结合自注意力特征融合的人脸模板重建技术

王亚男,董建民,孙冰阳,王宁博

(西藏民族大学 信息工程学院,陕西 咸阳 712082)

摘 要: 针对人脸识别系统模板反演重建方法存在的人脸模板特征数据单一,模板向量不足以捕捉多样性,生成的图像在细节上不够丰富的问题,提出一个新的人脸模板映射网络(FFMapNet)。首先,通过自注意力特征提取模块(SAFEM),结合反卷积层对初始人脸模版特征进行上采样,并引入高维通道特征;其次,利用自注意力机制的全局信息建模功能,从模板特征反演中提取并增强特征;最后,利用特征融合模块(FFM)进一步处理和融合增强后的特征,提升人脸模板到人脸生成器网络中间潜在空间(LS)映射的一致性,生成更具表达能力的特征映射。在MOBIO、LFW和AgeDB数据集上对最先进的人脸识别系统进行模板反演攻击,实验结果表明所提方法的重建质量与攻击成功率均优于现有方法。

关键词: 人脸识别;模板反演;黑盒攻击;自注意力机制;特征融合

DOI: 10. 11907/rjdk. 241996 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391 **文献标识码:** A **文章编号:** 1672-7800(2025)003-0177-08

Face Template Reconstruction Technology Combined With Self-Attention Feature Fusion

WANG Yanan, DONG Jianmin, SUN Bingyang, WANG Ningbo

(*College of Information Engineering*, *Xizang Minzu University*, *Xianyang* 712082, *China*)

Abstract: Aiming at the problem that the face template feature data of the template inversion and reconstruction method in face recognition system is single, the template vector is not enough to capture diversity, and the generated image is not rich in details, a new face template mapping network (FFMapNet) is proposed. Firstly, the self attention feature extraction module (SAFEM) is used in conjunction with the deconvolution layer to upsample the initial facial template features and introduce high-dimensional channel features; Secondly, utilizing the global information modeling function of self attention mechanism, features are extracted and enhanced from template feature inversion; Finally, the feature fusion module (FFM) is used to further process and fuse the enhanced features, improving the consistency of the latent space (LS) mapping from the face template to the face generator network, and generating more expressive feature maps. Template inversion attacks were conducted on state-of-the-art facial recognition systems on MOBIO, LFW, and AgeDB datasets, and experimental results showed that the proposed method outperformed existing methods in terms of reconstruction quality and attack success rate.

Key Words: face recognition; template inversion; blackbox attack; self-attention mechanism; feature fusion

0 引言

近年来,自动人脸识别(Automatic Face Recognition, FR)已广泛成为生物特征认证工具[1]。在FR系统数据库中存储着从用户人脸提取的特征,这些特征被称为人脸模板或嵌入,任何针对生物特征识别系统的攻击都可能威胁用户隐私和系统安全。

目前,已有文献研究了当前最先进的面部识别系统(State-of-the-Art Facial Recognition System, SOTA FR)中

收稿日期: 2024-11-25

扫描二维码阅读全文:

基金项目: 西藏自治区重点研发计划(XZ202401JD0009);西藏民族大学校内科研基金项目(22MDY013)

作者简介: 王亚男(2000-),女,CCF会员,西藏民族大学信息工程学院硕士研究生,研究方向为图形图像处理、三维重建;董建民(1977-),男,西藏民族大学信息工程学院副教授、硕士生导师,研究方向为图形图像处理与科学可视化、多维数据分析与处理;孙冰阳(1999-),女,西藏民族大学信息工程学院硕士研究生,研究方向为医学影像处理、深度学习;王宁博(1999-),男,西藏民族大学信息工程学院硕士研究生,研究方向为边缘检测。本文通讯作者:董建民。

不同类型的攻击方式，并评估了 FR 系统在面对这些攻击时的脆弱性[2-5]。其中，模板反转（Template Inversion，TI）是一种严重威胁用户隐私的攻击手段，攻击者利用对抗生成网络等技术分析人脸识别系统的输出或中间结果，试图通过系统数据库中的模板重建出能欺骗系统的合成人脸图像[6,7]；黑盒攻击是指在不了解人脸识别系统的内部结构和参数的情况下，只通过输入和输出来观察模型行为，能攻击多种人脸识别系统，因此具有较高的隐蔽性和灵活性，防御难度较大。

为此，本文针对黑盒攻击展开研究，从 FR 系统获得的人脸模板进行模板反演重建，提出一个新的基于自注意力的特征融合映射网络（Face Feature Mapping Network，FFMapNet）。首先，引入自注意力特征增强模块（Self-Attention Feature Extraction Module，SAFEM）与特征融合模块（Feature Fusion Model，FFM）的联合机制，利用自注意力机制中全局信息的建模能力捕捉长距离依赖关系，增强特征表达能力；其次，在真实人脸数据集上进行模板反演实验，揭示 SOTAFR 存在的安全威胁，为提升系统安全性提供重要参考与改进方向。

1 相关工作

现有人脸模板反演重建方法可根据重建方法（基于优化或基于学习）、重建的维度（二维或三维）、生成图像的分辨率（低分辨率 112×112 或高分辨率 512×512），以及攻击类型（黑盒或白盒）进行分类。其中，重建分辨率是直观影响重建面部细节的关键因素，本文基于该点分类讨论已有的重建方法。

在低分辨率人脸重建方法中，Zhmoginov 等[8]针对优化和学习两种方法提出反演攻击技术。当面对白盒（Threat Intelligence，TI）攻击场景时，通常采用带正则化的梯度上升算法。首先，从随机噪声或参考图像中，结合全变差（Total Variation，TV）和拉普拉斯金字塔梯度归一化技术生成平滑图像；其次，最小化中间层特征图的 ℓ_2 距离来控制姿态一致性。基于学习的方法则利用反卷积神经网络直接生成高分辨率图像，主要关注的是重建图像的视觉质量，未深入探讨其 TI 攻击效果。

