工程科学学报
Chinese Journal of Engineering

ViTAU：基于Vision transformer和面部动作单元的面瘫识别与分析

高嘉 蔡文浩 赵俊莉 段福庆

ViTAU: Facial paralysis recognition and analysis based on vision transformer and facial action units

GAO Jia, CAI Wenhao, ZHAO Junli, DUAN Fuqing

引用本文:
高嘉, 蔡文浩, 赵俊莉, 段福庆. ViTAU：基于Vision transformer和面部动作单元的面瘫识别与分析[J]. 北科大：工程科学学报, 2025, 47(2): 351–363. doi: 10.13374/j.issn2095–9389.2024.05.06.003
GAO Jia, CAI Wenhao, ZHAO Junli, DUAN Fuqing. ViTAU: Facial paralysis recognition and analysis based on vision transformer and facial action units[J]. *Chinese Journal of Engineering*, 2025, 47(2): 351–363. doi: 10.13374/j.issn2095–9389.2024.05.06.003

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2024.05.06.003

工程科学学报, 第 47 卷, 第 2 期: 351−363, 2025 年 2 月
Chinese Journal of Engineering, Vol. 47, No. 2: 351−363, February 2025
https://doi.org/10.13374/j.issn2095-9389.2024.05.06.003; http://cje.ustb.edu.cn

ViTAU: 基于 Vision transformer 和面部动作单元的面瘫识别与分析

高　嘉[1)], 蔡文浩[1)], 赵俊莉[1)✉], 段福庆[2)]

1) 青岛大学计算机科学技术学院, 青岛 266071　2) 北京师范大学人工智能学院, 北京 100875
✉通信作者, E-mail: zhaojl@yeah.net

摘　要　面部神经麻痹(Facial nerve paralysis, FNP), 通常称为贝尔氏麻痹或面瘫, 对患者的日常生活和心理健康产生显著影响, 面瘫的及时识别和诊断对于患者的早期治疗和康复至关重要. 随着深度学习和计算机视觉技术的快速发展, 面瘫的自动识别变得可行, 为诊断提供了一种更准确和客观的方式. 目前的研究主要集中关注面部的整体变化, 而忽略了面部细节的重要性. 面部不同部位对识别结果的影响力并不相同, 这些研究尚未对面部各个区域进行细致区分和分析. 本项研究引入结合 Vision transformer(ViT)模型和动作单元(Action unit, AU)区域检测网络的创新性方法用于面瘫的自动识别及区域分析. ViT 模型通过自注意力机制精准识别是否面瘫, 同时, 基于 AU 的策略从 StyleGAN2 模型提取的特征图中, 利用金字塔卷积神经网络分析受影响区域. 这一综合方法在 YouTube Facial Palsy(YFP)和经过扩展的 Cohn Kanade(CK+)数据集上的实验中分别达到 99.4% 的面瘫识别准确率和 81.36% 的面瘫区域识别准确率. 通过与最新方法的对比, 实验结果展示了所提的自动面瘫识别方法的有效性.

关键词　Transformer; 面部动作单元; 多分辨率特征图; 生成器; 热力图回归
分类号　TP391

ViTAU: Facial paralysis recognition and analysis based on vision transformer and facial action units

GAO Jia[1)], *CAI Wenhao*[1)], *ZHAO Junli*[1)✉], *DUAN Fuqing*[2)]

1) College of Computer Science & Technology, Qingdao University, Qingdao 266071, China
2) School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China
✉Corresponding author, E-mail: zhaojl@yeah.net

ABSTRACT　Facial nerve paralysis (FNP), commonly known as Bell's palsy or facial paralysis, significantly affects patients' daily lives and mental well-being. Timely identification and diagnosis are crucial for early treatment and rehabilitation. With the rapid advancement of deep learning and computer vision technologies, automatic recognition of facial paralysis has become feasible, offering a more accurate and objective diagnostic approach. Current research primarily focuses on broad facial changes and often neglects finer facial details, which leads to insufficient analysis of how different areas affect recognition results. This study proposes an innovative method that combines the vision transformer (ViT) model with an action unit (AU) facial region detection network to automatically recognize and analyze facial paralysis. Initially, the ViT model utilizes its self-attention mechanism to accurately determine the presence of facial paralysis. Subsequently, we analyzed the AU data to assess the activity of facial muscles, allowing for a deeper evaluation of the affected areas. The self-attention mechanism in the transformer architecture captures the global contextual information required to

收稿日期: 2024−05−06
基金项目: 山东省自然科学基金资助项目(ZR2024MF087); 国家自然科学基金资助项目(62172247)

recognize facial paralysis. To accurately determine the specific affected regions, we use the pixel2style2pixel (pSp) encoder and the StyleGAN2 generator to encode and decode images and extract feature maps that represent facial characteristics. These maps are then processed through a pyramid convolutional neural network interpreter to generate heatmaps. By optimizing the mean squared error between the predicted and actual heatmaps, we can effectively identify the affected paralysis areas. Our proposed method integrates ViT with facial AUs, designing a ViT-based facial paralysis recognition network that enhances the extraction of local area features through its self-attention mechanism, thereby enabling precise recognition of facial paralysis. Additionally, by incorporating facial AU data, we conducted detailed regional analyses for patients identified with facial paralysis. Experimental results demonstrate the efficacy of our approach, achieving a recognition accuracy of 99.4% for facial paralysis and 81.36% for detecting affected regions on the YouTube Facial Palsy (YFP) and extended Cohn Kanade (CK+) datasets. These results not only highlight the effectiveness of our automatic recognition method compared to the latest techniques but also validate its potential for clinical applications. Furthermore, to facilitate the observation of affected regions, we developed a visualization method that intuitively displays the impacted areas, thereby aiding patients and healthcare professionals in understanding the condition and enhancing communication regarding treatment and rehabilitation strategies. In conclusion, the proposed method illustrates the power of combining advanced deep learning techniques with a detailed analysis of facial AUs to improve the automatic recognition of facial paralysis. By addressing previous research limitations, the proposed method provides a more nuanced understanding of how specific facial areas are affected, leading to improved diagnosis, treatment, and patient care. This innovative approach not only enhances the accuracy of facial paralysis detection but also contributes to facial medical imaging.

KEY WORDS Transformer; action units; multi-resolution feature maps; generator; heatmap regression

面部神经麻痹, 常被称为贝尔氏麻痹或面瘫, 是面部肌肉受损所导致的面肌瘫痪, 会使患者在面部表情和运动方面遭受显著困难. 这种状况不仅影响日常生活的基本功能, 如咀嚼、说话和吞咽, 还可能对患者的社交交往和心理健康产生深远的影响. 面瘫的及时识别和诊断对于患者的早期治疗和康复至关重要.

传统的面瘫诊断方法主要依赖于医生的主观观察和评估, 但这种方法可能因医生经验和判断的不同而存在一定的主观性和误差. 因此, 发展更加客观和精确的面瘫诊断方法成为了医疗界的一大挑战. 近年来, 深度学习和计算机视觉技术的发展为面瘫自动识别提供了新的解决方案. 一些工作[1−5] 结合机器学习方法中支持向量机 (Support vector machine, SVM)、多层感知器 (Multilayer perceptron, MLP) 算法的去捕获面部特征来判别, 或者通过结合神经网络和手工特征对面瘫进行识别[6−8]. 一些工作通过对图像序列进行识别[9−11], 或者通过考虑基于区域的非对称面部特征[12−16] 定量评估面神经麻痹等级. 通过应用这些先进技术, 可以对大量面部表情图像进行深入分析, 自动提取关键特征, 实现对面瘫更加准确和客观的诊断. 这一方法不仅提高了诊断的准确性, 还大大降低了对专业医生的依赖, 使得面瘫诊断更加高效和普及. 面瘫的自动识别和分级已成为医学图像处理领域的一个重要研究方向.

然而, 面瘫自动识别面临着诸多挑战. 最主要的挑战之一是要求算法能够精确地识别和区分各种不同的面部肌肉活动. 现有的技术能够有效地对面瘫进行识别, 并区分出不同的面瘫等级. 然而, 这些方法通常只能对整个面部进行评估, 而无法提供关于面部具体受影响区域的详细信息. 这一限制意味着患者无法直观地了解面瘫的具体发病区域. 面部动作单元识别技术的融合不仅优化了面瘫检测的精度和效率, 而且为面瘫的深入研究和个性化治疗开辟了新的途径, 展现了该技术在提升医疗诊断质量和患者生活质量方面的巨大潜力.

