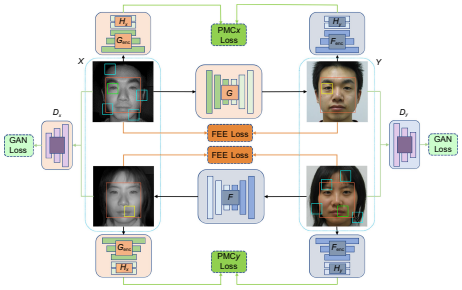
Opto-Electronic Engineering
光 电 工 程

Article
2022 年, 第 49 卷, 第 4 期

DOI: 10.12086/oee.2022.210317

双重对比学习框架下近红外-可见光人脸图像转换方法



孙　锐[1,2]，单晓全[1,2*]，孙琦景[1,2]，韩春军[3]，张旭东[1]

[1]合肥工业大学计算机与信息学院，安徽 合肥 230009；
[2]工业安全与应急技术安徽省重点实验室，安徽 合肥 230009；
[3]安徽省蚌埠市公安局科技信息科，安徽 蚌埠 233040

摘要：随着可见光-红外双模相机在视频监控中的广泛应用，跨模态人脸识别也成为计算机视觉领域的研究热点，而将近红外域人脸图像转化为可见光域人脸图像是跨模态人脸识别中的关键问题，在刑侦安防领域有着重要研究价值。针对近红外人脸图像在着色过程中面部轮廓易被扭曲、肤色还原不真实等问题，本文提出了一种双重对比学习框架下的近红外-可见光人脸图像转换方法。该方法构建了基于 StyleGAN2 结构的生成器网络并将其嵌入到双重对比学习框架下，利用双向的对比学习挖掘人脸图像的精细化表征。同时，本文设计了一种面部边缘增强损失，利用从源域图像中提取的面部边缘信息进一步强化生成人脸图像中的面部细节、提高人脸图像的视觉效果。最后，在 NIR-VIS Sx1 和 NIR-VIS Sx2 数据集上的实验表明，与近期的主流方法相比，本文方法生成的可见光人脸图像更加贴近真实图像，能够更好地还原人脸图像的面部边缘细节和肤色信息。

关键词：跨模态人脸识别；人脸图像转换；对比学习；StyleGAN2

中图分类号：TP391　　**文献标志码：**A

孙锐，单晓全，孙琦景，等. 双重对比学习框架下近红外-可见光人脸图像转换方法 [J]. 光电工程，2022，**49**(4): 210317
Sun R, Shan X Q, Sun Q J, et al. NIR-VIS face image translation method with dual contrastive learning framework[J]. *Opto-Electron Eng*, 2022, **49**(4): 210317

NIR-VIS face image translation method with dual contrastive learning framework

Sun Rui[1,2], Shan Xiaoquan[1,2*], Sun Qijing[1,2], Han Chunjun[3], Zhang Xudong[1]

[1]School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China;
[2]Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei, Anhui 230009, China;
[3]Science and Technology Information Section of Bengbu Public Security Bureau, Bengbu, Anhui 233040, China

Abstract: With the wide application of visible-infrared dual-mode cameras in video surveillance, cross-modal face recognition has become a research hotspot in the field of computer vision. The translation of NIR domain face images into VIS domain face images is a key problem in cross-modal face recognition, which has important research value in the fields of criminal investigation and security. Aiming at the problems that facial contours are easily distorted and skin color restoration is unrealistic during the coloring process of NIR face images, this paper proposes a NIR-VIS face images translation method under a dual contrastive learning framework. This method constructs a generator network based on the StyleGAN2 structure and embeds it into the dual contrastive learning

收稿日期：2021-09-30；收到修改稿日期：2022-01-20
基金项目：国家自然科学基金面上项目 (61471154, 61876057)；安徽省重点研发计划-科技强警专项 (202004d07020012)
*通信作者：单晓全，2334321350@qq.com。

framework to exploit the fine-grained characteristics of face images using bidirectional contrastive learning. Meanwhile, a facial edge enhancement loss is designed to further enhance the facial details in the generated face images and improve the visual effects of the face images using the facial edge information extracted from the source domain images. Finally, experiments on the NIR-VIS Sx1 and NIR-VIS Sx2 datasets show that, compared with the recent mainstream methods, the VIS face images generated by this method are closer to the real images and possesses more facial edge details and skin color information of the face images.

Keywords: cross-modal face recognition; face image translation; contrastive learning; StyleGAN2

1 引 言

近红外 (Near-infrared，NIR) 图像传感器由于可以很好地克服自然光的影响，能在各种光照条件不佳以及夜间场景下工作而受到广泛应用[1-2]。在刑侦安防领域，近红外人脸图像通常不能直接用于人脸检索与识别[3-5]，因为近红外传感器获取的单通道图像缺失了原始图像的自然色彩，对人眼视觉很不友好。与真实的可见光 (visible, VIS) 人脸图像相比，近红外人脸图像的人脸识别性能也较差。因此将近红外人脸图像转化为可见光人脸图像，还原人脸图像的色彩信息，有助于进一步提高人脸图像的主观视觉效果和跨模态识别性能，为构建全天候的视频监控系统提供技术支撑。

近年来，近红外-可见光图像转换与近红外图像的彩色化[6-11]引起了人们的广泛关注与研究。Limmer 等人[8]基于深度学习的方法提出了一种利用深度多尺度卷积神经网络对近红外图像进行着色的方法，然而该方法往往不能还原清晰的细节。生成对抗网络 (Generative adversarial network，GAN)[12]出现后在灰度图像的彩色化中得到了广泛的应用，因为它可以产生丰富且较清晰的细节。Liu 等人结合了变分自编码器和 GAN，构建了基于共享潜在空间假设和循环损失的无监督图像到图像转换网络 UNIT[13]，随后将其拓展至多模态，提出了 MUNIT[14]。Isola 等人提出的 Pix2pix GAN[15]使用 UNet[16]作为生成器，并提出了 PatchGAN 结构作为判别器，可以在生成的彩色图像中保留更多细节，较大程度提升了生成图像的质量。Wang 等人在 Pix2pix GAN 的基础上提出了升级算法 Pix2pix HD[17]，该算法采用多级生成的方式，先生成低分辨率的图像再将其输入到另一个网络中生成更高分辨率、更高质量的图像。然而 Pix2pix GAN 与 Pix2pix HD 算法都是针对已配对的数据集设计的，人脸的近红外-可见光图像对的采集非常困难，因为像素级匹配的近红外-可见光人脸数据集比未配对的数据集成本更高。所以，非配对的图像转换模型更适合于近红外-可见光人脸图像转换任务。

Zhu 等人提出的 CycleGAN[18]，是一种流行的非配对图像到图像的转换模型。CycleGAN 通过引入循环一致性损失，可以同步实现图像到图像的双向转换。与 UINT、MUNIT 等模型相比，CycleGAN 鲁棒性更好，更易于训练。Wang 等人基于 CycleGAN 结构提出了 FFE-CycleGAN[19]，采用通用的面部特征提取器来代替 CycleGAN 原始生成器中的编码器，在保持可见光域和近红外域的共同面部特征的同时，学习近红外域的特征。Dou 等人[20]提出了一种具有不同大小生成器的非对称 CycleGAN 方法。近红外域到可见光域的转换使用复杂网络，可见光域到近红外域的转换使用简单网络，与 CycleGAN 相比更加适用于非对称转换任务。Kancharagunta 等人提出了一种循环合成生成对抗网络 CSGAN[21]，该算法在一个域合成图像和另一个域循环图像之间使用了一种新的目标函数循环合成损失。随后 Kancharagunta 等人又提出了一种新的循环判别生成对抗网络 CDGAN[22]，通过结合用于循环图像的附加判别器网络来生成高质量和更逼真的图像。Taesung 等人提出了一种基于对比学习的图像到图像转换网络 CUT[23]，该方法创新性地将对比学习的思想应用到图像转换领域，并引入了多层对比损失，实现了一种轻量级的图像转换模型。基于 CUT 网络，Han 等人提出的 DCLGAN[24]使用两套对比学习的设置实现了图像到图像的双向转换。

