

循环迭代优化在面部关键点提取中的应用

俞 奇

(上海理工大学 健康科学与工程学院, 上海 200093)

摘要: 人脸关键点检测在需要精确定位面部特征的任务中具有关键作用。然而, 现有方法大多忽略了对时序特征相关性的探索, 而缺乏明确的时序约束特征会导致不同帧之间人脸关键点产生一定程度的闪烁, 即推理结果在连续帧之间快速振荡, 进而对连续视频帧中人脸特征点的稳定提取造成影响。为进一步提高视频流中面部关键点提取的稳定性和精确度, 提出一种基于时序与空间依赖性的面部关键点检测算法。该方法利用不同帧的所有像素构建多尺度多层特征相关性金字塔, 通过向特征图中添加相关性特征作为两帧之间的时序约束, 以预测时间上的运动关系。同时, 采用Swin Transformer来捕获人脸关键点之间的空间依赖关系。通过从粗特征到精特征的自适应迭代更新, 实现了视频流中关键点的稳定检测。实验结果显示, 所提方法在300VW数据集上的平均NME达到3.266%, 在300W和WFLW数据集上可达到目前先进方法的精度。

关键词: 人脸检测; Swin Transformer; 循环迭代优化; 特征相关性

DOI: 10.11907/rjtk.241351

开放科学(资源服务)标识码(OSID):

中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2025)006-0058-07



Application of Cyclic Iterative Optimization in Facial Landmark Extraction

YU Qi

(School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Facial landmark detection plays a crucial role in tasks that require precise localization of facial features. However, most existing methods neglect the exploration of temporal feature correlation, and the lack of clear temporal constraint features can lead to a certain degree of flickering of facial landmark between different frames, that is, the inference results oscillate rapidly between consecutive frames, which in turn affects the stable extraction of facial feature points in consecutive video frames. To further improve the stability and accuracy of facial landmark extraction in video streams, a facial landmark detection algorithm based on temporal and spatial dependencies is proposed. This method utilizes all pixels from different frames to construct a multi-scale and multi-layer feature correlation pyramid, and adds correlation features to the feature map as temporal constraints between two frames to predict temporal motion relationships. Meanwhile, Swin Transformer is used to capture the spatial dependencies between facial keypoints. Through adaptive iterative updates from coarse features to fine features, this method achieves stable detection of key points in video streams. The experimental results show that the proposed method achieves an average NME of 3.266% on the 300VW dataset, and can achieve the accuracy of current advanced methods on the 300W and WFLW datasets.

Key Words: facial detection; Swin Transformer; cyclic iterative optimization; feature correlation

0 引言

面部关键点检测是计算机视觉领域的重要研究方向, 在人机交互、情感计算、医疗诊断、增强现实以及虚拟现实等多个领域具有重要的应用价值^[1-3]。其主要目的是定位人脸上的关键点, 如眼角、嘴角、鼻尖等, 以辅助人脸识别、表情分析、三维人脸重建以及面部动画等任务的实现。早

期面部关键点检测方法主要依赖于手工设计的特征, 如Haar特征、HOG(Histogram of Oriented Gradient)特征、LBP(Local Binary Patterns)特征等, 通过级联回归的方法进行关键点预测。然而, 该类方法在复杂场景下的泛化能力较弱, 尤其是面对遮挡、光照变化、姿态变化等挑战时容易出现检测失败或定位误差较大的问题。

随着深度学习技术的快速发展, 基于神经网络的回归模型和热图回归技术逐渐成为主流, 后者更是通过预测关

收稿日期: 2024-04-09

扫描二维码阅读全文:

作者简介: 俞奇(1998-), 男, 上海理工大学健康科学与工程学院硕士研究生, 研究方向为深度学习在人脸关键点提取中的应用。



键点的概率分布显著提升了定位精度和鲁棒性^[4-5]。尽管这些方法在静态图像的关键点检测效果上已有较大提升,但在视频序列中的检测稳定性仍然面临挑战,特别是在有遮挡、模糊或剧烈位移的情况下,模型容易出现关键点抖动或丢失的现象,从而影响实际应用效果。因此,开发定位精度高和鲁棒性强的视频序列面部关键点检测方法十分必要。

1 相关研究

面部关键点检测模型可大致分为坐标回归模型和热图回归模型两类。坐标回归模型是一种通过训练数据学习输入特征与输出坐标之间映射关系的模型,在面部关键点检测任务中,其输入通常为面部图像,输出则为关键点坐标。例如Wu等^[6]提出一种基于边界感知的人脸关键点检测算法,该方法利用人脸热图捕捉边界信息,并通过边界线回归出其他关键点位置。Guo等^[7]引入欧拉角度作为辅助信息,通过加入角度约束提高关键点回归的准确性。然而,由于坐标回归模型在特征图上的连接特性,端到端回归容易导致较大误差,尤其是在复杂场景下难以保持高精度。相比之下,热图回归模型以每个关键点的概率分布图作为预测结果,能够更精准地捕捉局部信息,提高定位精度。例如,Deng等^[8]首次将热图回归方法应用于关键点检测任务,有效提升了不同场景下的鲁棒性。