为此, 本研究引入一种基于 Transformer 结构和注意力机制的面瘫区域检测方法. Transformer 架构和注意力机制在视觉任务中的应用表现出色, 尤其在处理高分辨率图像和复杂场景理解方面. 许多研究将 ViT 及其变体应用于图像分类、分割、检测和异常检测等领域, 取得了显著成果.

在图像分类领域, ViT 的变体已经被广泛应用于多种任务, 例如宠物分类、作物疾病检测和胸部 X 光图像分类等[17]. 此外, CTNet[18] 通过结合 CNN 与 ViT, 进一步提升了分类性能. 其他研究则针对遥感图像分类, 设计了小型网络和多实例 ViT[19−20], 并且 ViT 的应用范围还扩展到高光谱和激光雷达数据的分类任务[21]. 在特定应用中, LeViT[22] 被提出用于沥青路面图像的分类, 而 ViT 同样被应用

于股骨骨折分类[23], SeedViT[24] 则用于玉米种子的分类, 此外, 双输出 ViT[25] 也在空气质量分类任务中得到了开发与应用.

在医学图像处理方面, Transformer 技术同样展现了强大的应用潜力. 多实例增强 ViT[26] 被用于眼底图像的分类, 同时, Transformer 也被用于 CT 图像的去噪处理[27−29], 而 ViT 与 U-Net 结合[30] 的模型更是实现了医学图像分割任务. 进一步的研究也表明, Transformer 在医学图像处理中的应用[31−33] 具有显著的价值.

此外, Transformer 在物体检测和异常检测领域同样取得了进展. 基于 Transformer 的 3D 目标检测框架[34−35] 被应用于相关任务中, 无监督学习技术[36] 则被开发用于卫星图像的物体检测. 此外, ViT 在图像异常检测中的应用[37−39] 也得到了广泛研究, 在捕捉长距离依赖关系方面展现出了强大的优势.

针对面瘫检测这一具体应用, ViT 的引入有助于模型更有效地处理和理解面部的局部细节与整体结构之间的复杂关系, 这对于准确识别和定位面部不同区域的瘫痪状态至关重要. 因此, 我们设计了一个基于 ViT 的面瘫识别网络, 通过在面瘫人脸上进行自注意力机制增强对局部区域特征的提取, 以实现面瘫的精准识别. 接着, 通过集成面部动作单元信息, 对被判断为有面瘫的患者进一步进行面瘫区域的分析. 我们的方法在实现面瘫精准识别的基础上, 通过集成面部动作单元信息, 对面部进行细致的分区域分析, 从而提供更为精确和全面的面瘫评估, 这将有助于更好地理解面瘫的具体影响区域. 我们的网络能够实现自动化的面瘫检测, 提供快速、可靠的面瘫检测结果, 减轻医生的工作负担, 缩短诊断时间, 并根据具体的面瘫部位, 康复专业人员可以设计个性化的康复计划, 有针对性地锻炼和治疗患者的面部肌肉.

论文中的创新点主要集中于以下三点:

(1)开发了一种新方法, 结合面部动作单元信息, 对面瘫进行细致的分区域分析, 实现更为精确和全面的面瘫评估.

(2)引入 ViT 和面部动作单元, 通过自注意力机制增强局部特征提取, 实现精准识别, 并结合先验知识进一步分析面瘫区域, 提高识别效果.

(3)设计了一种通过热图展示面瘫部位的可视化方法, 便于康复人员直观判断病灶位置, 制定个性化康复计划.

1　相关工作

1.1　面瘫检测

一些方法通过使用面部特征点识别来评估面瘫, 如对面部图像进行分析并提取关键特征点, 通过比对左右两侧脸部特征点的位置和形状, 可以评估面部的对称性, 来判断是否为面瘫. Yoshihara 等[12] 提出通过深度卷积网络 (Deep convolutional neural networks, DCNN) 对面部特征点进行精确检测, 从而定量评估面部麻痹的严重程度. Guo 等[40] 提出通过卷积神经网络进行面部不对称度的分类, 以此定量评估单侧外周性面瘫. Li 等[13] 提出通过目标检测网络和语义分割网络提取鼻唇沟, 并通过计算鼻唇沟的长度、深度和方向评估面部不对称性, 从而定量评估面部麻痹的严重程度. Song 等[41] 提出使用单个卷积神经网络 (Convolutional neural network, CNN) 对面神经麻痹进行分类, 以此达到定量评估面神经麻痹的目的.

当进行面瘫的识别时, 除了通过计算左右两侧脸部的对称性和面部标记点的距离, 还可以通过定位面部的具体区域来获取更准确的结果. Storey 等[42] 提出 3DPalsyNet 框架, 使用 3D CNN 架构对面部麻痹进行定级和口部运动识别. Hsu 等[43] 提出深层分层网络 (Deep hierarchical network, DHN), 结合 YOLO2(You Only Look Once2) 检测器、标志点学习网络和目标检测网络, 定量分析面神经麻痹. Tan 等[44] 提出基于正则化互熵准则半监督极限学习机和级联卷积神经网络的面神经麻痹评估方法 (Facial nerve paralysis assessment based on regularized correntropy criterion SSELMvc and cascade CNN, FNPA-RCELM-CCNN), 通过提取嘴和眼部区域的特征, 使用半监督极限学习机对面神经麻痹进行分类.

此外, 现有的面瘫数据集通常数量有限且不平衡, 缺乏充足和多样化的训练样本, 这限制了深度学习模型的训练效果和泛化能力. 为了解决这一问题, 研究者们采用多种数据增强方法来提高识别的精度. 通过增加样本的多样性和数量, 可以改善模型的泛化能力和鲁棒性. Abayomi-Alli 等[1] 提出基于 Voronoi 分解的随机区域擦除 (Voronoi decomposition-based random region erasing, VDRRE) 数据增强方法, 提高面部麻痹检测精度和模型性能. Sakai 等[45] 提出一种产生伪面部图像的方法, 以此解决由于病人隐私无法在医生培训和教育中共享病人图像的问题. Parra-Dominguez 等[6] 提出一

种方法, 结合神经网络和手工特征, 识别面部麻痹患者的六种手势, 以此对面部麻痹进行评估.

从图像序列的角度出发, 可以提高面瘫判定的准确性和可靠性. 通过对多幅图像进行分析和比较, 可以观察面部表情的动态变化, 从而更好地判断人脸是否患有面瘫. Yang 等[46] 提出三流长短期记忆 (Triple-stream long short term memory) 网络, 通过提取局部和整体面部运动特征, 定量评估面神经麻痹的严重程度. Xu 等[47] 提出双路径长短期记忆网络与深度差异化网络 (Dual-path LSTM with deep differentiated network, DP-LSTM-DDN), 通过提取面部运动特征和分析面部不对称性, 定量评估面神经麻痹的严重程度. Liu 等[48] 提出并行层次卷积神经网络 (Parallel hierarchy convolutional neural network, PHCNN), 结合了长短期记忆 (Long short-term memory, LSTM) 网络结构, 考虑区域特征和图像序列的时间变化, 定量评估面神经麻痹的等级.

1.2　面部动作单元 (Action unit，AU) 检测

现有 AU 检测方法试图通过强调重要区域来识别面部部位, 但这些方法并不能有效地将专家先验知识纳入区域定义. 并且当前的 AU 检测方法未使用具有专业先验知识的区域卷积神经网络来适应性地关注与 AU 相关的区域. Ma 等[9] 提出的面部动作单元区域卷积神经网络 (Action units regional convolutional neural networks, AU R-CNN) 模型, 通过直接观察各个 AU 所在的不同面部区域, 将专家先验知识纳入区域定义来解决问题.

面部动作单元检测方法难以应对不同受试者之间的 AU 外观差异, 导致在训练和测试数据集不同的跨域场景中性能低下. 为了解决 AU 检测中基于身份的类内差异问题, Tu 等[10] 提出了 IdenNet 算法, 该算法利用带有身份标签的人脸图像来解决这个问题. Ntinou 等[49] 通过使用热图回归方法估计面部表情中动作单元的定位和强度.

