JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
04
VOL.27

ISSN1006-8961
CN11-3758/TB

口罩人脸姿态分类 P1125

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第4期（总第312期）
2022年4月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年4月 第27卷 第4期（总第312期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

残差密集结构的东巴画渐进式重建(第1084页)

结合双模板融合与孪生网络的鲁棒视觉目标跟踪(第1191页)

注意力机制的曲面沉浸式投影系统补偿(第1238页)

Volume 27, Number 4
Published April 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Gradual model reconstruction of Dongba painting based on residual dense structure(P1084)

Double template fusion based siamese network for robust visual object tracking(P1191)

An attention mechanism based inter-reflection compensation network for immersive projection system(P1238)

中图法分类号:TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)04-1097-13

论文引用格式: Zhang H Y and Bao W J. 2022. The cross-view gait recognition analysis based on generative adversarial networks derived of self-attention mechanism. Journal of Image and Graphics,27(04):1097-1109(张红颖,包雯静. 2022. 融合自注意力机制的生成对抗网络跨视角步态识别. 中国图象图形学报,27(04):1097-1109)[DOI:10.11834/jig.200482]

融合自注意力机制的生成对抗网络跨视角步态识别

张红颖[1,2*],包雯静[2]

1. 中国民航大学天津市智能信号与图像处理重点实验室, 天津　300300; 2. 中国民航大学电子信息与自动化学院, 天津　300300

摘　要: **目的**　针对目前基于生成式的步态识别方法采用特定视角的步态模板转换、识别率随视角跨度增大而不断下降的问题,本文提出融合自注意力机制的生成对抗网络的跨视角步态识别方法。**方法**　该方法的网络结构由生成器、视角判别器和身份保持器构成,建立可实现任意视角间步态转换的网络模型。生成网络采用编码器—解码器结构将输入的步态特征和视角指示器连接,进而实现不同视角域的转换,并通过对抗训练和像素级损失使生成的目标视角步态模板与真实的步态模板相似。在判别网络中,利用视角判别器来约束生成视角与目标视角相一致,并使用联合困难三元组损失的身份保持器以最大化保留输入模板的身份信息。同时,在生成网络和判别网络中加入自注意力机制,以捕捉特征的全局依赖关系,从而提高生成图像的质量,并引入谱规范化使网络稳定训练。**结果**　在 CASIA-B(Chinese Academy of Sciences' Institute of Automation gait database——dataset B)和 OU-MVLP(OU-ISIR gait database-multi-view large population dataset)数据集上进行实验,当引入自注意力模块和身份保留损失训练网络时,在 CASIA-B 数据集上的识别率有显著提升,平均 rank-1 准确率比 GaitGAN(gait generative adversarial network)方法高 15%。所提方法在 OU-MVLP 大规模的跨视角步态数据库中仍具有较好的适用性,可以达到 65.9% 的平均识别精度。**结论**　本文方法提升了生成步态模板的质量,提取的视角不变特征更具判别力,识别精度较现有方法有一定提升,能较好地解决跨视角步态识别问题。

关键词:机器视觉;步态识别;跨视角;自注意力;生成对抗网络(GANs)

The cross-view gait recognition analysis based on generative adversarial networks derived of self-attention mechanism

Zhang Hongying[1,2*], Bao Wenjing[2]

1. Tianjin Key Laboratory of Advanced Signal and Image Processing, Civil Aviation University of China, Tianjin 300300, China;

2. College of Electronic Information and Automation, Civil Aviation University of China, Tianjin 300300, China

Abstract: **Objective** Gait is a sort of human behavioral biometric feature, which is clarified as a style of person walks. Compared with other biometric features like human face, fingerprint and iris, the feature of gait is that it can be captured at a long-distance without the cooperation of the subjects. Gait recognition has its potential in surveillance security, criminal investigation and medical diagnosis. However, gait recognition is changed clearly in the context of clothing, carrying status,

收稿日期:2020-08-26;**修回日期**:2021-01-07;**预印本日期**:2021-01-14

*通信作者:张红颖　carole_zhang0716@163.com

基金项目:国家重点研发计划资助(2018YFB1601200);中国民航大学天津市智能信号与图像处理重点实验室开放基金资助项目(2019ASP-TJ06)

Supported by: National Key R&D Program of China (2018YFB1601200); Open Foundation of Tianjin Key Laboratory for Advanced Signal Processing of Civil Aviation University of China (2019ASP-TJ06)

view variation and other factors, resulting in strong intra gradient changes in the extracted gait features. The relevant view change is a challenging issue as appearance differences are introduced for different views, which leads to the significant decline of cross view recognition performance. The existing generative gait recognition methods focus on transforming gait templates to a specific view, which may decline the recognition rate in a large variation of multi-views. A cross-view gait recognition analysis is demonstrated based on generative adversarial networks (GANs) derived of self-attention mechanism. **Method** Our network structure analysis is composed of generator $\boldsymbol{G}$, view discriminator $\boldsymbol{D}$ and identity preserver $\boldsymbol{\Phi}$. Gait energy images (GEI) is used as the input of network to achieve view transformation of gaits across two various views for cross view gait recognition task. The generator is based on the encoder-decoder structure. First, the input GEI image is disentangled from the view information and the identity information derived of the encoder $\boldsymbol{G}_{\text{enc}}$, which is encoded into the identity feature representation $f(x)$ in the latent space. Next, it is concatenated with the view indicator v, which is composed of the one-hot coding with the target view assigned 1. To achieve different views of transformation, the concatenated vector as input is melted into the decoder $\boldsymbol{G}_{\text{dec}}$ to generate the GEI image from the target view. In order to generate a more accurate gait template in the target view for view transformation task, pixel-wise loss is introduced to constrain the generated image at the end of decoder. In the discriminant network, the view discriminator learning distinguishes the true or false of the input images and classifies them to its corresponding view domain. It is composed of four Conv-LeakyReLU blocks and in-situ two convolution layers those are real/fake discrimination and view classification each. For the constraint of the generated images inheriting identity information in the process of gait template view transformation, an identity preserver is introduced to bridge the gap between the target and generated gait templates. The input of identity preserver are three generated images, which are composed of anchor samples, positive samples from other views with the same identity as anchor samples, and negative samples are from the same view in related to different identities. The following Tri-Hard loss is used to enhance the discriminability of the generated image. The GAN-based gait recognition method can achieve the view transformation of gait template but it cannot capture the global, long-range dependency within features in the process of view transformation effectively. The details of the generated image are not clear result in blurred artifacts. The self-attention mechanism can efficiently sort the long-range dependencies out within internal representations of images. We yield the self-attention mechanism into the generator and discriminator network, and the self-attention module is integrated into the up-sampling area of the generator, which can involve the global and local spatial information. The self-attention module derived discriminator can clarify the real image originated from the generated. We update parameters of one module while keeping parameters of the other two modules fixed, and spectral normalization is used to increase the stable training of the network. **Result** In order to verify the effectiveness of the proposed method for cross-view gait recognition, several groups of comparative experiments are conducted on Chinese Academy of Sciences' Institute of Automation gait database——dataset B (CASIA-B) as mentioned below: 1) To clarify the influence of self-attention module plus to identify positions of the generator on recognition performance, the demonstrated results show that it is prior to add self-attention module to the feature map following de-convolution in the second layer of decoder; 2) Ablation experiment on self-attention module and identity preserving loss illustrates that the recognition rate is 15% higher than that of GaitGAN method when self-attention module and identity preserving loss are introduced simultaneously; 3) The frame-shifting method is used to enhance the GEI dataset on CASIA-B, and the improved recognition accuracy of the method is significantly harnessed following GEI data enhancement. Our illustration is derived of the OU-MVLP (OU-ISIR gait database-multi-view large population dataset) large-scale cross-view gait database, which has a rank-1 average recognition rate of 65.9%. The demonstrated results based on OU-MVLP are quantitatively analyzed, and the gait templates synthesized at four views (0°, 30°, 60°, and 90°) are visualized in the dataset. The results show that the generated gait images are highly similar to the gait images with real and target views even when the difference of views is large. **Conclusion** A generative adversarial network framework derived of self-attention mechanism is implemented, which can achieve view transformation of gait templates across two optioned views using one uniform model, and retain the gait feature information in the process of view transformation while improving the quality of generated images. The effective self-attention module demonstrates that the generated gait templates in the target view is incomplete, and improves the matching of the generated images; The identity preserver based on Tri-Hard loss constrains the generated gait templates inheriting identity information via input gaits and the discrimination of the generated

images is enhanced. The integration of the self-attention module and the Tri-Hard loss identity preserver improves the effect and quality of transformation of gaits, and the recognition accuracy is improved for qualified cross-view gait recognition. As the GEI input of the model, the quality of pedestrian detection and segmentation will intend to the quality loss of the synthesized GEI images straightforward in real scenarios. The further research will focus on problem solving of cross-view gait recognition in complex scenarios.