文献[9-12]只基于学习方法重建低分辨率人脸。Shahreza 等[9]提出一种专为白盒攻击设计的反卷积神经网络重建方法，综合优化了多个损失项，包括一个通过白盒面部识别模型的特征提取器，以保持重建图像身份一致性的损失项，但仅依赖反卷积上采样可能会引入额外的噪声和模糊问题，从而导致生成的图像不够清晰。Cole 等[10]提出一种结合多层感知机（Multi-Layer Perceptron，MLP）和卷积神经网络（Convolutional Neural Network，CNN）分别训练 MLP 和 CNN，并在推理阶段结合形变函数进行图像重建，通过多项损失函数进行端到端训练，也适用于白盒重建。然而，在安全性评估中该方法仅通过相似度直方图评估黑盒重建效果，未充分验证其在不同攻击场景下的鲁棒性和实际攻击效果。Mai 等[11]提出两个新型的基于 CNN 的黑盒人脸重建网络——NbNet-A 和 NbNet-B，他们均由反卷积层和卷积层构成，但连接方式不同。同时，为了优化重建效果，还针对两个损失函数设计了 NBNetA-M、NBNetA-P、NBNetB-M 和 NBNetB-P 共 4 种模型，并在不同重建场景进行实验比较，实验发现基于 CNN 重建的低分辨率人脸存在模糊伪影、面部细节缺失现象。

生成对抗网络（Generative Adversarial Network，GAN）在低分辨率人脸模板重建中的表现也十分出色。Duong 等[12]提出一个基于 GAN 的人脸模板黑盒重建框架，采用双注入学习机制，并使用姿势条件生成对抗网络的生成器结构和多种损失函数进行训练。虽然，该方法在黑盒攻击中的表现接近白盒攻击，但由于缺乏公开的网络细节和源代码，从而影响了结果的可复现性。此外，漏洞分析部分仅通过匹配精度评估 TI 攻击效果，未全面探讨 FR 系统在不同阈值时的脆弱性。

目前，已有的高分辨率模板反演重建方法均实现了黑盒攻击场景下的模板反演。Vendrow 等[13]提出一种基于贪婪随机优化的方法，在条件生成对抗网络（Style Generative Adversarial Network，StyleGAN）的潜在空间中搜索最优潜在向量，以生成与目标人脸模板最匹配的图像。然而，该方法仅基于真实人脸数据集中的 20 个样本进行实验，未在更大规模的数据集上计算模板差异，因此有效性存在质疑。Dong 等[14]尝试使用均方误差（Mean Squared Error，MSE）损失函数训练 MLP 回归模型，以重建高分辨率（1 024×1 024）图像，还在安全性分析中考虑了两种攻击类型，但该方法过度依赖将朴素的人脸模板特征映射至潜在空间（Latent Space，LS），导致重建的人脸图像存在大量伪影，面部细节不足。在此基础上，Dong 等[15]首先通过优化 StyleGAN 的潜在空间重建人脸，其次运用标准遗传算法在 StyleGAN 输入中找到最优噪声数据，最后借助 3 种商用活体检测系统评估重建人脸图像的安全性。尽管如此，由于上述方法完全依赖噪声质量和 StyleGAN 生成器映射网络，从而限制了特征表示能力，使得反演重建的人脸面部细节仍然较少。

综上，虽然基于人脸模板可重建出相似度更高的人脸图像，但在单一人脸模板特征数据中仅利用卷积与反卷积组合的生成器网络或 StyleGAN 生成器中的全连接层映射进行网络重建，在准确还原细微面部特征时较为困难、特征表示能力不足，进而导致重建人脸图像质量下降。因此，本文结合两种方法的优势，基于 Wasserstein GAN（WGAN）算法创新性地提出一种新的映射网络 FFMapNet[16]。首先，设计注意力特征提取模块 SAFEM，通过反卷积与卷积网络将模板特征映射到高维特征图；其次，利用自注意力机制建模特征之间的长距离依赖关系，从模板

特征反演中提取并增强特征；最后，根据特征融合模块FFM进一步处理和融合增强后的特征，生成更具表达能力的特征映射，提升模板反演重建质量。文献[17-19]利用合成的人脸数据集训练人脸识别模型，其优势在于能提供更丰富的数据多样性与真实的潜在空间映射。因此，本文不仅使用真实的人脸数据集训练，还通过合成的人脸数据来监督训练映射网络，以进一步提升模型性能，增强模型的泛化能力和鲁棒性。

2　人脸重建方法

2.1　威胁模型

本文针对FR系统的TI黑盒攻击，基于先前黑盒重建方法的知识统一使用以下攻击者属性：①攻击者的目标为冒充已注册用户，通过反转存储在目标FR系统数据库中的d维完整人脸模板$t_d \in \mathbb{R}^d$，生成伪造的身份图像并入侵系统；②攻击者的知识包括攻击者获得人脸识别系统中嵌入数据库的访问权限，窃取在系统数据库中登记用户的目标人脸模板t_d，目标人脸模板中已知元素的索引集合$\{1, 2, ..., d\}$，目标FR系统的特征提取模型F_{template}的黑盒知识，任何其他特征提取器F_{loss}的白盒知识，以从人脸图像$\boldsymbol{I}$中生成人脸模板$t' = F_{\text{loss}}(\boldsymbol{I})$。