为了解决检测图像序列数据中的面部动作单元的问题, Akay 等[50] 通过组合从不同级别的 AU 探测器中提取的多个线索来提高检测性能. 李学翰等[51] 提取面部图像序列, 引入迁移学习的方法, 并将视频序列的提取特征输入长短期记忆网络处理时域特征. 利用基于深度学习的方法和运动历史图像来捕捉面部表情的时间变化可以定位微小的 AU 并跟踪面部肌肉运动的微小变化.

为了解决联合面部动作单元检测和面部对齐问题, Shao 等[52] 强调了这两项任务之间的相关性, 因为面部标志可以提供精确的 AU 位置, 从而便于提取有意义的局部特征进行 AU 检测. 现有方法将面部对齐视为预处理步骤, 并对每个 AU 使用固定区域或注意力, 这可能无法有效地捕获 AU 的不规则区域. 将局部特征与面部对齐和全局特征集成以进行 AU 检测来克服这些局限性. Ge 等[53] 提出了多级图关系推理网络通过在区域级、像素层面和信道层面执行多级特征学习来解决局部与全局特征之间的动态互动问题. Li 等[54] 采用改进的光流法提取视频图像序列中相邻两帧图像的面部表情运动特征, 实现面部表情关键帧捕捉.

现有基于面部特征点的判断方法虽然广泛应用于多种面部分析任务中, 但在面瘫检测中, 特征点的准确定位受到面瘫本身影响而可能出现偏差, 从而影响最终的判断结果. 引入 Transformer 结构和注意力机制可以为面瘫检测带来显著的改进, 其内置的自注意力机制可以自动识别和关注输入数据中最相关的部分, 这在面瘫检测中意味着模型能够自动聚焦于面部的关键区域, 而非全脸. 这将提高检测的准确性, 也可以显著提升模型的效率. 此外, 面瘫往往影响面部的特定区域, 如眼睛、鼻子和嘴巴附近, 通过集成面部动作单元信息, 能够进行细致的面部区域分析, 从而提供更全面的面瘫评估.

2　方法

面瘫对整个面部的影响往往在面部的特定区域更为明显, 面瘫症状影响的不同区域如图 1 所示, 面部动作单元识别方法能够对这些特定区域进行精确分析.

本文通过结合面部动作单元识别技术, 以提高对面瘫影响区域的定位精度和细节分析. 我们

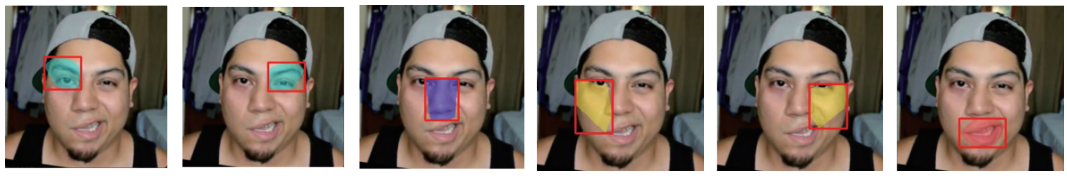

图 1　面瘫影响的不同区域

Fig.1　Different areas affected by facial paralysis

的面瘫检测网络模型分为面瘫识别模块和区域检测模块, 如图 2 所示. 首先, 输入的面瘫图像经过人脸定位并裁剪出人脸区域, 面瘫识别模块通过提取面瘫图像的特征, 实现对面瘫的精准识别, 判断出是否属于面瘫. 然后, 面瘫检测模块接收被判断为面瘫的患者图像, 通过检测 AU 的变化, 来推测出面瘫的发生区域. 最终, 我们通过输出一张区域热力图来表示出面瘫部位.

2.1　面瘫识别模块

我们引入 Vision transformer(ViT)模型来实现面瘫识别任务, 如图 3 所示. ViT 是一种基于自注意力机制的深度神经网络模型, 由 Google Brain 团队提出. 相比于传统的卷积神经网络(CNN)模型, ViT 不需要使用卷积层, 而是采用了 Transformer 架构中的自注意力机制, 通过将图像中的像素视为序列来处理图像, 其自注意力机制能够捕获图像的全局上下文信息, 这对于精确识别受面瘫影响的特定面部区域至关重要.

2.1.1　ViT 模型的基本架构

ViT 模型的基本架构主要包括输入预处理、位置编码、Transformer 编码器和分类头等部分. 首先, 输入的图像被切分成多个固定大小的小块. 每个小块被展平为一维向量, 然后通过一个线性层(或全连接层)转换成固定维度的嵌入向量, 这个维度通常与 Transformer 的隐藏层维度一致, 以便于后续处理. 由于 Transformer 本身不具备处理序列顺序信息的能力, 因此需要为每个图像块引入位置编码, 这些编码与图像块的嵌入向量相加, 以引入空间位置信息, 确保模型能够理解图像中各个部分的空间关系.

接下来, 经过预处理和位置编码的嵌入向量序列被输入到 Transformer 编码器中. 编码器由多个相同的层堆叠组成, 每层包括多头自注意力机制和前馈神经网络. 多头自注意力机制允许模型在计算每个图像块表示时, 考虑到所有其他图像块的信息, 从而捕获全局的图像特征. 每个 Transformer 层中还包含层归一化和残差连接, 这有助于稳定训练过程, 促进梯度的有效传播, 避免梯度消失或爆炸的问题.

在 Transformer 的输入序列中, ViT 模型引入了一个特殊的“分类标记”, 它与图像块的嵌入向量一起作为输入. 经过编码器的处理后, 分类标记的嵌入向量会聚合整个图像的信息, 用于最终的分类任务. 模型的输出通常是针对各个类别的概率分布, 这些概率由最后一个 Transformer 层输出的分类标记通过一个全连接层(即分类头)和 Softmax

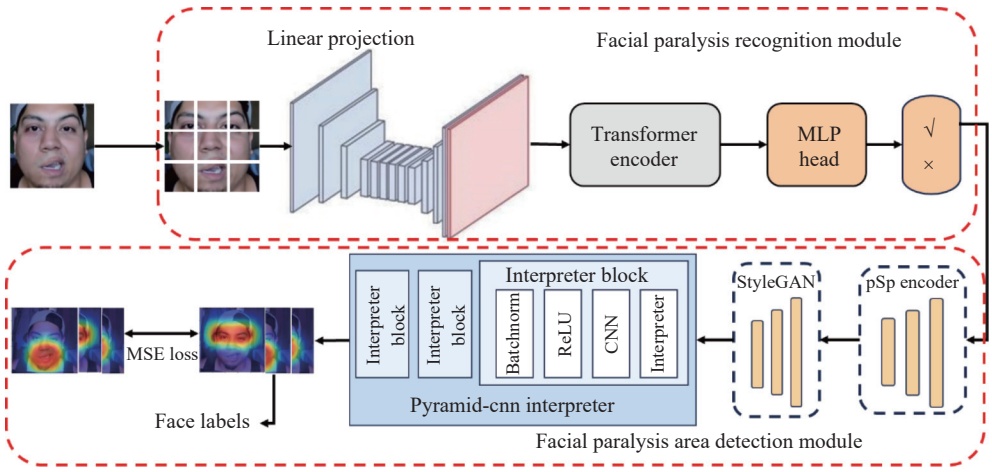

图 2　面瘫识别总流程图

Fig.2　Overall flowchart for facial paralysis recognition

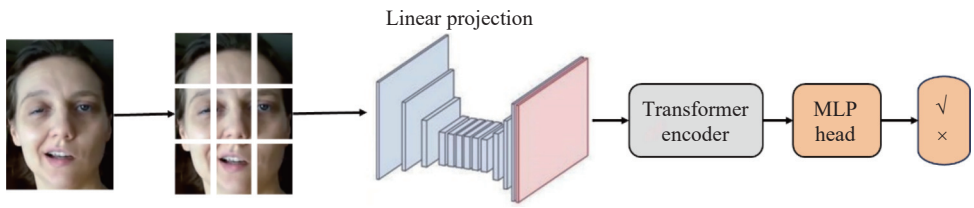

图 3　基于 ViT 架构的分类网络

Fig.3　Classification network based on ViT architecture

激活函数得到. 通过这种架构, ViT 模型成功地将图像分类任务转换为序列建模问题, 充分利用了 Transformer 在建模全局依赖关系方面的优势, 实现了对图像的有效理解和分类.