然而,在视频帧中进行关键点检测比静态图像更具挑战性,需要充分利用时序信息,以减少帧间抖动。为了在视频序列中保持关键点检测的稳定性,部分研究利用时序信息进行优化,例如通过前一帧的检测结果对当前结果施加约束,以减少关键点抖动。然而,这种简单的约束方式在发生较大位移或遮挡时可能会导致滞后,并且难以从错误的预测中恢复。针对这一问题,文献[9]提出TDCRNN模型,采用循环神经网络(Recurrent Neural Network, RNN)学习视频帧之间的时序信息,通过将前后帧特征嵌入当前帧来提高关键点检测的稳定性。然而,这种方法仍然存在滞后性问题,尤其是在快速运动场景下可能会造成关键点定位的误差积累。

Transformer最初是为自然语言处理任务设计的,但自注意力机制的优势使其在计算机视觉领域得到了广泛应用^[10]。Swin Transformer作为对传统Transformer的改进,特别适用于图像处理任务,并在人脸形状约束中展现出显著优势^[11]。Swin Transformer引入窗口化交换机制,可有效捕捉图像的局部和全局信息。与传统的图像分割方法相比,这种新的分区机制不仅显著减少了参数量,而且保持了对图像各部分的建模能力。最重要的是,Swin Transformer能对序列中每个元素之间的关系进行建模,从而更好地理解人脸的刚性特征,更好地理解人脸的结构与姿态变化,从而构建更准确的关键点约束模型。

关键点检测任务的另一大挑战是面部遮挡问题。当人脸被部分遮挡时,模型难以从缺失的表现信息中推理出关键点位置。为解决这一问题,部分研究引入循环优化策略,采用模拟热图法回归关键点,提高对遮挡区域的预测能力。事实上,计算机视觉领域中的许多问题都可以理解为优化问题,一种常用的优化方法为直接从数据中迭代更新^[12]。一阶优化器通过数据驱动的迭代更新方式,为这些问题提供了高效实用的解决方案,例如原始一对偶混合梯度(Primal Dual Hybrid Gradient, PDHG)优化可以表示为一系列迭代更新的步骤^[13];全局运动聚合(Global Motion Aggregation, GMA)通过模拟一阶算法更新来对粗特征进行局部与全局匹配,进而更精细化地匹配运动信息^[14]。

目前,基于神经网络的人脸关键点检测方法存在一定局限,受限于卷积核的固定感受野,难以全面建模关键点之间的空间结构关系,导致缺乏对人脸整体形状约束的能力,降低了检测的准确性。为此,本文提出一种结合时序与空间依赖性的面部关键点检测方法,引入具有全局感知能力的Swin Transformer结构,结合交叉注意力机制,充分建模人脸的刚性特征与关键点间的长距离依赖关系,提升特征表达能力。同时,引入循环优化策略,借助热图回归方法对人脸关键点进行逐步优化,并通过捕捉视频帧之间的特征相关性与运动线索,有效提升在连续帧场景下的检测稳定性。此外,该方法利用面部关键点在时间维度和空间结构中的关联性推断被遮挡的关键点位置,进一步提升检测算法的鲁棒性与精度。本文主要贡献点为:①通过建立多尺度相关性矩阵,从高分辨率到低分辨率捕捉不同程度的帧间位移,提高模型对大位移场景的适应性,并增强对快速运动环境的鲁棒性;②构建了基于Swin Transformer的人脸形状约束网络,进一步提高了模型预测人脸关键点的准确性;③利用门控循环单元(Gated Recurrent Unit, GRU)构建热图推理网络,通过多次迭代和堆叠细化特征,模拟热图回归关键点的过程,提高了模型精度。

2 方法构建

本文方法可表示为:输入一对连续的RGB图像 I_0, I_1 , 模型估计与人脸关键点数量相同的高斯概率分布图,只保留每张特征图中信号值最大的点。 Z 为人脸关键点数量,但通常不是直接将 Z 作为最终结果,而是表示为 N 个大小为 $H \times W$ 的热图,其中 H, W 通常为64。本文方法可以用函数形式表示为 $Z = F((I_0, I_1); \theta)$, θ 为可学习优化的参数。本文网络整体结构如图1所示,主要分为4个模块:①特征提取模块。该模块为使用残差块搭建的粗特征提取网络;②人脸关系约束模块。该模块使用带有交叉注意力的Swin Transformer构造;③特征相关计算模块。该模块使用来自粗特征提取器的特征计算特征相关性,同时构造4层