Key words: machine vision; gait recognition; cross-view; self-attention; generative adversarial networks (GANs)

0 引　言

步态识别是通过人走路的姿势进行身份识别。与人脸、指纹或虹膜等其他生物特征相比,步态的优势在于无需受试者的配合即可进行远距离身份识别(支双双 等,2019)。因此,步态识别在视频监控、刑事侦查和医疗诊断等领域具有广泛的应用前景。然而,步态识别易受衣着、携带物和视角等因素的影响,提取的步态特征呈现很强的类内变化(王科俊等,2019),其中视角变化从整体上改变步态特征,从而导致跨视角识别性能明显下降。

针对跨视角步态识别问题,提出了许多先进方法,这些方法通常分为基于模型的方法和基于外观的方法两类。其中,基于外观的方法可以更好地处理低分辨图像并且计算成本低,表现出很大优势。Makihara 等人(2006)提出以步态能量图(gait energy image,GEI)(Han 和 Bhanu,2006)为步态模板的视角转换模型(view transformation model,VTM),利用奇异值分解来计算 GEI 的投影矩阵和视角不变特征。Hu 等人(2013)提出视角无关判别投影(view-invariant discriminative projection,ViDP)方法,在无需知道视角情况下使用线性变换将步态模板投影到特征子空间中,但在视角变化大时识别率较低。近年来,深度学习应用于解决步态识别问题已成为主流方向。Wu 等人(2017)提出基于卷积神经网络(convolutional neural network,CNN)的方法从任意视角中自动识别具有判别性的步态特征,在跨视角和多状态识别中效果显著。Shiraga 等人(2016)提出基于 CNN 框架的 GEINet 应用于大型步态数据集,将 GEI 作为模型输入,其在视角变化范围较小时有较好表现。基于 CNN 提取视角不变特征进行跨视角步态识别方法表现出卓越的性能,但 CNN 是一个黑盒模型,缺乏视角变化的可解释性。生成对抗网络(generative adversarial network,GAN)(Goodfellow 等,2014)对数据分布建模具有强大性能,在人脸旋转(Tran 等,2017)和风格转换(Zhu 等,2017)等应用中取得显著效果。目前,基于 GAN 的方法重构目标视角的身份特征进行步态识别,可提供良好的可视化效果。Yu 等人(2017a)提出步态生成对抗网络(gait generative adversarial network, GaitGAN),将不同视角的步态模板标准化为侧面视角的步态模板进行匹配。He 等人(2019)提出多任务生成对抗网络(multi-task generative adversarial network,MGAN)用于学习特定视角的步态特征表示。Wang 等人(2019)提出双通道生成对抗网络(two-stream generative adversarial network, TS-GAN)进行步态模板的视角转换以学习标准视角的步态特征。尽管目前基于 GAN 的步态识别方法通过合成图像提供了良好的可视化效果,但这些方法只能进行特定视角的步态转换,误差随视角跨度增大而不断累积,而且在视角转换过程中未能充分利用特征间的全局依赖关系进行建模,生成图像的细节信息仍然不够清晰。而自注意力机制能更好地建立像素点远近距离依赖关系并且在计算效率上表现出良好性能,在图像生成(Zhang 等,2018)和图像超分辨率重建(欧阳宁 等,2019)上有较好表现。

为了实现任意视角间的步态模板转换并提升生成图像的质量,本文提出融合自注意力机制的生成对抗网络的跨视角步态识别方法。通过设计带有自注意力机制的生成器和判别器网络,学习更多全局特征的相关性,进而提高生成图像的质量并增强提取特征的区分度,同时在网络结构中引入谱规范化,提高训练过程的稳定性。本文网络框架由生成器 $\boldsymbol{G}$、视角判别器 $\boldsymbol{D}$ 和身份保持器 $\boldsymbol{\Phi}$ 构成,采用计算简单且有效的步态能量图作为步态模板,从而更好地实现跨视角步态识别。生成网络中使用具有编码器—解码器结构的生成器 $\boldsymbol{G}$ 以学习不同视角步态模板间的潜在关系,引入像素级损失以生成更准确的目标视角步态模板;在判别网络中使用两个独立

判别器 $\boldsymbol{D}$ 和 $\boldsymbol{\Phi}$,在视角转换的同时保留身份信息,并引入视角分类损失和身份保留损失来保持步态结构信息和身份特征,使生成的步态模板更加逼真并具有判别力。

1 本文方法

1.1 网络模型的整体结构

本文网络模型的整体框架如图 1 所示,包括生成器 $\boldsymbol{G}$、视角判别器 $\boldsymbol{D}$ 和身份保持器 $\boldsymbol{\Phi}$。假设数据集有 N_s 个行人,每个行人 $\boldsymbol{x}_i$ 都标注了身份标签 y_i,$i \in \{1,2,\cdots,N_s\}$,数据集有 N_v 个视角数,其中,$\boldsymbol{x}_i^p$ 表示验证集中第 i 个行人在视角为 p 的步态模板,$p \in \{1,2,\cdots,N_v\}$;$\boldsymbol{x}_i^g$ 表示注册集中第 i 个行人在视角为 g 的步态模板,$g \in \{1,2,\cdots,N_v\}$。对于一幅未知身份和视角的步态模板,本文网络模型训练有两个目标:1)学习视角不变的身份特征表示;2)在视角指示器 v 下,将验证集视角为 p 的步态模板 $\boldsymbol{x}_i^p$ 转换成具有相同身份的目标视角为 g 的步态模板 $\boldsymbol{x}_i^g$。

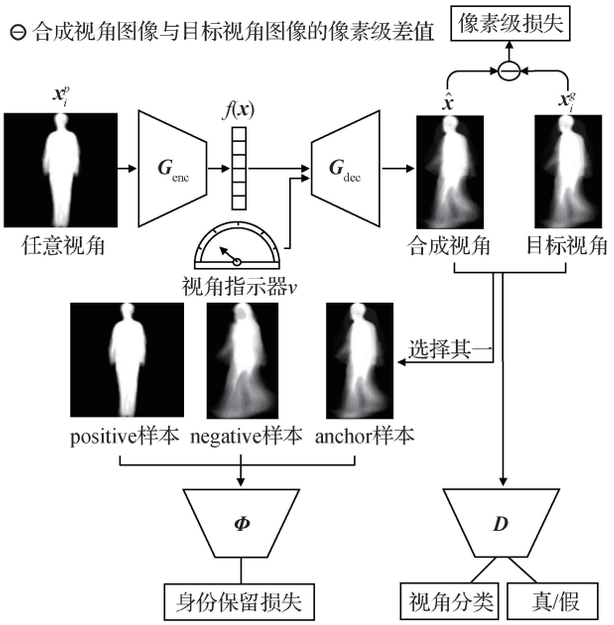

图 1 网络模型整体框架

Fig. 1 The framework of network

生成器 $\boldsymbol{G}$ 由编码器 $\boldsymbol{G}_{\text{enc}}$ 和解码器 $\boldsymbol{G}_{\text{dec}}$ 组成。步态模板 $\boldsymbol{x}_i^p$ 首先经过编码器 $\boldsymbol{G}_{\text{enc}}$ 分离出视角信息和身份信息,编码到潜在空间的身份特征表示 $f(\boldsymbol{x}) = \boldsymbol{G}_{\text{enc}}(\boldsymbol{x}_i^p)$,再与视角指示器 $\boldsymbol{v}$ 连接,然后将连接的向量作为输入送入解码器 $\boldsymbol{G}_{\text{dec}}$ 生成注册集中目标视角为 g 的步态模板 $\hat{\boldsymbol{x}}$。之后,利用视角判别器 $\boldsymbol{D}$ 和身份保持器 $\boldsymbol{\Phi}$ 对视角和身份信息进行判别。视角判别器 $\boldsymbol{D}$ 负责鉴定输入图像的真假,并将生成的步态模板分类到其对应的视角域,即 $\boldsymbol{D} \in \mathbf{R}^{N_v+1}$。身份保持器 $\boldsymbol{\Phi}$ 会接受以输入 anchor 样本 $\boldsymbol{x}_i^p$、与输入样本相同身份的其他视角的 positive 样本 $\boldsymbol{x}_i^k$ 和与 anchor 样本不同身份的相同视角的 negative 样本 $\boldsymbol{x}_j^g$ 组成的 3 幅图像,anchor 样本等概率在 $(\hat{\boldsymbol{x}}, \boldsymbol{x}_i^p)$ 中选择,并引入身份保留损失训练身份保持器以最大化保留身份信息。