2.1.2　输入和 Patch 嵌入

我们先将输入图像 $\boldsymbol{X}\in\mathbf{R}^{H\times W\times C}$ 分成 N 个 patch 块, 每个 patch 大小为 $P\times P$, 进行图像块的嵌入, 对于每个图像块 $\boldsymbol{P}_i$, 我们通过一个嵌入层将其映射到一个 D 维的向量空间, 公式如下:

$$z_i^0=[\boldsymbol{P}_i\boldsymbol{W}_P]+\boldsymbol{E}_{\mathrm{pos},i}$$

其中, $\boldsymbol{W}_P\in\mathbf{R}^{(P^2\cdot C)\times D}$ 是可学习的投影矩阵, $\boldsymbol{E}_{\mathrm{pos},i}\in\mathbf{R}^D$ 是位置编码.

2.1.3　位置编码

对于位置编码 $\boldsymbol{E}_{\mathrm{pos}}$, 将位置编码添加到 patch 的嵌入向量中, 确保模型能捕捉到位置信息, 公式如下:

$$\boldsymbol{Z}^0=[z_{\mathrm{cls}},z_1^0,z_2^0,\cdots,z_N^0]+\boldsymbol{E}_{\mathrm{pos}}$$

其中, z_{cls} 是一个特殊的可学习的分类 token, 代表整个人脸图像的全局信息.

2.1.4　Transformer 编码器

对于给定一个输入面瘫图像序列, 通过自注意力机制可以计算每个位置对应的注意力权重. 使用多头自注意力机制, 可以学习到不同 Attention 的注意力表示. 自注意力层的计算为:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V}$$

其中, $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别是查询(Query)、键(Key)和值(Value)矩阵, d_{k} 是键向量的维度, 目的是为了缩放点积.

多头注意力允许模型在不同的表示子空间中并行捕获信息, 公式如下:

$$\mathrm{Multihead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)\boldsymbol{W}^0$$

其中, $\mathrm{head}_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}})$, $\boldsymbol{W}_i^{\boldsymbol{Q}}$, $\boldsymbol{W}_i^{\boldsymbol{K}}$, $\boldsymbol{W}_i^{\boldsymbol{V}}$ 和 $\boldsymbol{W}^0$ 是可学习的权重矩阵, h 是头的数量.

在每个 Transformer 块中, 经过自注意力机制之后的输出会经过一个前馈神经网络. 这个前馈神经网络由两个全连接层和一个激活函数组成:

$$\mathrm{FFN}(\boldsymbol{x})=\max(0,\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1)\cdot\boldsymbol{W}_2+\boldsymbol{b}_2$$

其中, $\boldsymbol{x}$ 是经过注意力机制计算序列中不同位置之间的关系后输出的新的特征表示, $\boldsymbol{W}_1$ 和 $\boldsymbol{b}_1$ 是第一个全连接层的权重矩阵和偏置向量, $\boldsymbol{W}_2$ 和 $\boldsymbol{b}_2$ 是第二个全连接层的权重矩阵和偏置向量.

网络的输出层是一个全连接层, 它将最后一个 Transformer 块的输出映射到最终的分类标签或预测结果, 使用特殊的分类 token z_{cls} 的最终状态进行分类, 公式如下:

$$y=\mathrm{softmax}(z_{\mathrm{cls}}\boldsymbol{W}_{\mathrm{cls}})$$

其中, z_{cls} 是在 Transformer 编码器的最后一层输出的隐藏状态向量, $\boldsymbol{W}_{\mathrm{cls}}$ 是用于分类任务的线性层的权重.

2.1.5　损失函数

对面瘫检测图像分析任务, 准确区分至关重要, 使用交叉熵损失函数有助于加速模型在面瘫检测任务上的学习过程, 公式如下:

$$L_{\mathrm{CE}}(y,y_{\mathrm{true}})=-\sum_{i=1}^{C}y_{\mathrm{true},i}\log(y_i)$$

其中, y_i 表示模型对第 i 类别的预测概率. $y_{\mathrm{true},i}$ 是真实标签中第 i 个类别的概率.

这一流程充分利用 Transformer 架构的自注意力机制, 通过逐步处理图像序列中的信息来预测面部神经麻痹的状况.

2.2　面瘫人脸区域检测网络

尽管通常情况下, AU 识别用于面部表情分析, 我们认为 AU 信息同样可以用于面部瘫痪的识别. 通过分析 AU 信息以确定面部肌肉的活动情况, 能够更深入地评估面部瘫痪的发生区域. 由于面瘫主要影响眼睛、鼻子和嘴巴等区域, 这与面部表情所涉及的区域有一定的重叠. 因此, 为了解决对不同面部区域的面瘫判断问题, 我们在本研究中提出一种基于 AU 的面瘫区域检测网络.

图 4 展示了网络的整个流程, 我们先对面瘫人脸图像使用 pixel2style2pixel(pSp)编码器[55]进行编码, 然后再经过 StyleGAN2 生成器对图像进行解码, 从而获得图像的特征图, 提取的特征图经过金字塔卷积神经网络解释器生成热力图, 通过对预测和真实的热图进行均方误差优化, 完成对面瘫区域检测.

2.2.1　面部特征编码

具体来说, 我们从经过大规模和多样化的人脸图像数据集预训练的 StyleGAN2(Style-based generator architecture for generative adversarial networks)模型中提取特征图.

为了从 StyleGAN2 生成器中提取特征, 我们首先将输入图像编码到潜在空间, 然后对潜在代码进行解码. 我们利用 pSp 编码器对输入图像进行编码, 得到隐空间编码. 尽管 pSp 编码器进行了有效编码, 但生成器特征可能无法捕获用于面瘫检

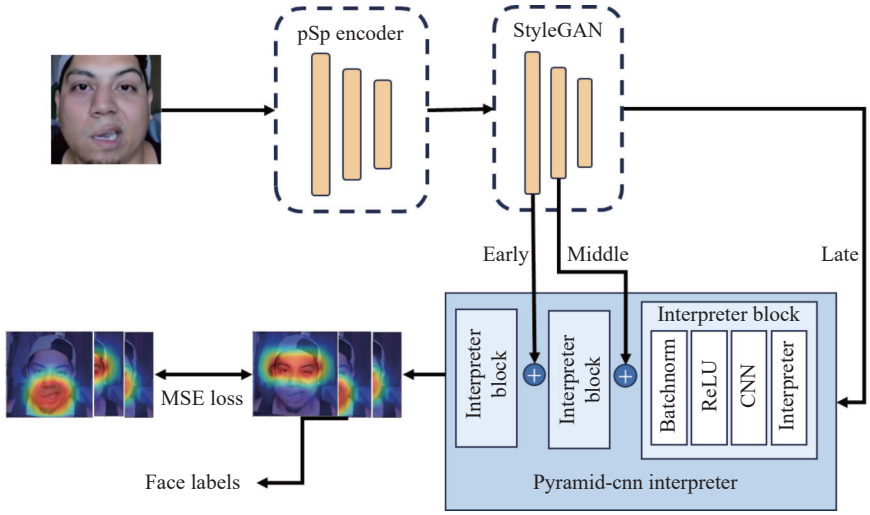

图 4　添加 AU 先验知识的面瘫区域识别

Fig.4　Facial paralysis area recognition incorporating AU prior knowledge

测的局部面部特征. 为了解决这个问题, 需要在训练期间对编码器和生成器进行微调. 然后用 StyleGAN2 生成器 G 对潜在编码进行解码, 以获得图像. 为了更好地理解模型的生成过程, 并且在训练过程中引导模型捕获到图像的关键特征, 从而生成更高质量的图像, 我们在解码过程中从生成器中提取了不同阶段的中间激活. 这些中间激活代表了生成图像时模型经历的各个处理步骤, 通过这些步骤, 模型逐渐生成出最终的特征图.

$$\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_k\}=G'(\boldsymbol{w}^+)=G'(E(\boldsymbol{x}))$$

其中, $\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_k$表示从生成器的多个中间层提取的特征映射, G'表示 StyleGAN2 的生成器, $\boldsymbol{w}^+$表示输入到 StyleGAN2 的潜在空间中的向量, E是 pSp 编码器, 用于将输入图像 $\boldsymbol{X}$ 是映射到潜在空间, 生成 $\boldsymbol{w}^+$.