然而，近红外人脸图像不同于其它的近红外图像，如图 1 所示，若人脸轮廓以及面部肤色等细节在着色的过程中被扭曲将会很大程度地影响生成人脸图像的视觉效果与图像质量。因此，有必要根据人脸图像的特点设计算法来强化近红外人脸图像在着色过程中细节信息的保留。

针对近红外人脸图像在着色过程中存在的挑战，

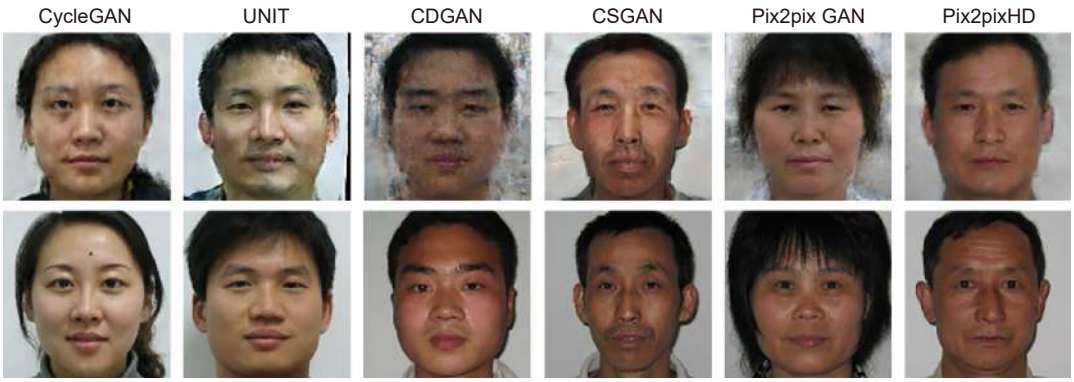


图 1 部分算法由近红外生成的可见光图像 (首行) 与真实可见光图像 (末行) 对比

Fig. 1 Comparison of the VIS image (the first row) generated by some algorithms from NIR domain with the real visible image (the last row)

本文提出了双重对比学习框架下的近红外-可见光人脸图像转换方法。该方法以双重对比学习网络为基础，采用双重对比学习的方式从图像局部细节出发增强生成图像的质量并且能够实现图像到图像的双向转换。同时，由于 StyleGAN2[25] 网络相较 ResNets[26] 能够提取人脸图像更深层次的特征，本文构建了基于 StyleGAN2 结构的生成器网络并将其嵌入到双重对比学习网络中替换原始的 ResNets 生成器，进一步提升生成人脸图像的质量。此外，本文设计了新的面部边缘增强损失，确保面部边缘信息在图像转换的过程中不被扭曲，提高生成人脸图像的视觉效果。主要贡献如下：

1) 本文提出了一种基于 StyleGAN2 网络的双重对比学习框架，构建了基于 StyleGAN2 结构的生成器网络并将其嵌入到双重对比学习网络中，利用双向的对比学习挖掘人脸图像的精细化表征。

2) 针对近红外域图像中人像外部轮廓模糊、边缘缺失的特点，本文设计了一种面部边缘增强损失，利用从源域图像中提取的面部边缘信息进一步强化生成的人脸图像中的面部细节。该损失与传统的边缘损失相比误差更小，更加贴合近红外-可见光人脸图像的转换任务。

3) 实验表明本文方法在 NIR-VIS Sx1 和 NIR-VIS Sx2 两个数据集上的生成效果明显优于近期的主流方法。本文方法生成的可见光人脸图像更加贴近真实图像，能够更好地还原人脸图像的面部边缘细节和肤色信息。

2 本文方法

对比学习作为一种常用的自监督学习方法，其指导原则是：通过自动构造相似实例和不相似实例，学习一个表示学习模型，通过这个模型，使得相似的实例在投影空间中比较接近，而不相似的实例在投影空间中距离较远。应用在图像转换领域的对比学习核心思想是通过构造正负样本使输入和输出图像的对应图像块之间的互信息最大化[23]。如图 2 所示，生成的人脸图像中黄色图像块应与输入图像中绿色图像块之间互信息最大，与输入图像中其他蓝色图像块之间的互信息较小。在 GAN 做图像转换时应用对比学习可以很好地增强生成图像局部质量，进而提升生成图像整体的质量。双重对比学习网络即使用两套对比学习网络，能够实现图像到图像的双向转换。

然而，仅使用双重对比学习网络在近红外-可见光人脸图像转换任务中并不能生成令人满意的人脸图像。因为 StyleGAN2 网络通过将图像的潜在特征在潜在空间进行解纠缠变换，能够提取人脸图像更深层次的特征，所以本文构建了基于 StyleGAN2 结构的生成器网络并将其嵌入到双重对比学习框架下。最终本文方法的网络框架图如图 2 所示。

2.1 网络概述

如图 2 所示，其中近红外域人脸图像记作 X 域，可见光域人脸图像记作 Y 域，G、F 分别为 $X \to Y$、$Y \to X$ 这两个方向的生成器，D_x、D_y 分别为 X 域和 Y 域的判别器，最终的目标是不断学习优化 G 和 F 这两个方向的映射。生成器前半部分被定义为编码器，后半部分为译码器。即 G 由 $G_{\rm enc}$ 与 $G_{\rm dec}$ 组成，F 由 $F_{\rm enc}$ 与 $F_{\rm dec}$ 组成，它们被依次应用于生成目标域图像 $\hat{y} = G(x) = G_{\rm dec}(G_{\rm enc}(x))$、$\hat{x} = F(y) = F_{\rm dec}(F_{\rm enc}(y))$。