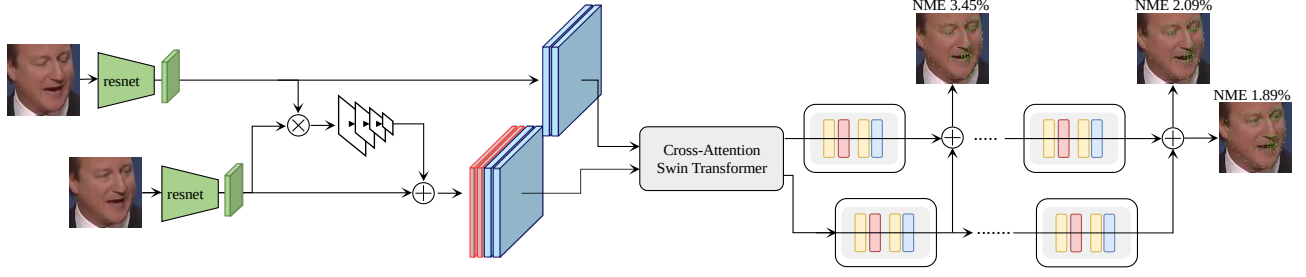


Fig. 1 Network structure

图1 网络结构

相关性金字塔;④迭代模块。该模块使用GRU构造的迭代块对特征和误差同时进行迭代优化,并多次计算loss。

2.1 粗特征提取

使用卷积网络从输入的视频帧中提取特征。特征提取器将输入图像映射到低分辨率的多通道特征图中,特征提取网络 $g_{ex}\theta$ 以 1/8 输出特征 $g_{ex}\theta: \mathbb{R}^{3 \times H \times W} \rightarrow \mathbb{R}^{D \times H/8 \times W/8}$, 其中 D 设置为 256。特征提取器由 6 个残差块组成,其中 2 个为 1/2 分辨率,2 个为 1/4 分辨率,2 个为 1/8 分辨率。

2.2 面部刚性约束

Swin Transformer 在捕获特征之间的空间依赖性和全局感受野方面具有优势,是处理面部关键点提取问题的最佳选择。首先,需要对关键点之间的位置关系进行建模并构建一个准确的面部刚性约束模型。由于 F_0 和 F_1 只是两组特征,没有空间位置的概念,需将正弦和余弦位置编码添加到特征中。添加位置信息有助于解决关键点标注存在的语义歧义问题。添加位置信息后,将 F_0 和 F_1 输入 6 个堆叠的自注意力中,通过交叉注意力和前馈网络来提高初始特征的质量。对于普通的自注意机制来说, Q, K, V 为相同的特征向量;但是对于交叉注意力来说, K, V 是相同的,而 Q 不同,以便引入它们的相互依赖关系。 F_0 和 F_1 存在连续关系。表示为:

$$\hat{F}_0 = T(F_0 + P, F_1 + P) \quad (1)$$

$$\hat{F}_1 = T(F_1 + P, F_0 + P) \quad (2)$$

式中: T 为 Transformer, P 为位置编码; T 的第一个输入为 Q , 第二个输入为 K 和 V 。标准的 Transformer 架构在窗口操作中计算复杂性较高,本文采用 Swin Transformer 中的滑动局部窗口注意力策略。与使用固定窗口大小的 Transformer 不同, Swin Transformer 将特征拆分为固定数量的本地窗口,使窗口大小可以自适应调整。具体来说,将输入大小为 $H \times W$ 的特征拆分为 $K \times K$ 个窗口,每个窗口大小为 $(H/K \times W/K)$,并在每个局部窗口内使用自注意力和交叉注意力机制,这种方式可达到良好的速度与精度。

2.3 特征相似性计算

特征相关性计算在计算机视觉任务中发挥着至关重要的作用^[15]。本文特征相关性计算方法参考文献[16]中提出的 RAFT (Recurrent All-Pairs Field Transforms) 方法,即为所有像素建立多尺度 4 维相关性特征,通过查找 4 维

相关性特征捕捉视频帧中的时序信息和运动关系。本文仅保留相关性计算模块和查找操作,主要作为视频帧时序信息使用,保证了在前后帧出现动作或遮挡时为模型提供稳定的人脸关键点推理。给定图像特征 $g_{ex}\theta(I_0) \in \mathbb{R}^{D \times H \times W}$ 和 $g_{ex}\theta(I_1) \in \mathbb{R}^{D \times H \times W}$,通过选取所有特征向量对之间的点积来形成相关性 C 。公式为:

$$C(g_{ex}\theta(I_0), g_{ex}\theta(I_1)) \in \mathbb{R}^{H \times W \times H \times W} \quad (3)$$

$$C_{ijkl} = \sum_h g_{ex}\theta(I_0)_{ijh} \cdot g_{ex}\theta(I_1)_{hkl} \quad (4)$$