在训练网络时,利用对抗损失来约束生成器和判别器,目标函数为

$$\min_{G}\max_{D} L_{\text{adv}} = E_{\boldsymbol{x}_i^g;P_{\text{data}}}[\log \boldsymbol{D}_{\text{dis}}(\boldsymbol{x}_i^g)] + E_{\boldsymbol{x}_i^p;P_{\text{data}},v;P_v}[\log(1 - \boldsymbol{D}_{\text{dis}}(\boldsymbol{G}(\boldsymbol{x}_i^p, \boldsymbol{v})))] \quad (1)$$

式中,$E[\,]$ 表示分布函数的期望值,行人的步态模板 $\boldsymbol{x}_i^p$ 和 $\boldsymbol{x}_i^g$ 服从样本分布 P_{data},视角指示器 v 服从视角指示分布 P_v,$\boldsymbol{G}(\boldsymbol{x}_i^p, \boldsymbol{v})$ 为步态模板 $\boldsymbol{x}_i^p$ 与目标视角指示器 v 连接送入生成器 $\boldsymbol{G}$ 生成的步态模板。生成器 $\boldsymbol{G}$ 要最小化该目标函数,而视角判别器中 $\boldsymbol{D}_{\text{dis}}$ 要最大化该目标函数。

1.2 生成器网络

为了在步态视角转换任务中获取高质量的生成图像,生成器基于 DRGAN(disentangled representation learning generative adversarial network)(Tran 等,2017)网络架构中的 generator 结构,并引入自注意力机制。如图 2 所示,生成器网络结构包括由卷积神经网络组成的下采样区 $\boldsymbol{G}_{\text{enc}}$ 和由反卷积神经网络与自注意力模块组成的上采样区 $\boldsymbol{G}_{\text{dec}}$。$\boldsymbol{G}_{\text{enc}}$ 和 $\boldsymbol{G}_{\text{dec}}$ 通过潜在空间的身份特征表示 $f(\boldsymbol{x}) \in \mathbf{R}^{N_f}$ 与目标视角为 1 的独热编码 $\boldsymbol{v}$ 的连接向量进行连接,其中 $\boldsymbol{v} \in \{0,1\}^{N_v}$,再经过一系列反卷积(deconvolution,Deconv)操作将 $(N_f + N_v)$ 维级联向量变换为合成目标视角的步态模板 $\hat{\boldsymbol{x}} = \boldsymbol{G}_{\text{dec}}(f(\boldsymbol{x}), \boldsymbol{v})$,使输出图像 $\hat{\boldsymbol{x}}$ 与输入图像 $\boldsymbol{x}$ 的尺寸相同。

生成器网络结构参数设置如表 1 所示,对于下采样区 $\boldsymbol{G}_{\text{enc}}$,在每个卷积层后均使用批标准化(batch normalization,BN)和 ReLU 激活函数;对于上采样区 $\boldsymbol{G}_{\text{dec}}$,除了输出层使用 Tanh 激活函数外,在每个反卷积层后均使用谱规范化(spectral normalization,SN)(Miyato 等,2018)、BN 和 ReLU 激活函数。

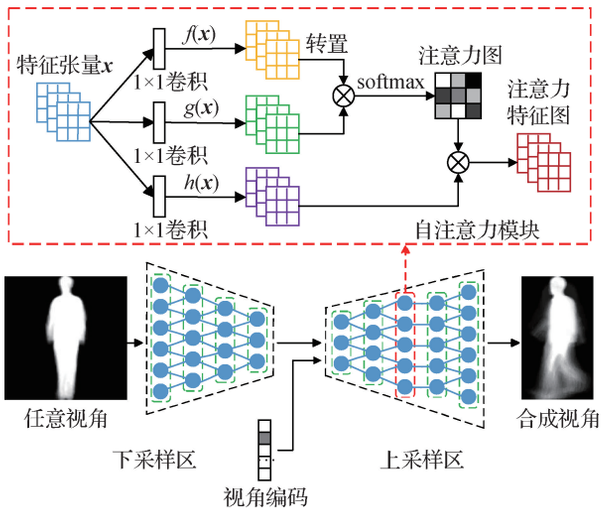

图 2　生成器网络

Fig. 2　Generator

表 1　生成器网络结构参数设置

Table 1　The parameter setting of generator

区域划分	层类型	卷积核	步长	深度	归一化	激活函数
下采样	卷积层	4×4	2	96	BN	ReLU
下采样	卷积层	4×4	2	192	BN	ReLU
下采样	卷积层	4×4	2	384	BN	ReLU
下采样	卷积层	4×4	2	768	BN	ReLU
特征连接	全连接层	–	–	–	–	–
上采样	反卷积层	4×4	2	768	SN、BN	ReLU
上采样	反卷积层	4×4	2	384	SN、BN	ReLU
上采样	自注意力模块	–	–	–	–	–
上采样	反卷积层	4×4	2	192	SN、BN	ReLU
上采样	卷积层	4×4	2	96	–	Tanh

注：“–”代表无具体参数。

生成器 $\boldsymbol{G}$ 试图在验证集视角为 p 的步态模板 $\boldsymbol{x}_i^p$ 与视角指示器 $\boldsymbol{v}$ 的引导下生成注册集中目标视角为 g 的步态模板 $\hat{\boldsymbol{x}}$。为了在步态视角转换任务中生成更准确的目标视角的步态模板，在 $\boldsymbol{G}_{\text{dec}}$ 末端引入像素级损失来约束生成图像，有助于更准确的步态模板的视角转换并使生成器在训练过程中更稳定。本文采用像素级 L1 损失来最小化生成图像和真实图像之间的像素级误差，具体为

$$\min_{\boldsymbol{G}} L_{\text{pixel}} = E_{\boldsymbol{x}_i:P_{\text{data}},v:P_v}\left[\left\|\boldsymbol{G}(\boldsymbol{x}_i^p,\boldsymbol{v}) - \boldsymbol{x}_i^g\right\|_1\right] \quad (2)$$

式中，$\|\cdot\|_1$ 表示像素级 L1 损失。

1.3　判别器网络

本文构建了两个判别器：视角判别器 $\boldsymbol{D}$ 和身份保持器 $\boldsymbol{\Phi}$，从而对真实的步态图像和生成器生成的步态图像进行区分，并在视角转换过程中保持身份特征。

1.3.1　视角判别器

对于给定的注册集视角的步态模板 $\boldsymbol{x}$，本文目标之一是将其转换为视角指示器 $\boldsymbol{v}$ 指定的目标视角的步态模板 $\hat{\boldsymbol{x}}$。因此，采用视角判别器学习判别输入样本是否真实并将其分类到对应的视角域，即 $\boldsymbol{D}:\boldsymbol{x}\rightarrow\{\boldsymbol{D}_{\text{dis}}(\boldsymbol{x}),\boldsymbol{D}_{\text{cls}}(\boldsymbol{x})\}$。视角判别器使用 PatchGAN 结构对图像进行真伪判别，并引入谱规范化（SN）来缓解梯度消失问题，从而提高模型的稳定性。视角分类器由 4 个 Conv-LeakyReLU 块和另外两个平行的卷积层组成，其中一个用于真/伪图像鉴别，另一个用于视角分类。视角判别器网络结构参数设置如表 2 所示。

表 2　视角判别器网络结构参数设置

Table 2　The parameter setting of view classifier

层类型	卷积核	步长	深度	归一化	激活函数
卷积层	4×4	2	64	–	LeakyReLU
卷积层	4×4	2	128	SN	LeakyReLU
卷积层	4×4	2	256	SN	LeakyReLU
自注意力模块	–	–	–	–	–
卷积层	4×4	2	512	SN	LeakyReLU
输出层($\boldsymbol{D}_{\text{dis}}$)	3×3	1	1	–	–
输出层($\boldsymbol{D}_{\text{cls}}$)	4×4	1	Nv	–	–