2.2.2　面部 AU 划分

在进行面瘫检测的研究中, 准确划分面部区域至关重要. 由于面瘫症状主要影响面部的特定区域, 如眼睛、鼻子、脸颊和嘴巴, 因此我们的研究专注于这些区域的详细分析. 为了系统地识别和分类面部表情变化, 我们采用了面部动作单元 (Action unit, AU)[56] 作为划分规则.

如图 5 所展示, 我们选择了以下动作单元来标识关键的面部区域:

AU1(额头), AU2(额头)和 AU4(眉毛)主要关联眼睛周围的表情变化.

AU6(眼角)涉及眼睛的动态, 对识别眼部瘫痪尤为关键.

AU9(鼻)和 AU12(唇角)与鼻子和嘴巴的动态紧密相关, 对分析面瘫影响的微笑等表情至关重要.

AU25(嘴巴)和 AU26(颚部)则直接影响嘴巴区域, 对于诊断下半部面瘫症状尤为重要.

需要注意的是, 传统的 AU 通常用于检测面部表情, 而我们在此工作中的定义有所不同. 我们的 AU 与面瘫区域一一对应, 不受面部表情变化的影响, 相同的个体对应相同的 AU 组合.

2.2.3　基于金字塔卷积神经网络的面部热图

为了获得热力图检测从而输出对应的面瘫区域, 对于从 StyleGAN2 生成器生成的不同时期的特征图经由金字塔卷积神经网络 (Pyramid CNN) 生成 n 个基本热图 $\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_n\in\mathbf{R}^{w\times h}$和面瘫 AU 标签, 其中 n 为标签数量, w 和 h 代表图像大小.

我们利用金字塔卷积神经网络解释这些特征图, 以检测面部区域, 从而实现高效的训练并捕获必要的局部特征. 金字塔卷积神经网络解释器 H 包括 k 个层级, k 代表从生成器提取的特征图数量. 在每个层级中, 先将前一层级的隐藏状态 $\boldsymbol{c}_{i-1}$ 与生成器的特征图 $\boldsymbol{f}_i$ 相加, 随后通过解释器块 C_i 处

图 5　AU 描述

Fig.5　Description of AU

理. 每个解释器块包含插值、卷积、ReLU 及批归一化层. $\boldsymbol{m}=\boldsymbol{c}_k$代表最终的面部的热图.

$$\boldsymbol{c}_0=\boldsymbol{0},\ \boldsymbol{c}_i=C_i(\boldsymbol{c}_{i-1}+\boldsymbol{f}_i),\ i=1,2,\cdots,k;$$

$$\boldsymbol{m}=\boldsymbol{c}_k=H(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_k)$$

2.2.4　损失函数

为了能够有效地优化预测热图与真实热图之间的误差, 帮助模型准确地定位和量化面部各个部分, 引入均方误差作为损失函数, 公式为

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

其中, n 是样本数量, y_i是真实值, $\hat{y}_i$是模型预测值.

2.3　可视化方法

为了直观地展示面瘫诊断中症状对应的面部区域, 我们采用了梯度加权类激活映射(Grad-CAM, Gradient-weighted class activation mapping)技术[57]. Grad-CAM 利用卷积神经网络的特征图和输出层的梯度信息, 通过关注模型决策过程中权重较高的特征区域, 生成热力图, 突出显示对分类决策最重要的区域. 这样, 我们可以可视化模型对特定类别的响应, 深入理解模型的决策过程, 并直观地显示出面瘫影响的面部区域.

2.3.1　选择卷积层

我们选择了网络中最后一个卷积层作为提取特征图的层. 之所以选择该层, 是因为它能够提供足够的空间细节, 同时保留目标类别的语义信息. 这对于定位面瘫症状对应的面部区域至关重要, 因为需要在高层语义和空间分辨率之间取得平衡.

2.3.2　计算梯度

对于目标类别 u, 计算该类别得分y^{u}对于特征图$\boldsymbol{A}^k$的梯度, 即:

$$\mathrm{grad}\boldsymbol{s}=\frac{\partial y^{\mathrm{u}}}{\partial \boldsymbol{A}^k}$$

其中, y^{u}是模型对类别 u 的输出得分, $\boldsymbol{A}^k$表示第 k 个特征图, 这个梯度表示特征图$\boldsymbol{A}^k$中的每个像素对最终类别得分的影响程度.

2.3.3　全局平均梯度

接下来, 我们对每个特征图的梯度进行全局平均池化, 获得对应的权重$\boldsymbol{\alpha}_k^{\mathrm{u}}$.

$$\boldsymbol{\alpha}_k^{\mathrm{u}}=\frac{1}{Z}\sum_i\sum_j\mathrm{grad}\boldsymbol{s}_{ijk}$$

其中, Z 是特征图$\boldsymbol{A}^k$中元素总数, i 和 j 分别是特征图的高度和宽度索引. 权重$\boldsymbol{\alpha}_k^{\mathrm{u}}$反映了特征图$\boldsymbol{A}^k$是对类别 u 的重要性.

2.3.4　权重特征图生成

利用计算得到的权重$\boldsymbol{\alpha}_k^{\mathrm{u}}$, 对对应的特征图进行加权求和, 并应用 ReLU 激活函数, 生成类别激活图$\boldsymbol{L}_{\text{Grad-CAM}}^{\mathrm{u}}$:

$$\boldsymbol{L}_{\text{Grad-CAM}}^{\mathrm{u}}=\mathrm{ReLU}\left(\sum_k\boldsymbol{\alpha}_k^{\mathrm{u}}\boldsymbol{A}^k\right)$$

ReLU 函数确保我们只保留对类别判断有正向贡献的特征, 过滤掉负贡献的部分, 从而突出显示关键面瘫区域.

2.3.5　后处理与可视化

最后, 我们将得到的类别激活图$\boldsymbol{L}_{\text{Grad-CAM}}^{\mathrm{u}}$上采样到与输入图像相同的尺寸. 这样, 我们就可以将激活图与原始输入图像叠加, 直观地展示模型关注的面部区域, 具体显示出面瘫症状影响的部位.

3　实验与讨论

3.1　数据集与预处理

我们使用公开的面部麻痹数据集 YouTube (YouTube Facial Palsy, YFP)和经过扩展的 Cohn Kanade(CK+)数据集. YFP 数据集包含了从 YouTube 收集的 22 名面部麻痹患者的 32 个视频剪辑. 我们将视频转换为图像序列, 将其中的 26000 张图片作为面瘫图像参与训练. 而正常人脸的数据则来自扩展的 CK+数据集, 其中的 29000 张图片作为正常人脸参与训练. 总共 55000 张图片参与进行训练. 在进行面瘫自动识别与定位研究时, 我们构建了专门的数据集, 以支持不同阶段的训练和测试. 为了识别面部是否存在面瘫现象, 我们整理了包含 25000 张正常人脸图像和 20000 张面瘫人脸图像的训练集, 以及一套测试集, 其中包括 2000 张正常人脸图像和 6100 张面瘫人脸图像. 在面瘫具体部位识别的任务中, 仅使用面瘫图像, 其中 13000 张用于训练, 10000 张用于测试.

首先对数据集中的图像进行预处理, 利用 dlib 库提取面部标记, 保存下这些标记用作后续神经网络的输入. 同时, 用检测到的标记裁剪和对齐人脸, 将原始图像裁剪到 256×256 的大小, 以便作为后续网络的输入. 为了提高诊断的准确性, 确保网络集中分析面部区域而不受背景因素的干扰, 我们采取了对原始图像进行裁剪的策略, 仅保留面部关键区域. 通过这种方法, 能够显著提高网络对面瘫状态判断的准确率, 进而提升整体诊断结果的可靠性.

