**中图法分类号**:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-14

**论文引用格式**:Li Kaiyao, Zhang Fan, Li Qing, Peng Bao, Peng Xiaojiang. Collaborative Optimization of ViT Pre-training and Post-training for Facial Expression Recognition[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260250.(黎凯耀, 章帆, 李青, 彭保, 彭小江. 面向人脸表情识别的ViT预训练-后训练协同优化方法[J/OL]. 中国图象图形学报, XXXX:1-14. DOI: 10.11834/jig.260250.)[DOI:10.11834/jig.260250]

# 面向人脸表情识别的ViT预训练–后训练协同优化方法

黎凯耀[1],章帆[2],李青[1],彭保[3],彭小江[1,4*]

1. 深圳技术大学,深圳　518118; 2. 香港中文大学,香港　999077; 3. 深圳信息职业技术大学,深圳　518172; 4. 广东省边缘智能工程技术研究中心,深圳　518118

**摘　要**:**目的**　人脸表情识别(Facial Expression Recognition, FER)是人机交互与情感计算领域的核心技术,在智能监控、教育评估和医疗诊断等方面具有广阔的应用前景。近年来,Vision Transformer(ViT)架构凭借其强大的建模能力以及与多模态大模型良好的兼容性,正逐步取代传统卷积神经网络,成为FER领域的研究热点。然而,现有基于ViT的FER方法在预训练与后训练策略的选择上缺乏系统性探究,不同方法间性能差异显著,且原生ViT模型通常难以直接取得优于CNN或混合架构的识别效果。**方法**　针对上述问题,本文首先系统梳理并实验对比了监督预训练、自监督预训练及多种后训练策略在原生ViT上的适用性。进一步,结合人脸表情识别需捕捉面部细微形变的特性,本文提出一种混合监督对比学习后训练方法,通过在监督对比学习中引入样本混合策略,有效扩展正样本多样性,增强模型对细粒度特征的判别能力。基于系统评估,本文提出掩码图像预训练与混合监督对比学习后训练相结合的原生ViT训练范式MIMIC(Masked Image Pre-training with Mixed Contrastive Learning)。**结果**　在RAF-DB、AffectNet和FERPlus三个基准数据集上的大量实验表明,MIMIC在不引入任何卷积模块或复杂辅助模块的条件下,显著优于现有训练范式,展现出更强的特征学习能力,并随着模型规模增大,该方法未出现明显性能饱和现象,**结论**　MIMIC能够有效提升原生ViT在面部表情识别任务中的特征学习能力,在多个数据集上达到当前最优或相当的性能,验证了预训练与后训练协同优化策略对于提升ViT模型FER性能的有效性。(本文相关代码可见https://github.com/zfkarl/MIMIC/tree/master/Code_for_MIMIC)

**关键词**:人脸表情识别;自监督学习;视觉Transformer;掩码图像建模;对比学习

# Collaborative Optimization of ViT Pre-training and Post-training for Facial Expression Recognition

Li Kaiyao[1], Zhang Fan[2], Li Qing[1], Peng Bao[3], Peng Xiaojiang[1,4*]

*1. Shenzhen Technology University, Shenzhen 518118, China; 2. The Chinese University of Hong Kong, Hong Kong 999077, China; 3. Shenzhen University of Information Technology, Shenzhen 518172, China; 4. Guangdong Provincial Engineering Technology Research Center for Edge Intelligence, Shenzhen 518118, China*

**收稿日期**:2026-05-05;**修回日期**:2026-07-24

*通信作者:彭小江 pengxiaojiang@sztu.edu.cn

**基金项目**:广东省普通高校创新团队项目(2025KCXTD040);国家重点研发计划项目(2025YFE0103200);国家自然科学基金项目(62406200)

**Supported by**:The Innovative Team Project in Guangdong Province (2025KCXTD040); National Key Research and Development Program of China (2025YFE0103200); National Natural Science Foundation of China (62406200)

**Abstract**: **Objective** Facial expression recognition (FER) is a fundamental technology for understanding human emotions and human-computer interaction, possessing extensive application value across diverse fields such as intelligent surveillance, educational assessment, and healthcare diagnostics. Despite significant advancements driven by deep learning, developing robust FER systems remains a highly challenging endeavor. In uncontrolled real-world environments, the high visual and macroscopic similarity between different emotion categories—where distinguishing features are often confined to subtle local deformations—complicates accurate classification. Historically, the dominant paradigm in FER has heavily relied on Convolutional Neural Networks (CNNs) pretrained in a fully supervised manner on massive facial datasets (e. g. , MS-Celeb-1M). This conventional paradigm incurs exorbitant costs due to its strict reliance on millions of manually annotated labels. Recently, Vision Transformer (ViT) architectures have emerged as a powerful alternative due to their superior structural capacity and ease of integration with other modalities. However, a critical gap remains: current ViT-based FER methods often exhibit highly variable performance, and purely "plain ViT" models typically underperform compared to hybrid CNN-ViT architectures. We identify that this discrepancy primarily stems from a lack of systematic investigation into the pre-training and post-training strategies specifically tailored for plain ViTs. Therefore, the primary objective of this study is to systematically explore whether plain ViT architectures are inherently suitable for FER tasks, and to formulate an optimal, highly efficient pre-training and post-training paradigm that mitigates domain gaps and enhances fine-grained feature discrimination without relying on complex auxiliary modules. **Methodology** To address the aforementioned challenges, we first conduct a comprehensive empirical review of prevalent supervised pre-training, self-supervised pre-training, and post-training paradigms under a unified experimental setting. Our preliminary analysis reveals that simply adopting general visual training paradigms (e. g. , supervised pre-training on ImageNet followed by standard cross-entropy fine-tuning) fails to adequately capture the subtle facial deformations critical for FER. Consequently, we propose an innovative, unified two-stage training framework termed Masked Image Pre-training with Mixed Contrastive Learning (MIMIC) designed exclusively for plain ViT architectures. The first stage focuses on self-supervised representation learning to eliminate annotation reliance. We employ a plain ViT encoder and pre-train it on a mid-scale general image dataset (ImageNet-1K) utilizing a Masked Image Modeling (MIM) task. By randomly masking a high proportion of image patches and optimizing a reconstruction loss to recover the missing pixels, the network is forced to learn highly fine-grained and discriminative visual representations. This self-supervised approach successfully shifts the pre-training data source from domain-specific face datasets to general images, significantly reducing manual annotation costs. The second stage—post-training (fine-tuning)—is specifically designed to bridge the severe domain gap between general images and facial expressions, while simultaneously addressing the high inter-class similarity of human emotions. After discarding the MIM decoder, we adapt the pre-trained ViT encoder for the downstream FER task. Standard cross-entropy loss or traditional supervised contrastive learning (SCL) tends to apply rigid, "hard" positive/negative sample divisions. This rigid grouping inadvertently ignores the valuable discriminative information hidden in samples from different emotion classes that share high visual similarity (e. g. , "happiness" and "neutrality"). To overcome this, we introduce a novel Mixed Supervised Contrastive Learning (MSCL) strategy. Based on the standard SCL framework, MSCL incorporates a data mixing strategy (combining Mixup and CutMix techniques with blending coefficients sampled from a Beta distribution) to synthesize augmented images and their corresponding blended labels. These mixed samples are projected into a dense hidden space. Depending on the distance between the synthesized labels and the anchor labels relative to a predefined threshold, the mixed samples are dynamically categorized as soft positive or negative pairs. This substantially expands the diversity of positive samples and enhances the model′s capacity to scrutinize subtle facial nuances. The final optimization objective is a balanced combination of the standard classification loss and the MSCL loss. **Results** Extensive experiments were conducted on three widely recognized and highly challenging benchmark datasets: RAF-DB, FERPlus, and AffectNet-7. To ensure fair evaluation, a plain ViT-B/16 was employed as the primary backbone. The experimental results conclusively demonstrate that the proposed MIMIC framework consistently exhibits superior feature learning capabilities. When evaluated against strong baselines—including fully supervised approaches, standard contrastive learning methods such as MoCo V3, and conventional masked image modeling techniques like MAE and LocalMIM—MIMIC achieved the highest accuracy across all three datasets. Furthermore, the MIMIC paradigm demonstrates remarkable scalability. When scaling the model size to ViT-L/16, no performance saturation was