对于每个方向的映射，在原图中随机选择若干图像块，用编码器提取图像块的特征，再通过一个两

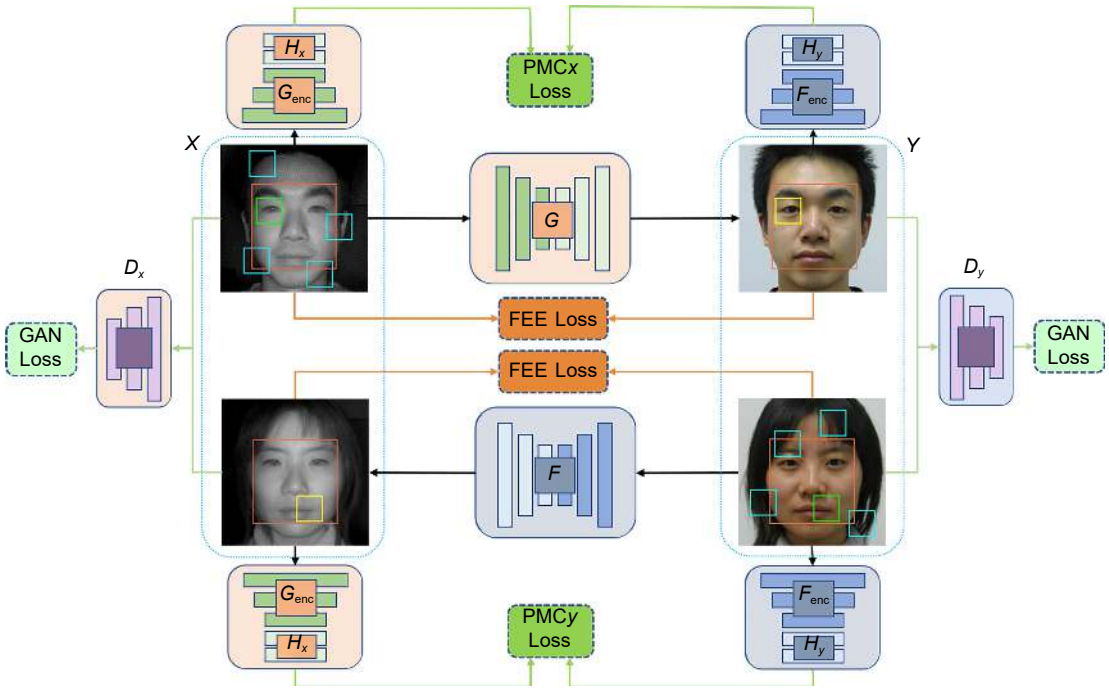


图 2 本文网络框架图。

为简化网络结构，同一性损失在图中未标示，详见 2.4.4 节

Fig. 2 The structure diagram of the proposed method.

To simplify the network structure, the identity loss is not indicated in the figure, see Section 2.4.4 for details

层 MLP 网络把提取的特征投影到共享的嵌入空间，在此基础上计算图像块多层对比损失，利用对比学习的思想使生成图像对应位置的图像块特征更加贴近原图。此外，在原图与生成的目标域人脸图像中同时裁剪出面部区域，并对此区域使用 Sobel 算子[27] 提取边缘并计算面部边缘增强损失，进一步强化生成的人脸图像中的面部细节。

2.2 生成器结构

在各式生成对抗网络中，StyleGAN[28] 网络由于显著地提升了生成图像的分辨率和质量，并且在多个不同数据集上的生成效果都很稳定，一经提出就引起了广泛关注。然而 StyleGAN 生成图像中存在类似水滴的斑状伪影，在生成器网络的中间特征图中此类伪影更加明显。升级后的 StyleGAN2 网络通过权重解调、延迟正则化与路径长度正则化等方法重点解决了初代网络中存在的伪影问题，进一步提高了图像的生成质量。同时，StyleGAN2 网络也成为了人脸生成领域较为先进的模型。相比双重对比学习网络中原始的 ResNets 生成器，StyleGAN2 网络通过将图像的潜在特征在潜在空间进行解纠缠变换，能够提取图像更深层次的特征。此外，在网络结构上 StyleGAN2 吸收了 ResNets 网络的部分优点，探索了残差连接设计和其它与 ResNets 类似的残差概念。因此在人脸图像生成任务中，使用基于 StyleGAN2 的生成器网络的生成效果要优于 ResNets 网络。

本文的生成器由编码器和译码器构成，结构如图 3 所示。原始的 StyleGAN2 网络实现的是从向量到图像的转换过程，即通过将潜在向量或随机噪声输入到生成模型中可以输出高质量的生成图像。在此基础上，本文构建的生成器则完成了从输入图像到潜在向量再到输出图像的完整转换过程，通过编码器部分的多个样式块 (Style block) 将256×256大小的输入人脸图像转换为 512 维潜在向量 $\boldsymbol{z}\in Z$，译码器来自于 StyleGAN2 网络的生成模型，负责将潜在向量 $\boldsymbol{z}$ 归一化后通过 8 个全连接层映射为潜在向量$\boldsymbol{w}\in W$进行特征解纠缠，最终潜在向量$\boldsymbol{w}$再经多个样式块生成256×256大小的目标域人脸图像。图 3 中绿色样式块包括调制、3×3卷积、解调与实例归一化等操作，StyleGAN2 网络利用样式块进行权重解调以简化模型设计。

2.3 判别器结构

本文方法中判别器D_x、D_y均为70×70 PatchGAN

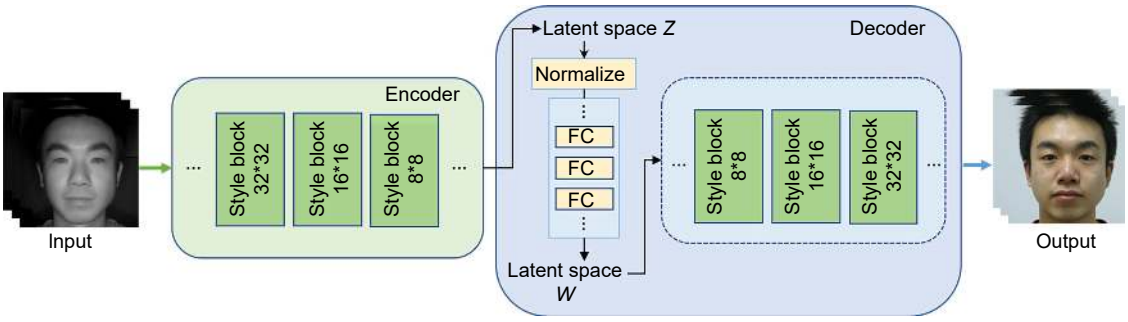


图 3　本文生成器结构图

Fig. 3　The structure diagram of generator in the proposed method

结构。该网络包含五个卷积层，其中第一层由卷积-激活函数 (LeakyReLU) 构成，中间三层均由卷积-实例归一化-激活函数构成，最后一层只由一个卷积构成。该判别器每次从原图中选取70×70大小的图像补丁判别真假，最终输出一个30×30大小的矩阵，输出矩阵的均值将作为对图像的评价。

一般的 GAN 判别器是针对整张图像输出一个真或假的矢量作为评价，而 PatchGAN 通过逐次叠加的卷积层最终输出一个矩阵，其中每个元素实际代表原图中70×70大小的图像补丁。这样的补丁级判别器架构比全图像判别器的参数更少，并且对输入图像尺寸的适应性更强。

2.4 损失函数

本文方法共结合了四种损失，包括面部边缘增强损失、图像块多层对比损失、对抗性损失和同一性损失，具体细节如下文所述。

2.4.1 面部边缘增强损失

图像转换领域传统的边缘损失[20]是直接对生成的目标域图像和源域图像提取边缘，计算两张图像边缘之间的损失。然而，这种直接提取图像边缘的方法在近红外人脸图像中并不可取。以图 4 所示 CASIA NIR-VIS 2.0 数据库[29]为例，其近红外图像中人像的头发与背景几乎融为一体，头发的外围轮廓很难分辨，在此情况下若使用传统的边缘损失直接对近红外图像提取边缘，则与可见光图像提取的边缘相比会产生极大的误差。于是本文提出了针对近红外人脸图像特点的面部边缘增强 (facial edge enhancement，FEE) 损失，对近红外域和生成的可见光域人脸图像仅裁剪出面部区域，在面部区域上提取边缘并计算损失。如图 4 所