注：“–”代表无具体参数。

对于给定的训练图像 $\boldsymbol{x}_i$，通过在视角判别器 $\boldsymbol{D}$ 上施加辅助的视角分类损失，实现在视角指示器下步态模板的视角转换。当输入图像是注册集中真实的步态模板 $\boldsymbol{x}$ 时，$\boldsymbol{D}$ 将其鉴别为真实图像，并将真实图分类到对应的视角域；当输入图像是生成器 $\boldsymbol{G}$ 生成目标视角的步态模板 $\hat{\boldsymbol{x}}$ 时，$\boldsymbol{D}$ 将其判别为伪图像，并将其分类到视角指示器 $\boldsymbol{v}$ 指定的视角域。为了优化 $\boldsymbol{D}$，最小化如下目标函数

$$L_{\text{cls}}^{\text{D}} = E_{\boldsymbol{x}_i:P_{\text{data}}}\left[\log \boldsymbol{D}_{\text{cls}}(\boldsymbol{x}_i)\right] \quad (3)$$

式中，$\boldsymbol{D}_{\text{cls}}(\boldsymbol{x}_i)$ 是输入目标视角中真实的步态模板 $\boldsymbol{x}$ 在视角域的概率分布。优化 $\boldsymbol{G}$ 时，输入生成的步态模板及相应的视角指示器，目标函数为

$$L_{\text{cls}}^{\text{D}} = E_{\boldsymbol{x}_i^p:P_{\text{data}}}\left[\log \boldsymbol{D}_{\text{cls}}(\boldsymbol{G}(\boldsymbol{x}_i^p,\boldsymbol{v}))\right] \quad (4)$$

通过最小化该目标函数，生成器 $\boldsymbol{G}$ 试图合成可以分类到视角指示器 $\boldsymbol{v}$ 指定视角的步态模板。

1.3.2　身份保持器

传统的 GAN 模型生成的样本缺乏多样性，生成器会在某种情况下重复生成完全一致的图像，而对于跨视角步态识别任务，在步态模板视角转换过程中保持身份信息是至关重要的。因此，在本文模型中引入身份保持器 $\boldsymbol{\Phi}$，缩减目标视角与生成视角的步态模板间差距，进而保持身份信息。身份保持器 $\boldsymbol{\Phi}$ 基于 GaitGAN 中的身份判别器 $\boldsymbol{D}_A$ 的结构，与视角判别器 $\boldsymbol{D}$ 类似，引入谱规范化来增加模型的稳定性。如图 3 所示，身份保持器 $\boldsymbol{\Phi}$ 以（$\boldsymbol{x}_{\text{anc}}$，$\boldsymbol{x}_{\text{pos}}$，$\boldsymbol{x}_{\text{neg}}$）3 个图像作为输入，输出 $\boldsymbol{x}_{\text{anc}}$ 相关性标签 。

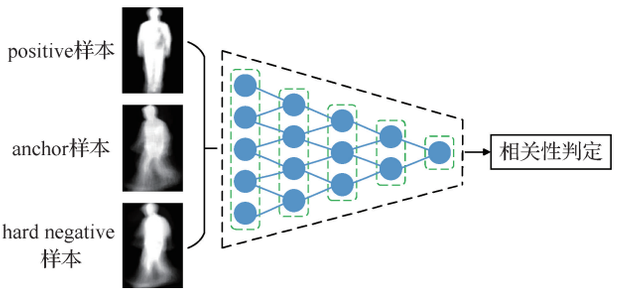

图 3　身份保持器网络

Fig. 3　Identity preserving discriminator

为了使生成的步态模板更好地保持身份信息，本文引入困难样本三元组 Tri-Hard 损失（Hermans 等，2017）作为身份保留损失来增强生成图像的可判别性。以（$\boldsymbol{x}_{\text{anc}}$，$\boldsymbol{x}_{\text{pos}}$，$\boldsymbol{x}_{\text{neg}}$）3 个图像作为输入，最小化如下身份保持损失

$$L_{\text{trihard}}(\boldsymbol{x}_{\text{anc}},\boldsymbol{x}_{\text{pos}},\boldsymbol{x}_{\text{neg}}) = \frac{1}{P\times K}\times \sum_{x_{\text{anc}}\in \boldsymbol{batch}}\left(\max_{x_{\text{pos}}\in \boldsymbol{A}} d(\boldsymbol{x}_{\text{anc}},\boldsymbol{x}_{\text{pos}}) - \min_{x_{\text{neg}}\in \boldsymbol{B}} d(\boldsymbol{x}_{\text{anc}},\boldsymbol{x}_{\text{neg}}) + \delta\right) \tag{5}$$

式中，$\boldsymbol{x}_{\text{anc}}$ 和 $\boldsymbol{x}_{\text{pos}}$ 是正样本对，它们的身份标签相同，$\boldsymbol{x}_{\text{pos}}$ 所属的图片集为 $\boldsymbol{A}$；$\boldsymbol{x}_{\text{anc}}$ 和 $\boldsymbol{x}_{\text{neg}}$ 是负样本对，它们的身份标签不同，$\boldsymbol{x}_{\text{neg}}$ 所属的图片集为 $\boldsymbol{B}$。困难样本三元组指对于每一个训练 ***batch***，随机挑选 P 个不同身份的行人，每个行人随机选 K 幅不同的图像，对于训练批次中每一个 $\boldsymbol{x}_{\text{anc}}$，选取类内距离最远的样本作为 $\boldsymbol{x}_{\text{pos}}$，在训练批次内所有负样本中选取距离最近的作为 $\boldsymbol{x}_{\text{neg}}$。$d(\cdot,\cdot)$ 表示两个输入元素的欧氏距离，而 $\delta \geqslant 0$ 表示三元组损失的边界。在实验中根据经验将 δ 设为 0.2。通过使 Tri-Hard 损失最小化，$d(\boldsymbol{x}_{\text{anc}},\boldsymbol{x}_{\text{pos}})$ 趋于 0，而 $d(\boldsymbol{x}_{\text{anc}},\boldsymbol{x}_{\text{neg}})$ 以一定的边界距离大于 $d(\boldsymbol{x}_{\text{anc}},\boldsymbol{x}_{\text{pos}})$。当损失变为 0 时，不会向后传播梯度。

身份保持器 $\boldsymbol{\Phi}$ 需要根据身份特征 $f(\boldsymbol{x})$ 来区分输入图像的身份信息，而生成器 $\boldsymbol{G}$ 需要从源视角生成满足视角指示器所分类视角的步态模板。因此，本文将身份保留损失分为两部分：用于 $\boldsymbol{\Phi}$ 优化的保留真实步态图像的相似度损失 L_{tri}^{Φ}，以及用于 $\boldsymbol{G}$ 优化的保留生成的步态图像的相似度损失 L_{tri}^{G}。当优化身份保持器 $\boldsymbol{\Phi}$ 时，利用目标视角的真实的步态模板 GEI 作为 $\boldsymbol{x}_{\text{anc}}$；当优化生成器 $\boldsymbol{G}$ 时，将生成视角指示器 $\boldsymbol{v}$ 指定的步态模板 $\hat{\boldsymbol{x}}$ 作为 $\boldsymbol{x}_{\text{anc}}$。$L_{\text{tri}}^{\Phi}$ 与 L_{tri}^{G} 的主要区别在于 $\boldsymbol{x}_{\text{anc}}$ 样本的选取，即真实的步态模板或生成的步态模板。

1.4　自注意力机制

虽然基于 GAN 的步态识别方法可实现步态模板的视角转换，但在视角转换过程中未能有效捕获特征间的全局依赖关系，生成图像的细节信息不够清晰，而且会伴随模糊的伪影。这是由于卷积核大小受限，无法在有限的网络层次结构中直接获取图像所有位置特征间的关联关系；而自注意力机制可以更好地处理图像中长范围、多层次的依赖关系，有助于增强步态特征的表达能力，提高步态识别的性能。因此，本文将自注意力机制（Zhang 等，2018）引入到生成器和判别器网络中，在生成器的上采样区引入自注意力模块能更好地整合全局和局部的空间信息，提高生成图像的协调性和质量；在判别器引入自注意力模块可以更准确地将真实图像和生成图像进行区分。

如图 2 所示，自注意力模块将前一层提取的特征图 $\boldsymbol{x}\in \mathbf{R}^{C\times N}$ 分别送入两个卷积核为 1、输出通道数是 $C/8$ 的特征空间 $f(\boldsymbol{x})$，$g(\boldsymbol{x})$ 和卷积核为 1、输出通道数为 C 的特征空间 $h(\boldsymbol{x})$，其中 $f(\boldsymbol{x}) = \boldsymbol{W}_f\boldsymbol{x}$，$g(\boldsymbol{x}) = \boldsymbol{W}_g\boldsymbol{x}$，$h(\boldsymbol{x}) = \boldsymbol{W}_h\boldsymbol{x}$，式中，$\boldsymbol{W}_f$、$\boldsymbol{W}_g$、$\boldsymbol{W}_h$ 分别为特征空间 $f(\boldsymbol{x})$、$g(\boldsymbol{x})$、$h(\boldsymbol{x})$ 对应的权重矩阵，且 $\boldsymbol{W}_f\in \mathbf{R}^{C/8\times N}$，$\boldsymbol{W}_g\in \mathbf{R}^{C/8\times N}$，$\boldsymbol{W}_h\in \mathbf{R}^{C\times N}$。通过对 $f(\boldsymbol{x})$ 和 $g(\boldsymbol{x})$ 进行张量相乘来计算两个特征空间相似度 s_{ij}，再使用 softmax 函数进行归一化，得到第 j 个区域对第 i 个位置所占权重的注意力图 $\beta_{j,i}$，具体为