observed. Specifically, without utilizing any CNN feature extraction modules, additional parameters, or complex relation-aware networks, our plain ViT-L/16 model achieved an outstanding accuracy of 91.26% on the RAF-DB dataset and 91.24% on the FERPlus dataset. These results successfully surpass a wide array of current state-of-the-art hybrid methods, strongly underscoring the untapped potential of pure attention-based architectures in FER. Comprehensive ablation studies further validated our specific architectural and training design choices. For the contrastive loss optimization, utilizing a dense projection head proved significantly more effective than a linear one. Moreover, for bridging cross-domain discrepancies, Global Average Pooling (GAP) was found to be superior to the standard class token. Interestingly, pre-training on the general ImageNet-1K dataset yielded better downstream performance than pre-training on the domain-specific MS-Celeb-1M face dataset, indicating that the superior image diversity of general datasets facilitates more robust representation learning for MIM tasks. Finally, parameter sensitivity analysis confirmed optimal stability at a batch size of 64 and a temperature parameter of 0.07, while t-SNE visualizations of global features intuitively confirmed that the MIMIC strategy achieves exceptionally clear class separation and significantly alleviates instance-level identity confusion compared to traditional supervised training. **Conclusion** This study thoroughly validates the innate suitability and immense potential of plain ViT architectures for facial expression recognition. By systematically analyzing the limitations of existing training paradigms, we successfully formulate the MIMIC methodology. By substituting supervised pre-training on massive face datasets with self-supervised masked image pre-training on general datasets, we significantly alleviate the reliance on costly manual annotations. Concurrently, the introduction of the mixed supervised contrastive post-training strategy effectively mitigates severe domain discrepancies and solves the challenge of inter-class similarity by expanding positive sample diversity. The proposed paradigm proves that combining self-supervised general image pre-training with advanced contrastive post-training is a highly viable, robust, and cost-effective pathway for constructing high-performing, pure-attention-based emotion recognition systems. Future research will focus on extending this highly effective training paradigm to broader affective computing domains, such as multimodal emotion recognition and video-based sentiment analysis.

**Key words**: Facial expression recognition; Self-supervised learning; Vision Transformer; Masked image modeling; Contrastive learning

论文引用格式:[DOI:10.11834/jig.260250]

## 0 引　言

通过面部表情表达情绪是人类最基本的沟通方式之一。自动人脸表情识别在监控(Clavel等,2008)、教育(Yang等,2018)、医疗(Pioggia等,2005)和市场营销(Ren等,2012)等领域具有广泛应用价值。然而,在非受控环境下,不同类别表情在整体外观上具有较高相似性,其差异往往集中于少数面部区域的细微形变;同时,有限的标注数据以及光照、姿态、分辨率和身份等非表情因素也容易导致模型过拟合或学习到虚假关联(Kopalidis等,2024)。因此,如何获得兼具细粒度判别能力和跨场景泛化能力的面部表征,仍是人脸表情识别的关键问题。

人脸表情识别经历了从人工特征到深度表征的发展。早期方法主要提取局部纹理、全局几何或二者的组合,并在CK+(Lucey等,2010)、Oulu-CASIA(Zhao等,2011)等实验室数据集上进行评估。随着大规模非受控数据集的出现(Barsoum等,2016;Mollahosseini等,2017),基于卷积神经网络(convolutional neural network,CNN)的深度方法逐渐成为主流,并通过局部特征学习(Li等,2017)、注意力机制(Wang等,2020)、遮挡处理(Li等,2018)和噪声标签学习(She等,2021;Zeng等,2018)提升识别性能。

近年来,视觉Transformer(vision transformer,ViT)凭借全局关系建模能力和多模态扩展潜力,逐渐被用于人脸表情识别。现有研究多通过结构设计弥补ViT对局部细节建模的不足,例如在CNN特征之上利用Transformer挖掘局部判别表示(Xue等,2021),采用双流局部注意力联合建模局部表情与全局上下文(Tian等,2024),或结合多尺度特征与patch dropping增强遮挡场景下的鲁棒性(Li等,2024)。另有研究借助大规模视觉语言模型特征改善跨数据集泛化(Zhang等,2024)。这些方法验证了Transformer架构的潜力,但通常依赖CNN、局部注意力、多尺度融合或外部大模型等额外模块。相比之下,原生ViT缺少卷积归纳偏置,又对数据规模