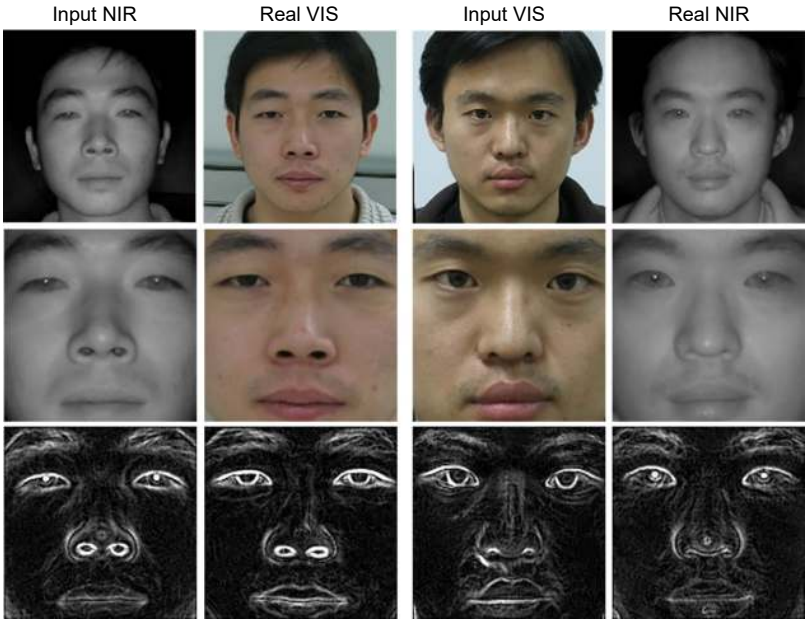


图 4　在近红外和可见光条件下分别对人脸图像裁剪出面部区域并提取边缘

Fig. 4　Crop out facial regions and extract edges from face images in NIR and VIS conditions respectively

示，裁剪后的面部区域在近红外与可见光条件下均可以提取到较为完整的边缘信息，以此指导人脸图像的生成，保证在人脸图像转换的过程中面部边缘不被扭曲。

最终，本文的面部边缘增强损失定义为源域图像与生成的目标域图像分别提取面部边缘得到的边缘图像之间的 L1 距离。其表达式如下所示：

$$L_{\mathrm{FEE}}(G,F,X,Y)=E_{x\sim X}\left[\|\widehat{G(x)}-\widehat{x}\|_1\right]+E_{y\sim Y}\left[\|\widehat{F(y)}-\widehat{y}\|_1\right], \quad (1)$$

其中：$\widehat{x}$、$\widehat{y}$分别为近红外域图像x与可见光域图像y经裁剪后提取的边缘图像，$\widehat{G(x)}$、$\widehat{F(y)}$分别为生成的目标域图像$G(x)$、$F(y)$裁剪后提取的边缘图像。

2.4.2 图像块多层对比损失

如图 2 所示，在生成的人脸图像中随机选取的黄色图像块称为查询样本，那么在输入图像中相同位置的绿色图像块即为相应的正样本，输入图像中除正样本位置外随机位置选取的蓝色图像块即为相应的负样本。先将查询样本、正样本和N个负样本映射为 K 维向量，分别记作v、$v^+\in R^K$和$v^-\in R^{N*K}$。对 K 维向量进行 L2 正则化后转换为一个$(N+1)$路分类问题，此时排除负样本选出正样本的概率在数学上可以表示为交叉熵损失，表达式如式 (2) 所示：

$$l(v,v^+,v^-)=-\log\left[\frac{\exp(v\cdot v^+/\tau)}{\exp(v\cdot v^+/\tau)+\sum_{n=1}^{N}\exp(v\cdot v_n^-/\tau)}\right], \quad (2)$$

其中τ是一个缩放查询样本和其他样本之间距离的参数，默认值为 0.07。

本文使用G_{enc}和H_x提取近红外域图像特征，选择$G_{\mathrm{enc}}(x)$中的 L 层发送到H_x后，图像可以投影为特征集$\{z_l\}_L=\{H_x^l(G_{\mathrm{enc}}^l(x))\}_L$，其中$G_{\mathrm{enc}}^l$表示第$l$层的输出。选中的每层中的空间位置记作$s\in\{1,\ldots,S_l\}$，其中$S_l$是每层的空间位置数。每次查询样本的正样本特征记为$z_l^s\in R^{C_l}$，负样本特征记为$z_l^{S\setminus s}\in R^{(S_l-1)C_l}$，其中$C_l$是每层通道数。对于生成的可见光域图像$G(x)$，使用$F_{\mathrm{enc}}$和$H_y$同样可以提取特征得到另一个特征集$\{\hat{z}_l\}_L=\left\{H_y^l(F_{\mathrm{enc}}^l(G(x)))\right\}_L$。那么$X\to Y$方向的图像块多层对比 (patchwise multilayer contrastive，PMC) 损失[22]可以表示为

$$L_{\mathrm{PMC}x}(G,H_x,H_y,X)=E_{x\sim X}\sum_{l=1}^{L}\sum_{s=1}^{S_l}l\left(\hat{z}_l^s,z_l^s,z_l^{S\setminus s}\right). \quad (3)$$

同理，$Y\to X$方向的图像块多层对比损失可以表示为：

$$L_{\mathrm{PMC}y}(F,H_x,H_y,Y)=E_{y\sim Y}\sum_{l=1}^{L}\sum_{s=1}^{S_l}l\left(\hat{z}_l^s,z_l^s,z_l^{S\setminus s}\right), \quad (4)$$

其中：$\{z_l\}_L=\left\{H_y^l(F_{\mathrm{enc}}^l(y))\right\}_L$，$\{\hat{z}_l\}_L=\{H_x^l(G_{\mathrm{enc}}^l(F(y)))\}_L$。

2.4.3 对抗性损失

对抗性损失[12]的目的是使生成器生成的图像在视觉上与目标域图像更加相似。对于从近红外域到可见光域的映射$G:X\to Y$，其对抗性损失为

$$L_{\mathrm{GAN}}(G,D_y,X,Y)=E_{y\sim Y}[\log D_y(y)]+E_{x\sim X}[\log(1-D_y(G(x)))], \quad (5)$$

其中：生成器G试图生成更加逼真的可见光域人脸图像$G(x)$，而判别器D_y则试图区分生成的可见光域人脸图像$G(x)$与真实的可见光域人脸图像。

同理，对于从可见光域到近红外域的映射$F:Y\to X$，其对抗性损失为

$$L_{\mathrm{GAN}}(F,D_x,X,Y)=E_{x\sim X}[\log D_x(x)]+E_{y\sim Y}[\log(1-D_x(F(y)))]. \quad (6)$$

2.4.4 同一性损失

生成器G负责从近红外域到可见光域图像的映射，然而若将可见光域图像输入到生成器G中，理想中的生成器对此时的输入应该不做任何的更改而输出。同理，理想中的生成器 F 对输入的近红外域图像也应该不做任何的更改而输出。这种情况下真实的输出图像与输入图像之间的 L1 损失被定义为同一性损失 (identity loss)[18]。通过同一性损失可以纠正生成器的色偏，更好地还原目标域图像的色彩信息。其表达式如下：

$$L_{\mathrm{IDT}}(G,F)=E_{x\sim X}[\|F(x)-x\|_1]+E_{y\sim Y}[\|G(y)-y\|_1]. \quad (7)$$

2.4.5 总损失函数

本文通过图像块多层对比损失引入对比学习的思想，使生成图像的整体质量得到了很好的增强；通过面部边缘增强损失保证人脸图像在转换的过程中面部边缘不被扭曲，强化面部细节的保留；通过对抗性损失和同一性损失来进一步优化生成器和判别器，使生成的图像更加贴近目标域真实图像。总损失函数表达