首先,构建一个 4 层金字塔 $\{C^1, C^2, C^3, C^4\}$,通过将相关性矩阵的后两个维度与内核大小设置为 1、2、4、8 进行最大池化。因此, C^k 的维度为 $H \times W \times H/2^k \times W/2^k$ 。这组相关性金字塔提供了像素级大小位移的信息,保持前两个维度不变,可以快速追踪到物体的移动。

其次,定义一个查找算子 L_c ,通过对相关性金字塔进行索引来生成特征图。给定当前特征流 (f^0, f^1) ,将 I_0 中的每个像素 $x = (\mu, \nu)$ 映射到 $I_1: (\mu + f^0(\mu), \nu + f^1(\nu))$,然后在 x' 周围定义一个局部网格。用公式表示为:

$$Nx'_r = \{x' + d_x | d_x \in \mathbb{Z}^2, \|d_x\|_1 \leq r\} \quad (5)$$

式中: r 为搜索半径; $N(x'_r)$ 为局部领域,从局部领域中的相关量进行索引。

对不同尺寸的相关性金字塔执行查找操作,然后将不同分辨率的值拼接成单个特征图。特征查找操作如图 2 所示。

2.4 迭代更新

更新算子会迭代更新一系列热图 $\{f_1, \dots, f_n\}$, n 为人脸关键点数目。更新算子将相关性、隐藏状态、特征图作为输入,并输出更新后的 $\{f_1, \dots, f_n\}$ 和隐藏状态。迭代更新模块旨在模仿优化算法的步骤,更新算子的核心组件为基于 GRU 的门控神经网络,其中全连接层替换为卷积。本文设置两条迭代更新路径,实现了误差和特征的同时优化。单个迭代模块的计算公式为:

$$z_t = \sigma(\text{Conv}_{3 \times 3}([h_{t-1}, x_t], W_z)) \quad (6)$$

$$r_t = \sigma(\text{Conv}_{3 \times 3}([h_{t-1}, x_t], W_r)) \quad (7)$$

$$\tilde{h} = \tanh(\text{Conv}_{3 \times 3}([r_t \odot h_{t-1}, x_t], W_h)) \quad (8)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h} \quad (9)$$

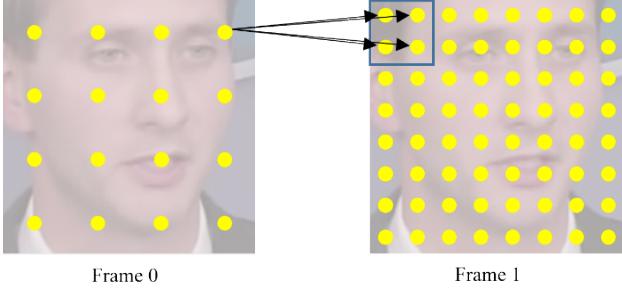


Fig. 2 Feature query operations

图2 特征查找操作

式中: $\text{Conv}_{3 \times 3}$ 为 3×3 的卷积; h_{t-1} 为上一个时刻的隐藏状态; σ 为激活函数; W_r 为权重参数; h_t 为当前时刻的隐藏状态。

2.5 损失函数

模型的每一次迭代都可以看作是对人脸关键点坐标的一次更新。第 n 次迭代的目标为确定一组新的人脸关键点坐标向量 z_n^* , 通过最小化当前待确定的关键点坐标向量 z 与上一帧已确定的关键点坐标向量 z_{n-1}^* 之间的 L_2 范数平方实现。即:

$$z_n^* = \arg \min_z \|z - z_{n-1}^*\|_2^2 \quad (10)$$

式中: $\arg \min$ 负责搜索使距离度量最小的 z ; $\|z - z_{n-1}^*\|_2^2$ 衡量的是坐标向量间的差异, 通过不断减小这个差异, 使关键点坐标在相邻帧间的变化尽可能平滑, 避免出现抖动等不稳定情况, 从而在视频序列中实现连贯、稳定的关键点定位。

同时, 为确保确定的关键点坐标符合模型内在的映射关系, 本文引入不动点约束。 $f(z, x_n; \theta^*)$ 表示基于当前帧的图像特征 x_n 和模型参数 θ^* 对关键点坐标向量 z 进行的变换。可得:

$$\text{subject to: } f(z, x_n; \theta^*) = z \quad (11)$$

式中: subject to : 表示约束条件; x_n 为与第 n 步迭代相关的输入向量; θ^* 为模型参数向量。该约束将距离最小化的迭代过程与不动点的定义紧密结合, 保证每一步更新的 z_n^* 均在满足不动点性质的前提下不断优化。