同时，假设攻击者不具备以下信息：①目标模板身份的附加信息或先验知识；②特征提取模型的训练集信息，无法使用相同或相似的数据集进行模板反演学习；③目标系统的比较和决策子模块的知识，例如相似度评分函数和系统的决策阈值。攻击者的能力包括：①攻击者能将反转模板生成的重建图像$\hat{\boldsymbol{I}}$直接注入特征提取器$F_{\text{loss}}()$，在绕过摄像头等传感器后，在特征提取阶段发起攻击，但每个目标模板攻击者仅有一次尝试机会；②攻击者策略为在上述假设下，攻击者训练一个人脸重建网络，通过反转人脸模板生成对应的图像，并使用该图像作为查询输入目标FR系统，尝试冒充注册用户，绕过身份验证来入侵系统。

本文基于以上攻击者属性，对模板反转攻击进行可行性分析。大量研究工作指出：模板反转攻击现已成为威胁人脸识别系统安全的主要手段之一，这得益于攻击者能利用目标系统可能存在的不完整保护机制，通过多种不正当途径获取人脸识别数据库的访问权限，进而获得系统用户的人脸模板数据。这些途径包括SQL注入、远程代码执行(Remote Code Execution，RCE)和跨站脚本攻击(Cross-site Scripting，通常称为XSS)等技术漏洞，通过实施网络攻击直接获取或篡改数据；采用钓鱼攻击或伪装成内部人员等社会工程学方法，诱骗员工泄露敏感信息；通过盗窃存储设备、尾随进入受限制区域接触关键基础设施、通过摄像偷拍获取部分登录信息等物理访问策略，结合暴力破解与字典攻击猜测弱密码。

当前，最先进的人脸识别模型的训练逻辑、内部参数等信息可通过网络搜集得知。首先，通过这些信息可轻松从一张人脸图像提取出对应的人脸模板数据，攻击者便具备模板反转攻击的前提条件(t_d、F_{loss})；其次，基于生成对抗网络等技术训练一个人脸模板重建网络，通过反转人脸模板生成对应的伪造用户人脸图像，并将该图像输入目标FR系统，尝试冒充注册用户，绕过身份验证，完成攻击入侵。

2.2　人脸重建网络框架

本文为了将映射网络学习到的潜在编码重建出人脸图像，选用EG3D作为基于生成神经辐射场(Generative Neural Radiance Fields，GNeRF)的预训练人脸生成器网络[20]。该模型通过两部分网络生成人脸，首先通过映射网络$M()$结合一个随机噪声生成真实中间潜码w，其次通过生成器、渲染网络$G()$结合相机参数c和相机姿态合成人脸数据$I = G\big(M(w, c)\big)$。

其中，相机参数c为相机的内参矩阵(包括焦距和主点位置)；焦距$f = \dfrac{1}{\tan\left(\dfrac{\text{FOV}}{2} \times \dfrac{\pi}{180}\right) \times \sqrt{2}}$由水平视场角(Field of View，FOV)在图像宽高比1∶1的前提下计算得到，构建的相机参数矩阵为$\begin{bmatrix} f & 0 & 0.5 \\ 0 & f & 0.5 \\ 0 & 0 & 1 \end{bmatrix}$；相机姿势为相机的外参矩阵，即从世界坐标系到相机坐标系的变换矩阵，涉及到相机的旋转中心(Pivot Point)和相机距离(Radius)。

本文采用与EG3D预训练模型一致的相机姿态，即默认生成的人脸图像为正面人脸图像，人脸重建方法的框架如图1所示。首先，使用固定的$G()$，并重新学习从F_{template}模型获取到的泄露人脸模板到中间潜空间的映射$\hat{w}$；其次，通过$G()$生成重建人脸$\hat{I}$，并使用不同的FR模型(F_{loss})提取重建人脸模板t'；最后，将t'与F_{template}数据库中的模板数据进行匹配，利用成功攻击率(Success Attack Rate，SAR)评估人脸重建网络的性能，以证明SOTA FR模型在面对此TI攻击的脆弱性。

本文为了进一步判别生成的潜在编码的真实性，使用WGAN模板反演方法使用的批评家网络$C()$来指导FFMapNet学习，具体网络结构如图2所示。该网络包含多层全连接层，每层后接Leaky ReLU激活函数，分别将生成的潜在编码和对应真实的潜在编码进行展平处理，最终输出一组评分来量化两者之间的差异，差异越小表明潜在编码越“真”。因此，本文将此差异作为批评家网络的损失项，通过交替训练FFMapNet与批评家网络最小化该损失项，以优化FFMapNet的生成性能，确保生成器能准确学习潜在空间中的分布特性，从而显著提升生成编码的质量和真实性。

在训练过程中，FFMapNet与批评家网络采用交替训练策略，即每更新两次FFMapNet参数就更新一次批评家网络，这种交替训练方式有助于维持生成器网络和批评家

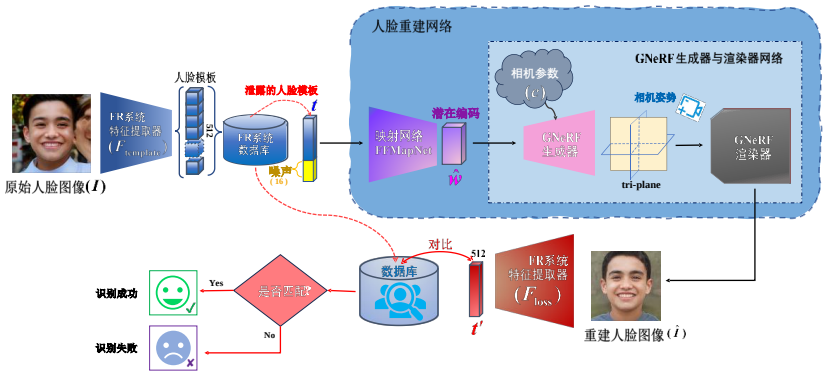


Fig. 1 Face template reconstruction network framework

图 1 人脸模板重建网络框架

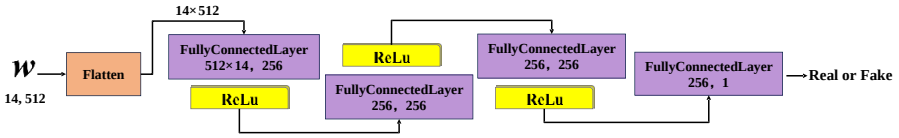


Fig. 2 Critic network structure

图 2 批评家网络结构

网络之间的平衡，防止其中一个网络过于强势而影响整体训练效果。具体而言，首先通过批评家网络评估生成编码与真实编码之间的差异，并将此差异作为损失项来优化FFMapNet；其次更新批评家网络以确保其评估能力的准确性，通过该方式协同进化两个网络来逐步提升生成编码的真实性和质量。