由于我们使用的面瘫数据集未包含 AU 信息

文件, 在处理数据之前, 我们必须自行创建图像对应包含 AU 信息的文件. 在创建这些文件时需确保相同的个体具有相同的 AU 结果, 以确保数据的一致性. 例如, 如果一个人的鼻子区域表现出面瘫症状, 那么所有这个人的图像的 AU9 标签应该标为 1(表示活动), 如果没有表现出症状, 则标为 0(表示无活动), 其他标签依此类推. 对于同一个人的所有图像的标签应保持一致.

模型采用 Pytorch 框架实现, 并在 NVIDIA V100 GPU 上进行训练. 在训练之前, 我们使用 dlib 方法对人脸图像进行识别, 并进行裁剪, 裁剪大小为 256×256. 训练中选择使用 AdamW 优化器, 该优化器结合了自适应学习率调整和权重衰减策略, 旨在提高训练过程中的稳定性并减轻过拟合问题. 模型在单个显卡上进行 15 轮训练, 批处理大小设置为 4, 学习率初始设定为 5×10^{-5}, 权重衰减设置为 5×10^{-4}, 梯度裁剪设为 0.1, Dropout 率设定为 0.1.

3.2　面瘫识别精度对比

我们与近年来基于网络学习方法进行了比较, 本节中主要对使用 YFP 面瘫数据集的其他工作与本工作进行准确率的对比. 与最先进的方法对比, 我们的方法与 improved SSD[58] 进行了对比实验, 表中结果显示, 我们的方法在同一数据集上的准确率提高了 11.5%, 这表明我们的方法能够更好地捕捉人脸图像中的全局特征, 特别是在处理图像中的长距离依赖关系方面, 相较于仅处理局部特征的传统方法有显著优势. 与 MobileNetV2[59] 对比, 由于此方法需要对人脸面部网格进行提取, 这导致识别结果依赖网格提取的质量, 而我们的方法不依赖此步骤, 因此能够避免由于提取过程中的误差对最终识别结果的影响, 表现出更高的鲁棒性. 与 FKA+GAN[60] 对比, 由于此方法是通过 GAN 生成高分辨率图像并对其进行关键点分析, 识别结果受制于 GAN 生成的图像质量以及关键点的选取. 相比之下, 我们的方法避免了 GAN 和关键点分析带来的误差累积, 在处理轻微的图像形变和噪声时展现出更好的鲁棒性. 与 autoFPR[61] 对比, 由于此方法使用的是不包含深度学习的传统机器学习方法. 相比之下, 我们的方法利用大规模预训练模型进行迁移学习, 能够更好地捕捉面部图像中的通用特征, 因此在 YFP 面瘫数据集上取得了更优的表现, 准确率比 autoFPR 高出 2.1%. 表 1 展示了在验证实验中, 展示了我们的方法与现有技术的准确率比较.

从数据中可以看出, 对于面瘫的总体判断, 我们的方法在 YFP 数据集上对面瘫的准确率判断达到了 99.4%, 显著高于其他方法, 这一结果证明了我们的方法在面瘫总体判断上的优越性和先进性.

表 1　在 YFP 和 CK+数据集上面瘫识别准确率与现有方法的比较

Table 1　Comparison of facial paralysis recognition accuracy on the YFP and CK+ datasets with existing methods

Method	Accuracy/%
LSTM[14]	79.29
SENet[57]	86.98
DHN[43]	91.2
PHCNN[48]	87.91
PHCNN-LSTM[48]	94.81
FNPA-RCELM-CCNN[44]	85.5
improved SSD[58]	87.9
MobileNetV2[59]	98.93
FKA+GAN[60]	98.2
autoFPR[61]	97.3
VGG16+Softmax[62]	87.64
ViTAU	99.4

3.3　面瘫区域检测

面瘫区域检测专注于识别面瘫影响的具体面部区域, 而非仅仅判断面瘫等级. 以往的研究多集中于评估面瘫的严重程度, 却往往忽视了对具体发病区域的详细识别. 这种精细化的区域检测方法能为面瘫患者的诊断和后续治疗计划制定提供更为精确的依据.

表 2 详细列出了在 5000 个样本上测试得到的面瘫 AU 的准确率结果, 而图 6 则通过柱状图形式直观展示了最终 AU 准确率的结果. 该表格反映了不同网络模型在面瘫 AU 识别上的性能差异, 显示出各模型的准确率对比. 如表 2 所示, 我们的方法在 YFP 数据集上的对面瘫区域识别的平均准确率达到 81.36%. 优于之前的方法.

从表 2 中可以看出, AU9 和 AU12 的识别上展示了较高的准确率, 说明对判断面瘫具有关键性作用. 特别是对于 AU12, 它主要关联面部的嘴巴区域, 我们的网络模型强调了该区域在面瘫判断中的重要性, 基于 AU 先验知识的面瘫区域识别能够更好区分面部的不同面瘫区域.

3.4　面瘫区域分析与可视化

我们采用了 Grad-CAM(Gradient-weighted class activation mapping)[57] 技术, 用于揭示网络在进行面瘫识别任务时, 对面瘫区域进行着重显示. Grad-CAM 利用网络中特定层的梯度信息, 生成的热图

表 2　对面部区域 AU 检测的准确率

Table 2　Accuracy of AU detection in facial regions　　%

Method	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Average
DRML[63]	47.22	38.12	65.50	84.55	88.50	51.38	37.63	58.99
JÂA-Net[52]	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.69
AU R-CNN[9]	25.90	59.80	55.30	39.80	67.70	77.40	52.60	54.07
ME-GraphAU[64]	67.54	66.84	95.84	59.13	82.93	54.61	61.04	69.70
AUNCE[65]	65.83	65.54	29.08	83.08	98.34	67.07	77.45	69.48
ViTAU	76.08	75.33	76.18	86.51	98.27	78.72	78.47	81.36

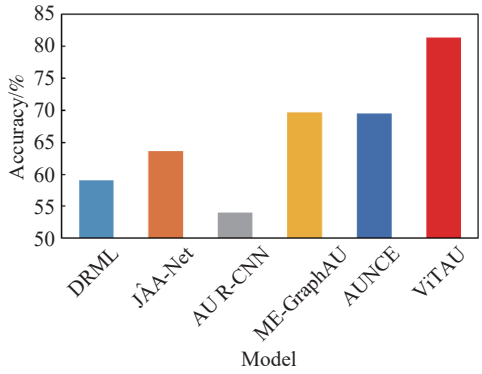

图 6　面部区域 AU 检测的准确率

Fig.6　Accuracy of facial action unit detection

直观地对应于面瘫的不同区域, 图 7 展示了可视化的效果. 这种方法的优势在于它的通用性和直观性, 可以应用于任何卷积神经网络而无需对模型架构进行修改或重新训练.

通过 Grad-CAM 生成的热力图, 我们可以直接观察到模型在面瘫区域识别中学习到的感兴趣的区域. 例如眼睛区域、鼻子脸颊区域和嘴巴区域. 这不仅有助于验证我们的模型是否正确地聚焦于面瘫区域, 而且还提供了对模型决策过程的深入理解. 相反, 如果热力图显示模型关注了与面瘫无关的区域, 就说明了模型学习过程中的偏差或潜在的改进方向.

我们提出的网络通过基于热图的检测展现出强大的性能, 这要归功于从预训练生成模型中提取的可泛化和语义丰富的特征. 通过对面瘫区域的分析, 能够为康复专业人员提供面瘫部位的重要信息, 从而制定更加个性化的康复计划, 使得面部肌肉的锻炼和治疗更加有针对性, 有助于提高康复效率和效果.

3.5　消融实验

为了探索最佳的面瘫检测效果, 我们对比了多种损失函数对诊断准确率的影响. 具体而言, 我们分别采用了交叉熵损失(Cross-entropy loss)、Kullback-Leibler 散度损失(KL divergence loss)、带逻辑斯特回归的二元交叉熵损失(Binary cross-entropy with logits loss)以及均方误差损失(Mean squared error loss, MSE loss)来训练模型. 通过表 3 和图 8 展示的实验结果显示, 使用均方误差损失函数能够实现最高的准确率.