在模型的优化过程中, 首先将约束问题转化为无约束优化问题, 通过构建拉格朗日函数并引入乘子 λ , 得到拉格朗日鞍点目标:

$$\min_z \max_\lambda \|z - z_{n-1}^*\|_2^2 + \lambda'(f(z, x_n; \theta^*) - z) \quad (12)$$

在这种对抗式的优化过程中, 求 z 最小化和求 λ 最大化同时进行, 这使得优化过程极不稳定, 容易出现梯度消失或爆炸等问题, 导致模型难以收敛。为解决拉格朗日鞍点目标优化不稳定的问题, 本文引入松弛优化, 即将鞍点目标转化为一个光滑可优化的目标。松弛形式表示为:

$$\mathcal{L}_{rw} = \min_z \|f(z, x_n; \theta^*) - z\|_2^2 + \alpha \|z - z_{n-1}^*\|_2^2 \quad (13)$$

式中: 第一项 $\|f(z, x_n; \theta^*) - z\|_2^2$ 用于衡量 z 是否为网络 f 的不动点, 即确保输入与输出一致, 这有助于维持模型内部映射关系的稳定性; 第二项 $\alpha \|z - z_{n-1}^*\|_2^2$ 则鼓励 z 不要

偏离上一帧的结果, 以实现跨时间帧的平滑过渡, 保证在视频序列中关键点检测的连贯性; α 为用于调节以上两者的权重。

在整个迭代优化序列上监督网络预测热图与真实热图之间的 L_1 距离, 使用 Adaptive Wing Loss 来约束网络学习到真实热图的概率分布^[17]; 同时使用 Smooth L1 Loss 使得最终结果更加平滑:

$$\mathcal{L}_1(y - \hat{y}) = \begin{cases} \omega \ln \left(1 + \left| \frac{y - \hat{y}}{\epsilon} \right|^{\alpha - y} \right) & \text{if } |y - \hat{y}| < \theta \\ A|y - \hat{y}| - C & \text{otherwise} \end{cases} \quad (14)$$

式中: y 和 $\hat{y}$ 分别为人脸关键点热图上的真实值和预测值; ω, θ, ϵ 和 α 为正值; $A = \omega(1/(1 + (\partial/\epsilon)(\alpha - y)))(\alpha - y)((\theta/\epsilon)\alpha - y - 1)/(1/\epsilon)$, $C = (\theta A - \omega \ln(1 + (\theta/\epsilon)\alpha - y))$, 用于使损失函数在 $|y - \hat{y}| = \theta$ 处连续平滑。

当使用 L_1 损失时, 模型会从简单样本中学到很多噪声而不是有效信息; 而当 L_2 应用于异常值时, 模型很容易出现梯度爆炸。随着 L_n 的发展, Smooth L1 被提出, 通过分段定义来解决这个问题:

$$\mathcal{L}_2(p_i, \hat{p}_i) = \begin{cases} 0.5 \times |p_i - \hat{p}_i|^2, & |p_i - \hat{p}_i| < 1 \\ |p_i - \hat{p}_i| - 0.5, & \text{otherwise} \end{cases} \quad (15)$$

式中: $p_i = \text{soft argmax}(y)$, $\hat{p}_i = \text{soft argmax}(\hat{y})$, y 为模型预测值, $\hat{y}$ 为从实际面部关键点生成的真实热图标签。

第一阶段 loss 为:

$$\mathcal{L}_{\text{early}} = \mathcal{L}_1 + \mathcal{L}_2 \quad (16)$$

第二阶段 loss 为:

$$\mathcal{L} = \mathcal{L}_{\text{early}} + \mathcal{L}_{rw} \quad (17)$$

3 实验方法与结果分析

3.1 实验环境与参数设置

本文实验在 Intel i7-13700KF CPU 和 NVIDIA GeForce GTX 2080Ti GPU*2 的硬件平台上进行, 操作系统为 Ubuntu22.04 LTS, 编程语言为 Python。在 CUDA11.4 环境下利用深度学习框架 PyTorch 对模型进行训练与测试, 其中每个 GPU 的 batch size 为 12, 模型训练次数 epoch 为 60, 前 10 个 epoch 使用第一阶段 loss, 后 50 个 epoch 使用第二阶段 loss。采用 Adam 优化器, 初始学习率为 0.000 1, 并在 15、30 和 45 个 epoch 时将学习率降低 1/10。

将输入图像裁剪出脸部区域并将大小调整为 256×256 , 通过随机旋转、随机变换 HSV 颜色、随机中心裁剪、随机添加遮挡进行数据增强。根据经验将热图生成中的高斯 sigma 超参数设置为 1.5, 边缘热图生成中的线宽设置为 1.0。