2.3 自注意力特征提取模块

本文针对人脸模板映射网络特征表现能力不足的问题，在原始全连接层映射网络特征映射之前对人脸模板数据进行深层次提取特征，提出一种结合反卷积层和自注意力机制的特征提取模块（Self-Attention Feature Extraction Module，SAFEM），以克服以往全连接层映射网络单纯依靠人脸模板特征数据映射特征表达的局限性，如图3所示。首先，通过反卷积层上采样为人脸模板数据 t 引入高维通道特征，保持模板特征的数量和维度；其次，基于8×8的卷积核捕捉空间信息和细节特征，生成的高空间分辨率的特征图为 (b,c,h,w)，b 为batch批量大小，c 为通道数，h 为特征图高度，w 为特征图宽度，为后续自注意力机制提供丰富的上下文信息，促进模板特征之间的相互关系建模，提升整体特征映射效果。

本文介于特征图不同部分的相互依赖关系，利用自注意力机制对特征全局关系的建模能力，计算特征图中不同位置之间的注意力权重，聚焦于重要的特征区域，动态调整各位置的特征表示，以有效捕捉人脸模板特征之间的长距离依赖关系，增强特征表示。首先，Self-Attention将上采样得到的特征图通过3个1×1的卷积核计算得到查询（Query）、键（Keys）和值（Values）特征，Query为 $(b,h\times w,c/8)$，Keys为 $(b,c/8,h\times w)$，Values为 $(b,c,h\times w)$；其次，将键的转置 Keys^{T} 与Query进行矩阵乘积计算出初始注意力能量矩阵；再次，为了避免能量矩阵过度关注某些特征，发生梯度消失问题，引入缩放因子 $\sqrt{d_k}$ 使计算出的注意力权重更平滑，从而生成聚焦于重要特征的能量矩阵（Attention Matrix），矩阵数值表示各输入特征彼此之间的重要性权重，这些权重经过Softmax激活函数处理可将较大的能量值转换为接近于1的权重，将较小的能量值转换为接近于0的权重，从而突出在当前上下文中最重要的特征；最后，应用Attention Matrix对Values特征进行加权求和，突出人脸模板数据中最相关的部分（见式（1）），以有效捕捉输入特征之间的相关性，增强模型对全局信息与局部细节的理解，生成增强的特征表示 $enh\_feature$。

$$enh\_feature = F_{\text{A}}(Q,K,V) = \text{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \qquad (1)$$

式中：F_{A} 为注意力层；缩放因子 $\sqrt{d_k}$ 中的 k 为人脸模板数据特征数；d_k 指输入到自注意力层中的每个注意力头的键向量的维度，即 $d_k = k/8$。

同时，引入自适应平均池化层（AvgPool2d）进一步压缩空间维度，在保留输入图像关键信息的同时，有效减少数据维度，并使用flatten展平操作将池化后的特征图转换为一维特征向量，用于特征融合模块进行处理。

2.4 特征融合模块

目前，在已有模板重建方法中基于全连接层网络映射方法通过一维模板数据学习映射，无法有效捕捉面部特征的空间结构，基于反卷积与卷积生成器网络则会导致重建人脸存在伪影。因此，本文首先将归一化的嵌入模板向量重塑为四维特征图用于SAFEM模块的输入；其次结合反卷积与自注意力机制捕捉特征之间的关系，提取重要的全局信息，使模型能更好地理解面部特征的空间结构；最后结合全连接层的特征映射能力，进而提升模板特征表达能力来构成本特征融合模块。

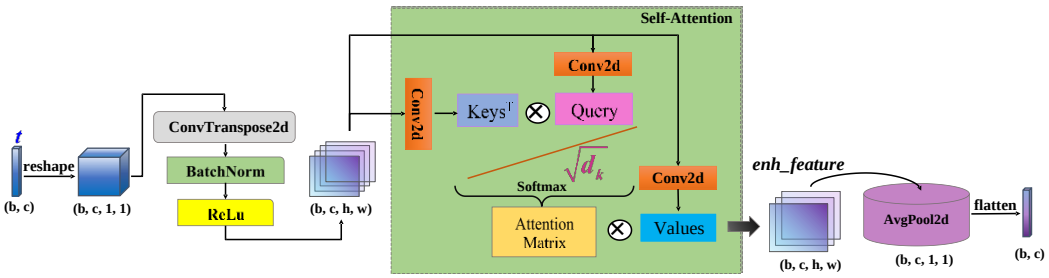


Fig. 3　SAFEM module structure

图 3　SAFEM 模块结构

图 4 为 FFM 具体的实现流程，首先对噪声向量 n 和人脸模板 t 分别进行二阶矩阵归一化，即 $t \cdot \frac{1}{\sqrt{\frac{1}{t_n}\sum_{i=1}^{t_n} t_i^2 + \epsilon}}$、$n \cdot \frac{1}{\sqrt{\frac{1}{N_n}\sum_{i=1}^{N_n} n_i^2 + \epsilon}}$，其中 t_n、N_n 分别代表模板特征和噪声数据的特征数量，ϵ 为一个小常数，用于避免除零错误。