和训练方式较为敏感，其在表情识别中的预训练与后训练策略尚缺少统一设置下的系统比较，真实潜力仍未得到充分揭示。

训练范式是影响ViT迁移性能的另一重要因素。纯视觉自监督学习主要包括对比学习和掩码图像建模两类：MoCo（He等，2020）、SimCLR（Chen等，2020）等对比学习方法通过正负样本关系学习实例级表征，但在表情类别较少时，实例级负样本中可能包含语义相同的样本；掩码图像建模则通过重建被遮挡的图像内容学习可迁移表征（He等，2022），避免显式划分负样本。已有研究将掩码重建预训练用于表情识别，并验证了无标注人脸数据、预训练数据来源及训练目标对下游性能的影响（Song等，2024；Li等，2025）。然而，获得预训练权重后，常规方法通常仅使用交叉熵进行全参数微调，难以充分适应细粒度分类需求。监督对比学习利用标签构造类内正样本和类间负样本，可增强类别可分性（Khosla等，2020；Gunel等，2021；Moukafih等，2023），但其硬性的正负样本划分难以描述不同表情之间的连续过渡。Liu等（2025）通过表情热图邻域扩展自监督预训练阶段的正样本，而如何在下游后训练阶段构造更丰富、更连续的监督关系，仍有待研究。

针对上述问题，本文首先在统一实验设置下系统比较监督预训练、自监督对比学习预训练和掩码图像预训练，以及与其组合的后训练策略，分析不同训练范式对原生ViT表情识别性能的影响。实验发现，直接沿用通用视觉任务中常见的ImageNet监督预训练与下游微调范式（Steiner等，2022），难以使原生ViT在以细微差异为主的表情识别任务中稳定获得优势，说明有必要协同设计适合该任务的预训练目标与后训练方法。

基于以上分析，本文提出掩码图像预训练与混合监督对比学习后训练范式（masked image pre-training with mixed contrastive learning，MIMIC）。该方法首先在通用图像数据上进行掩码图像重建式自监督预训练，以学习具有良好迁移性的视觉表征；随后在表情识别下游阶段引入混合监督对比分支，通过样本混合及其连续标签扩展正样本的多样性，并在新的投影空间中优化混合样本的深度特征，从而缓解传统监督对比学习中正负样本硬性划分的问题，增强模型对细微表情差异的判别能力。MIMIC聚焦原生ViT的训练策略，无需引入CNN特征模块或复杂辅助结构，且推理阶段不增加额外结构负担。

Anchor
Positive
Negative
Negative
Negative
Negative
Anchor
Positive
Positive
Negative
Negative
Negative
Anchor
Positive
Positive
Negative
Positive
Negative
Happiness
Happiness
Sadness
Surprise
Anger

（a）自监督对比学习 （b）监督对比学习 （c）混合监督对比学习

（a）Self-supervised contrastive learning （b）Supervise contrastive learning （c）Mixed supervision contrast learning

图1 三种对比学习的比较

Fig. 1 Comparison of three formats of contrastive learning

本文的主要贡献如下：1）在统一设置下系统比较监督预训练、自监督预训练及不同后训练策略，分析原生ViT在人脸表情识别中的性能差异，为训练范式选择提供实验依据；2）提出混合监督对比学习后训练方法，利用混合样本及其连续标签构造更丰富的监督关系，提升原生ViT对细粒度表情差异的判别能力；3）提出MIMIC原生ViT表情识别训练范式，并在3个基准数据集上开展广泛实验。ViT-B/16在RAF-DB和FERPlus上分别取得91.26%和91.24%的准确率，且模型规模增大时性能仍有提升

空间。

## 1 方　法

本节首先介绍已有深度学习模型训练方法，包括大规模监督训练、自监督预训练等常用方法，然后结合人脸表情特征学习的性质进一步阐述了本文提出的掩码图像预训练与混合监督对比学习后训练新训练方法。

### 1.1 常用深度学习训练方法

**大规模类别监督训练**。以往深度面部表情识别方法的成功很大程度上归功于在大规模自然图像或人脸识别数据集上预训练所得的权重。在传统监督预训练中，训练数据表示为实例与标签对$(x_i, y_i)$，网络参数通常通过最小化分类交叉熵损失进行更新。如下所示：

$$\mathcal{L}_{cls}^{p} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \log\left(p_c(x_i,\theta)\right), \#(1)$$

式中，$N$表示训练样本数量，$C$表示类别数量，$y_i^c$表示第$i$个样本属于第$c$类的真实标签，$p_c(x_i,\theta)$表示模型参数$\theta$下样本属于第$c$类的预测概率。在ImageNet数据集上利用类别标签训练模型为该类方法典型的例子。这种范式的目标是学习能够识别通用图像的强判别表征。由于其依赖大规模标注数据，训练成本通常较高。

**自监督对比学习预训练**。给定包含$B$个样本$\{x_k\}_{k=1}^{B}$的批次，每张图像经数据增强后得到$2B$个增强视图，增强过程定义为$\tilde{x} = Aug(x)$。增强样本经编码器映射为表示$r = Enc(\tilde{x})$，再经投影头映射为投影向量$z = Proj(r)$。在$2B$个样本中，选定一个锚点后，其余$2B-1$个样本构成正负样本候选集。

自监督对比学习旨在无人工标注干预的前提下，通过数据增广的方式构建正负样本对在特征空间中进行表征学习，如图1(a)所示。给定多视图批次数据，对比学习主要依赖于数据增强技术(例如随机裁剪、颜色抖动等)为同一张图像生成不同的增强视图。属于同一图像的不同视图被定义为正样本对，来自不同图像的视图则被划分为负样本对。模型通过最小化InfoNCE损失，在潜在空间中拉近正样本的距离，同时推远负样本的距离：