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})},\quad s_{ij} = f(\boldsymbol{x}_i)^{\mathrm{T}}g(\boldsymbol{x}_j) \tag{6}$$

随后，将特征图 $\boldsymbol{x}$ 经过特征空间 $h(\boldsymbol{x})$，再与 $\beta_{j,i}$

构成的注意力权重矩阵相乘，注意力层的输出为

$$o_i = \sum_{i=1}^{N} \beta_{j,i} h(\boldsymbol{x}_i), \quad h(\boldsymbol{x}_i) = \boldsymbol{W}_h \boldsymbol{x}_i \tag{7}$$

式中，o_i 为注意力层的输出，$h(\boldsymbol{x}_i)$ 为输入信息 $\boldsymbol{x}$ 与权重矩阵 $\boldsymbol{W}_h \in \mathbf{R}^{C \times N}$ 的乘积。

最后，将注意力层的输出与比例系数 γ 相乘，并添加回输入特征图 $\boldsymbol{x}$，最终输出为

$$\boldsymbol{y}_i = \gamma o_i + \boldsymbol{x}_i \tag{8}$$

式中，γ 是初始值为 0 的比例系数，$\boldsymbol{y}_i$ 表示最终的输出。输出的注意力特征图会进入下一个网络中继续特征提取与学习的过程。随着网络训练的进行，注意力特征图逐渐为非局部区域分配更多的权重。

2　网络训练

本文采用 Goodfellow 等人(2014)提出的交替迭代训练的策略，当更新一方的参数时，另一方的参数固定住不更新。网络的训练过程如下：

输入：训练集 $\boldsymbol{X}$。

输出：网络 $\boldsymbol{D}$，$\boldsymbol{G}$，$\boldsymbol{\Phi}$。

1)判别过程：

(1)训练集 $\boldsymbol{X}$ 选取任意视角的步态模板 $\boldsymbol{x}_i^p$ 输入 $\boldsymbol{G}$ 网络生成目标视角的步态模板 $\hat{\boldsymbol{x}}$；

(2)视角判别器 $\boldsymbol{D}$ 网络输出图像真/伪标签并分类到相应的视角域，计算 L_D；

(3)身份判别器 $\boldsymbol{\Phi}$ 网络以注册集中真实的步态模板作为 $\boldsymbol{x}_{\text{anc}}$，计算 L_{tri}^{Φ}；

(4) L_{D} 和 L_{tri}^{Φ} 反向传递至 $\boldsymbol{D}$、$\boldsymbol{\Phi}$ 网络更新参数。

2)生成过程：

(1)验证集任意视角的步态模板 $\boldsymbol{x}_i^p$ 输入 $\boldsymbol{G}_{\text{enc}}$ 网络提取步态特征 $f(\boldsymbol{x})$；

(2)对目标视角以等概率来随机采样目标视角指示器 $\boldsymbol{v}$；

(3)步态特征 $f(\boldsymbol{x})$ 与视角指示器 $\boldsymbol{v}$ 向量连接送入 $\boldsymbol{G}_{\text{dec}}$ 网络生成目标视角的步态模板 $\hat{\boldsymbol{x}}$；

(4)视角判别器 $\boldsymbol{D}$ 网络输出图像真/伪标签并分类到相应的视角域，计算 L_{D}；

(5)身份判别器 $\boldsymbol{\Phi}$ 网络以生成视角指示器 $\boldsymbol{v}$ 所指定的步态模板 $\hat{\boldsymbol{x}}$ 作为 $\boldsymbol{x}_{\text{anc}}$，计算 $L_{\text{tri}}^{\text{G}}$；

(6)反向传递损失至 $\boldsymbol{G}$ 网络并计算 L_{pixel}；

(7)根据 L_{D}、$L_{\text{tri}}^{\text{G}}$ 和 L_{pixel} 计算 L_{G}，并更新 $\boldsymbol{G}$ 网络的权重。

3)重复步骤 1)和 2)，直至网络收敛。

本文的目标是将步态模板从验证集中的任意视角转换至注册集中的目标视角，同时保留身份信息。为了实现这个目标，联合上述损失函数协同训练，总体目标函数为

$$L = \lambda_1 L_{\text{adv}} + \lambda_2 L_{\text{cls}}^{\text{D}} + \lambda_3 L_{\text{tri}}^{\text{G}} + \lambda_4 L_{\text{pixel}} \tag{9}$$

式中，$\lambda_t, t \in \{1,2,3,4\}$ 是超参数，用来平衡不同的损失。随着模型训练次数增加，视角判别器区分真/伪和视角分类性能越来越强，身份保持器更准确地保留输入步态图像的身份标签，而生成器更好地生成具有目标视角并保持身份信息的步态图像。整个训练过程得益于 4 个方面：1)$\boldsymbol{G}_{\text{enc}}$ 学习输入步态图像的特征表示 $f(\boldsymbol{x})$，将保留更多具有鉴别性的身份信息；2)$\boldsymbol{D}$ 中视角分类可引导步态图像的视角转换更加准确；3)视角指示器和身份特征连接向量作为 $\boldsymbol{G}_{\text{dec}}$ 的输入，可引导生成器生成不同视角的步态图像；4)引入自注意力机制，提高了生成图像的协调性和质量。

3　实验结果及分析

3.1　数据集

3.1.1　公共数据集

CASIA-B(Chinese Academy of Sciences' Institute of Automation gait database——dataset B)步态数据集(Yu 等，2006)是广泛用于评估跨视角步态识别效果的公共数据集，包含 124 人、3 种行走状态和 11 个不同视角(0°，18°，…，180°)。每个人在正常状态下有 6 个序列(NM #01—06)，穿着外套状态下有 2 个序列(CL#01—02)，携带背包状态下有 2 个序列(BG#01—02)，所以，每个人有 11 × (6 + 2 + 2) = 110 个序列。

OU-MVLP(multi-view large population dataset)步态数据集(Takemura 等，2018)是迄今为止世界上最大的跨视角步态数据库，包含 10 307 人、14 个不同视角(0°，15°，…，90°；180°，195°，…，270°)以及每个角度有 2 个序列(#00—01)，步行状态没有变化。官方将数据库分为 5 153 人的训练集和 5 154 人的测试集。在测试阶段，序列#01 作为注册集，序列#00 作为测试集。

3.1.2　帧移式合成 GEI 数据集

本文方法是基于 CNN 实现的 GAN 网络，其性

能在一定程度上取决于训练样本的数据规模。考虑到 CASIA-B 数据量较少,而 OU-MVLP 数据量大,因此通过对 CASIA-B 的 GEI 数据集进行数据增强来评估对步态识别准确率的影响。

本文采用帧移式方法来增加合成 GEI 的数量,帧移式生成 GEI 的原理如图 4 所示。输入步态序列为 N 帧,根据轮廓的宽高比,得到步态周期为 k 帧($k \leqslant N$),首先将前 k 帧的步态序列图合成一幅 GEI,再以 i 帧间隔抽取第 i 帧到第 $i+k$ 帧的步态序列图合成下一幅 GEI,以此类推,直至 $c \times i + k$ 为 N,则合成完该序列所有 GEI,本文设置 i 为 2。大多学者是将所有周期内的轮廓图合成最终一幅 GEI,数据量略显不足。本文利用步态序列的前后循环性和连贯性,将步态序列按照周期帧移方式合成更多 GEI。