相较于其他归一化方式，二阶矩归一化能更好地保留数据之间的相对关系，使模型更有效地捕捉特征的结构信息。首先，将归一化的模板特征输入 SAEFM 模块进行特征提取处理，并将生成增强后的特征与随机噪声数据拼接（concat）形成新的特征表示，进而整合噪声信息的多样性及自注意力动态特征数据建模的优势，提升模型特征表达能力。其次，通过两层全连接层（Fully Connected Layer）对特征进行映射，每个全连接层分别学习特定维度的特征，并配合 ReLU 激活函数进行非线性变换，使模型学习到更复杂的特征，进而广播（BC）得到最终的潜在编码 $\hat{w}$。

上述特征融合映射方法不仅融入了反卷积上采样到的模板特征中的空间特征信息，还利用了自注意力计算提取模板特征图中的关键信息。此外，全连接层的高效特征映射能力使其能灵活适应不同的人脸特征，因此在人脸重建面部特征时的表现更好。

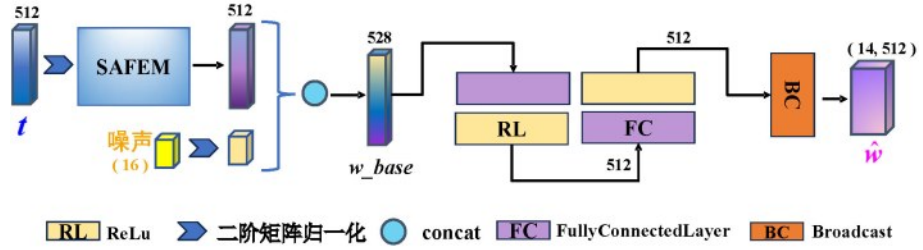


Fig. 4　FFM module structure

图 4　FFM 模块结构

2.5　损失函数

本文在真实人脸数据集下，针对 WGAN 框架采用以下多项损失函数来训练映射网络 FFMapNet，并将其设为 $M_{FFM}()$，具体训练过程如图 5 所示。首先，$M_{FFM}()$ 使用人脸模板 t 与随机噪声 $n \in N$ 以无监督学习到潜在空间 W 的映射代码 $\hat{w} = M_{\mathrm{FFM}}([n,t])$；其次，生成重建人脸图像来无监督训练映射网络。其中，多项损失函数为：

$$\mathcal{L}_{real}^{rec} = \mathcal{L}_{Pixel} + \mathcal{L}_{ID} \tag{2}$$

式中：$\mathcal{L}_{Pixel}$ 为像素损失，像素损失使像素级重构误差最小化。

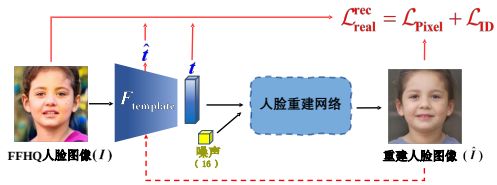


Fig. 5　Unsupervised training of mapping networks using real data

图 5　使用真实数据无监督训练映射网络

本文为了防止图像避免的异常值造成的影响，提升训练稳定性，先将原始人脸与重建图像均标准化映射到[0，1]，再计算像素损失。

$$\mathcal{L}_{\mathrm{Pixel}} = \mathbb{E}_{\hat{w}\sim M_{\mathrm{FFM}}([n,t])}\left[\left\| I - G(\hat{w},c)\right\|_2^2\right] \tag{3}$$

$$\mathcal{L}_{\mathrm{ID}} = \mathbb{E}_{\hat{w}\sim M_{FFM}([n,t])}\left[\left\| F_{\mathrm{loss}}(I) - F_{\mathrm{loss}}(G(\hat{w},c))\right\|_2^2\right] \tag{4}$$

式中：$\mathcal{L}_{\mathrm{ID}}$ 为 ID 损失，有助于网络生成与提取的人脸模板相似身份信息的人脸图像。

本文针对合成人脸数据，直接学习 GNeRF 中间潜码 $w = M(z) \in \mathrm{W}$ 来监督训练网络。假设 $\mathcal{L}_{\mathrm{W}}$ 为 w - loss，通过最小化合成人脸图像的真实潜在编码 w 与映射网络学习到的潜在编码 $\hat{w} = M_{\mathrm{FFM}}([n,t])$ 之间的均方误差来直接学习潜在空间的映射。具体流程如图 6 所示。

$$\mathcal{L}_{\mathrm{W}} = \mathbb{E}_{w\sim M(z)}\left[\left\| w - \hat{w}\right\|_2^2\right] \tag{5}$$

因此，合成人脸图像的训练总损失为：

$$\mathcal{L}_{\mathrm{syn}}^{\mathrm{rec}} = \mathcal{L}_{\mathrm{Pixel}} + \mathcal{L}_{\mathrm{ID}} + \mathcal{L}_{\mathrm{W}} \tag{6}$$

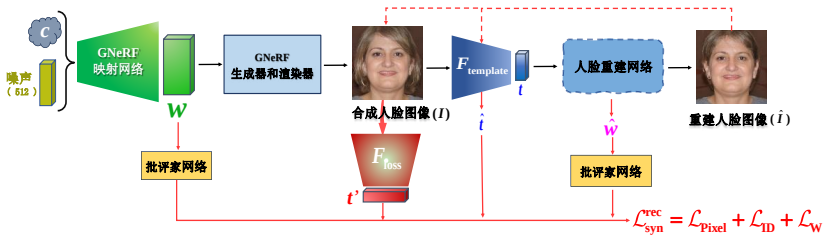


Fig. 6 Supervised training of mapping networks using synthetic data

图6 使用合成数据监督训练映射网络

3 实验结果与分析

3.1 实验环境与参数设置

本文研究的人脸识别模型 ArcFace 和 ElasticFace 均在 MS-Celeb-1M 数据集中进行训练，使用 FFHQ 数据集（由 70 000 张没有标签的人脸图像组成，包括年龄、种族、配饰和图像背景方面的变化）作为训练数据，按 9∶1 的比例随机分割为训练集与验证集，并使用 MOBIO、Labeled faces in the Wild（LFW）和 AgeDB 作为评估数据集[21-26]。

