中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-13

论文引用格式:Su Xueping, Chen Xiang, Guo Chenxi, Li Yunhong, Wang Libin. XXXX. Cross-Modal Identity Recognition with Temporal Context Fusion. Journal of Image and Graphics, XX(XX):0001-0013(苏雪平, 陈翔, 郭晨曦, 李云红, 王力斌. XXXX. 融合时序上下文的跨模态身份识别算法. 中国图象图形学报, XX(XX):0001-0013)[DOI:10.11834/jig.260201]

融合时序上下文的跨模态身份识别算法

苏雪平[1,3],陈翔[2],郭晨曦[2],李云红[2],王力斌[2]

1. 西安工程大学计算机科学学院,西安　710048; 2. 西安工程大学电子信息学院,西安　710048; 3. 陕西省服装设计智能化重点实验室,西安　710048

摘　要:**目的**　针对低质量视频和远距离场景下人脸与步态识别精度下降的问题,本文提出一种融合时序上下文的跨模态身份识别算法,以提升复杂航空场景下身份识别的准确性与鲁棒性。**方法**　首先,针对低质量视频人脸图像模糊、遮挡及时序不一致问题,提出基于动态空间-时间-通道模型(Dynamic Spatial-Temporal-Channel Model, DSTC)的低质量视频人脸恢复算法,实现跨帧细节补偿与身份特征增强;其次,针对步态识别中多尺度动态信息建模不足问题,提出融合空间与通道重构卷积(Spatial and Channel Reconstruction Convolution, SCConv)与多尺度动态增强模块(Multi-Scale Dynamic Augmentation Module, MS-DAM)的步态特征提取算法,提升模型对复杂行走条件的适应能力;最后,针对未充分融合人脸与步态异构信息的问题,提出跨模态上下文感知时序融合(Context-aware Temporal, CoT)特征融合的身份识别算法,实现人脸与步态信息的有效互补。**结果**　,DSTC方法在Labeled Faces in the Wild(LFW)数据集和自建无感考勤数据集上分别达到PSNR(27.43, 26.22)、SSIM(0.813, 0.801)、LPIPS(0.189, 0.197);步态识别在CASIA-B数据集的NM、BG和CL场景下Rank-1准确率分别为97.6%、94.6%和84.9%;跨模态CoT融合算法在CASIA-B人脸-步态数据集上正常行走(Normal Walking, NM)、携带背包(Bag, BG)和服装变化(Clothing, CL)场景下的Rank-1准确率分别达到97.8%、95.4%和86.9%,性能均优于现有主流算法。**结论**　本文提出的跨模态身份识别算法有效利用人脸与步态互补信息,显著提升航空领域下身份识别的精度。三个子模块形成递进式协同:DSTC为后续跨模态融合提供高质量人脸特征,与基于SCConv+MS-DAM的步态分支共同构成CoT融合的特征基础,在保证轻量化的同时实现复杂场景下鲁棒的身份识别。

关键词:跨模态身份识别;人脸恢复;人脸识别;步态识别;时序上下文融合

Cross-Modal Identity Recognition with Temporal Context Fusion

Su Xueping[1,3], Chen Xiang[2], Guo Chenxi[2], Li Yunhong[2], Wang Libin[2]

1. School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China; 2. School of Electronic Information, Xi'an Polytechnic University, Xi'an 710048, China; 3. Shaanxi Key Laboratory of Clothing Intelligence, Xi'an 710048, China

Abstract: **Objective** With the rapid development of intelligent perception and biometric recognition technologies, cross-modal identity recognition has become an important solution for contactless authentication in aviation applications, such as intelligent access control, security inspection, video surveillance, and attendance management. Among various biometric

收稿日期:XXXX-XX-XX;**修回日期**:XXXX-XX-XX;**预印本日期**:XXXX-XX-XX

基金项目:陕西省社科基金(2025M010);西安市科技局科学家+工程师队伍建设项目(25KGYB00029);西安市碑林区科技局项目(GX2405);陕西省教育厅青年团队项目(24JP071,25JP072);陕西高校青年创新团队

Supported by: Social Science Fund of Shaanxi Province (2025M010); Xi'an Municipal Bureau of Science and Technology (25KGYB00029); Science and Technology Bureau of Beilin District, Xi'an (GX2405); Shaanxi Provincial Department of Education (24JP071,25JP072); Youth Innovation Team of Shaanxi Universities.

traits, face and gait provide complementary identity information. Face recognition offers strong discriminative capability under favorable imaging conditions, whereas gait recognition enables long-distance and non-contact identification. However, practical aviation environments are often affected by low-quality video, motion blur, long-distance observation, occlusion, and appearance variations, which significantly reduce the performance of existing methods. Moreover, most studies optimize face restoration or gait recognition independently, while the complementary information between heterogeneous modalities and the temporal consistency in video sequences remain insufficiently exploited. To address these challenges, this paper proposes a unified cross-modal identity recognition framework incorporating temporal contextual information to improve accuracy and robustness in complex aviation scenarios. **Methods** The proposed framework consists of three progressive modules. First, a Dynamic Spatial-Temporal-Channel (DSTC) model is developed for low-quality video face restoration. By jointly modeling spatial structures, temporal consistency among adjacent frames, and channel-wise feature dependencies, the proposed model effectively compensates for missing facial details caused by image degradation while preserving identity-related characteristics. The restored facial representations provide reliable identity information for subsequent recognition tasks. Second, a lightweight gait feature extraction network is designed by integrating Spatial and Channel Reconstruction Convolution (SCConv) with a Multi-Scale Dynamic Augmentation Module (MS-DAM). SCConv reconstructs spatial and channel features to reduce feature redundancy and improve computational efficiency, while MS-DAM captures gait information at different temporal and spatial scales to enhance dynamic feature representation. The combination of these two modules improves the robustness of gait recognition under viewpoint variation, carrying conditions, and clothing changes. Finally, a Context-aware Temporal (CoT) fusion strategy is proposed to integrate heterogeneous face and gait features. Instead of directly concatenating multimodal representations, the proposed fusion module exploits temporal contextual relationships to adaptively combine complementary identity information from the two modalities. The three modules are connected in a progressive manner, where high-quality face restoration strengthens facial feature representation, enhanced gait modeling provides discriminative motion features, and context-aware fusion further constructs a unified identity representation for robust cross-modal recognition. **Results** Extensive experiments were conducted to evaluate the proposed framework from three aspects: low-quality video face restoration, gait recognition, and cross-modal identity recognition. For face restoration, experiments were performed on the public LFW dataset and a self-collected contactless attendance dataset. Three widely used image quality assessment metrics, namely Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), were adopted for quantitative evaluation. The proposed DSTC model achieved PSNR values of 27.43 dB and 26.22 dB, SSIM values of 0.813 and 0.801, and LPIPS scores of 0.189 and 0.197 on the two datasets, respectively, demonstrating its ability to recover degraded facial details while preserving identity-related information. For gait recognition, experiments on the CASIA-B dataset under the normal walking (NM), carrying bag (BG), and clothing variation (CL) protocols yielded Rank-1 accuracies of 97.6%, 94.6%, and 84.9%, respectively. These results indicate that the proposed SCConv-MS-DAM network effectively improves multi-scale dynamic feature representation and maintains robust recognition performance under challenging walking conditions. Furthermore, the proposed Context-aware Temporal fusion algorithm was evaluated on the CASIA-B face-gait dataset. By jointly exploiting restored facial features and discriminative gait representations, the proposed cross-modal framework achieved Rank-1 recognition accuracies of 97.8%, 95.4%, and 86.9% under the NM, BG, and CL scenarios, respectively. The experimental results consistently demonstrate that the proposed method outperforms representative mainstream approaches under different evaluation protocols and achieves more stable recognition performance in complex environments. **Conclusion** A A unified cross-modal identity recognition framework incorporating temporal context is proposed for intelligent aviation. It progressively integrates low-quality video face restoration, lightweight gait feature extraction, and context-aware multimodal fusion to improve identity representation reliability. The DSTC module enhances facial quality via spatial, temporal, and channel correlations, while SCConv-MS-DAM strengthens gait features through spatial-channel reconstruction and multi-scale dynamic modeling. Based on these enhanced unimodal representations, the CoT fusion strategy exploits complementary face and gait characteristics to generate more discriminative identity features. Experiments on public and self-collected datasets demonstrate superior performance in face restoration, gait recognition, and cross-modal identity recognition with a lightweight architecture. These findings verify the framework's effec-