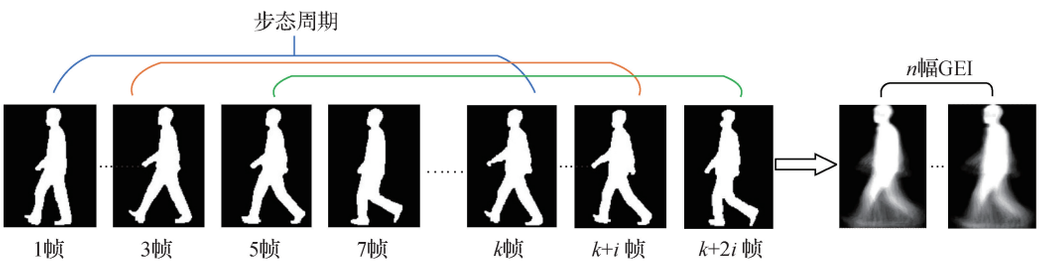

图 4　帧移式生成 GEI 的原理图

Fig. 4　Schematic diagram of frame-shift generation of GEI

3.2　评价指标及实验设置

3.2.1　评价指标

步态识别任务中常用的评价指标是 CMC(cumulative matching characteristic)曲线中的 rank-1 识别准确率,是由最近邻分类器计算得出的。对于行人 $\boldsymbol{x}_i$ 在验证集视角为 p 的步态特征 $f(\boldsymbol{x}_i^p)$,首先通过本文方法进行视角转换到注册集视角为 g 的步态特征 $f(\boldsymbol{x}_i^g)$,再采用欧氏距离作为距离度量标准,即

$$d = \sqrt{f(\boldsymbol{x}_i^p)^2 - f(\boldsymbol{x}_i^g)^2} \tag{10}$$

然后根据欧氏距离搜寻注册集中距离最近的步态特征,从而判断是否具有相同身份。

3.2.2　实验设置

实验基于深度学习框架 Pytorch 在显卡为 NVIDA RTX2080Ti × 2 的 Dell 工作站上进行训练。本文在 CASIA-B 数据集的实验设置是将数据集均匀划分为两组,前 62 人用于训练,后 62 人用于测试。网络输入和输出的 GEI 尺寸设置为 64 × 64 像素,批量大小 batch_size 设为 64。考虑到 CASIA-B 数据集训练人数较少,使用 GEI 数据增强进行实验。在 OU-MVLP 数据集的设置与官方(Takemura 等,2018)一致,由于 OU-MVLP 中 GEI 数据量远超 CASIA-B,故将 batch_size 设为 32。

如第 2 节所述,本文采用交替训练 $\boldsymbol{G}$、$\boldsymbol{D}$ 和 $\boldsymbol{\Phi}$ 网络的方式。由于判别器的学习能力强于生成器,为了保持两者同步,当判别器 $\boldsymbol{D}$ 和 $\boldsymbol{\Phi}$ 训练 5 次后,对生成器 $\boldsymbol{G}$ 更新 1 次。在训练过程中,所有网络模型的权重通过均值为 0、方差为 0.02 的高斯分布进行随机初始化。采用 Adam 优化器更新网络参数,$\beta_1 = 0.5$,$\beta_2 = 0.999$,生成器和判别器网络分别采用 0.000 1 和 0.000 4 的初始化学习率进行单独训练。对于 CASIA-B 数据集,本文训练模型 40 K 迭代次数,前 20 K 迭代时学习率保持不变,剩下 20 K 轮迭代采用 step 策略,每 5 K 轮迭代学习率下降为原来的 10%,直至衰减为 0。对于 OU-MVLP 数据集,本文训练模型 200 K,前 150 K 迭代时学习率保持不变,剩下 50 K 轮迭代,每 10 K 轮迭代学习率变为原来的 10%。在本文实验中,凭经验设置式(9)中的权重系数,$\lambda_1 = \lambda_2 = 1$、$\lambda_3 = \lambda_4 = 10$。

3.3　实验结果与分析

3.3.1　消融实验

为探究自注意力模块在网络中所处位置对识别性能的影响,本文将自注意力模块添加到生成器的不同位置,并在 CASIA-B 数据集进行对比实验,如表 3 所示。可以看出,自注意力模块添加到解码器第 2 层反卷积之后位置识别效果更好,而位置靠前、靠后或添加到编码器的识别效果均不理想。当添加位置较靠前时,采集到的信息较粗糙,噪声较大;而当对较小的特征图建立依赖关系时,其作用与局部卷积作用相似。因此在特征图较大情况下,自注意力能捕获更多的信息,选择区域的自由度也更大,从而使生成器和判别器能建立更稳定的依赖关系。自

注意力模块需在中高层特征图之间使用，所以本文将自注意力机制添加到解码器第 2 层反卷积后的特征图上。而同时在编码器中加入自注意力模块会导致部分生成的步态模板信息丢失，所以没有单独在解码器中加入自注意力模块的效果好。此外，通过对比生成器中添加自注意力模块与未使用自注意力模块的实验结果，前者识别率较高，进一步验证了自注意力模块的有效性。

表 3　自注意力模块处于生成器不同位置对识别率的影响

Table 3　The effect of different position of the generator of self-attention module on recognition performance

/%

网络模型	验证集角度			平均值
	54°	90°	126°	
$\boldsymbol{G}$（without self-attention）	79.5	70.2	78.4	76.0
$\boldsymbol{G}_{\mathrm{dec}}$（Deconv1 + self-attention）	80.7	71.4	80.1	77.4
$\boldsymbol{G}_{\mathrm{dec}}$（Deconv2 + self-attention）	**81.7**	**72.0**	**80.7**	**78.1**
$\boldsymbol{G}_{\mathrm{dec}}$（Deconv3 + self-attention）	81.2	71.7	80.4	77.8
$\boldsymbol{G}_{\mathrm{enc}}$（Conv2 + self-attention） + $\boldsymbol{G}_{\mathrm{dec}}$（Deconv2 + self-attention）	81.0	71.3	80.1	77.5
$\boldsymbol{G}_{\mathrm{enc}}$（Conv3 + self-attention） + $\boldsymbol{G}_{\mathrm{dec}}$（Deconv3 + self-attention）	81.1	71.5	80.2	77.6

注：加粗字体表示各列最优结果。$\boldsymbol{G}_{\mathrm{dec}}$（Deconv2 + self-attention）表示自注意力模块放置于解码器中第 2 层反卷积之后，其他依次类推。

通过上述实验，自注意力模块对步态模板生成具有较好的识别效果，为进一步提高生成图像的判别能力，在身份保持器中融合身份保留损失，为验证其对步态识别效果的影响，在 CASIA-B 数据集进行消融实验。实验结果如表 4 所示。

表 4　本文不同方案在 CASIA-B 的识别率对比

Table 4　Comparison of recognition performance among different schemes under proposed framework

/%

方法	验证集角度			平均值
	54°	90°	126°	
GaitGAN	64.5	58.2	65.7	62.8
GaitGAN + 自注意力模块	79.3	67.7	77.3	74.8
GaitGAN + 身份保留损失	79.5	70.2	78.4	76.0
GaitGAN + 自注意力模块 + 身份保留损失	**81.7**	**72.0**	**80.7**	**78.1**

注：加粗字体表示各列最优结果。

从表 4 可以看出，在网络模型中没有自注意力模块或身份保留损失的情况下，本文方法仍然比基准方法 GaitGAN 的识别率高。当引入自注意力模块和身份保留损失训练网络时，在 CASIA-B 数据集上的识别率有显著提升，平均 rank-1 准确率提升了 15%。实验结果表明，自注意力模块有效解决了目标视角步态模板生成的不完全的问题，提升了生成图像的协调性；身份保留损失使生成的步态模板更好地保持身份信息，增强了生成图像的可判别性。自注意力模块和身份保留损失两者结合有效提高了步态视角转换的效果与质量。

为进一步验证 GEI 数据增强对步态识别效果的影响，在 CASIA-B 数据集上进行实验，结果如图 5 所示。

从图 5 可以看出，经过 GEI 数据增强，达到了最佳识别精度。与 GaitGAN 方法相比，即使未经数据增强训练的方法也能取得较高的识别率。通过 GEI 数据增强，既避免了因生成的步态能量图过少导致的识别率不高问题，也避免了不同身份的 GEI 样本过于接近问题，有助于提高跨视角步态识别率。