3.2 实验数据集

300W 数据集由多个子数据集组成, 包括 LFPW、AFW、HELEN、IBUG 和 XM2VTS, 共包含 3 148 张训练图像

和 689 张测试图像^[18]。测试图像被分为一个通用子集, 包含 554 张图像; 以及一个困难子集, 包含 135 张图像。

300VW 是一个基于视频的人脸数据集, 训练集包含 50 个视频, 共 95 192 帧; 测试集分为 3 个子集, 分别为 A、B 和 C, 其中子集 C 是识别难度最大的^[19]。

WFLW 数据集包含 10 000 张面孔(7 500 张用于训练, 2 500 张用于测试)以及对面脸 98 个关键点的标注信息^[20]。除关键点标注外, 该数据集还包含丰富的属性标注, 如遮挡、姿态、妆容、光照、模糊和表情等。

3.3 评价指标

规范化均值误差(Normalization Mean Error, NME)表示模型预测的关键点与真实标签之间的平均欧氏距离, 通常

以双眼间距进行归一化。该值越低越好。

FR10% 用于衡量模型在检测人脸关键点时的准确性, 具体定义为在人脸关键点检测任务中, 误差超过 10% 的人脸数量所占的比例。该值越低越好。

AUC10% 用于衡量模型在检测人脸关键点时的整体性能, 具体定义为累积错误分布曲线下前 10% 部分面积。该值越高越好。

3.4 验证试验

图 3 为本文方法在视频和静态图片中的有效性展示。可以看出, 本文方法在静态图片中提取人脸关键点时具有较高精度, 在不同视频之间切换时检测结果没有出现较大失误。

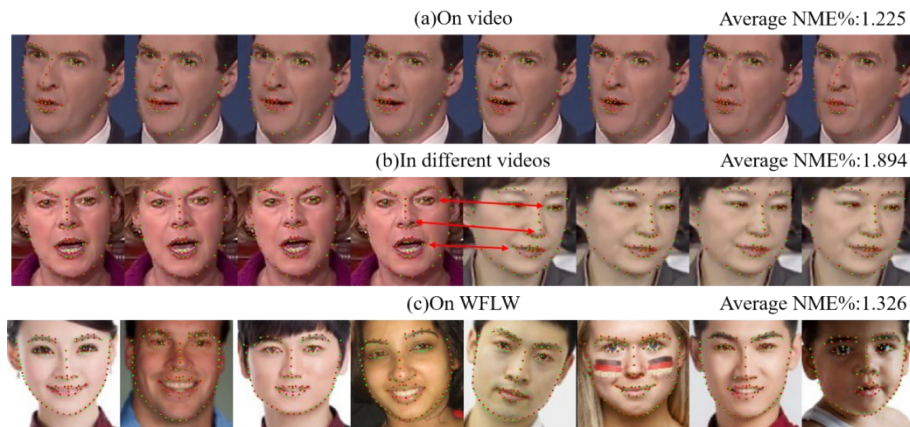


Fig. 3 Effectiveness demonstration of the proposed method on videos and static images

图 3 本文方法在视频和静态图片中的有效性展示

3.5 与其他方法的比较

为验证所提方法的有效性, 在不同数据集上将其与 MSKI^[15]、Awing^[17]、DeCaFa^[21]、HRNet^[22]、AVS+SAN^[23]、DAG^[24]、LUVLi^[25]、PIPNet^[26]、TSTN^[27]、FHR^[28]、FHR+STA^[28]、GAN^[29]、T-AOHR^[30]、SCPAN^[31]、SAAT^[32]、SimCC^[33]等方法进行比较。在 300W 和 WFLW 数据集上将本文方法与 DeCaFa、HRNet、Awing 等进行比较。结果见表 1。

可以看出, 本文方法在 300W 数据集 Full、Com、Chal 类别中的 NME 值分别为 3.11、2.61、4.67, 与其他方法相比处于较低水平, 表明其在面部关键点检测任务中的误差更小、精度更高。

本文方法在 WFLW 数据集上的 NME 值为 4.19, FR10% 值为 2.95, AUC10% 值为 0.596, 与其他方法相比 NME 值和 FR10% 值处于较低水平, AUC10% 值处于较高水平, 综合性能更优。

Table 1 Comparison of each method on 300W and WFLW dataset

表 1 各方法在 300W 和 WFLW 数据集上的比较

方法	300W (NME)			WFLW		
	Full	Comm.	Chal.	NME	FR10%	AUC10%
DeCaFa	3.39	2.93	5.26	4.62	4.84	0.563
HRNet	3.32	2.87	5.15	4.60	4.64	—
Awing	3.07	2.72	4.52	4.36	2.84	0.572
AVS + SAN	3.86	3.21	6.46	4.39	4.08	0.591
DAG	3.04	2.62	4.77	4.21	3.04	0.589
LUVLi	3.23	2.76	5.16	4.37	3.12	0.577
PIPNet	3.19	2.78	4.89	4.31	—	—
本文方法	3.11	2.61	4.67	4.19	2.95	0.596