tiveness and its potential for practical deployment in aviation-oriented contactless authentication systems under complex real-world conditions.

Key words: cross-modal person identification; face restoration; face recognition; gait recognition; temporal context fusion

论文引用格式：Su Xueping, Chen Xiang, Guo Chenxi, Li Yunhong, Wang Libin. Cross-Modal Identity Recognition with Temporal Context Fusion[J/OL]. Journal of Image and Graphics. DOI: 10.11834/jig.260201.（苏雪平，陈翔，郭晨曦，李云红，王力斌. 融合时序上下文的跨模态身份识别算法[J/OL]. 中国图象图形学报. DOI: 10.11834/jig.260201.）

0 引 言

随着生物特征识别技术的发展，人脸、步态等单模态方法已广泛应用于金融、安防等领域。然而，单模态识别易受噪声、环境变化及伪造攻击影响，复杂场景下稳定性不足。跨模态身份识别通过融合不同生物特征的互补信息，可有效提升识别鲁棒性与准确率，已成为当前研究热点。

在人脸恢复方面，早期插值方法对复杂退化场景处理能力有限；Liu 等（2018）和 Li 等（2018）分别利用多层卷积神经网络（convolutional neural networks，CNN）与参考图像引导提升重建精度；PULSE、GFP-GAN 等借助生成先验获得较好的恢复效果；Zhao 等（2023）的扩散模型进一步提高了图像真实性，但计算开销较大；Wang 等（2025）提出的 Spatial-Video Face Restoration（SVFR）构建了统一的视频恢复框架，但依赖多步推理，难以满足轻量化实时部署需求。总体而言，现有方法普遍结构复杂，计算代价较高，且对视频帧间时序关联利用不足；基于语义分割先验的两阶段方法还依赖分割精度，易受误差传播影响。针对上述问题，本文提出 DSTC 方法，通过 Spatial-Temporal-Channel（STC）模块联合建模空间、时间与通道信息，采用 Context Feature Aggregation Network（CFA-Net）替代传统自注意力降低计算复杂度，并引入自适应时序损失兼顾恢复质量与计算效率。

步态识别具有非接触、难伪装等优势。GaitSet、GaitPart、GaitGL 和 GPGait 分别通过集合学习、局部时序建模及姿态先验等方法提升了步态特征表达能力；GaitLLM 将步态序列映射至文本空间，引入大语言模型增强时序建模，但跨模态融合能力仍有提升空间。与此同时，现有 3D CNN 特征提取存在空间冗余，常规动态增强仅依赖单一帧差，难以有效建模不同时间尺度下的运动信息。为此，本文引入 SCConv 抑制空间与通道冗余，并设计 MS-DAM 通过多尺度差分提取时序动态特征，进一步提升复杂行走条件下的特征判别能力。

在跨模态融合方面，Liu 等（2020）、Xu 等（2023）和 Prakash 等（2023）分别采用注意力或无键注意力实现自适应融合，但现有方法多采用静态或弱自适应策略，难以根据场景动态调整模态权重，同时全矩阵注意力计算复杂度较高。针对上述问题，本文提出 CoT 融合方法，在局部时间窗口内通过轻量门控机制实现人脸与步态特征的自适应融合，在兼顾计算效率的同时提升跨模态身份识别性能。

针对上述问题，本文构建统一的跨模态身份识别框架，主要贡献如下：

1）提出基于 DSTC 的低质量视频人脸恢复算法，设计时空-通道联合注意力与轻量 CFA-Net，配合自适应时序损失，在增强纹理与判别特征的同时确保恢复稳定性；

2）提出融合 SCConv 与 MS-DAM 的步态特征提取算法，抑制卷积冗余并捕获多尺度动态，提升对复杂行走条件的适应能力；

3）提出时序上下文感知的 CoT 融合身份识别算法，以局部窗口轻量门控自适应融合人脸与步态异构信息，实现高效互补与判别。

1 本文方法

1.1 总体框架

本文构建跨模态 CoT 特征融合的身份识别总体框架，如图 1 所示。主要包括如下步骤：1）基于 DSTC 的低质量视频人脸恢复：设计空间、时间与通道三维联合注意力机制、构建轻量化上下文特征对齐网络 CFA-Net、定义自适应时序损失函数。2）融合 MS-DAM 的步态特征提取：引入 SCConv 模块和设

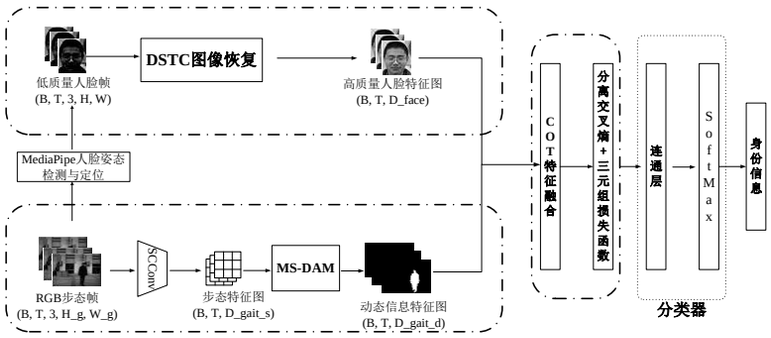


图1　融合时序上下文的跨模态身份识别算法框架图

Fig 1　Framework of a Cross-Modal Person Re-Identification Algorithm Integrating Temporal Context