MOBIO 数据集是使用移动设备拍摄的 150 人的音频和面部图像数据，实验使用 mobio-all 协议的开发子集；LFW 数据库包含 5 749 人的 13 233 张照片，1 680 人有两张或两张以上照片，实验使用 View 2 协议；AgeDB 数据集包含 568 人的 16 488 张图像，最小年龄与最大年龄分别为 1、101 岁，每个受试者的平均年龄范围为 50.3 岁，实验使用 30 年的协议（即每对人脸年龄差等于 30）。

在本文实验中，通过 Bob 工具箱构建 FR 系统，基于 PyTorch 包训练人脸重建网络，所有模型均在配备单个 NVIDIA RTX 3090 GPU 的系统上以 30 个 epoch 进行网络训练，batch_size=6[27]。FR 模型提取到的嵌入模板及 GNeRF 输入的噪声 z 都为 512 维，映射网络 FFMapNet 的输入噪声 n 为 16 维，优化器为 Adam[28]。根据大量实验并结合经验发现，当使用初始学习率为 10^{-2}，每 3 个 epoch 将学习率降为 0.8 倍时的重建效果最好。

3.2 比较实验

本文将所提方法的性能与文献中已有的黑盒攻击方法进行比较，包括 NBNetA-M，、NBNetA-P、NBNetB-M、NBNetB-P、Vendrow、文献[14]的方法、文献[15]等方法[11,13,14,15]，并针对 ArcFace（ElasticFace 为 F_{loss}）、ElasticFace（ArcFace 为 Floss）系统进行 SAR 评估。表 1、表 2 分别为本文方法与上述方法在 MOBIO、LFW 和 AgeDB 数据集上，在假匹配率（False Match Rate，FMR）分别为 10^{-2}、10^{-3} 时，针对不同 SOTA FR 系统的黑盒攻击 SAR 的性能。其中，SAR 值越高代表 FR 系统对重建人脸图像识别成功率越高，反演重建效果越好，表中数值以百分比表示，比较方法中的数值均为该方法下 F_{target} 对应的 SAR 最大值，并对最大的两个值加粗显示，白盒攻击采用“-”代表。

由表 1、表 2 可知，当 FMR 为 10^{-3} 时，在 3 个数据集上进行反演攻击时，本文方法相较于上述黑盒人脸重建方法均有所提升，虽然有些数值与文献[15]相差不大，但重建的人脸分辨率为 512×512，而文献[15]重建的是 1 024×1 024 的高分辨率人脸，由此可得本文方法可在节省大量计算量、设备资源和模型训练时间的基础上，提升模板反演成功率。实验表明：①高特征映射能力的映射网络 FFMapNet 不仅能在较低分辨率下实现高效的特征表示，还能保证模板重建质量，证明了其在低资源消耗下的优越性和实际应用价值；②选用 ArcFace 损失训练的网络相较于 ElasticFace 损失训练的网络性能更好，如表 3 所示。在 Arc-

Table 1 SAR evaluation results of different models with FMR=10^{-2}

表 1 FMR=10^{-2}时不同模型的SAR评估结果

Method	MOBIO		LFW		AgeDB	
	ArcFace	Elastic-Face	ArcFace	Ealstic-Face	ArcFace	Elastic-Face
NBNetA-M[11]	2.85	10.00	14.30	37.13	2.56	8.44
NBNetA-P[11]	23.81	60.96	35.61	60.05	9.30	20.07
NBNetB-M[11]	20.95	30.00	26.90	52.99	5.40	14.56
NBNetB-P[11]	49.05	70.95	61.66	81.74	23.89	44.46
Vendrow and Vedrow[13]	69.52	74.29	77.00	79.37	44.74	25.17
Dong[14]	24.29	34.76	28.21	34.56	9.13	12.10
Dong[15]	**87.62**	**90.95**	**87.26**	**89.00**	**58.80**	**66.10**
[Ours](F_{loss}= ArcFace)	–	**93.42**	–	**93.54**	–	**70.33**
[Ours](F_{loss}= ElasticFace)	**88.84**	–	**88.21**	–	**59.41**	–

Table 2 SAR evaluation results of different models with FMR = 10^{-3}

表 2 FMR=10^{-3}时不同模型的SAR评估结果

Method	MOBIO		LFW		AgeDB	
	ArcFace	Elastic-Face	ArcFace	Elastic-Face	ArcFace	Elastic-Face
NBNetA-M[11]	0	2.38	4.32	10.90	0.81	2.55
NBNetA-P[11]	4.76	16.19	16.83	26.98	3.99	8.92
NBNetB-M[11]	1.90	3.80	10.98	21.44	1.88	6.72
NBNetB-P[11]	15.24	43.81	40.26	58.16	13.18	28.94
Vendrow and Vedrow[13]	29.05	43.81	57.70	53.03	29.64	34.89
Dong[14]	3.33	8.10	13.21	12.61	3.93	4.88
Dong[15]	**61.43**	**76.67**	**74.48**	**73.67**	**43.22**	**48.98**
[Ours](F_{loss}= ArcFace)	–	**82.53**	–	**74.30**	–	**52.11**
[Ours](F_{loss}= ElasticFace)	**61.99**	–	**75.23**	–	**44.63**	–

Face、ElasticFace模型FMR分别为10^{-2}、10^{-3}的阈值上的真匹配率(True Match Rate，TMR)表中，ArcFace的识别性能优于ElasticFace。