式如下：

$$\begin{aligned}L(G,F,D_x,D_y,H_x,H_y)=&\lambda_{\mathrm{FEE}}L_{\mathrm{FEE}}(G,F,X,Y)\\&+\lambda_{\mathrm{PMC}}(L_{\mathrm{PMC}x}(G,H_x,H_y,X)\\&+L_{\mathrm{PMC}y}(F,H_x,H_y,Y))\\&+\lambda_{\mathrm{GAN}}(L_{\mathrm{GAN}}(G,D_y,X,Y)\\&+L_{\mathrm{GAN}}(F,D_x,X,Y))\\&+\lambda_{\mathrm{IDT}}L_{\mathrm{IDT}}(G,F)\,.\end{aligned}\tag{8}$$

根据经验和多次实验，本文中各权重参数分别设置为λ_{FEE}=1，λ_{PMC}=2，λ_{GAN}=1，λ_{IDT}=1。

3 实验结果及分析

3.1 数据集预处理

本文在 NIR-VIS Sx1 和 NIR-VIS Sx2 两个数据集上分别建立了训练集和测试集。这两个数据集分别来自 CASIA NIR-VIS 2.0 数据库[29] 的 S1 和 S2 部分。S1 部分包含了来自 202 位受试者的 3002 张近红外图像和 2095 张可见光图像，本文从中选择了 1000 对近红外-可见光图像对组成了 NIR-VIS Sx1 数据集。S2 部分包含了来自 308 位受试者的 5455 张近红外图像和 1891 张可见光图像，本文从中选择了 1236 对近红外-可见光图像对组成了 NIR-VIS Sx2 数据集。两个数据集虽然同属亚洲人脸，但本质上有着较大的区别。NIR-VIS Sx1 数据集中近红外域人脸图像较为清晰，可见光域人脸图像均为纯白色背景。而 NIR-VIS Sx2 数据集中近红外域人脸图像相对较模糊，可见光域人脸图像因光照不足导致肤色普遍偏暗，并且少部分人像背景为杂乱的室外建筑。

本文在每个数据集中选择 75% 的图像对作为训练集，其余图像对作为每个数据集中相应的测试集。在实验前，根据眼睛、嘴巴和鼻子等面部器官坐标对每个数据集中的近红外和可见光人脸图像进行预对齐处理，随后以人脸位置为中心统一将图像裁剪并缩放到256×256大小。

3.2 实验设置

实验环境：PC 平台为 Ubuntu 18.04.5 LTS 系统，Intel Core i7-8700 CPU，Nvidia GeForce GTX 1070Ti GPU，8 GB 显存，使用的软件为 PyCharm 2021.1，CUDA 10.1，cuDNN 8.0.5。本文方法使用β_1=0.5、β_2=0.999 的 Adam 优化策略，学习率为 0.0001，训练周期为 400 轮并且学习率在训练总周期一半后线性衰减，批量训练样本数为 1。

3.3 定性实验分析

本文选取了 CycleGAN[18]、CSGAN[21]、CDGAN[22]、UNIT[13]、Pix2pixHD[17] 五种图像转换网络与本文方法作比较，分别在预处理的 NIR-VIS Sx1 和 NIR-VIS Sx2 数据集上进行了多重验证实验。实验结果如图 5 所示。

从图 5 中可以看出，原始 CycleGAN 方法的性能较弱，生成的人脸图像不够清晰，且人脸肤色还原度较差。CSGAN 方法在 NIR-VIS Sx1 数据集上生成的人脸图像较为清晰，但在 NIR-VIS Sx2 数据集上生成的人脸肤色信息失真严重。CDGAN 方法在 NIR-VIS Sx1 数据集上生成效果优于 CycleGAN 但人像外部轮廓不够清晰，在 NIR-VIS Sx2 数据集上表现较差，生成的人脸较为模糊且背景十分杂乱。UNIT 方法相较 CycleGAN 在人脸肤色重建方面性能有所提升，在两个数据集上生成的人脸均和真实可见光人脸较为贴近，但生成的人脸图像中头发等细节仍不够清晰。Pix2pixHD 作为 Pix2pix GAN 的升级算法，确实拥有较为出色的性能。在 NIR-VIS Sx1 数据集上，Pix2pixHD 生成的人脸图像有着清晰的面部细节和更加真实自然的人脸肤色。然而在 NIR-VIS Sx2 数据集上 Pix2pixHD 生成的效果不尽理想，例如在图 5 第Ⅵ行中对于人像额头上的刘海未能很好地还原，在第Ⅶ行中忽略了人脸皱纹细节导致与真实可见光图像相差较大。本文方法在 NIR-VIS Sx1 和 NIR-VIS Sx2 两个数据集上都表现出了稳定且更加优异的性能，尤其在图 5 第Ⅲ、Ⅵ和Ⅶ行生成的可见光人脸图像不仅保留了更完整的面部细节，还重建了更加真实的肤色信息。

3.4 定量实验分析

为了进一步衡量各种图像转换网络生成的人脸图像质量，本文引入了结构相似性 (structural similarity, SSIM[30]) 与峰值信噪比 (peak signal to noise ratio, PSNR[31]) 两项指标。SSIM 用于衡量两幅图像相似度，取值在 0~1 之间，数值越大表示相似度越高。同样地，PSNR 数值越大表示两张图片均方误差越小、图片越接近，单位是 dB。在 NIR-VIS Sx1 和 NIR-VIS Sx2 数据集上，CycleGAN、CSGAN、CDGAN、UNIT、Pix2pixHD 与本文方法所生成的人脸图像的平均结构相似性与平均峰值信噪比分别如表 1 和表 2 所示。

在 NIR-VIS Sx1 数据集上，原始 CycleGAN 方法生成图像的两项指标均为最低分、性能较差，本文方法取得了最高的 SSIM 分数，同时 Pix2pixHD 方法生

图 5　在两个数据集上的对比实验结果。

从左到右依次为：输入 NIR 人脸图像、CycleGAN、CSGAN、CDGAN、UNIT、Pix2pixHD、本文方法、真实 VIS 人脸图像。其中 Ⅰ~Ⅲ行来自 NIR-VIS Sx1 数据集，Ⅳ~Ⅶ行来自 NIR-VIS Sx2 数据集

Fig. 5　The comparison experimental results on two datasets.

From left to right: input NIR face image, CycleGAN, CSGAN, CDGAN, UNIT, Pix2pixHD, the proposed method, and real VIS face image. Where rows Ⅰ~Ⅲ are from NIR-VIS Sx1 dataset, and rows Ⅳ~Ⅶ are from NIR-VIS Sx2 dataset

表 1　NIR-VIS Sx1 数据集上各图像转换网络性能比较

Table 1　Performance comparison of image translation networks on the NIR-VIS Sx1 dataset

Method	Mean SSIM	Mean PSNR/dB
CycleGAN	0.7433	29.0987
CSGAN	0.7964	29.9471
CDGAN	0.7636	29.4922
UNIT	0.7935	29.8568
Pix2pixHD	0.8023	**31.6584**
Ours	**0.8096**	31.0976

成结果的两项指标与本文方法均较为接近。在 NIR-VIS Sx2 数据集上，CDGAN 方法生成图像的两项指标比 CycleGAN 方法更低、表现更差，本文方法生成图像的质量显著优于其他方法，SSIM 与 PSNR 两项指标均获得了最高分并远超其他方法。

由于 SSIM，PSNR 这两项评价指标都是对图像

表 2 NIR-VIS Sx2 数据集上各图像转换网络性能比较

Table 2 Performance comparison of image translation networks on the NIR-VIS Sx2 dataset