计多尺度动态增强模块MS-DAM。3)跨模态CoT特征融合:构建时序上下文感知的CoT融合机制,自适应调整不同模态的融合权重。

1.2　基于DSTC的低质量视频人脸恢复

低质量人脸图像因模糊、光照等退化及帧间错位,常导致语义特征不一致,制约跨模态融合。现有单帧恢复方法(如DeblurGAN、MS PL等)缺乏时序建模能力,难以利用相邻帧的互补信息;基于语义分割先验的两阶段方法(如Shen等、Yasarla等)则引入了额外标注依赖,且易受分割误差传播的影响。相比之下,本文方法直接从退化帧序列出发,通过STC模块的空间-时间-通道联合注意力实现跨帧上下文建模,并采用线性复杂度的CFA Net替代传统自注意力运算,在不依赖额外先验的条件下实现高效、高质量的人脸恢复。方法结构如图2所示。

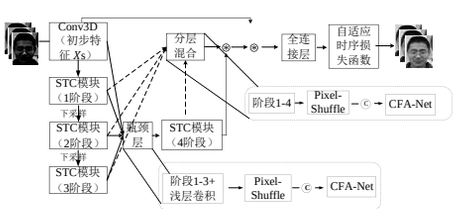


图2　基于DSTC低质量视频人脸恢复算法框架图

Fig 2　Framework of a DSTC-Based Low-Quality Video Face Restoration Algorithm

1.2.1　STC模块

针对视频序列中帧间错位、运动模糊及遮挡导致的局部信息缺失与跨模态语义不一致问题,本文设计了STC(Spatial - Temporal - Channel)特征补偿模块。该模块对空间、时间与通道三个维度的上下文进行联合建模,充分利用前后帧互补信息推断并补全当前帧缺失细节,有效缓解时序错位与特征不完整,从而提升低质量人脸的恢复精度与语义一致性。其架构如图3所示。

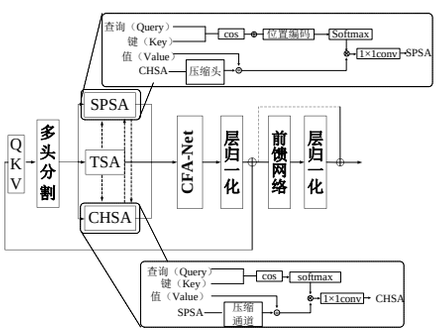


图3　STC框架图

Fig 3　STC Framework

1)　空间自注意力

空间自注意力(Spatial Self-Attention, SPSA)用于捕获捕获图像空间中的局部纹理与结构信息,采用多头机制,将通道划分为L_{sp}个注意力头:

$$SPSA = P_{sp}(Concat[head_{sp}^{1}, L, head_{sp}^{L_{sp}}]) \quad (1)$$

式中单个head定义为:

$$head_{sp}^{1} = Softmax\left(\frac{\cos(Q_{sp}^{1}, K_{sp}^{1})}{\tau} + B_{sp}\right)V_{sp}^{1} \quad (2)$$

式中τ为可学习温度参数,B_{sp}为空间位置偏置,P_{sp}表示拼接后的线性映射。

2)　时间自注意力

时间自注意力(Temporal Self-Attention, TSA)用于对齐并补偿当前帧缺失信息。设空间位置(h, w)处的时间序列特征为$X^{(h,w)} = [x_{t-k}, \cdots, x_t, \cdots, x_{t+k}]$,时间自注意力计算为:

$$TSA = P_t\left(Concat[head_t^1, \cdots, head_t^{L_t}]\right) \quad (3)$$

单个时间head定义为:

$$head_i^1 = Softmax\left(\frac{Q_i^1 (K_i^1)^T}{\sqrt{d}} + B_t\right)V_t^1 \quad (4)$$

式中 B_t 为时间相对位置编码，d 为每个 head 的维度。

3） 通道自注意力

通道自注意力（Channel Self-Attention，CHSA）用于提取高层语义特征并补全纹理信息，其形式为：

$$CHSA = P_{ch}\left(Concat\left[head_{ch}^1,\cdots,head_{ch}^{L_{ch}}\right]\right) \quad (5)$$

单个通道 head 为：

$$head_{ch}^1 = Softmax\left(\frac{\cos(Q_{ch}^1,K_{ch}^1)}{\tau} + B_{ch}\right)V_{ch}^1 \quad (6)$$

在通道维度计算余弦相似度，B_{ch} 为通道位置偏置。

1.2.2 CFA-Net

传统卷积难以捕获长程全局依赖，自注意力在高分辨率特征图上的计算开销较大。为此，本文设计轻量级上下文特征对齐模块 CFA-Net，由上下文生成模块（CGM）和特征调制单元（FMU）组成（如图 4 所示）。CGM 通过线性映射与空间归一化构建紧凑的全局上下文表示，降低计算复杂度；FMU 采用元素级特征调制实现语义引导重构，增强特征表达能力。两模块协同实现高效的全局语义对齐与细节增强。

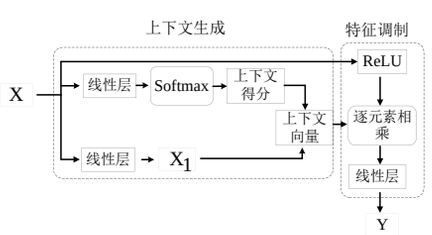


图 4 CFA-Net 框架图

Fig 4 Framework of CFA-Net

1.2.3 自适应时序损失函数 L_{ATL}

低质量帧容易引入较大的梯度噪声，导致决策边界偏移、类内特征离散度增加以及整体分布不稳定，本文定义自适应时序损失函数 L_{ATL}（Learnable Adaptive Temporal Loss），计算公式表示为：

$$L_{ATL} = -\frac{1}{T}\sum_i^T logP_i \quad (7)$$

$$P_i = \frac{exp[s(cos(\theta_i + m_i))]}{exp[s(cos(\theta_i + m_i))] + \sum_{j\neq i} exp[s(cos(\theta_j))]} \quad (8)$$

式中 T 为视频帧数，s 为尺度因子，θ_{y_t} 当前帧特征与真实类别中心的夹角，θ_j 表示与其他类别权重的夹角，m_i 为自适应角度间隔。m_i 由帧质量分数 $q_i \in (0,1]$ 经映射函数 $m_i=\alpha\cdot(1-qi)+\beta$ 计算得到，其中 q_i 由基于浅层 MLP 和 Sigmoid 激活的轻量化质量评估网络在线预测，α 和 β 分别控制质量敏感性和基础间隔。L_{ATL} 根据帧质量自适应调整监督强度，强化高质量及关键帧的贡献，抑制低质量帧干扰，从而提升视频特征的时序一致性和判别能力。

1.3 融合 SCConv 与 MS-DAM 的步态特征提取

步态特征易受背景干扰、外观变化及动态噪声影响，导致特征表达不稳定。传统三维卷积存在空间冗余，DAM 模块仅依赖单一时间间隔的帧差，难以有效建模不同步速下的时序变化。为此，本文提出融合 SCConv 与 MS-DAM 的步态特征提取算法：SCConv 通过空间与通道重构抑制特征冗余，提高特征提取效率；MS-DAM 引入多尺度差分单元，联合建模短时动作细节与长时运动信息，并采用可学习加权融合生成动态特征表示。算法结构如图 5 所示。