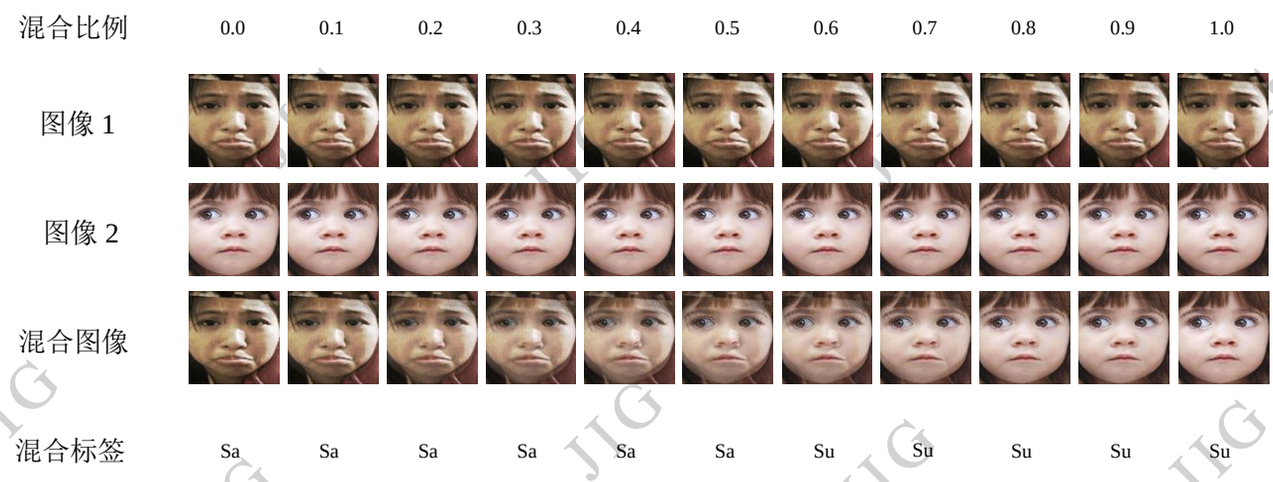



图2 混合策略示意图。Sa和Su分别表示悲伤和惊讶

Fig. 2 An illustration of mixing strategy. Sa and Su denotes Sadness and Surprise, respectively.

$$\mathcal{L}_{sscl} = -\sum_{i=1}^{2B}\log\left(\frac{\exp\left(z_i \cdot z_{j(i)}/\tau\right)}{\sum_{a\in A(i)}\exp\left(z_i \cdot z_a/\tau\right)}\right) \#(2)$$

式中$z_{j(i)}$为锚点$i$对应的正样本特征投影，$A(i)$为除了$z_i$的其余样本，$\tau$为调节对负样本惩罚粒度的温度参数。典型的案例为在ImageNet数据集上使用SimCLR进行预训练(Chen等，2020)。该范式具备区分样本个体差异的强大能力，但在处理类别数较少的面部表情数据时，随机采样的负样本候选集中往往会包含大量与锚点同类的样本，极易引发负样本划分错误的现象。

**监督对比学习训练**。针对自监督对比学习容易

将同类样本错误推远的问题，监督对比学习（SCL）引入了真实类别标签作为先验知识来定义正样本，如图1（b）所示。SCL放宽了正样本的定义标准，其判断依据转换为标签一致性，即将批次内与锚点标签相同的所有样本均视为正样本。通过这一调整，监督对比学习能够有效引导网络最大化类内特征相似度并最小化类间特征相似度：

$$\mathcal{L}_{scl}=-\sum_{i=1}^{2B}\frac{1}{\left|P(i)\right|}\sum_{p\in P(i)}\log\left(\frac{\exp\left(z_i\cdot z_p/\tau\right)}{\sum_{a\in A(i)}\exp\left(z_i\cdot z_a/\tau\right)}\right)\#(3)$$

式中$P(i)$代表批次内与锚点$i$标签相同的所有样本集合。尽管SCL从根本上缓解了负样本选择偏差，但面对表情识别任务中高度相似的跨类别样本（例如局部微表情一致的“高兴”与“中性”样本），其非此即彼的硬性正负样本划分逻辑依然略显僵化，未能充分利用跨类样本的细粒度特征。

**掩码图像预训练**。掩码图像建模（MIM）作为ViT架构下的前沿自监督范式，为网络设计了更为底层的像素级前置任务。该方法通过随机遮挡输入图像中极高比例的图像块（patch），强制编码器结合未被遮挡的有限上下文信息来推断并重建缺失区域。网络参数通过最小化重建预测值与真实信号（如像素值或特征标记）之间的均方误差进行更新：

$$\mathcal{L}_{rec}^{p}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}\left(x_i^d-p_d\left(x_i,\theta\right)\right)^2\#(4)$$

式中$\boldsymbol{D}$表示图像被遮挡的patch数量，$p_d\left(x_i,\theta\right)$表示重建的patch，$\theta$代表网络（通常为ViT）参数。该类策略典型案例为MAE（Xie等，2022）。相较于以输出单一概率分布为导向的全局分类任务，图像重建要求模型深刻理解局部纹理细节与全局空间拓扑结构。该过程完全摆脱了对类别标签的依赖，为利用海量通用无标签数据进行高效预训练铺平了道路。

### 1.2 MIMIC人脸表情识别方法

通过以上分析，本文提出一套面向表情识别的ViT预训练-后训练新方法，即掩码图像预训练与混合监督对比学习后训练方法。

**掩码图像预训练（MIM）与表情识别特征分析。**面部表情识别高度依赖于对局部区域（如眼角、嘴角等面部动作单元）微小肌肉形变的精细判别。选择掩码图像预训练模型对于表情识别特征的好处包括：1）常规监督预训练的信号仅为全局图像级标签，这种粗维度的指导难以促使骨干网络充分挖掘局部精细特征。MIM的密集重建任务迫使ViT编码器在像素级别上捕捉细微的局部纹理与空间上下文关联，这与表情识别强烈关注微小形变的刚性需求完美对应；2）MIM训练方式为无监督模式，直觉上讲，它与带监督的后训练互补性比监督预训练要强；3）MIM训练降低领域标注依赖并提升表征泛化能力。该策略解除了人脸表情识别模型对人脸识别数据集人工标注的依赖。