Table 3　TMR of ArcFace and ElasticFace under different FMRs

表3　ArcFace、ElasticFace在不同FMR下的TMR

Dataset	ArcFace		ElasticFace	
	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
MOBIO	100.00	99.98	100.00	100.00
LFW	97.60	96.40	96.87	94.70
AgeDB	98.33	98.07	98.20	97.57

3.3　消融实验

3.3.1　网络结构有效性验证

为了验证自注意力机制的有效性，将去除注意力机制的映射网络在MOBIO、LFW和AgeDB数据集上进行消融实验，并针对ElasticFace模型进行评估，如表4所示。实验结果表明，使用自注意力机制对模板反演性提升较大，原因为自注意力模块SAFEM能计算特征图中每个位置之间的相关性生成注意力权重，并利用这些权重加权合并特征图中的信息，动态调整每个位置的特征表示，以增强模型对全局和局部特征的建模能力。

Table 4　Influence of attention mechanism on the performance of facial template reconstruction

表4　注意力机制对人脸模板重建性能的影响

Methods	MOBIO		LFW		AgeDB	
	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
Without Self-attention	77.56	56.23	79.11	67.36	41.55	33.56
Ours	**88.84**	**61.99**	**88.21**	**75.23**	**59.41**	**44.63**

图7展示了在FFHQ数据集上基于以上两个方法重建的人脸。其中，图7(a)为不使用注意力方法重建出的图像；图7(b)为本文方法重建出的图像；图中数值为原始与重建人脸的余弦相似度。由此可见，注意力机制可提升模板重建人脸图像的质量，人脸更相似、细节较丰富。

(a)　Without Self-attention　　(b)　Ours

Fig. 7　Reconstruction samples on FFHQ dataset using method without self-attention and the proposed method

图7　未使用自注意力方法与本文方法在FFHQ数据集上的重建样本

3.3.2　多项损失函数有效性验证

为了评估损失函数中每个损失项对重建性能的影响，将上述损失函数按照不同损失项进行线性组合来训练网络，并针对ArcFace模型在MOBIO、LFW数据集上进行评估。表5为不同损失函数训练的模型在不同系统FMR值下的SAR重建性能。由此可知，每一项损失都能有效提升重建性能，比较$\mathcal{L}_1$、$\mathcal{L}_4$及$\mathcal{L}_3$、$\mathcal{L}_4$可得潜在编码损失和身份损失均对SAR重建性提升较大，尤其是在$\mathcal{L}_1$中不使用$\mathcal{L}_W$训练出的模型反演攻击识别成功率几乎为0，说明将WGAN用于没有中间潜在代码真正值的真实数据，有助于网络学习GNeRF中间潜在空间w的分布情况，否则映射网络生成的潜代码将不能按照GNeRF中间潜在空间的规则分布，会导致GNeRF生成器网络生成人脸图像失败。此外，像素损失$\mathcal{L}_{Pixel}$作为辅助损失，与潜在编码损失和身份损失任一项结合均能提升人脸模板的反演性能。

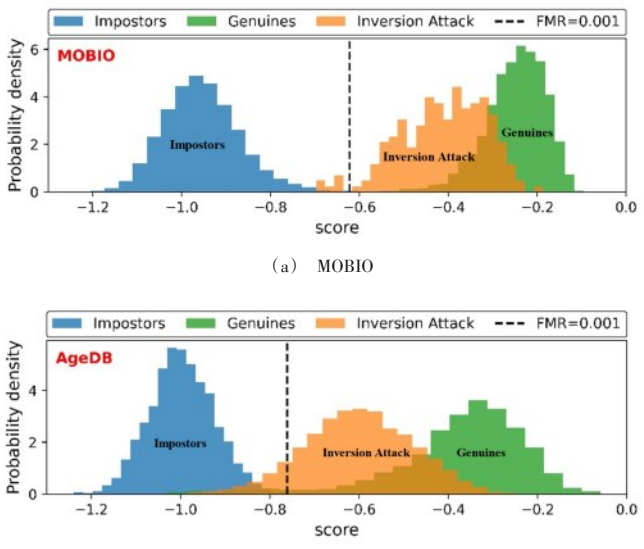
Table 5　Influence of different loss terms on the performance of face template reconstruction

表5　不同损失项对人脸模板重建性能的影响

Loss function	MOBIO		LFW	
	FMR=10^{-2}	FMR=10^{-3}	FMR=10^{-2}	FMR=10^{-3}
$\mathcal{L}_1=\mathcal{L}_{ID}+\mathcal{L}_{Pixel}$	0	0	0.12	0.02
$\mathcal{L}_2=\mathcal{L}_W$	33.51	12.60	42.02	22.32
$\mathcal{L}_3=\mathcal{L}_W+\mathcal{L}_{Pixel}$	39.25	18.66	44.41	24.32
$\mathcal{L}_4=\mathcal{L}_W+\mathcal{L}_{ID}+\mathcal{L}_{Pixel}$	**88.84**	**61.99**	**88.21**	**75.23**

3.4　实验结果

图8显示了在MOBIO、AgeDB数据集上评估黑盒攻击时，从真实和零努力冒充者对中提取的ArcFace嵌入模版之间的得分直方图，以及原始和重建人脸图像之间的得分直方图(负余弦距离)。由此可见，重建图像的得分接近真实图像得分，说明本文方法反演重建出的人脸能成功伪装成真实人脸，验证了本文方法在黑盒攻击场景中的实用性