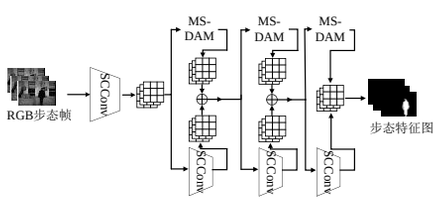


图 5 融合 SCConv 与 MS-DAM 的步态特征提取框架图

Fig 5 Framework of a Gait Feature Extraction Method Integrating SCConv and MS-DAM

1.3.1 SCConv

传统三维卷积在步态特征提取中易产生空间冗余，影响特征表达效率。为此，本文采用 SCConv 替代传统 Conv3D 作为步态分支的核心卷积模块，结构如图 6 所示。SCConv 通过自校准机制对特征进行分离与重构，在抑制冗余的同时增强特征表达能力。该模块由空间重构单元（SRU）和通道重构单元（CRU）组成：SRU 突出关键姿态区域、抑制冗余响应；CRU 在低维空间融合上下文信息，并映射回原始特征域，生成更具判别力的步态表示。

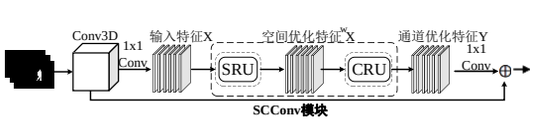


图 6 SCConv 结构图

Fig 6 Structure of SCConv

1.3.2 MS-DAM

传统方法难以有效捕获帧间时序动态，而此类动态差异恰是区分个体步态模式的关键判别特征。传统DAM模块虽通过相邻帧差分突出运动区域，但仅依赖单一时间间隔，难以兼顾不同速度与频率的运动变化，限制了复杂时序建模能力。鉴于肢体部分相比躯干运动幅度更大，其动态变化反映个体独特的运动节奏与行为模式，构成身份识别最具代表性的判别来源。为此，本文设计多尺度动态增强模块MS-DAM（如图7所示），在动态特征提取分支中引入多尺度差分单元（MDU），通过多时间间隔的帧间差分操作（如 $\triangle_1 = f_t - f_{t-1}$、$\triangle_2 = f_t - f_{t-2}$、$\triangle_3 = f_t - f_{t-3}$），捕获不同时间尺度下的运动变化。随后，采用可学习加权融合策略整合各尺度差分，得到综合动态特征表示 F_{MS}，在保留细粒度动作细节的同时强化长时间依赖的运动趋势。

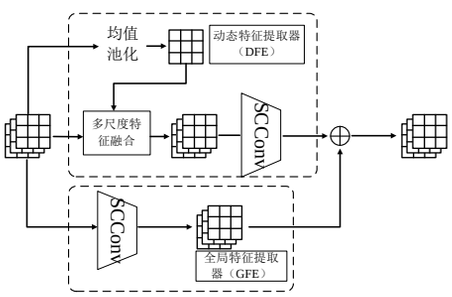


图7 MS-DAM结构图

Fig 7 Structure of MS-DAM

给定输入步态序列 $X_o = \{f_1, f_2, \cdots, f_T\}$，计算不同时间间隔下的差分特征图 $X_d^{\Delta 1}$, $X_d^{\Delta 2}$, $X_d^{\Delta 3}$。对不同尺度的差分特征图进行加权融合，得到多尺度特征 F_{MS}，计算公式如下：

$$X_d^{\Delta 1} = (f_t - f_{t_1}), X_d^{\Delta 2} = (f_t - f_{t_2}), X_d^{\Delta 3} = (f_t - f_{t_3}) \quad (9)$$

$$F_{MS} = w_1 X_d^{\Delta 1} + w_2 X_d^{\Delta 2} + w_3 X_d^{\Delta 3} \quad (10)$$

式中 w_1, w_2, w_3 是可学习权重。利用动态特征提取器（Dynamic Feature Extractor, DFE），建立空间-时间特征表示，并利用全局特征提取器（Global Feature Extractor, GFE），提取步态序列的全局特征。MS-DAM融合DFE和GFE特征生成增强的动态特征表示 Y_{MS-DAM}，公式如下：

$$Y_{DFE} = C^{3\times 3\times 3}(F_{MS}), Y_{GFE} = C^{1\times 3\times 3}(X_o) \quad (11)$$

$$Y_{MS-DAM} = \sigma(Y_{GFE} + Y_{DFE}) \quad (12)$$

式中 σ 表示Leaky ReLU激活函数。

1.4 跨模态CoT特征融合

针对人脸与步态异构信息融合不足的问题，现有融合方法难以兼顾自适应性与计算效率。为此，本文提出跨模态CoT特征融合算法，如图8所示。该方法首先分别构建人脸与步态特征的时序上下文表示，再通过轻量门控机制自适应分配模态权重，实现人脸与步态特征的高效互补。

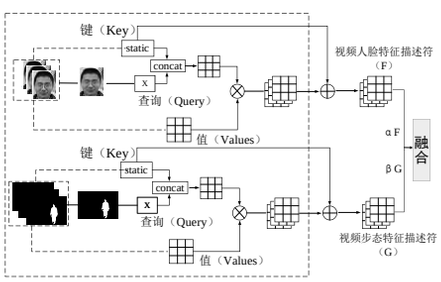


图8 跨模态CoT特征融合的身份识别算法框架图

Fig 8 Framework of a Cross-Modal Identity Recognition Algorithm Based on CoT Feature Fusion

假设视频序列长度为T，人脸与步态两种模态的帧级特征分别表示为：

$$F = (f_i)_{i=1}^T, G = (g_i)_{i=1}^T \quad (13)$$

为捕获时间邻域内的动态演化信息，在时间窗口 $\Omega(t)$ 内对特征进行加权聚合，构建时序增强表示：

$$H_t = \sum_{z \in \Omega(t)} \alpha_{\tau,z} x_\tau \quad (14)$$

式中 x_τ 表示对应模态的帧级特征（人脸或步态），$\alpha_{\tau,x}$ 为在时间维度上学习得到的权重系数，满足归一化约束。引入门控融合机制，根据场景变化动态调整模态权重，计算公式如下：

$$\lambda_t = \sigma(MLP([\tilde{F}_t, \tilde{G}_t])) \quad (15)$$

式中 σ 为Sigmoid函数，$\lambda_t \in (0,1)$ 表示第t帧中人脸模态的可信度权重。最终融合特征表示为：

$$z_t = \lambda_t F_t + (1 - \lambda_t) G_t \quad (16)$$

通过这种动态调节机制，CoT模块能充分利用人脸与步态的互补特性，减弱瞬时遮挡、光照变化等干扰因素，实现自适应的跨模态身份识别。

2 实 验

2.1 数据集

LFW（Labeled Faces in the Wild）数据集包含5749名个体的13233张自然场景人脸图像，涵盖姿

态、光照、表情和分辨率等复杂变化。本文从中随机选取约1000张图像，并采用与训练阶段一致的退化策略生成测试样本，用于评估低质量人脸恢复性能。

CASIA-B数据集包含124名受试者，前74人为训练集，后50人为测试集。每名受试者包含普通行走（NM）、背包行走（BG）和换装行走（CL）三种条件，步态剪影覆盖11个视角（0°－180°）。按照常用评测协议，分别选取NM#5－6、BG#1－2和CL#1－2作为测试序列。自制数据集为本文构建的面向无感考勤场景的多模态数据集，共采集124名受试者，训练集与测试集划分方式与CASIA-B保持一致。数据集包含人脸图像和步态视频，覆盖正常行走、背包行走和穿大衣三种条件，以及6个水平视角（0°－90°，间隔18°）。步态剪影由Mask R-CNN提取，人脸图像经MTCNN检测与对齐，并通过硬件时间戳实现人脸与步态样本的帧级同步，所有数据均完成身份、视角和序列标注。