为了寻找效果最好的提取的特征图, 由于 StyleGAN2 生成器引入了渐进式增强的训练策略, 通过逐渐增加生成器和鉴别器的复杂度和分辨率, 使模型逐步学习生成高分辨率图像的能力. 在训练的早期阶段, 生成器主要负责生成低分辨率的图像, 随着训练的进行, 逐渐增加生成器的分辨率, 直到达到目标分辨率. 这里我们从 StyleGAN2 生成器的不同阶段提取不同时期的特征图, 我们

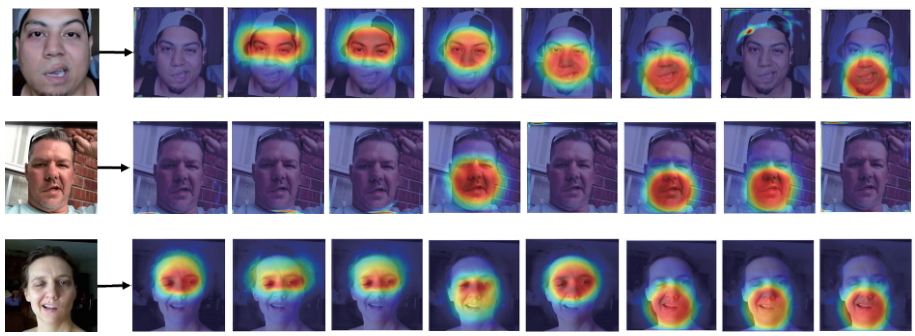

图 7　面瘫区域的可视化

Fig.7　Visualization of facial paralysis areas

表 3　对不同损失函数的消融实验

Table 3　Ablation experiments on different loss functions　%

Loss	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Accuracy
Cross-entropy loss	71.63	72.65	57.80	65.31	86.31	76.53	46.28	68.07
KL divergence loss	66.46	66.45	71.25	82.09	99.35	68.50	79.39	76.21
Binary cross-entropy with logits loss	36.87	61.75	61.31	67.96	97.50	69.33	60.75	65.07
MSE loss	76.08	75.33	76.18	86.51	98.27	78.72	78.47	81.36

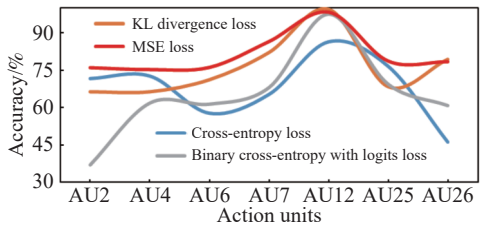

图 8　不同损失函数对判断结果准确率的影响

Fig.8　Impact of different loss functions on the accuracy of judgment results

将此部分分成早期、中期和晚期提取的特征图. 如表 4 所示, 对于提取到的完整的特征信息可以获得最高的准确率结果.

表 4　对不同阶段提取的特征对结果的影响

Table 4　Impact of features extracted from different models and stages on the results

Stage	Accuracy/%
Early	80.41
Middle	79.59
Late	79.59
ViTAU	81.36

4　结论

在本研究中, 我们针对面瘫序列图像, 开发了一种新的基于注意力网络架构, 并结合面部动作单元进行面瘫检测与分析. 通过基于 ViT 的面瘫识别网络, 实现面瘫的精准识别; 并通过集成面部动作单元信息和 Grad-CAM 可视化技术, 对被判断为有面瘫的患者进一步进行面瘫区域的分析. 我们在 YFP 和 CK+数据集上进行的对比实验结果证明了所提出方法各个模块的有效性. 特别是, 引入注意力机制和 AU 信息不仅提高了面瘫识别的准确性, 还为理解和解释模型决策提供了新的视角. 这项工作不仅推进了面瘫识别技术, 也为面瘫分析研究提供了一个有力的框架, 亦可应用于利用深度学习技术进行医学图像分析和诊断的其他领域.

参　考　文　献

[1] Abayomi-Alli O O, Damaševičius R, Maskeliūnas R, et al. Few-shot learning with a novel voronoi tessellation-based image augmentation method for facial palsy detection. *Electronics*, 2021, 10(8): 978

[2] Vletter C V, Burger H L, Alers H, et al. Towards an automatic diagnosis of peripheral and central palsy using machine learning on facial features. *arXiv preprint* (2022-01-27) [2024-05-05]. https://arxiv.org/abs/2201.11852

[3] Jiang C Q, Wu J H, Zhong W Z, et al. Automatic facial paralysis assessment *via* computational image analysis. *J Healthc Eng*, 2020, 2020: 2398542

[4] Guarin D L, Yunusova Y, Taati B, et al. Toward an automatic system for computer-aided assessment in facial palsy. *Facial Plast Surg Aesthetic Med*, 2020, 22(1): 42

[5] Hochreiter J, Hoche E, Janik L, et al. Machine-learning-based detecting of eyelid closure and smiling using surface electromyography of auricular muscles in patients with postparalytic facial synkinesis: A feasibility study. *Diagnostics (Basel)*, 2023, 13(3): 554

[6] Parra-Dominguez G S, Sanchez-Yanez R E, Garcia-Capulin C H. Towards facial gesture recognition in photographs of patients with facial palsy. *Healthcare (Basel)*, 2022, 10(4): 659

[7] Lou J W, Yu H, Wang F Y. A review on automated facial nerve function assessment from visual face capture. *IEEE Trans Neural Syst Rehabil Eng*, 2020, 28(2): 488

[8] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks *via* gradient-based localization // 2017 *IEEE International Conference on Computer Vision* (*ICCV*). Venice, 2017: 618

[9] Ma C, Chen L, Yong J H. AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection. *Neurocomputing*, 2019, 355: 35

[10] Tu C H, Yang C Y, Hsu J Y J. IdenNet: Identity-aware facial action unit detection // 2019 14*th IEEE International Conference on Automatic Face & Gesture Recognition* (*FG* 2019). Lille, 2019: 1

[11] Fu R S, Zhou G H. Automatic evaluation of facial paralysis with transfer learning and improved ResNet34 neural network // 2023 15*th International Conference on Intelligent Human-Machine Systems and Cybernetics* (*IHMSC*). Hangzhou, 2023: 218

[12] Yoshihara H, Seo M, Ngo T H, et al. Automatic feature point

detection using deep convolutional networks for quantitative evaluation of facial paralysis // 2016 9*th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics* (*CISP-BMEI*). Datong, 2016: 811

[13] Li H, Wu K J, Cheng H, et al. Nasolabial Folds Extraction based on Neural Network for the Quantitative Analysis of Facial Paralysis // 2018 2*nd International Conference on Imaging*, *Signal Processing and Communication* (*ICISPC*). Kuala Lumpur. IEEE, 2018: 54

[14] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9(8): 1735

[15] Ikezawa N, Okamoto T, Yoshida Y, et al. Toward an application of automatic evaluation system for central facial palsy using two simple evaluation indices in emergency medicine. *Sci Rep*, 2024, 14: 3429

[16] Vrochidou E, Papić V, Kalampokas T, et al. Automatic facial palsy detection-From mathematical modeling to deep learning. *Axioms*, 2023, 12(12): 1091

[17] Chen C F R, Fan Q F, Panda R. CrossViT: Cross-attention multi-scale vision transformer for image classification // 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Montreal, 2021: 347

[18] Deng P F, Xu K J, Huang H. When CNNs meet vision transformer: A joint framework for remote sensing scene classification. *IEEE Geosci Remote Sens Lett*, 2022, 19: 8020305

[19] Xu K J, Deng P F, Huang H. Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification. *IEEE Trans Geosci Remote Sens*, 2022, 60: 5618715

[20] Sha Z Y, Li J F. MITformer: A multiinstance vision transformer for remote sensing scene classification. *IEEE Geosci Remote Sens Lett*, 2022, 19: 6510305

[21] Xue Z X, Tan X, Yu X C, et al. Deep hierarchical vision transformer for hyperspectral and LiDAR data classification. *IEEE Trans Image Process*, 2964, 31: 3095

[22] Chen Y H, Gu X Y, Liu Z, et al. A fast inference vision transformer for automatic pavement image classification and its visual interpretation method. *Remote Sens*, 2022, 14(8): 1877

[23] Tanzi L, Audisio A, Cirrincione G, et al. Vision Transformer for femur fracture classification. *Injury*, 2022, 53(7): 2625

[24] Chen J Q, Luo T, Wu J H, et al. A Vision Transformer network SeedViT for classification of maize seeds. *J Food Process Eng*, 2022, 45(5): e13998