(a)　MOBIO

(b)　AgeDB

Fig. 8　Score histogram of the ArcFace embedding template

图8　ArcFace嵌入模版的得分直方图

和可靠性。此外,这也暗示了未保护的最先进的人脸识别系统在面对此类攻击时的脆弱性,为进一步研究和提升防御策略提供了重要依据。

4 结语

本文针对SOTA FR系统的人脸模板重建提出一种新的映射网络FFMapNet,旨在解决模板特征映射存在的细节表示能力不足的问题。首先通过结合反卷积层与自注意力机制优势的SAFEM,捕捉空间信息和人脸模板特征之间的长程依赖关系,增强特征表示的上下文信息;其次利用FFM全连接层高效特征映射能力将增强特征与多样性噪声信息拼接后进行高效的特征映射,使其能灵活适应不同的人脸特征,提升重建人脸网络的多样性和鲁棒性,从而在人脸重建面部特征时取得更好的表现。同时,本文为人脸模板映射网络提供了新思路,通过反卷积网络与自注意力机制来改善特征映射的能力,进而提升人脸模板反演的成功攻击率。

本文人脸模板重建主要针对正面人脸图像,针对单一FR系统获得人脸模板进行重建,可得到不同FR系统对人脸图像的特征提取能力不同。由于黑盒重建具有的多个FR模型知识,未来可考虑从多个FR系统获得同一人脸模板数据进行分析处理,并考虑如何组合成更具表现力的模板特征。在伦理方面,尽管本文方法可能对未加保护的人脸识别系统带来潜在风险,但坚决反对将本文研究思路用于攻击真实人脸识别系统,以期推动计算机行业朝着更安全的识别技术迈进。

参考文献:

[1] BOUTROS F, STRUC V, FIERREZ J, et al. Synthetic data for face recognition: current state and future prospects[J]. Image and Vision Computing, 2023, 135: 104688.

[2] SARKAR E, KORSHUNOV P, COLBOIS L, et al. Are GAN-based morphs threatening face recognition?[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing, 2022: 2959-2963.

[3] CHATTERJEE A. Biometric presentation attack detection: towards securing biometric authentication systems[M]. Bristol: IOP Publishing, 2023.

[4] DAMER N, FANG M, SIEBKE P, et al. Mordiff: recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders[C]// 11th International Workshop on Biometrics and Forensics, 2023: 1-6.

[5] FANG M, DAMER N, KIRCHBUCHNER F, et al. Real masks and spoof faces: on the masked face presentation attack detection[J]. Pattern Recognition, 2022, 123: 108398.

[6] RUSIA M K, SINGH D K. A comprehensive survey on techniques to handle face identity threats: challenges and opportunities[J]. Multimedia Tools and Applications, 2023, 82(2): 1669-1748.

[7] YU Z, QIN Y, LI X, et al. Deep learning for face anti-spoofing: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5): 5609-5631.

[8] ZHMOGINOV A, SANDLER M. Inverting face embeddings with convolutional neural networks[DB/OL]. https://arxiv.org/abs/1606.04189.

[9] SHAHREZA H O, HAHN V K, MARCEL S. Face reconstruction from deep facial embeddings using a convolutional neural network[C]// 2022 IEEE International Conference on Image Processing, 2022: 1211-1215.

[10] COLE F, BELANGER D, KRISHNAN D, et al. Synthesizing normalized faces from facial identity features[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3703-3712.

[11] MAI G, CAO K, YUEN P C, et al. On the reconstruction of face images from deep face templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(5): 1188-1202.

[12] DUONG C N, TRUONG T D, LUU K, et al. Vec2face: unveil human faces from their blackbox features in face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 6132-6141.

[13] VENDROW E, VENDROW J. Realistic face reconstruction from deep embeddings[EB/OL]. https://openreview.net/forum?id=-WsBmzW-wPee.

[14] DONG X, JIN Z, GUO Z, et al. Towards generating high definition face images from deep templates[C]// 2021 International Conference of the Biometrics Special Interest Group, 2021: 1-11.

[15] DONG X, MIAO Z, MA L, et al. Reconstruct face from features based on genetic algorithm using GAN generator as a distribution constraint[J]. Computers & Security, 2023, 125: 103026.

[16] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]// International Conference on Machine Learning, 2017: 214-223.

[17] BAE G, DE L G M, BALTRUŠAITIS T, et al. Digiface-1m: 1 million digital face images for face recognition[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2023: 3526-3535.

[18] BOUTROS F, HUBER M, SIEBKE P, et al. Sface: privacy-friendly and accurate face recognition using synthetic data[C]// 2022 IEEE International Joint Conference on Biometrics, 2022: 1-11.

[19] KIM M, LIU F, JAIN A, et al. Dcface: synthetic face generation with dual condition diffusion model[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023: 12715-12725.

[20] CHAN E R, LIN C Z, CHAN M A, et al. Efficient geometry-aware 3D generative adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022: 16123-16133.

[21] BOUTROS F, DAMER N, KIRCHBUCHNER F, et al. Elasticface: elastic margin loss for deep face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022: 1578-1587.

[22] GUO Y, ZHANG L, HU Y, et al. Ms-celeb-1m: a dataset and benchmark for large-scale face recognition[C]// 14th European Conference on Computer Vision, 2016: 87-102.

[23] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4401-4410.

[24] MCCOOL C, WALLACE R, MCLAREN M, et al. Session variability modelling for face authentication[J]. IET Biometrics, 2013, 2(3): 117-129.

[25] HUANG G B, MATTAR M, BERG T, et al. Labeled faces in the wild: a database for studying face recognition in unconstrained environments[EB/OL]. https://inria.hal.science/inria-00321923/file/Huang_long_eccv2008-lfw.pdf.

[26] MOSCHOGLOU S, PAPAIOANNOU A, SAGONAS C, et al. AgeDB: the first manually collected, in-the-wild age database[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017: 51-59.

[27] ANJOS A, GÜNTHER M, DE F P T, et al. Continuously reproducing toolchains in pattern recognition and machine learning experiments[EB/OL]. https://openreview.net/pdf?id=BJDDItGX-.

[28] KINGMA D P. Adam: a method for stochastic optimization[DB/OL]. https://arxiv.org/abs/1412.6980.

(责任编辑:刘嘉文)