2.2 评价指标

PSNR用于衡量恢复图像与真实图像之间的重建质量，其值越大，表示恢复图像与真实图像越接近。

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{17}$$

SSIM综合考虑图像的亮度、对比度和结构信息，用于评价恢复图像与真实图像的结构相似性，其取值范围为0到1，数值越大表示恢复效果越好。

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{18}$$

LPIPS（Learned Perceptual Image Patch Similarity）是一种基于深度学习特征的感知质量评价指标，用于衡量恢复图像与真实图像在视觉感知层面的差异，数值越小表示恢复图像越接近真实图像。

$$LIPS(x,y) = \sum_1 \frac{1}{H_1 W_1} \sum_{h,w} \left\| w_1 \odot \left(\hat{f}_1(x)_{hw} - \hat{f}_1(y)_{hw} \right) \right\|_2^2 \tag{19}$$

为衡量步态模型的识别能力，本文选用Rank-1准确率作为评价指标。该指标数值越高，说明模型的识别效果越好。

2.3 实验设置

优化器采用Adam，初始学习率1×10^{-4}，每30个epoch衰减为原来的0.5倍，训练150个epoch。人脸恢复批量大小为16，步态识别与跨模态融合为32。输入方面，人脸恢复模块将视频帧对齐至112×112像素后输入DSTC网络；步态识别模块输入为64×44像素步态剪影图，序列长度30帧。所有模型均采用随机水平翻转进行数据增强。

2.4 对比试验

2.4.1 人脸恢复对比实验

为了验证本文人脸恢复算法的有效性，本文将所提出方法与Kupyn（2019）、Lee（2020）、Lee（2020）以及Tao（2024）等典型恢复方法进行定量对比。评价指标采用PSNR、SSIM与LPIPS，实验结果如表2所示。

表1 LFW和自制人脸数据集上不同方法对比结果

Table 1 Comparison of Different Methods on the LFW and Self-Collected Face Datasets

数据集	LFW数据集			自制人脸数据集		
方法	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Kupyn	21.84	0.673	0.2716	20.67	0.643	0.2954
Yasarla	25.83	0.761	0.2742	24.66	0.737	0.2836
Lee	26.29	0.786	0.2232	25.31	0.756	0.2387
Tao	<u>26.58</u>	<u>0.801</u>	<u>0.2112</u>	<u>24.98</u>	<u>0.761</u>	<u>0.2334</u>
Ours	**27.43**	**0.813**	**0.1894**	**26.22**	**0.801**	**0.1968**

注：表中加粗表示最优结果，下划线表示次优结果。

如表1所示，如表1所示，本文提出的DSTC模型在LFW和自建数据集上的图像恢复性能均优于对比方法。在LFW数据集上，相比Tao方法，PSNR提升0.85 dB、SSIM提升0.012、LPIPS降低0.0218；相比Lee方法，PSNR提升1.14 dB、SSIM提升0.027。在自建数据集上，DSTC同样取得最佳结果

(PSNR 26.22、SSIM 0.801、LPIPS 0.1968)。性能提升主要得益于STC模块对空间、时间和通道信息的联合建模,有效增强了跨帧细节补偿能力。典型人脸恢复结果如图9、图10所示。

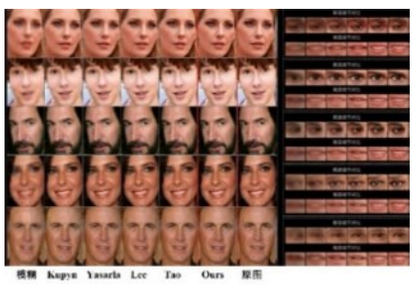


图9 LFW数据集中不同算法的人脸恢复对比效果

Fig 9 Comparison of Face Restoration Results of Different Algorithms on the LFW Dataset

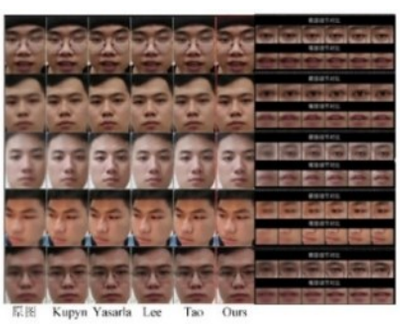


图10 自制人脸数据集上不同方法恢复对比效果图

Fig 10 Comparison of different face recovery methods on a self-made face dataset

如表2所示,DSTC人脸恢复算法在各视角下均取得较优识别性能。0°正面视角准确率达到94.37%,较Tao方法提高2.25个百分点;90°侧脸视角仍达到67.00%,领先1.35个百分点。随着视角增大,各方法准确率均有所下降,但DSTC下降幅度最小,在0°–36°范围内准确率均超过89%,54°–72°大视角下优势更加明显。最终平均准确率达到83.65%,较Tao和Lee方法分别提高2.01和3.01个百分点。

表2 CASIA-B数据集上人脸识别准确率性能对比

Table 2 Face recognition accuracy performance comparison on CASIA-B dataset

方法	0°	18°	36°	54°	72°	90°
Blur	72.5	71.3	69.2	63.2	55.1	44.2
Kupyn	82.2	81.2	79.1	73.3	66.1	56.8
Yasarla	88.4	87.0	82.2	74.2	65.5	51.8
Lee	90.4	90.2	86.4	81.9	72.8	62.2
Tao	92.1	90.9	86.5	81.2	73.5	65.7
Ours	**94.4**	**93.5**	**89.6**	**83.2**	**74.3**	**67.0**

注:表中加粗表示最优结果,下划线表示次优结果。

2.4.2 步态识别对比实验

为验证所提方法在步态识别任务中的有效性,本文在CASIA-B数据集上与GaitSet(2019)、GaitPart(2020)、MT3D(2020)、GaitGL(2024)、GPGait(Fu等,2023)及GaitHeat(Fu等,2024)进行了对比实验。实验采用Rank-1准确率(%)作为评价指标,统计不同视角下的识别性能,结果如表3所示。

表3 CASIA-B数据集上不同方法步态识别Rank-1准确率对比

Table 3 Comparison of Rank-1 Accuracy of Gait Recognition Methods on the CASIA-B Dataset