Method	Mean SSIM	Mean PSNR/dB
CycleGAN	0.6317	28.7974
CSGAN	0.6891	28.8176
CDGAN	0.5283	28.1679
UNIT	0.6986	29.0634
Pix2pixHD	0.7894	30.5449
Ours	**0.8135**	**31.2393**

逐像素计算的，当输入图像与目标域图像并非像素级对齐的情况下，使用 Fréchet Inception Distance (FID)[32] 指标更为有效。FID 是计算生成图像分布和真实图像分布之间距离的一种度量。FID 分数越低意味着生成图像分布与真实图片分布之间越接近，图像质量越好；反之，分数越高则意味着生成图像质量越差。本文同样计算了各图像转换网络分别在 NIR-VIS Sx1 和 NIR-VIS Sx2 数据集上生成图像的 FID 分数，结果如表 3 所示。本文方法在两个数据集上均获得了最低的 FID 分数、生成的图像质量最好。CycleGAN 方法在 NIR-VIS Sx1 数据集上 FID 分数最高、表现最差。CDGAN 方法在 NIR-VIS Sx2 数据集上生成的人脸最为模糊且图像背景杂乱，因而 FID 分数最高。

同时，在表 3 中本文还计算了各图像转换网络在测试阶段平均处理单张图像所用时间。CSGAN 与 UNIT 方法平均处理单张图像用时和本文方法相近，而 Pix2pixHD 与 CDGAN 方法平均处理单张图像用时均未超过 0.1 s，速度上明显优于本文方法，所以模型的轻量化也将成为本文方法进一步优化的方向。

3.5 消融实验

本文分别在 NIR-VIS Sx1 和 NIR-VIS Sx2 数据集上做了多项消融实验，进一步验证本文添加的基于 StyleGAN2 结构的生成器与各项损失函数的有效性。实验结果如图 6 所示，其中“Baseline”为 Han 等人提出的原始 DCLGAN[24] 方法。

从图 6 中可以看出，使用 ResNets 生成器的原始基线方法在近红外-可见光人脸图像转换任务中表现极差，生成的可见光图像几乎为同一人脸作微调的结果，无法还原真实的人脸细节。本文方法在去除基于 StyleGAN2 结构生成器后生成图像的面部细节相比原始基线方法更加清晰，但生成的图像与真实图像仍然相差较大。去除对抗性损失的方法已经不能有效区分近红外域与可见光域的图像，无法正常地训练网络模型，所以生成的图像与输入图像相近。去除同一性损失的方法生成的图像整体上质量较好，但不能很好地纠正生成器的色偏，导致生成的人脸肤色不够真实。去除图像块多层对比损失的方法在生成的图像面部出现了冗余的细节，且肤色有一定的偏差。去除面部边缘增强损失的方法生成的人脸图像整体上与真实图像较为接近，但在眉毛、眼眶和鼻子底部等边缘细节上重建得不够清晰。于是本文方法在基于 StyleGAN2 结构生成器的基础上进一步综合各项损失函数，能够在生成的人脸图像中保持清晰的面部细节，有效提升了人脸图像的视觉质量。

本文同样计算了在 NIR-VIS Sx1 数据集上各项消融实验生成图像的平均结构相似性与平均峰值信噪比指标，实验结果如表 4 所示。原始基线方法的性能较

表 3 各图像转换网络在不同数据集上 FID 性能与平均单张测试耗时比较

Table 3 Comparison of FID performance and average single test time of each image translation network on different datasets

Method	FID (NIR-VIS Sx1)	FID (NIR-VIS Sx2)	Time/s
CycleGAN	142.2574	171.3596	0.181
CSGAN	70.2146	102.6718	0.344
CDGAN	123.7183	212.4299	0.098
UNIT	74.8315	95.7638	0.358
Pix2pixHD	67.1044	106.3615	**0.079**
Ours	**58.5286**	**46.9364**	0.337

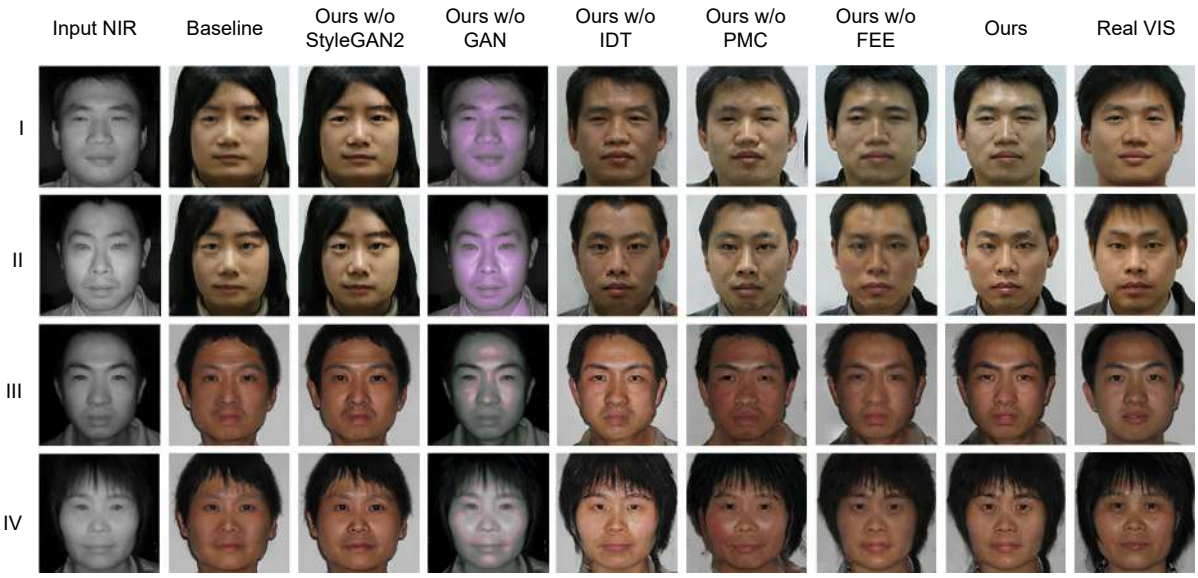


图 6　在两个数据集上的消融实验结果。

从左到右依次为：输入 NIR 人脸图像，基线方法，分别去除 StyleGAN2、L_{GAN}、L_{IDT}、L_{PMC}、L_{FEE}的本文方法，本文方法，真实 VIS 人脸图像。其中 Ⅰ~Ⅱ 行来自 NIR-VIS Sx1 数据集，Ⅲ~Ⅳ 行来自 NIR-VIS Sx2 数据集

Fig. 6　Results of the ablation experiments on two datasets.

