF-B 类期刊)

[3] **金玺**, 孙广玲, 吴汉舟, 王莹桂. 通用对抗扰动、人脸对抗样本生成方法及装置: 202410039635.6[P]. 2024-1-11. (本人第一作者，已受理)

致 谢

随着毕业论文的最终落笔，我的七年上大求学时光也迎来尾声。在此，我向所有在研究生期间给予我支持和帮助的人表达我深深的感激之情。

在研究生阶段，我很荣幸能加入人工智能安全实验室，在这里我的学术水平得到了提高，学术视野得到了拓展，每一次报告和交流都是我学习和成长的机会，这里的点点滴滴组成了我科研生涯不可或缺的篇章。

首先我要特别感谢我的导师吴汉舟老师，在硕士的三年学习和生活中，吴老师以其严谨的学术态度和深厚的专业知识，给予了我精心的指导和无私的帮助，对我的影响深远。同时，我还要感谢孙广玲老师，从科研项目的开题到前期准备，再到中期讨论方案以及后期小论文撰写，每一个环节孙老师都深度参与，给了我许多宝贵的意见。在毕业论文的选题、实验设计到最终成文的过程中，两位老师都给予了我极大的帮助，我向他们表示深深的敬意和感谢。

我还要感谢实验室的师兄师姐和同门们。感谢他们在科研方面和生活方面给我提供的帮助，无论是学术上的交流，还是实验上的合作，因为有一群志同道合的伙伴，我的科研生涯变得丰富多彩，我也衷心祝愿他们未来一切顺利。

此外，我要感谢我的家人和朋友们的关心与支持。在整个研究生生涯中，他们是我最坚强的后盾。在我遇到困难与挑战时，是他们一直陪伴着我，给了我坚持前进的力量，让我在科研的道路上更加坚定，充满信心。

最后，我要真挚地感谢评阅本论文的专家老师们，感谢你们提出的宝贵建议，让这篇论文不断完善。愿你们在今后的科研道路上硕果累累，学术生涯一帆风顺！

金玺

上海大学