Method		Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5-6	GaitSet	90.8	97.9	**99.4**	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart	94.1	**98.6**	99.3	**98.5**	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	MT3D	95.7	98.2	99.0	97.5	95.1	93.9	96.1	98.6	99.2	98.2	92.0	96.7
	GaitGL	96.0	98.3	99.0	97.9	**96.9**	95.4	97.0	98.9	99.3	**98.8**	94.0	97.4
	GPGait	95.2	98.4	99.1	98.3	96.8	95.1	97.0	98.7	99.2	98.0	93.1	97.2
	Ours	**96.4**	98.1	99.2	98.0	96.7	**96.2**	**97.3**	**99.2**	**99.4**	98.6	**94.8**	**97.6**

注:**续表3 CASIA-B数据集上不同方法步态识别Rank-1准确率对比**

从表3可以看出,本文方法在三种测试条件下均取得了较优的识别性能。NM条件下平均识别率达到97.6%,优于GaitSet(95.0%)、GaitPart(96.2%)、MT3D(96.7%)、GPGait(97.2%)和GaitGL

BG#1-2	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	MT3D	91.0	95.4	**97.5**	94.2	92.3	86.9	91.2	95.6	97.3	96.4	86.6	93.0
	GaitGL	92.6	**96.6**	96.8	**95.5**	93.5	**89.3**	92.2	**96.5**	98.2	96.9	**91.5**	94.5
	GPGait	90.8	95.2	96.4	95.0	92.5	87.6	91.5	95.3	97.0	95.8	88.0	92.3
	Ours	**93.2**	96.4	97.2	95.4	**93.9**	89.1	**92.6**	96.3	**98.5**	**97.1**	91.3	**94.6**
CL#1-2	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	MT3D	76.0	87.6	89.8	85.0	81.2	75.7	81.0	84.5	85.4	82.2	68.1	81.5
	GaitGL	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GPGait	73.5	87.2	89.0	86.1	82.8	77.5	82.0	85.9	86.8	83.7	68.4	82.1
	Ours	**78.4**	**90.3**	**91.5**	**88.6**	**85.0**	**81.0**	**84.6**	**88.3**	**88.9**	**85.6**	**71.2**	**84.9**

注：表中加粗表示最优结果，下划线表示次优结果。

(97.4%)。BG 条件下，各方法受背包影响性能均有所下降，本文方法仍取得 94.6% 的平均识别率，较 GPGait 提升 2.3 个百分点，并略优于 GaitGL。CL 条件下，受服装变化影响，本文方法平均识别率达到 84.9%，分别较 GaitGL 和 GPGait 提高 1.3 和 2.8 个百分点，较 GaitSet、GaitPart 和 MT3D 也具有明显优势。相比 GaitHeat，本文方法在 NM 条件下性能接近，在 BG 和 CL 条件下分别提高 2.46 和 10.75 个百分点，同时避免了热图表示带来的额外计算与存储开销，更适合轻量化部署。上述结果表明，融合 SCConv 与 MS-DAM 的步态特征提取框架在多种行走条件下保持高精度与鲁棒性。注意力能量图可视化结果如图 11 所示。

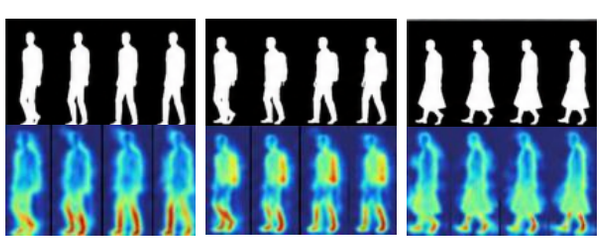


图 11　MS-DAM 动态特征增强后的注意力能量图

Fig 11　Attention Energy Map after Dynamic Feature Enhancement by MS-DAM

从图 11 可见，经 MS-DAM 处理后，模型注意力集中于腿部摆动与躯干轮廓等关键运动区域，背景响应显著减弱。注意力分布随步态序列动态跟随人体运动轨迹，保持良好时序一致性，能够有效捕捉判别性动态特征、增强关键结构表达，从而提升复杂场景下的识别性能。

2.5　跨模态融合身份识别方法对比

为验证跨模态融合方法的有效性，本文在 CASIA-B 数据集上开展了单模态与多种融合策略的对比实验，结果如表 4 所示。

从表 4 可以看出，本文提出的 CoT 融合方法在三种行走条件下的大多数视角均取得了最佳识别性能。相比简单融合方法（Geng 等，2008）和注意力融合方法（Prakash 等，2023），CoT 融合能够更有效地利用人脸与步态的互补信息。NM 条件下，CoT 融合在 0° 视角达到 98.8%，侧面视角仍保持较高准确率；BG 条件下，各方法性能均有所下降，但 CoT 融合在多数视角仍保持领先；CL 条件下，步态识别受服装变化影响明显，而 CoT 融合在 0° 和 180° 视角分别达到 94.5% 和 69.5%。结果表明，所提方法在复杂场景下具有更好的跨模态身份识别性能。

2.6　消融实验

以不包含 STC、CFA-Net 和 LATL 的基础恢复网络作为基线，逐步引入各模块进行消融实验，LFW 数据集上的实验结果如表 5 所示。

2.6.1　人脸恢复模块消融实验

人脸恢复分支含 STC、CFA-Net 与 LATL 三个模块。以去除上述模块的基础恢复网络为基线，逐步添加各组件进行实验，在 LFW 数据集上的结果如表 5 所示。

由表 5 可见，加入 STC 后，PSNR 由 24.61 dB 提

表4　CASIA-B跨模态融合方法rank-1准确率结果

Table 4　Rank-1 Accuracy Results of Cross-Modal Fusion Methods on the CASIA-B Dataset

Method		Probe View											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5-6	人脸识别	94.4	93.5	89.6	83.2	74.3	67.0	/	/	/	/	/	83.7
	步态识别	96.4	98.1	99.2	98.0	96.7	96.2	**97.3**	**99.2**	**99.4**	**98.6**	**94.8**	97.6
	简单融合	97.2	97.5	96.8	93.5	92.0	90.2	86.5	88.4	87.6	86.3	85.2	91.0
	注意力机制融合	98.5	98.7	98.4	96.6	95.7	96.1	94.4	95.7	95.2	95.1	93.2	96.2
	CoT融合	**98.8**	**99.2**	**99.4**	**98.5**	**97.6**	**96.8**	96.2	97.8	97.8	97.3	92.4	**97.8**
BG#1-2	人脸识别	94.4	93.5	89.6	83.2	74.3	67.0	/	/	/	/	/	83.7
	步态识别	93.2	96.4	97.2	95.4	93.9	89.1	**92.6**	**96.3**	**98.5**	**97.1**	**91.3**	94.6
	简单融合	95.1	96.2	97.1	91.5	89.5	86.0	81.4	83.4	86.7	85.2	81.2	88.5
	注意力机制融合	96.8	97.3	96.8	94.7	93.6	88.7	91.2	93.8	96.5	96.1	90.2	94.2
	CoT融合	**97.4**	**98.6**	**98.9**	**97.4**	**94.5**	**92.2**	91.5	94.7	97.3	96.4	90.6	**95.4**