**混合监督对比学习**。为了应对面部表情识别中不同情绪类别间存在的极高视觉相似性，并解决传统后训练策略中硬性正负样本划分过于僵化的问题，本文提出一种混合监督对比学习方法（Mixed Supervised Contrastive Learning，MSCL）进行后训练阶段的表征优化，如图1（c）所示。传统的微调策略往往仅依赖交叉熵损失，缺乏对潜在空间中类间距离的有效几何度量。即便引入标准的监督对比学习，由于其严格的“同类即正、异类即负”划分规则，仍然会遗漏跨类别相似样本中蕴藏的丰富判别信息。

为了突破这一限制，本文在后训练阶段引入基于数据增强的样本混合策略（结合Mixup与CutMix技术，如图2所示）。算法在每次迭代时从Beta分布中采样混合系数$\lambda$，对输入图像及标签对进行插值与拼接：

$$\lambda\sim Beta\left(\alpha,\beta\right)\#(5)$$

式中$\alpha$和$\beta$默认均为2，混合操作定义为：

$$Mix\left(\tilde{x}_a,\tilde{x}_b\right)=\lambda\tilde{x}_a+(1-\lambda)\tilde{x}_b\#(6)$$

根据混合结果，样本被划分为以下三种情况：

情况1：给定锚点标签，如果两个样本与锚点标签相同，混合样本为锚点标签的正样本。

情况2：给定锚点标签，至少一个样本标签与锚点不同，但混合标签与锚点标签距离低于阈值$t$，混合样本为锚点标签的正样本。

情况3：给定锚点标签，至少一个样本标签与锚点不同，混合标签距离超过阈值$t$，混合样本为锚点标签的负样本。

这种混合操作产生的新样本不仅极大地扩充了特征流形的分布密度，还生成了带有连续语义的软标签。根据混合标签与锚点标签之间的相对距离阈值，MSCL动态地将这些混合样本重新归类为软正

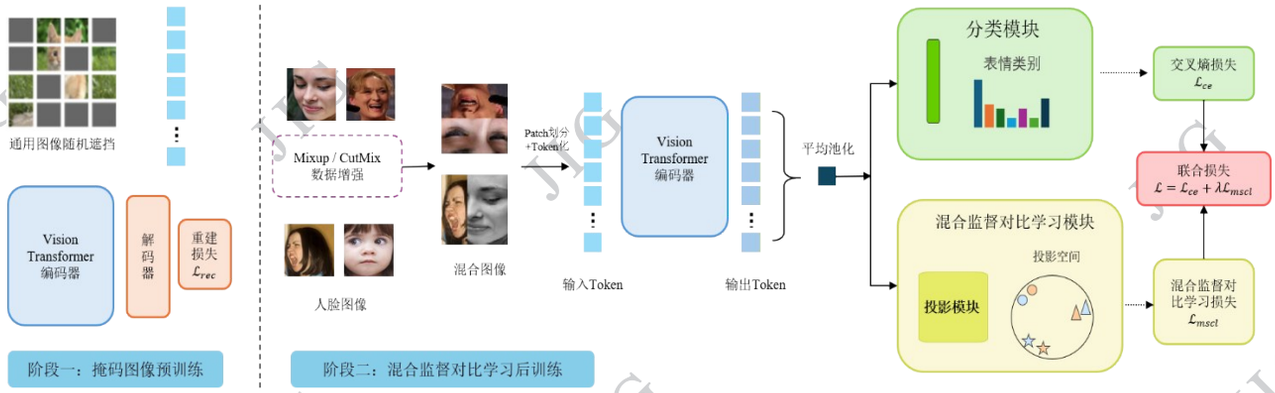



图 3　MIMIC 方法整体框架图。

Fig. 3　The framework of our MIMIC method.

样本或负样本。这一机制显著扩展了对比学习中正样本的拓扑多样性，迫使模型跳出宏观轮廓的捷径，转而敏锐地甄别决定面部表情的核心局部特征。混合监督对比损失表述如下：

$$\mathcal{L}_{mscl}=-\sum_{i=1}^{2B}\frac{1}{\left|P(i)\right|}\sum_{p\in P(i)}\log\frac{\exp\left(z_i\cdot z_p/\tau\right)}{\sum_{a,b\in A(i)}\exp\left(z_i\cdot z_{a,b}/\tau\right)}\#(7)$$

$$z_{a,b}=Proj\left(Enc\left(Mix\left(\tilde{x}_a,\tilde{x}_b\right)\right)\right)\#(8)$$

在端到端后训练框架下，最终的总体优化目标将标准的全局分类损失与混合监督对比损失予以动态加权融合：

$$\mathcal{L}_{all}^{f}=\mathcal{L}_{cls}^{f}+w\mathcal{L}_{Mscl}^{f}\#(8)$$

式中，$w$ 表示分类损失与 MSCL 损失之间的权衡系数。通过此种联合优化机制，网络在保留强大判别性分类性能的同时，在潜在特征空间中塑造出具有更高内聚度与更清晰决策边界的视觉表征。

图4预训练方法的比较。

Fig. 4 Comparison with other pre-training methods.

整体框架如图3所示，本文方法整体采用两阶段训练流程。首先，在通用图像数据上通过掩码图像建模对ViT编码器进行自监督预训练；随后，丢弃解码器，仅保留编码器作为下游任务的初始化模型。在微调阶段，将经过数据增强与混合策略处理的表情图像（及混合样本）输入编码器，对得到的各个编码进行平均池化，并在其输出特征上同时接入分类模块与混合监督对比学习模块。其中分类模块是一个线性分类层，用于完成表情识别任务；混合监督对比学习模块中包含一个投影模块，用于构建对比学习的表征，实验部分将对投影模块进行不同结构评估。

## 2　实　验

### 2.1　数据集

**RAF-DB**（Li 等，2017）包含约 3 万张面部图像，由 315 名标注员标注了基本或复合表情。本文仅使用其中七种基本表情（中性、高兴、惊讶、悲伤、厌恶和恐惧）的子集，共包含 12,271 张训练图像和 3,068 张测试图像。

**FERPlus**（Barsoum 等，2016）是 FER2013 数据集（源自 ICML 2013 挑战赛）的改进版本，通过众包方式对原始图像重新标注以纠正错误标签，并在原有基础上增加了“轻蔑”类别，共涵盖 8 类表情，包含 28,709 张训练图像、3,589 张验证图像和 3,589 张测试图像。