在 300VW 数据集上将本文方法与 TSTN、FHR、GAN 等方法进行比较。结果见表 2。可以看出, 本文方法在 Category-A、Category-B、Category-C 类别中的 NME 值分别为 2.83、2.85、4.12, 在所有方法中处于较低水平, 验证了其人

脸检测的准确性。

3.6 消融实验

本文方法主要包含以下 3 个方面的改进: ①在网络训练和推理阶段增加前后帧时序特征; ②加入了使用交叉注

Table 2 Comparison of each method on 300VW dataset

表 2 各方法在 300VW 数据集的比较

方法	300VW (NME)		
	Category-A	Category-B	Category-C
TSTN	5.36	4.51	12.84
FHR	4.82	4.23	7.09
FHR+STA	4.21	4.02	5.64
GAN	3.50	3.67	4.43
T-AOHR	3.06	3.17	4.12
SPAN	4.49	4.23	5.87
SAAT	3.46	3.41	5.23
SimCC	4.07	4.12	5.13
MSKI	3.89	3.94	5.07
本文方法	2.83	2.85	4.12

意力的 Swin Transformer;③使用 GRU 迭代块作为最终的热图推理网络。通过消融实验分别探讨各项改进对模型推理人脸关键点准确性的影响,使用 300VW 的训练集和测试集以及相同的数据增强方法。结果见表 3。其中,T 表示只包含前后帧相关性特征匹配和提取模块的网络,S+T 表示引入了使用交叉注意力的 Swin Transformer 的模块,S+T+G 表示本文方法。可以看出,引入使用交叉注意力的 Swin Transformer 模块(即从 T 到 S+T)后,模型在 3 个类别下的 NME 值均显著下降,证实了该模块的有效性。加入 GRU 迭代块(从 S+T 到 S+T+G)后,模型在 3 个类别下的 NME 值再次降低。表明该模块可进一步优化模型,减少关键点提取误差。从 T 到 S+T,再到 S+T+G,模型的 NME 值持续降低,表明各项改进措施叠加后能有效提高人脸关键点的推理准确性。

Table 3 Result of ablation experiment on 300VW dataset

表 3 在 300VW 数据集上的消融实验结果

模型	300VW (NME)		
	Category-A	Category-B	Category-C
T	5.49	5.63	6.01
S+T	3.89	3.76	5.17
S+T+G	2.83	2.85	4.12

3.7 失败案例

虽然本文方法在面面对齐方面取得了一些改进成果,但仍然存在一些局限性。如图 4 所示,本文方法在极端姿势和严重遮挡时检测结果不尽如人意,主要原因是训练数据中的极端姿势和严重遮挡案例较少,导致模型在处理这些情况时的鲁棒性不足。

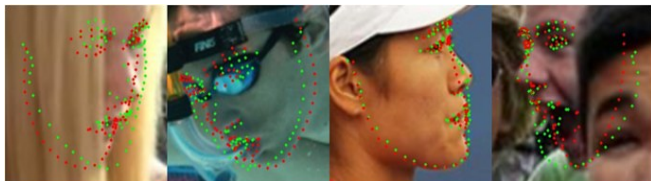


Fig. 4 Failure cases

图 4 失败案例

4 结语

针对人脸关键点推理在连续帧中出现场景切换和遮挡时会抖动的现象,本文提出一种应用特征相关性、Swin Transformer 与迭代更新策略的人脸关键点检测方法。该方法引入连续帧的相关性,为网络添加了新的信息,与大多数人脸关键点检测方法通过单纯提高准确度来缓解由于遮挡或姿势变换使得关键点来回快速震荡的问题有本质区别。本文提出的特征相关性计算模块能为模型提供有价值的特征,以便进行高质量的视频帧人脸关键点推理。

然而,由于数据集的缺乏,无法进行更广泛的训练与测试,影响了模型在极端姿势和严重遮挡情况下的表现。未来计划扩展数据集、结合多模态信息并优化模型架构,以提高其在各种复杂场景中的鲁棒性和精度。

参考文献:

- [1] GOGIĆ I, AHLBERG J, PANDŽIĆ I S Regression-based methods for face alignment: a survey[J]. Signal Processing, 2021, 178: 107755.
- [2] REN F, HUANG Z. Automatic facial expression learning method based on humanoid robot XIN-REN [J]. IEEE Transactions on Human-Machine Systems, 2016, 46(6): 810-821.
- [3] ZHU T, ZHANG C, WU T, et al. Research on a real-time driver fatigue detection algorithm based on facial video sequences[J]. Applied Sciences, 2022, 12(4): 2224.
- [4] DONG H Z, XIE M. Real-time driver fatigue detection based on simplified landmarks of AAM[C]//Chengdu: The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding, 2010.
- [5] BULAT A, TZIMIROPOULOS G. Human pose estimation via convolutional part heatmap regression[DB/OL]. <http://arxiv.org/abs/1609.01743>.
- [6] WU W, QIAN C, YANG S, et al. Look at boundary: a boundary-aware face alignment algorithm[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2129-2138.
- [7] GUO X, LI S, YU J, et al. PFLD: a practical facial landmark detector[DB/OL]. <https://arxiv.org/abs/1902.10859>.
- [8] DENG J, TRIGEORGIS G, ZHOU Y, et al. Joint multi-view face alignment in the wild[J]. IEEE Transactions on Image Processing, 2019, 28(7): 3636-3648.
- [9] ZHANG Z, LUO P, LOY C C, et al. Learning deep representation for face alignment with auxiliary attributes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(5): 918-930.
- [10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [DB/OL]. <https://arxiv.org/abs/1706.03762>.
- [11] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.
- [12] ADLER J, ÖKTEM O. Solving ill-posed inverse problems using iterative deep neural networks[J]. Inverse Problems, 2017, 33(12): 124007.
- [13] CHAMBOLLE A, POCK T. A first-order primal-dual algorithm for con-

- vex problems with applications to imaging[J]. Journal of Mathematical Imaging and Vision, 2011, 40: 120–145.
- [14] JIANG S, CAMPBELL D, LU Y, et al. Learning to estimate hidden motions with global motion aggregation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 9772–9781.
- [15] ZHU C, LI X, LI J, et al. Multi-sourced knowledge integration for robust self-supervised facial landmark tracking[J]. IEEE Transactions on Multimedia, 2022, 25: 6616–6628.
- [16] TEED Z, DENG J. Raft: recurrent all-pairs field transforms for optical flow[C]//16th European Conference on Computer Vision, 2020: 402–419.
- [17] WANG X, BO L, LI F X. Adaptive wing loss for robust face alignment via heatmap regression[DB/OL]. <https://arxiv.org/abs/1904.07399>.
- [18] SAGONAS C, ANTONAKOS E, TZIMIROPOULOS G, et al. 300 faces in-the-wild challenge: database and results[J]. Image and Vision Computing, 2016, 47: 3–18.
- [19] SHEN J, ZAFEIRIOU S, CHRYSOS G G, et al. The first facial landmark tracking in-the-wild challenge: benchmark and results[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015: 50–58.
- [20] WATCHAREERUETAI U, SOMMANA B, JAIN S, et al. LOTR: face landmark localization using localization transformer[J]. IEEE Access, 2022, 10: 16530–16543.
- [21] DAPOGNY A, BAILLY K, CORD M. Decafa: deep convolutional cascade for face alignment in the wild[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6893–6901.
- [22] WANG J, SUN K, CHENG T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(10): 3349–3364.
- [23] QIAN S, SUN K, WU W, et al. Aggregation via separation: boosting facial landmark detector with semi-supervised style translation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 10153–10163.
- [24] LI W, LU Y, ZHENG K, et al. Structured landmark detection via topology-adapting deep graph learning[C]//16th European Conference on Computer Vision, 2020: 266–283.
- [25] KUMAR A, MARKS T K, MOU W, et al. Luvli face alignment: estimating landmarks' location, uncertainty, and visibility likelihood[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 8236–8246.
- [26] JIN H, LIAO S, SHAO L. Pixel-in-pixel net: towards efficient facial landmark detection in the wild[J]. International Journal of Computer Vision, 2021, 129(12): 3174–3194.
- [27] LIU H, LU J, FENG J, et al. Two-stream transformer networks for video-based face alignment[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(11): 2546–2554.
- [28] TAI Y, LIANG Y, LIU X, et al. Towards highly accurate and stable face alignment for high-resolution videos[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 8893–8900.
- [29] YIN S, WANG S, PENG G, et al. Capturing spatial and temporal patterns for facial landmark tracking through adversarial learning[C]//International Joint Conference on Artificial Intelligence, 2019: 1010–1017.
- [30] YIN S, WANG S, CHEN X, et al. Attentive one-dimensional heatmap regression for facial landmark detection and tracking[C]//Proceedings of the 28th ACM International Conference on Multimedia, 2020: 538–546.
- [31] WAN J, XI H, ZHOU J, et al. Robust and precise facial landmark detection by self-calibrated pose attention network[J]. IEEE Transactions on Cybernetics, 2021, 53(6): 3546–3560.
- [32] ZHU C, LI X, LI J, et al. Reasoning structural relation for occlusion-robust facial landmark localization[J]. Pattern Recognition, 2022, 122: 108325.
- [33] LI Y, YANG S, LIU P, et al. Simcc: a simple coordinate classification perspective for human pose estimation[C]//European Conference on Computer Vision, 2022: 89–106.

(责任编辑:尹晨茹)