注：**续表4 CASIA-B跨模态融合方法rank-1准确率结果**

CL#1-2	人脸识别	94.4	93.5	89.6	83.2	74.3	67.0	/	/	/	/	/	83.6
	步态识别	78.4	90.3	91.5	88.6	85.0	81.0	**84.6**	**88.3**	**88.9**	**85.6**	**71.2**	84.8
	简单融合	86.5	93.0	**92.4**	87.7	82.5	77.5	76.4	77.9	77.8	75.3	64.2	81.0
	注意力机制融合	91.2	94.5	91.7	85.3	82.7	79.2	82.3	84.7	85.3	82.1	67.6	84.2
	CoT融合	**94.5**	**95.7**	**92.4**	**91.0**	**87.5**	**84.2**	83.4	86.9	87.3	84.4	69.5	**86.9**

注：表中加粗表示最优结果，下划线表示次优结果。

表5　人脸恢复模块消融实验

Table 5　Face restoration module ablation experiment

STC	CFA-Net	LATL	PSNR ↑	SSIM ↑	LPIPS ↓
			24.61	0.693	0.2742
✓			26.18	0.751	0.2367
	✓		24.95	0.709	0.2661
		✓	25.73	0.736	0.2489
✓	✓		26.34	0.758	0.2286
✓		✓	27.12	0.794	0.2065
✓	✓	✓	**27.43**	**0.813**	**0.1894**

注：表中加粗表示最优结果，下划线表示次优结果。

升至26.18 dB，SSIM由0.693提升至0.751，LPIPS由0.2742降至0.2367，表明STC能够有效建模跨帧时空信息。单独引入CFA-Net性能提升较小（PSNR提升0.34 dB），但与STC结合后PSNR进一步提升至26.34 dB，说明轻量化全局上下文建模具有良好的协同作用。LATL单独使用时PSNR提升1.12 dB，与STC结合后PSNR和SSIM分别达到27.12 dB和0.794，验证了时序建模与自适应损失的互补性。三者联合使用时取得最佳性能，PSNR、SSIM和LPIPS分别达到27.43 dB、0.813和0.1894，较基线分别提升2.82 dB、0.120和降低0.0848，说明各模块协同作用有效提升了人脸恢复性能。

2.6.2　步态识别模块消融实验

步态识别分支包括SCConv与MS-DAM。以纯Conv3D网络为基线，逐步替换/添加各模块，在CASIA-B三种行走条件的11个视角上进行消融实验，结果如表6所示。

由表6可见，SCConv替换Conv3D后，NM、BG和CL条件下的平均准确率分别提升2.1%、5.1%和6.1%，说明SCConv能够有效抑制空间冗余，增强步态特征表达能力。单独引入MS-DAM后，NM、BG和CL条件下分别提升2.5%、5.8%和9.7%，表明多尺度动态建模能够有效提升复杂行走条件下的特征判

表 6　步态识别模块消融实验

Table 6　Gait recognition module ablation experiment

Conv3D	SCConv	MS-DAM	NM	BG	CL	Mean
✓			94.2	87.2	70.7	84.0
	✓		96.3	<u>92.3</u>	76.8	88.5
✓		✓	<u>96.7</u>	93.0	<u>80.4</u>	<u>90.0</u>
	✓	✓	**97.6**	**94.6**	**84.9**	**92.4**

注：表中加粗表示最优结果，下划线表示次优结果。

别能力。两者联合使用时，NM、BG 和 CL 条件下的准确率分别达到 97.6%、94.6% 和 84.9%，取得最佳性能，说明 SCConv 与 MS-DAM 具有良好的协同作用，可进一步提升步态识别性能。

2.6.3　CoT 跨模态融合消融实验

为直观展示跨模态融合方法的有效性，从 CASIA-B 跨模态数据集中随机选取 10 个身份样本，将其高维特征映射到二维空间进行展示。采用 t-SNE（t-distributed Stochastic Neighbor Embedding）方法对特征分布进行可视化分析，结果如图 12 所示。

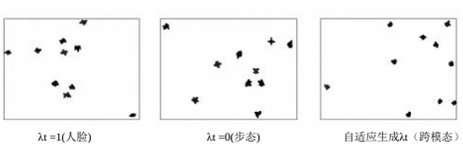


图 12　t-SNE 可视化结果

Fig 12　t-SNE visualization results

图 12 表明，单模态方法在特征分布上存在一定程度的类间混叠现象，而 CoT 跨模态融合方法能够使同一身份的特征分布更加紧凑，显著增大不同身份之间的分布间隔。

2.7　计算复杂度分析

为全面评估本文方法的部署可行性，本节从参数量、FLOPs 与推理时间三个维度对各模块进行复杂度分析。

从表 7 可知，本文方法参数量仅 12.7 M，较 Lee 方法降低约 31%；FLOPs 为 137 G，仅为 Kupyn 方法的约 1/3；推理时间为 77 ms，显著快于 Yasarla 和 Kupyn 方法，体现出较好的轻量化优势。性能提升主要得益于 CFA-Net 采用线性上下文建模降低全局特征提取开销，SCConv 有效抑制空间冗余，MS-DAM 以轻量化多尺度动态建模增强时序特征表达，CoT 融合通过局部时序加权实现高效跨模态特征融合，在保证识别性能的同时进一步降低了计算复杂度。

表 7　各方法计算复杂度对比

Table 7　Comparison of computational complexity of various methods

方法	参数量 (M)	FLOPs (G)	推理时间 (ms)
Kupyn	60.0	411	850
Yasarla	<u>14.4</u>	—	160
Lee (MSPL)	18.5	—	<u>80</u>
Ours	**12.7**	**137**	**77**

注：表中加粗表示最优结果，下划线表示次优结果。

3　结　论

针对复杂环境下单一模态身份识别精度低、适应性不足的问题，本文提出融合时序上下文信息的跨模态身份识别算法，构建了低质量人脸恢复、步态特征提取和跨模态融合的统一框架。实验结果表明，DSTC 算法有效提升了低质量视频人脸恢复效果，融合 MS-DAM 的步态识别算法增强了复杂行走条件下的特征表达能力，CoT 融合算法充分利用人脸与步态信息互补性，在多个公开数据集上均取得优于现有方法的识别性能。未来将进一步构建航空场景多模态数据集，研究面向多目标、遮挡及跨视角场景的轻量化跨模态身份识别方法，提升算法在实际航空场景中的应用能力。

参考文献

Chen C, Li X, Luo Z and Loy C C. 2018. FSRNet: End-to-end learning face super-resolution with facial priors//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR′18). Salt Lake City: IEEE: 2492-2501 [DOI: 10.1109/CVPR.2018.00264]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. 2014. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems (NeurIPS′14). Montreal: 2672-2680 [DOI: 10.48550/arXiv. 1406.2661]

Li X, Liu M Y, Kautz J and Yang M H. 2018. Learning warping for GFRNet: Guided face restoration//Proceedings of the European Conference on Computer Vision (ECCV′18). Munich: Springer:

272-289 [DOI: 10.1007/978-3-030-01237-3_17]

冯建江,贾伟,李琦,崔哲,赵才荣,雷震,等.生物特征识别学科发展报告(2021—2025)[J].中国图象图形学报,2026,31(6):1689-1740). DOI:10.11834/jig.260069.