**AffectNet**（Mollahosseini 等，2017）是目前规模最大的同时提供分类标签与效价-唤醒度标注的数据集，通过搜索引擎关键词爬取超过一百万张图像，其中约 440,000 张经人工标注了八种表情，类别分布存在明显不平衡。按照主流研究惯例，AffectNet 分为 AffectNet-7（不含轻蔑）和 AffectNet-8 两个版本。本文采用 AffectNet-7 子集，包含 3,667 张训练图像和 500 张测试图像。

### 2.2　实现细节

为确保公平比较，实验默认以 ViT-B/16 作为骨干网络，以 ImageNet-1K 作为预训练数据集。预训练阶段的掩码策略和训练轮数遵循 He 等人（2022）MAE 论文中的设定，混合标签与锚点标签距离阈值 t 设为 0.5，$w$ 默认设置为 0.1。所有输入图像经对齐

后调整至224×224分辨率，遵循Xie等人(2022)文中的数据增强策略，对每张图像生成两个裁剪视图。三个数据集统一采用Adam优化器，权重衰减为5×$10^{-4}$，初始学习率为1×$10^{-4}$，层衰减率为0.65。针对AffectNet的类别不平衡问题，采用平衡采样策略。微调阶段，RAF-DB和FERPlus训练50个轮次，AffectNet训练10个轮次。所有实验基于PyTorch框架，在NVIDIA A100 GPU上完成。

### 2.3 与基线方法的比较

本文选取基于三种训练策略的四种方法作为基线，均以ViT-B/16为骨干网络并搭配简单分类头，

涵盖全监督方法、对比学习方法(MoCo V3(Chen等，2021))以及掩码图像建模方法(MAE(He等，2022))和LocalMIM(Wang等，2023))。图4所有方法均在ImageNet-1K上预训练，然后在对应FER数据集上进行微调。除MIMIC外，其他方法均只用交叉熵损失微调后训练，其结果显示，本方法在三个数据集上均优于所有基线。在RAF-DB和AffectNet上，对比学习方法和MIM方法均超越监督预训练策略。本方法结合了MIM的强表征学习能力与对比学习的领域自适应能力，在这两个数据集上取得了最优性能。然而在FERPlus上，LocalMIM和MoCo V3的表现甚至不及监督预训练，反映出FERPlus与ImageNet之间的领域差异显著。对比微调有效弥合了这一差距，使本方法超越了全部基线方法，验证了其在表征空间中校正类间相似性的有效性。

### 2.4 与最先进方法的比较

表1汇总了现有方法在RAF-DB、FERPlus和AffectNet上报告的测试结果，并依据模型架构将其大致划分为CNN、CNN-ViT混合架构和纯ViT三类。需要指出的是，各方法在骨干网络、预训练数据和训练策略等方面存在差异，因此表中结果主要用于展示本文方法在现有研究中的相对位置，而不作为严格控制变量下的公平比较。本文对预训练与后训练策略有效性的验证，主要依据图4和表2中统一实验设置下的对比与消融结果。

在传统CNN方法中，多数工作采用ResNet系列作为骨干网络，并依赖在MS-Celeb-1M的等大规模人脸数据集上的监督预训练。这类方法在RAF-DB和FERPlus等数据集上取得了稳定性能，然而模型能力主要受限于卷积结构的局部建模机制。

为进一步提升性能，部分研究引入CNN-ViT混合架构，通过卷积提取局部特征，再结合Transformer建模全局关系，这类方法在一定程度上融合了两种架构的优势，取得了较高准确率，但其性能提升依赖额外结构设计与模块堆叠，模型复杂度较高。

相比之下，纯ViT方法由于缺乏卷积归纳偏置，在面部表情识别任务中长期处于劣势，通常需要借助专用人脸数据或复杂训练策略才能取得可接受性能。例如，部分基于ViT-B/16的工作在标准ImageNet预训练条件下表现不及CNN方法，这也是当前研究普遍倾向采用混合架构的重要原因。

本文结果表明，在不引入任何卷积模块或复杂结构的前提下，通过针对性的预训练与后训练策略设计，原生ViT同样可以获得具有竞争力甚至领先的性能。具体而言，基于ImageNet-1K自监督预训练的ViT-L/16在RAF-DB和FERPlus数据集上分别达到91.26%和91.24%，超过了多数CNN及混合方法。这一结果说明，限制纯ViT性能的关键因素并非模型结构本身，而在于训练范式的设计。

综上，本文从训练策略角度重新审视了不同架构在面部表情识别任务中的表现，证明了在合理的预训练与后训练设计下，纯ViT可以摆脱对卷积结构的依赖，成为一种具有潜力的统一建模方案。

### 2.5 消融实验

**预训练与后训练策略评估。**以ImageNet-1K上预训练的ViT-B/16为基线，掩码图像建模的超参数遵循MAE的默认配置，实验结果如表2所示。首先，固定微调方式后，对比第一至三行与第四至六行可以看出，掩码图像预训练在三个数据集上均优于监督预训练，且无需任何标签监督。其次，在两种预训练设定下，引入SCL和MSCL均能进一步提升性能，且本文提出的MSCL对比微调策略优于单纯基于交叉熵损失和基于交叉熵联合SCL的微调策略。最后，结合两个阶段(表2最后一行)，本方法在ViT架构下达到最优性能。

**投影模块评估。**投影头之前还是之后的潜空间更适合微调，是对比学习中的常见问题。通常认为投影头之后的空间具有更强且更稳定的表征能力。投影头分为两类：线性投影头(单层线性层)和稠密投影头(线性-ReLU-线性结构)。表3结果表明，线性投影头表现最差；稠密投影头相较于不使用投影头能带来稳定的性能提升。