From left to right: input NIR face image, Baseline method, the proposed method without StyleGAN2、L_{GAN}、L_{IDT}、L_{PMC}、L_{FEE} respectively, the proposed method and real VIS face image. Where rows Ⅰ~Ⅱ are from NIR-VIS Sx1 dataset and rows Ⅲ~Ⅳ are from NIR-VIS Sx2 dataset

表 4　NIR-VIS Sx1 数据集上各消融方法性能比较

Table 4　Performance comparison of ablation methods on the NIR-VIS Sx1 dataset

Method	Mean SSIM	Mean PSNR/dB
Baseline	0.5279	28.3419
Ours w/o StyleGAN2	0.5293	28.4381
Ours w/o GAN	0.3617	11.5007
Ours w/o IDT	0.6864	29.2308
Ours w/o PMC	0.6359	28.6156
Ours w/o FEE	0.7982	30.2057
Ours	**0.8096**	**31.0976**

差，本文方法在去除基于 StyleGAN2 结构的生成器后相较基线方法性能提升较为有限。去除对抗性损失的方法性能最差，SSIM 与 PSNR 指标均为最低。最后，本文方法明显优于其它消融方法，达到了最高的 SSIM 与 PSNR 指标。

3.6 讨论

本文设计的面部边缘增强损失，利用从源域图像中提取的面部边缘信息进一步强化生成人脸图像中的面部细节。所以，选择合适的边缘提取方法确保在可见光和近红外条件下都能提取到准确完整的面部边缘十分重要。本文选择了 Roberts 算子、Prewitt 算子、Sobel 算子、Laplacian 算子和 Canny 算子分别对可见光和近红外人脸图像提取面部边缘，结果如图 7 所示。使用 Roberts 算子和 Laplacian 算子得到的图像边缘较为微弱，使用 Canny 算子得到的二值图像轮廓过于粗犷、人脸细节损失较多，所以这三种算子均不适用于本文的面部边缘提取任务。使用 Prewitt 算子得到的边缘图像整体较为接近 Sobel 算子得到的边缘图像，但 Sobel 算子能够提取到更加完整的边缘细节，如第一行图像中的鼻翼边缘与第二行图像中的嘴唇边缘。

为了进一步比较使用 Prewitt 算子和 Sobel 算子对本文方法生成效果的影响，本文分别使用这两种算子

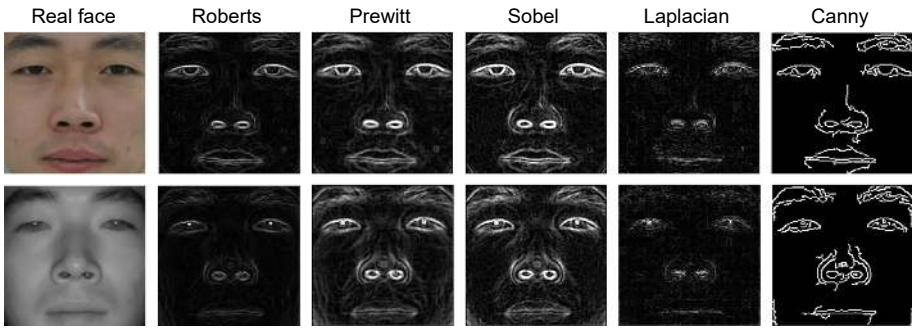


图 7 分别使用各边缘提取方法得到的边缘图像对比。

从左到右依次为：真实人脸图像、Roberts 算子、Prewitt 算子、 Sobel 算子、Laplacian 算子、Canny 算子

Fig. 7 Comparison of edge images obtained by using each edge extraction method separately.

From left to right: real face image, Roberts operator, Prewitt operator, Sobel operator, Laplacian operator, Canny operator

应用到面部边缘增强损失中，在 NIR-VIS Sx1 数据集上进行对比实验并计算生成图像的平均结构相似性与平均峰值信噪比，实验结果如表 5 所示。使用 Sobel 算子的方法在 SSIM 和 PSNR 性能上均优于使用 Prewitt 算子的方法，所以本文最终选定 Sobel 算子作为面部边缘损失中的边缘提取方法。

表 5 NIR-VIS Sx1 数据集上分别应用 Prewitt 算子与 Sobel 算子的性能比较

Table 5 Performance comparison of applying the Prewitt operator and Sobel operator respectively on the NIR-VIS Sx1 dataset

Method	Mean SSIM	Mean PSNR/dB
Ours (Prewitt)	0.7924	30.2815
Ours (Sobel)	**0.8096**	**31.0976**

本文在确定各项损失函数权重参数时，参考了基线模型 DCLGAN[24] 中关于 λ_{PMC}、λ_{GAN}和 λ_{IDT}这三项权重的设置，于是本文重点评估了 λ_{FEE}的不同取值对本文方法性能的影响。如图 8 所示，横轴为面部边缘增强损失的权重参数 λ_{FEE}，纵轴为在 NIR-VIS Sx1 数据集上本文方法生成图像的 FID 分数。从图 8 中可以看出，λ_{FEE}取值的变化确实会影响本文方法的性能，当 λ_{FEE}=1 时，生成图像的 FID 分数最低、图像质量最好，故本文中设置 λ_{FEE}=1。

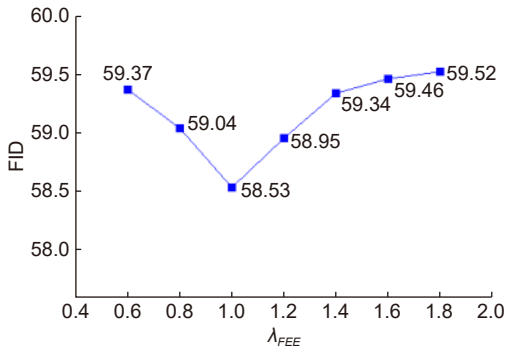


图 8 NIR-VIS Sx1 数据集上不同 λ_{FEE}取值对本文方法性能的影响

Fig. 8 The effect of different values of λ_{FEE} on the performance of our method on the NIR-VIS Sx1 dataset

4 结 论

本文提出了一种新的双重对比学习框架下的近红外-可见光人脸图像转换方法。该方法构建了基于 StyleGAN2 结构的生成器网络并将其嵌入到双重对比学习框架下，使用基于 StyleGAN2 结构的生成器网络提取人脸图像更深层次的特征，同时利用双向的对比学习挖掘人脸图像的精细化表征。此外，由于近红外域图像中人像外部轮廓模糊、边缘缺失，本文提出了施加在源域图像与生成的目标域图像之间的面部边缘增强损失，确保面部边缘信息在图像转换的过程中不被扭曲，进一步提高生成人脸图像的视觉质量。最后，在 NIR-VIS Sx1 和 NIR-VIS Sx2 两个数据集上的实验结果验证了本文方法的有效性和优越性。与近期主流的方法相比，本文方法生成的人脸图像不仅保留了更完整的面部细节，还重建了更加真实的肤色信息。

参考文献

[1] Dutta A K. Imaging beyond human vision[C]//*2014 8th International Conference on Electrical and Computer Engineering (ICECE)*, 2014: 224–229.

[2] Cao Z C, Schmid N A, Bourlai T. Composite multilobe descriptors for cross-spectral recognition of full and partial face[J]. *Opt Eng*, 2016, **55**(8): 083107.

[3] Sun Y, Wang X G, Tang X O. Deep learning face representation from predicting 10, 000 classes[C]//*IEEE Conference on Computer Vision & Pattern Recognition*, 2014: 1891–1898.

[4] He R, Wu X, Sun Z N, et al. Wasserstein CNN: learning invariant features for NIR-VIS face recognition[J]. *IEEE Trans Pattern Anal Mach Intell*, 2019, **41**(7): 1761−1773.

[5] Hu S W, Short N, Riggan B S, et al. Heterogeneous face recognition: recent advances in infrared-to-visible matching[C]//*2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017: 883–890.

[6] Mori A, Wada T. Part based regression with dimensionality reduction for colorizing monochrome face images[C]//*2013 2nd IAPR Asian Conference on Pattern Recognition*, 2013: 506–510.

[7] Cheng Z Z, Yang Q X, Sheng B. Deep colorization[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2015: 415–423.

[8] Limmer M, Lensch H P A. Infrared colorization using deep convolutional neural networks[C]//*2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016: 61–68.

[9] Larsson G, Maire M, Shakhnarovich G. Learning representations for automatic colorization[C]//*14th European Conference on Computer Vision*, 2016: 577–593.