Peng Z J. Review of face recognition technology[J]. China Computer & Communication (Theory Edition), 2023, 35(15): 168-171.

Liu H, Zheng X, Han J, Feng H and Chen L. 2019. Survey on GAN-based face hallucination with its model development. IET Image Processing, 13 (14): 2662-2672 [DOI: 10.1049/iet-ipr. 2018. 6644]

唐鑫,张飞飞,李光辉,马宇轩,董正阳.面向轻量级人脸识别的二阶段知识蒸馏方法[J].中国图象图形学报,2025,1-14). DOI:10.11834/jig.250454.

Wang Z W, Lu Z Y. Cross-modal visual question answering based on fine-grained feature enhancement[J]. Journal of Nanjing University of Information Science & Technology, 2026, 18(1): 35-47.

Menon S, Damian A, Hu S, Ravi N, Rudin C and Chellappa R. 2020. PULSE: self-supervised photo upsampling via latent space exploration of generative models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20). Seattle: IEEE: 2437-2445 [DOI: 10.1109/CVPR42600.2020.00251]

Jain A K, Ross A and Pankanti S. 2006. Biometrics: a tool for information security. IEEE Transactions on Information Forensics and Security, 1(2): 125-143 [DOI: 10.1109/TIFS.2006.873653]

Song Y F, Zhang Z, Shan C and Wang L. 2020. Stronger, faster and more explainable: a graph convolutional baseline for skeleton-based action recognition//Proceedings of the 28th ACM International Conference on Multimedia (MM'20). Seattle: ACM: 1625-1633 [DOI: 10.1145/3394171.3413802]

Wang X, Li Y, Zhang H, Yu J, Liu Y, Wang Y, et al. 2021. Towards real-world blind face restoration with generative facial prior//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21). Nashville: IEEE: 9168-9178 [DOI: 10.1109/CVPR46437.2021.00905]

Xu P, Zhu X and Clifton D A. 2023. Multimodal learning with transformers: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10): 12113-12132 [DOI: 10.1109/TPAMI.2023.3246806]

Yan S, Xiong Y and Lin D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18). New Orleans: AAAI: 7444-7452 [DOI: 10.1609/aaai.v32i1.12328]

Hori C, Hori T, Lee T Y, Zhang Z, Hershey J R and Sainath T N. 2017. Attention-based multimodal fusion for video description//Proceedings of the IEEE International Conference on Computer Vision (ICCV'17). Venice: IEEE: 4193-4202 [DOI: 10.1109/ICCV.2017.448]

Kupyn O, Martyniuk T, Wu J, Wang Z and Shi H. 2019. DeblurGAN-v2: deblurring (orders-of-magnitude) faster and better//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'19). Seoul: IEEE: 8878-8887 [DOI: 10.1109/ICCV.2019.00897]

Lee T B, Jung S H and Heo Y S. 2020. Progressive semantic face deblurring. IEEE Access, 8: 223548-223561 [DOI: 10.1109/ACCESS.2020.3033890]

Yang T, Ren P, Xie X, Zhang L, Wang X and Cao X. 2021. GAN prior embedded network for blind face restoration in the wild//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21). Nashville: IEEE: 672-681 [DOI: 10.1109/CVPR46437.2021.00073]

Yasarla R, Perazzi F and Patel V M. 2020. Deblurring face images using uncertainty guided multi-stream semantic networks. IEEE Transactions on Image Processing, 29: 6251-6263 [DOI: 10.1109/TIP.2020.2989230]

Chao H, He Y, Zhang J, Wang J and Feng J. 2019. GaitSet: regarding gait as a set for cross-view gait recognition//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'19). Honolulu: AAAI: 8126-8133 [DOI: 10.1609/aaai.v33i01.33018126]

Fan C, Peng Y, Cao C, Liu X, Hou S and Chi J. 2020. GaitPart: temporal part-based model for gait recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20). Seattle: IEEE: 14225-14233 [DOI: 10.1109/CVPR42600.2020.01423]

Fu Y, Hou S, Meng S, Hu X, Cao C, Liu X, et al. 2024. Cut out the middleman: revisiting pose-based gait recognition//Computer Vision – ECCV 2024. Cham: Springer: 112-128 [DOI: 10.1007/978-3-031-72751-1_7]

Geng X, Wang L, Li M and Chen Y. 2008. Adaptive fusion of gait and face for human identification in video//Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV'08). Copper Mountain: IEEE: 1-6 [DOI: 10.1109/WACV.2008.4544006]

Hao Y, Hou T, Su Y C, Chan K C K, Wang X and Loy C C. 2023. Towards authentic face restoration with iterative diffusion models and beyond//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23). Paris: IEEE: 4390-4399 [DOI: 10.1109/ICCV51070.2023.00406]

Lin B, Zhang S and Bao F. 2020. Gait recognition with multiple-temporal-scale 3D convolutional neural network//Proceedings of the 28th ACM International Conference on Multimedia (MM'20). Seattle: ACM: 3054-3062 [DOI: 10.1145/3394171.3413861]

Lin B, Zhang S, Wang M and Yu Y. 2024. GaitGL: learning discriminative global-local feature representations for gait recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46 (7): 4213-4229 [DOI: 10.1109/TPAMI.2023.3302352]

Prakash A, Nambiar A and Bernardino A. 2023. Multimodal adaptive fusion of face and gait features using keyless attention based deep neural networks for human identification [EB/OL]. 2023-03-24.

https://arxiv. org/abs/2303.13814 ［DOI： 10.48550/arXiv. 2303. 13814］

Fu Y, Meng S, Hou S, Zhang X, Li Y and Wang L. 2023. GPGait: generalized pose-based gait recognition//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23). Paris: IEEE: 19538-19547［DOI: 10.1109/ICCV51070.2023.01795］

Yang S, Wang J, Hou S, Zhang X, Li Y and Yu Y. 2025. Bridging gait recognition and large language models for sequence modeling//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'25). Nashville: IEEE: 3460-3469［DOI: 10.1109/CVPR52733.2025.00338］

Wang Z, Chen X, Xu C, Zhu J, Hu X, Zhang J, et al. 2025. SVFR: a unified framework for generalized video face restoration//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'25). Nashville: IEEE: 7406-7415［DOI: 10.1109/CVPR52734.2025.00694］

作者简介

陈翔，男，硕士研究生，主要研究方向为人工智能。E-mail：240411048@stu.xpu.edu.cn

郭晨曦，男，硕士研究生，主要研究方向为生物识别。E-mail：2394980093@qq.com

李云红，女，教授，主要研究方向图像处理、信号与信息处理，E-mail：hitliyunhong@163.com

王力斌，男，硕士研究生，主要研究方向为生物识别。E-mail：1349717408@qq.com