**预训练数据评估。**考虑到面部表情识别数据集

**表 1　现有方法在测试集上的实验结果（%）对比，AffectNet默认为7类评测结果。**

**Table 1　Experimental results (%) of existing methods on the test dataset.**

| 方法 | 骨干网络 | 预训练数据集 | RAF-DB | FERPlus | AffectNet |
|---|---|---|---|---|---|
| DDA（Farzaneh和Qi等，2020） | ResNet-18 | MS-Celeb-1M | 86.90 | – | 62.34 |
| SCN（Wang等，2020） | ResNet-18 | MS-Celeb-1M | 87.03 | 88.01 | 63.40 |
| SCAN（Gera 和 Balasubramanian 等，2021） | ResNet-50 | MS-Celeb-1M | 89.02 | 89.42 | 65.14 |
| DMUE（She等，2021） | ResNet-18 | MS-Celeb-1M | 88.76 | 88.64 | – |
| RUL（Zhang等，2021） | ResNet-18 | MS-Celeb-1M | 88.98 | 88.75 | 61.43 |
| DACL（Farzaneh和Qi等，2021） | ResNet-18 | MS-Celeb-1M | 87.78 | 88.39 | 65.20 |
| MA-Net（Zhao等，2021） | ResNet-18 | MS-Celeb-1M | 88.40 | – | 64.53 |
| EAC（Zhang等，2022） | ResNet-18 | MS-Celeb-1M | 89.99 | 89.64 | 65.32 |
| RC（Feng等，2020） | ResNet-18 | MS-Celeb-1M | 88.92 | 88.53 | 64.87 |
| LW（Wen等，2021） | ResNet-18 | MS-Celeb-1M | 88.70 | 88.10 | 64.53 |
| PICO（Wang等，2022） | ResNet-18 | MS-Celeb-1M | 88.53 | 87.93 | 64.35 |
| Twinned-Att（Devasena 和 Vidhya，2025） | ResNet-34 | – | 86.92 | 85.64 | – |
| LFNSB（Chen等，2024） | MFN | MS-Celeb-1M | 91.07 | – | 66.57 |
| Ada-DF（Liu等，2023） | ResNet-50 | MS-Celeb-1M | 90.35 | – | 65.34 |
| FerNeXt（El-Khashab等，2023） | ConvNeXt | ImageNet-1K | 88.56 | – | 64.77 |
| MGFA(Yu等，2025) | ResNet-50 | ImageNet-1K | 88.90 | – | 60.60 |
| HFE-Net（Song等，2025） | EfficientNetV2 | ImageNet-21K | 87.29 | 88.72 | 58.55 |
| APVIT（Xue等，2022） | ResNet-50+ViT | MS-Celeb-1M | 90.53 | 89.35 | 63.92 |
| MVT（Li等，2021） | DeIT-S/16 | ImageNet-1K | 88.62 | 89.22 | 64.57 |
| VTFF（Ma等，2021） | ResNet-18+ViT-B/32 | MS-Celeb-1M | 88.14 | 88.81 | 64.80 |
| MRAN（Chen等，2023） | Hybrid ViT | MS-Celeb-1M | 90.03 | 89.59 | 66.31 |
| FG-AGR（Li等，2023） | Hybrid ViT | MS-Celeb-1M | 90.81 | 91.09 | 64.91 |
| MTAC（Liu等，2022） | Swin-S | ImageNet-1K | 90.52 | 90.4 | – |
| MAE-ViT（Li等，2025） | ViT-B/16 | AffectNet-270K | 91.07 | 90.18 | **66.09** |
| MAE-ViT（Li等，2025） | ViT-B/16 | ImageNet-1K | 88.49 | 88.29 | – |
| SSFER（Song等，2024） | ViT-B/16 | CASIA-WebFace | 88.23 | 85.82 | 64.02 |
| Pazandeh和Fatemizadeh，2025 | ViT-B/16 | ImageNet-1K | 88.14 | 89.71 | – |
| MIMIC | ViT-B/16 | ImageNet-1K | 89.02 | 89.74 | 65.35 |
| MIMIC | ViT-L/16 | ImageNet-1K | **91.26** | **91.24** | 65.62 |

注：加粗字体表示各列最优结果。“–”表示对应方法未提供相关信息。表中结果来自不同文献报道。由于各方法在骨干网络、预训练数据、训练策略和附加模块方面存在差异，本表主要用于展示不同方法的文献结果和方法定位，不作为严格控制变量下的公平比较。本文关于预训练与后训练策略有效性的主要结论来自图4和表2中的统一设置实验。

表 2　预训练与微调阶段所提策略的消融研究

Table 2　Ablation on the proposed strategies for pre-training and fine-tuning stages

| 预训练 | | 微调 | | | RAF-DB | FERPlus | AffectNet |
|---|---|---|---|---|---|---|---|
| Supervised | MIM | CE | SCL | MSCL | | | |
| ✓ | | ✓ | | | 87.33 | 87.91 | 60.05 |
| ✓ | | ✓ | ✓ | | 87.52 ↑ 0.19 | 88.06 ↑ 0.15 | 63.10 ↑ 3.05 |
| ✓ | | ✓ | | ✓ | 87.60 ↑ 0.27 | 88.21 ↑ 0.30 | 63.62 ↑ 3.57 |
| | ✓ | ✓ | | | 88.14 ↑ 0.81 | 88.42 ↑ 0.51 | 65.05 ↑ 5.00 |
| | ✓ | ✓ | ✓ | | 88.87 ↑ 1.54 | 89.43 ↑ 1.52 | 65.26 ↑ 5.21 |
| | ✓ | ✓ | | ✓ | 89.02 ↑ 1.69 | 89.74 ↑ 1.83 | 65.35 ↑ 5.30 |

注：CE、SCL 和 MSCL 分别表示交叉熵损失、监督对比损失和混合监督对比损失

≈

与人脸数据集在图像分布上较为接近，在大规模人脸数据集上进行 MIM 预训练的效果同样值得关注。然而表 4 结果显示，在 MS-Celeb-1M 上预训练的效果并未优于在 ImageNet-1K 上的预训练。本文推断，原因在于人脸识别数据集图像多样性不足，导致网络易收敛至平凡解，文献 MAE-ViT（Li 等，2025）发现在大规模人脸表情数据上面进行掩码预训练效果是最好。为了公平和通用，本文在其它实验中默认采用先在 ImageNet-1K 上预训练、再在面部表情识别数据集上微调的两阶段流程。