3.3.2　与最新方法对比

1）在 CASIA-B 数据集实验结果。为验证本文

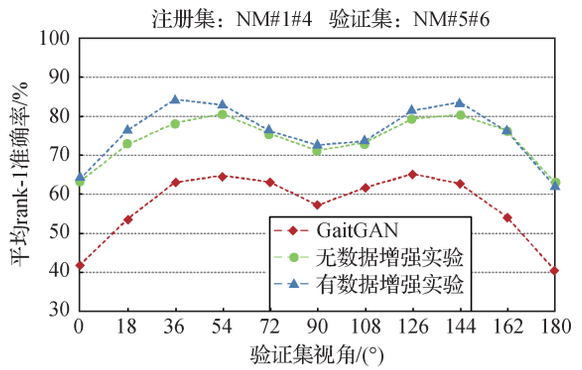

图 5　GEI 数据增强在 CASIA-B 数据集的实验结果

Fig. 5　Impact of GEI data augmentation evaluated on CASIA-B

方法的有效性，与 C3A（complete canonical correlation analysis）（Xing 等，2016）、SPAE（stacked progressive auto-encoders）（Yu 等，2017b）、GaitGAN（Yu 等，2017a）和 MGAN（He 等，2019）等最新方法进行比较，选择验证集视角为 54°、90°、126°进行跨视角步态识别的对比实验。图 6 显示了排除相同视角的所有注册集视角的跨视角步态识别率。

从图 6 可以看出，本文方法在 3 个验证集视角下的识别率均优于其他 4 种方法。当注册集视角与验证集视角相邻近时，识别效果最好；且当视角跨度逐渐增大时，其他方法识别率下降幅度剧烈，本文方法仍能表现良好。为了更明显展示跨视角步态识别性能，在排除相同视角下，对验证集视角为 54°、90°、126°时的平均识别率进行比较，结果如表 5 所示。可以看出，本文方法优于 GaitGAN、TS-GAN、MGAN 和 DV-GEIs（Liao 等，2020），平均 rank-1 识别率达到 79.8%，比 MGAN 方法提升了 2.1%。DV-GEIs 采用线性插值方法以 1°为间隔生成 0°～180°的密集视角步态特征，但会使生成的 GEI 出现边缘模糊和振铃等瑕疵，引入不必要的噪声；而本文方法加入自注意力机制并扩充了 GEI 数据集，所以识别率较高于 DV-GEIs。实验结果表明，注册集视角为 54°和 126°时的识别率明显高于 90°。这是因为除了 0°和 180°与摄像头拍摄方向几乎相同或相反的视角，侧面 90°视角与其他角度的步态外观差别最大，而视角为 54°和 126°时的样本更具有区分性的步态特征，所以识别率更高。

2）在 OU-MVLP 数据集实验结果。本文对 4 个在 OU-MVLP 数据集实验的方法不多，所以选择与 GEINet（Shiraga 等，2016）、3in＋2diff（Takemura 等，

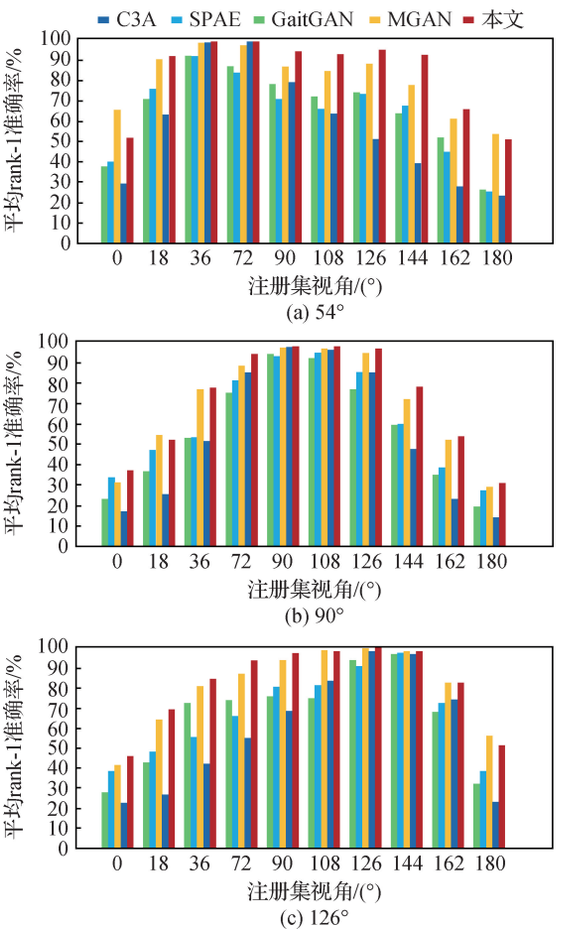

图 6　在 Probe NM 的 3 个代表性视角 54°、90°和 126°下与最新方法比较结果（排除相同视角）

Fig. 6　Comparison with the state-of-the-art methods under the probe views 54°, 90° and 126° excluding identical view（（a）54°；（b）90°；（c）126°）

表 5　排除相同视角下，在 CASIA-B 数据集中 3 种验证集视角的平均识别率比较

Table 5　Comparison of average identification rates among three probe views excluding identical view on CASIA-B dataset

/%

方法	验证集角度			平均值
	54°	90°	126°	
GaitGAN	64.5	58.2	65.7	62.8
TS-GAN	67.3	63.1	68.2	66.2
MGAN	81.6	71.2	80.3	77.7
DV-GEIs	80.8	72.6	78.9	77.4
本文	**83.7**	**73.2**	**82.4**	**79.8**

注：加粗字体表示最优结果。

典型视角（0°、30°、60°、90°）进行实验，由于近几年

2019)和 GaitSet(Chao 等,2019)等 3 种方法进行比较,结果如表 6 所示,所有结果都是在排除相同视角的注册集视角下取平均值得到的识别率。从表 6 可以看出,GEINet 和 3in + 2diff 方法在 OU-MVLP 这种大规模的跨视角步态识别评估实验中识别性能较差,而本文方法可以达到 65.9% 的平均识别精度,远高于这两种方法。由于 GaitSet 采用人体轮廓序列作为输入特征,比 GEI 包含更多的时空特征信息,所以识别率更高。实验结果表明,与采用 GEI 步态模板的其他方法相比,本文方法在大规模的跨视角步态数据库中仍具有较好的适用性。

表 6　排除相同视角下,在 OU-MVLP 数据集中 4 种典型视角的平均识别率比较

Table 6　Comparison of average identification rates among four representative probe views on OU-MVLP dataset

/%

方法	验证集角度				平均值
	0°	30°	60°	90°	
GEINet	8.2	32.3	33.6	28.5	25.7
3in + 2diff	25.5	50.0	45.3	40.6	40.4
GaitSet	**77.7**	**86.9**	**85.3**	**83.5**	**83.35**
本文	58.4	70.6	67.9	66.7	65.9

注:加粗字体表示最优结果。

3.3.3　实验结果定性分析

目前基于 GAN 的步态识别方法中,MGAN 需要事先对视角进行估计才能实现特定视角的步态图像生成,GaitGAN 和 TS-GAN 则是将任意视角的步态模板标准化到侧面视角进行识别,如果要将某一视角的步态模板转换到任意视角,则需构建多个模型,而本文方法建立的统一模型可将步态模板从任意视角转换到目标视角。本文将 OU-MVLP 数据集中的 4 个典型视角(0°,30°,60°,90°)合成的步态模板进行可视化,如图 7 所示。其中,左侧图像为验证集中

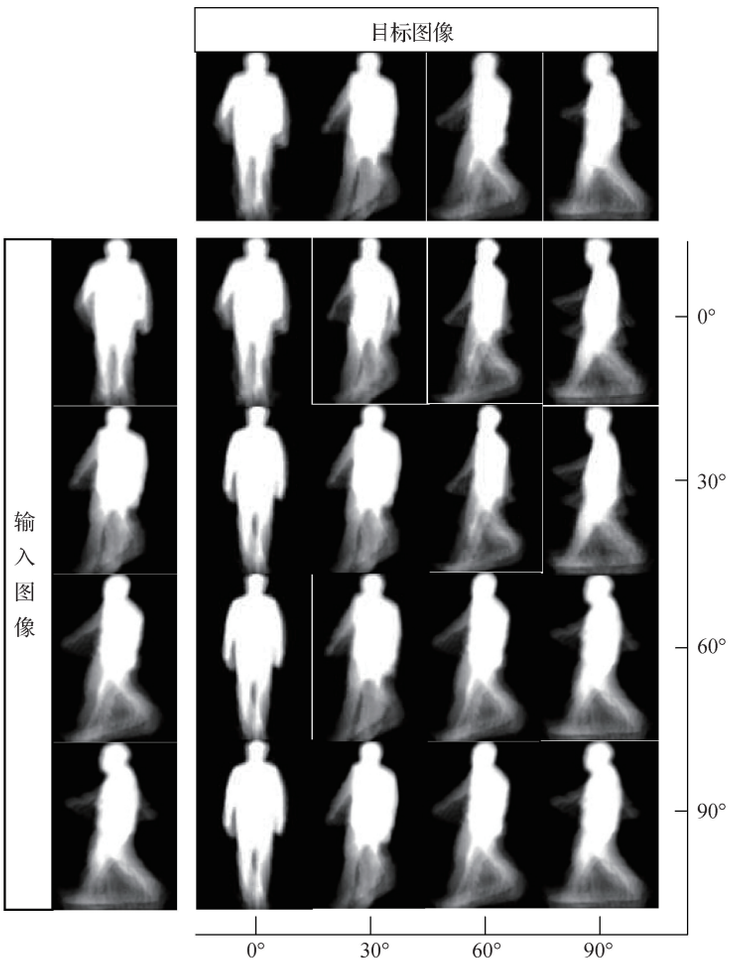

图 7　输入视角为 0°,30°,60°和 90°的步态模板合成的可视化结果

Fig. 7　Visualization of generated gait templates at 0°, 30°, 60° and 90° with different input views