[25] Wang Z Y, Yang Y D, Yue S L. Air quality classification and measurement based on double output vision transformer. *IEEE Internet Things J*, 2022, 9(21): 20975

[26] Yu S, Ma K, Bi Q, et al. MIL-VT: Multiple instance learning enhanced vision transformer for fundus image classification // *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2021. Strasbourg, 2021: 45

[27] Wang D Y, Wu Z, Yu H Y. TED-Net: Convolution-free T2T vision transformer-based encoder-decoder dilation network for low-dose CT denoising // *Machine Learning in Medical Imaging—MICCAI* 2021. Strasbourg, 2021: 416

[28] Luthra A, Sulakhe H, Mittal T, et al. Eformer: Edge enhancement based transformer for medical image denoising. *arXiv preprint* (2021-09-16) [2024-05-05]. https://arxiv.org/abs/2109.08044

[29] Fan C M, Liu T J, Liu K H. SUNet: Swin transformer UNet for image denoising. *arXiv preprint* (2021-02-08) [2024-05-05]. https://arxiv.org/abs/2202.14009

[30] Chen J N, Lu Y Y, Yu Q H, et al. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint* (2021-02-08) [2024-05-05]. https://arxiv.org/abs/2102.04306

[31] Sagar, A. Vitbis: Vision transformer for biomedical image segmentation // *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID*-19 *and Secure and Privacy-Preserving Machine Learning*. Strasbourg, 2021

[32] Cheng B W, Schwing A, Kirillov A. Per-pixel classification is not all you need for semantic segmentation. *Adv Neural Inf Process Syst*, 2021, 34: 17864

[33] Hatamizadeh A, Xu Z Y, Yang D, et al. UNetFormer: A unified vision transformer model and pre-training framework for 3D medical image segmentation. *arXiv preprint* (2021-02-08) [2024-05-05]. https://arxiv.org/abs/2204.00631

[34] Wang Y K, Ye T Q, Cao L L, et al. Bridged transformer for vision and point cloud 3D object detection // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 12104

[35] Liu Z, Zhang Z, Cao Y, et al. Group-free 3D object detection *via* transformers // 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Montreal, 2021: 2929

[36] Horváth J, Baireddy S, Hao H X, et al. Manipulation detection in satellite images using vision transformer // 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*). Nashville, 2021: 1032

[37] Mishra P, Verk R, Fornasier D, et al. VT-ADL: A vision transformer network for image anomaly detection and localization // 2021 *IEEE* 30*th International Symposium on Industrial Electronics* (*ISIE*). Kyoto, 2021: 1

[38] Lee Y, Kang P. AnoViT: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access*, 2022, 10: 46717

[39] Yuan H C, Cai Z Y, Zhou H, et al. TransAnomaly: Video anomaly detection using video vision transformer. *IEEE Access*, 2021, 9: 123977

[40] Guo Z X, Shen M M, Duan L, et al. Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis *via* deep convolutional neural network // 2017 *IEEE* 14*th International Symposium on Biomedical Imaging* (*ISBI* 2017). Melbourne, 2017: 135

[41] Song A P, Wu Z Y, Ding X H, et al. Neurologist standard classification of facial nerve paralysis with deep neural networks.

Future Internet, 2018, 10(11): 111

[42] Storey G, Jiang R, Keogh S, et al. 3DPalsyNet: A facial palsy grading and motion recognition framework using fully 3D convolutional neural networks. *IEEE Access*, 2019, 7: 121655

[43] Hsu G S J, Kang J H, Huang W F. Deep hierarchical network with line segment learning for quantitative analysis of facial palsy. *IEEE Access*, 2019, 7: 4833

[44] Tan X Y, Yang J, Cao J W. Facial nerve paralysis assessment based on regularized correntropy criterion SSELMvc and cascade CNN // 2021 55*th Asilomar Conference on Signals, Systems, and Computers*. Pacific Grove, 2021: 1043

[45] Sakai T, Seo M, Matsushiro N, et al. Simulation of facial palsy using conditional generative adversarial networks and face shape normalization // 2021 *IEEE* 10*th Global Conference on Consumer Electronics* (*GCCE*). Kyoto, 2021: 793

[46] Yang C L, Kang J L, Xue X T, et al. Automatic Degree Evaluation of Facial Nerve Paralysis Based on Triple-stream Long Short Term Memory // *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*. Xi'an, 2019: 7

[47] Xu P F, Xie F, Su T S, et al. Automatic evaluation of facial nerve paralysis by dual-path LSTM with deep differentiated network. *Neurocomputing*, 2020, 388: 70

[48] Liu X, Xia Y F, Yu H, et al. Region based parallel hierarchy convolutional neural network for automatic facial nerve paralysis evaluation. *IEEE Trans Neural Syst Rehabil Eng*, 2020, 28(10): 2325

[49] Ntinou I, Sanchez E, Bulat A, et al. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Trans Affect Comput*, 2023, 14(1): 436

[50] Akay S, Arica N. Stacking multiple cues for facial action unit detection. *Vis Comput*, 2022, 38(12): 4235

[51] Li X H, Hu S Q, Shi Z G, et al. Micro-expression recognition algorithm based on separate long-term recurrent convolutional network. *Chin J Eng*, 2022, 44(1): 104
(李学翰, 胡四泉, 石志国, 等. 基于 S-LRCN 的微表情识别算法. 工程科学学报, 2022, 44(1): 104)

[52] Shao Z W, Liu Z L, Cai J F, et al. JÂA-net: Joint facial action unit detection and face alignment *via* adaptive attention. *Int J Comput Vis*, 2021, 129(2): 321

[53] Ge X, Jose J M, Xu S, et al. MGRR-Net: multi-level graph relational reasoning network for facial action units detection [J/OL]. *arXiv preprint* (2022-04-04) [2024-05-05]. https://arxiv.org/abs/2204.01349v1

[54] Li D, Xie L, Lu T, et al. Capture of microexpressions based on the entropy of oriented optical flow. *Chin J Eng*, 2017, 39(11): 1727

[55] Richardson E, Alaluf Y, Patashnik O, et al. Encoding in style: A StyleGAN encoder for image-to-image translation // 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Nashville, 2021: 2287

[56] Ekman P, Friesen W V. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. CA: Consulting Psychologists Press, 1978

[57] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell*, 2020, 42(8): 2011

[58] Shi H P, Fan Y Q, Zhang Y, et al. Intelligent Bell facial paralysis assessment: A facial recognition model using improved SSD network. *Sci Rep*, 2024, 14(1): 12763

[59] Baig Z M, Van Der Haar D. Facial paralysis recognition using face mesh-based learning // 12*th International Conference on Pattern Recognition Applications and Methods*. 2023: 881

[60] Naganjaneyulu S, SrinivasaRao N, Sesha S, et al. Efficient composite model to detect facial paralysis using generative adversarial network and facial key points analysis [J/OL]. (2024-03-19) [2024-05-05]. https://doi.org/10.21203/rs.3.rs-4111256/v1

[61] Gogu S R, Sathe S R. AutoFPR: An efficient automatic approach for facial paralysis recognition using facial features. *Int J Artif Intell Tools*, 2023, 32(2): 2340005

[62] Khalifa A N, Ali H R, Abdulazeez Jebur S, et al. Automate facial paralysis detection using vgg architectures. *Int J Curr Innov Adv Res*, 2024: 1. https://www.ijciar.com/index.php/journal/article/view/158

[63] Li Dan, Xie Lun, Lu Ting, et al. Capture of microexpressions based on the entropy of oriented optical flow. *Chin J Eng*, 2017, 39(11): 1727
(李丹, 解仑, 卢婷, 等. 基于光流方向信息熵统计的微表情捕捉. 工程科学学报, 2017, 39(11): 1727)

[64] Luo C, Song S Y, Xie W C, et al. Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition. *arXiv preprint* (2022-05-03) [2024-05-05]. https://arxiv.org/abs/2205.01782v1

[65] Ziqiao Shang, Bin Liu, Fei Teng, Tianrui Li. Learning contrastive feature representations for facial action unit detection. *arXiv preprint* (2024-02-03) [2024-05-05]. https://arxiv.org/abs/2402.06165v3