[10] Limmer M, Lensch H P A. Improved IR-colorization using adversarial training and estuary networks[C]//*British Machine Vision Conference*, 2017.

[11] Suárez P L, Sappa A D, Vintimilla B X, et al. Near InfraRed imagery colorization[C]//*Proceedings of 2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018: 2237–2241.

[12] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//*Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, **2**: 2672–2680.

[13] Liu M Y, Breuel T, Kautz J. Unsupervised image-to-image translation networks[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 700–708.

[14] Huang X, Liu M Y, Belongie S, et al. Multimodal unsupervised image-to-image translation[C]//*Proceedings of the 15th European Conference on Computer Vision*, 2018: 179–196.

[15] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks[C]//*IEEE Conference on Computer Vision & Pattern Recognition*, 2017: 5967–5976.

[16] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C]//*Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015: 234–241.

[17] Wang T C, Liu M Y, Zhu J Y, et al. High-resolution image synthesis and semantic manipulation with conditional GANs[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018: 8798–8807.

[18] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//*Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017: 2242–2251.

[19] Wang H J, Zhang H J, Yu L, et al. Facial feature embedded Cyclegan for Vis-Nir translation[C]//*ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020: 1903–1907.

[20] Dou H, Chen C, Hu X Y, et al. Asymmetric cyclegan for unpaired NIR-to-RGB face image translation[C]//*ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: 1757–1761.

[21] Babu K K, Dubey S R. CSGAN: cyclic-synthesized generative adversarial networks for image-to-image transformation[J]. *Expert Syst Appl*, 2021, **169**: 114431.

[22] Babu K K, Dubey S R. CDGAN: cyclic discriminative generative adversarial networks for image-to-image transformation[J]. *J Vis Commun Image Represen*, 2022, **82**: 103382.

[23] Park T, Efros A A, Zhang R, et al. Contrastive learning for unpaired image-to-image translation[C]//*Proceedings of the 16th European Conference on Computer Vision*, 2020: 319–345.

[24] Han J L, Shoeiby M, Petersson L, et al. Dual contrastive learning for unsupervised image-to-image translation[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021: 746–755.

[25] Karras T, Laine S, Aittala M, et al. Analyzing and improving the image quality of StyleGAN[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 8107–8116.

[26] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 770–778.

[27] Gao W S, Zhang X G, Yang L, et al. An improved Sobel edge detection[C]//*Proceedings of the 3rd IEEE International Conference on Computer Science & Information Technology*, 2010: 67–71.

[28] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019: 4396–4405.

[29] Li S Z, Yi D, Lei Z, et al. The CASIA NIR-VIS 2.0 face database[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013: 348–353.

[30] Sheikh H R, Sabir M F, Bovik A C. A statistical evaluation of recent full reference image quality assessment algorithms[J]. *IEEE Trans Image Process*, 2006, **15**(11): 3440−3451.

[31] Ma J Y, Yu W, Liang P W, et al. FusionGAN: a generative adversarial network for infrared and visible image fusion[J]. *Inf Fusion*, 2019, **48**: 11−26.

[32] Heusel M, Ramsauer H, Unterthiner T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017: 6629–6640.

作者简介

孙锐 (1976-)，男，博士，教授，主要从事计算机视觉的研究。

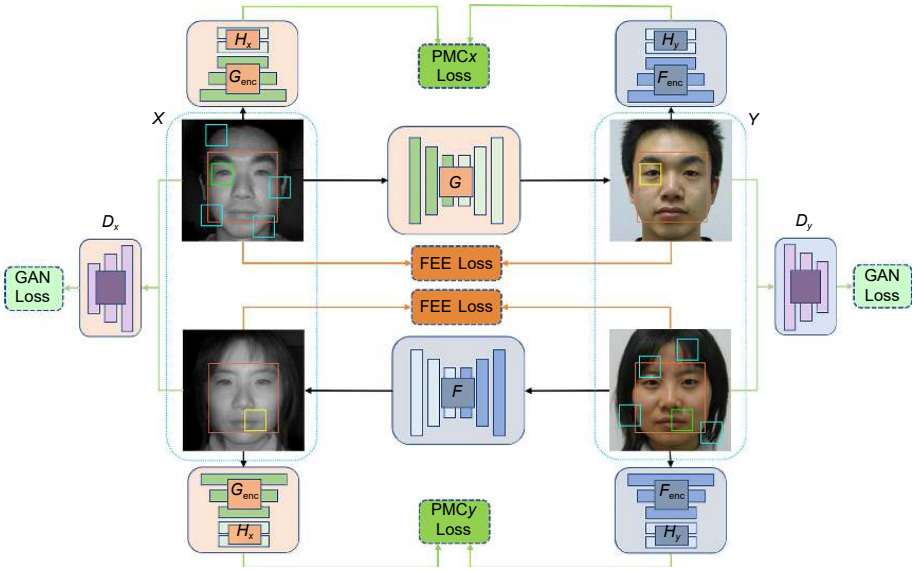
E-mail: sunrui@hfut.edu.cn

【通信作者】单晓全 (1998-)，男，硕士研究生，主要从事计算机视觉的研究。

E-mail: 2334321350@qq.com

NIR-VIS face image translation method with dual contrastive learning framework

Sun Rui[1,2], Shan Xiaoquan[1,2*], Sun Qijing[1,2], Han Chunjun[3], Zhang Xudong[1]



Overview: Near-infrared image sensors are widely used because they can overcome the effects of natural light and work in various lighting conditions. In the field of criminal security, NIR face images are usually not directly used for face retrieval and recognition because the single-channel images acquired by NIR sensors miss the natural colors of the original images. Therefore, converting NIR face images into VIS face images and restoring the color information of face images can help further improve the subjective visual effect and cross-modal recognition performance of face images, and provide technical support for building a 24/7 video surveillance system. However, NIR face images are different from other NIR images. If the details of face contours and facial skin tones are distorted in the coloring process, the visual effect and image quality of the generated face images will be greatly affected. Therefore, it is necessary to design algorithms to enhance the retention of detailed information in the coloring process of NIR face images. We propose a NIR-VIS face image translation method under a dual contrastive learning framework. The method is based on the dual contrastive learning network and uses contrastive learning to enhance the quality of the images generated from the image localization. Meanwhile, since the StyleGAN2 network can extract deeper features of face images compared with ResNets, we construct a generator network based on the StyleGAN2 structure and embed it into the dual contrastive learning network to replace the original ResNets generator to further improve the quality of the generated face images. In addition, for the characteristics of blurred external contours and missing edges of human figures in NIR domain images, a facial edge enhancement loss is designed in this paper to further enhance the facial details of the generated face images by using the facial edge information extracted from the source domain images. Experiments show that the generation results on two public datasets based on our method are significantly better than those of recent mainstream methods. The VIS face images generated by our method are closer to the real images and possesses more facial edge details and skin tone information of face images.

Sun R, Shan X Q, Sun Q J, et al. NIR-VIS face image translation method with dual contrastive learning framework[J]. *Opto-Electron Eng*, 2022, **49**(4): 210317; DOI: 10.12086/oee.2022.210317

Foundation item: National Natural Science Foundation of China (61471154,61876057) and the Key Research Plan of Anhui Province - Strengthening Police with Science and Technology (202004d07020012).

[1]School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230009, China; [2]Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei, Anhui 230009, China; [3]Science and Technology Information Section of Bengbu Public Security Bureau, Bengbu, Anhui 233040, China

* E-mail: 2334321350@qq.com