的输入 GEI,上侧图像是注册集中真实的目标 GEI,右下 4 × 4 矩阵中的图像是生成的 GEI。由图 7 可以看出,本文训练的任意视角间步态模板转换模型即使在视角变化较大情况下,生成的步态图像也与真实的目标视角的步态图像高度相似。

4 结 论

针对步态识别中的跨视角问题,本文提出融合自注意力机制的生成对抗网络框架,建立可实现任意视角间的步态模板转换模型,由生成器、视角判别器和身份保持器构成,解决了目前生成式方法只能进行特定视角的步态转换并且生成图像的特征信息容易丢失问题,达到了使用统一模型生成任意视角的步态模板的效果,并在视角转换过程中保留步态特征信息,提升了生成图像的质量。

为验证本文方法对跨视角步态识别的有效性,在 CASIA-B 步态数据库上分别进行对比、消融和增强实验,设计将自注意力模块添加到生成器的不同位置进行对比实验,结果表明在解码器第 2 层反卷积后加入自注意力模块效果更好;对自注意力模块和身份保留损失进行消融实验,相比于 Gait GAN 方法,两者结合时的步态识别率有显著提升;采用帧移式方法对 CASIA-B 数据集进行 GEI 数据增强实验,进一步提升了识别率。在 OU-MVLP 大规模的跨视角步态数据库中进行对比实验,与 GEINet、3in + 2diff 两种方法相比,所提方法仍具有较好的适用性,可以达到 65.9% 的平均识别精度。

本文方法以步态能量图为模型输入,计算简单有效,但在实际场景中,行人检测与分割的好坏会直接影响合成步态能量图的质量;同时在实际应用中,视角变化会与其他协变量(如衣着、携带物)结合。因此,如何建立功能更强大的网络模型来解决复杂场景的步态识别问题,仍是未来步态识别研究的技术难点。

参考文献(References)

Chao H Q, He Y W, Zhang J P and Feng J F. 2019. GaitSet: regarding gait as a set for cross-view gait recognition//Proceedings of 2019 AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press: 8126-8133 [DOI: 10.1609/aaai.v33i01.33018126]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: ACM: 2672-2680 [DOI: 10.5555/2969033.2969125]

Han J and Bhanu B. 2006. Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(2): 316-322 [DOI: 10.1109/TPAMI.2006.38]

He Y W, Zhang J P, Shan H M and Wang L. 2019. Multi-task GANs for view-specific feature learning in gait recognition. IEEE Transactions on Information Forensics and Security, 14(1): 102-113 [DOI: 10.1109/TIFS.2018.2844819]

Hermans A, Beyer L and Leibe B. 2017. In defense of the triplet loss for person re-identification [EB/OL]. [2020-08-10]. https://arxiv.org/pdf/1703.07737.pdf

Hu M D, Wang Y H, Zhang Z X, Little J J and Huang D. 2013. View-invariant discriminative projection for multi-view gait-based human identification. IEEE Transactions on Information Forensics and Security, 8(12): 2034-2045 [DOI: 10.1109/TIFS.2013.2287605]

Liao R J, An W Z, Yu S Q, Li Z and Huang Y Z. 2020. Dense-view GEIs set: view space covering for gait recognition based on dense-view GAN//Proceedings of 2020 IEEE International Joint Conference on Biometrics. Houston, USA: IEEE: 1-9 [DOI: 10.1109/IJCB48548.2020.9304910]

Makihara Y, Sagawa R, Mukaigawa Y, Echigo T and Yagi Y. 2006. Gait recognition using a view transformation model in the frequency domain//Proceedings of the 9th European Conference on Computer Vision. Graz, Austria: ACM: 151-163 [DOI: 10.1007/11744078_12]

Miyato T, Kataoka T, Koyama M and Yoshida Y. 2018. Spectral normalization for generative adversarial networks [EB/OL]. [2020-08-10]. https://arxiv.org/pdf/1802.05957.pdf

Ouyang N, Liang T and Lin L P. 2019. Self-attention network based image super-resolution. Journal of Computer Applications, 39(8): 2391-2395 (欧阳宁, 梁婷, 林乐平. 2019. 基于自注意力网络的图像超分辨率重建. 计算机应用, 39(8): 2391-2395) [DOI: 10.11772/j.issn.1001-9081.2019010158]

Shiraga K, Makihara Y, Muramatsu D, Echigo T and Yagi Y. 2016. GEINet: view-invariant gait recognition using a convolutional neural network//Proceedings of 2016 International Conference on Biometrics (ICB). Halmstad, Sweden: IEEE: 1-8 [DOI: 10.1109/ICB.2016.7550060]

Takemura N, Makihara Y, Muramatsu D, Echigo T and Yagi Y. 2018. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Transactions on Computer Vision and Applications, 10(4): #4 [DOI: 10.1186/s41074-018-0039-6]

Takemura N, Makihara Y, Muramatsu D, Echigo T and Yagi Y. 2019.

On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Transactions on Circuits and Systems for Video Technology, 29(9): 2708-2719 [DOI: 10.1109/TCSVT.2017.2760835]

Tran L, Yin X and Liu X M. 2017. Disentangled representation learning GAN for pose-invariant face recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1283-1292 [DOI: 10.1109/CVPR.2017.141]

Wang K J, Ding X N, Xing X L and Liu M C. 2019. A survey of multi-view gait recognition. Acta Automatica Sinica, 45(5): 841-852 (王科俊, 丁欣楠, 邢向磊, 刘美辰. 2019. 多视角步态识别综述. 自动化学报, 45(5): 841-852) [DOI: 10.16383/j.aas.2018.c170599]

Wang Y Y, Song C F, Huang Y, Wang Z Y and Wang L. 2019. Learning view invariant gait features with Two-Stream GAN. Neurocomputing, 339: 245-254 [DOI: 10.1016/j.neucom.2019.02.025]

Wu Z F, Huang Y Z, Wang L, Wang X G and Tan T N. 2017. A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(2): 209-226 [DOI: 10.1109/TPAMI.2016.2545669]

Xing X L, Wang K J, Yan T and Lyu Z W. 2016. Complete canonical correlation analysis with application to multi-view gait recognition. Pattern Recognition, 50: 107-117 [DOI: 10.1016/j.patcog.2015.08.011]

Yu S Q, Chen H F, Reyes E B G and Poh N. 2017a. GaitGAN: invariant gait feature extraction using generative adversarial networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA: IEEE: 532-539 [DOI: 10.1109/CVPRW.2017.80]

Yu S Q, Chen H F, Wang Q, Shen L L and Huang Y Z. 2017b. Invariant feature extraction for gait recognition using only one uniform model. Neurocomputing, 239: 81-93 [DOI: 10.1016/j.neucom.2017.02.006]

Yu S Q, Tan D L and Tan T N. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition//Proceedings of the 18th International Conference on Pattern Recognition. Hong Kong, China: IEEE: 441-444 [DOI: 10.1109/ICPR.2006.67]

Zhang H, Goodfellow I, Metaxas D and Odena A. 2018. Self-attention Generative Adversarial Networks [EB/OL]. [2020-08-10]. https://arxiv.org/pdf/1805.08318.pdf

Zhi S S, Zhao Q H and Tang J. 2019. Semantic segmentation of gait body with multilayer semantic fusion convolutional neural network. Journal of Image and Graphics, 24(8): 1302-1314 (支双双, 赵庆会, 唐琎. 2019. 多层语义融合 CNN 的步态人体语义分割. 中国图象图形学报, 24(8): 1302-1314) [DOI: 10.11834/jig.180597]

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

作者简介

张红颖，1978 年生，女，教授，主要研究方向为图像工程与计算机视觉。
E-mail: carole_zhang0716@163.com

包雯静，女，硕士研究生，主要研究方向为步态识别与深度学习。E-mail: wjbao_cauc@hotmail.com