表 3　投影模块评估

Table 3　Evaluation on Projection Module

| 投影头 | RAF-DB | FERPlus | AffectNet |
|---|---|---|---|
| 无 | 87.68 | 88.65 | 65.14 |
| 线性 | 87.18 ↓ 0.50 | 88.43 ↓ 0.22 | 64.62 ↓ 0.52 |
| 稠密 | 89.02 ↑ 1.34 | 89.74 ↑ 1.09 | 65.35 ↑ 0.21 |

表 4　预训练数据评估

Table 4　Evaluation on Pre-training Dataset

| 预训练数据集 | RAF-DB | FERPlus | AffectNet |
|---|---|---|---|
| MS-Celeb-1M | 86.18 | 87.35 | 61.62 |
| ImageNet-1K | 89.02 ↑ 2.84 | 89.74 ↑ 2.39 | 65.35 ↑ 3.73 |

## 2.6　参数敏感性

图 5 展示了本文的方法在两个数据集上，针对不同投影维度、批次大小、权衡系数 $w$ 以及温度参数 $\tau$ 的性能表现。

在 RAF-DB 数据集上，投影维度为 128 时表现最佳；而在 FERPlus 数据集上，维度为 512 时效果最好。批次大小指的是正样本和负样本的数量。在自监督对比学习中，通常批次大小越大越好，因为这样能提供更多的负样本。然而，在面部表情识别任务中，混合监督对比微调的情况有所不同。由于利用了标签信息，正样本和负样本的标注更加准确，因此无需通过增大批次来获取更多负样本。此外，面部表情识别任务的类别数量特别烧，通常只有 7 类或 8 类，过大的批次大小可能导致过拟合。实验结果显示，在两个数据集上，批次大小为 64 时均取得了最佳性能。

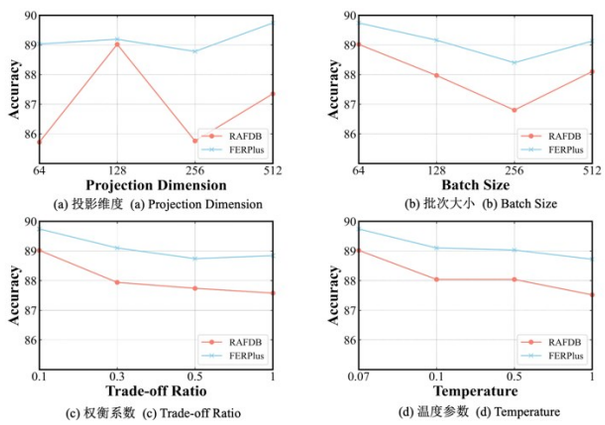


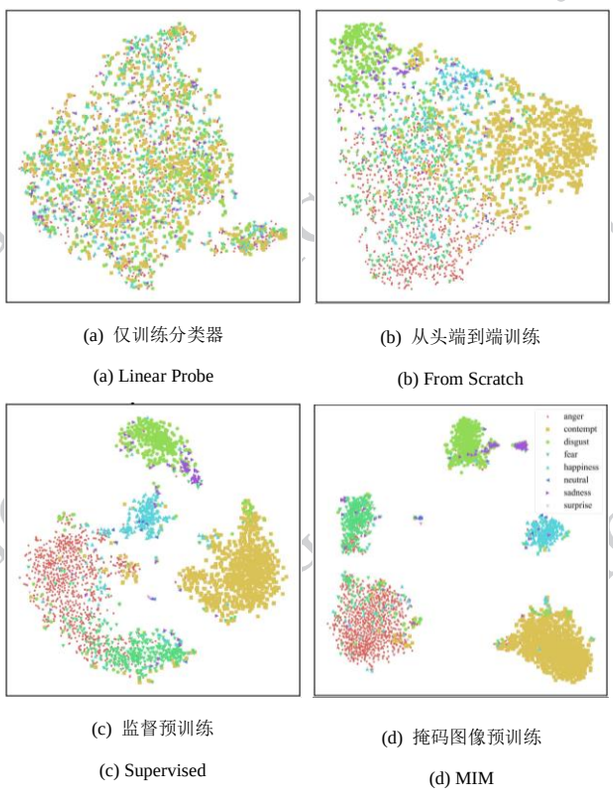

图 5　参数对投影维度、批次大小、权衡系数 $w$ 和温度参数 $\tau$ 的敏感性

Fig. 5　Parameter sensitivity on projection dimension, batch size, trade-off ratio $w$, and temperature $\tau$.

权衡系数 $w$ 旨在训练分类器时，既提升表征能力，又缩小两个领域之间的差距。实验中，$w$ 取值为 0.1 时效果最佳。温度参数 $\tau$ 用于增加对负样本的惩罚粒度，通常设置为小于 1 的值，经典设定为

0.07。在两个数据集上的测试结果表明，对于面部表情识别任务，温度参数$\tau$的最优值与传统图像分类任务相同。

### 2.7 定性分析




(a) 仅训练分类器 (a) Linear Probe

(b) 从头端到端训练 (b) From Scratch

(c) 监督预训练 (c) Supervised

(d) 掩码图像预训练 (d) MIM

图6 FERPlus验证集上全局视角的t-SNE可视化

Fig. 6 t-SNE visualizations in a global view on FERPlus validation set

图6通过t-SNE(Van der Maaten和Hinton，2008)可视化了四种方法所得ViT-B/16编码器最后一层特征的二维分布。其中Linear Probe指冻结MIM编码器、仅训练分类器；From Scratch指随机初始化后直接端到端训练。两者分别缺失预训练或微调阶段，所学表征类别区分能力较弱(图6第一行)。在固定微调策略的条件下，对比监督预训练与本文掩码图像预训练(图6第二行)可以看出，本方法实现了更清晰的类别分离，有效缓解了实例级身份混淆问题。

## 3 结 论

本文针对原生ViT在面部表情识别中性能不稳定、通常依赖复杂结构增强的问题，提出了掩码图像预训练与混合监督对比学习相结合的MIMIC训练范式。实验结果表明，该方法能够有效提升原生ViT的表征能力与跨域适应能力，在不引入卷积模块和额外复杂结构的情况下，取得了具有竞争力的识别性能。尤其在ViT-L/16设置下，本文方法在RAF-DB和FERPlus数据集上分别达到91.26%和91.24%的准确率，验证了原生ViT在合理训练策略下用于面部表情识别的潜力。

总体来看，本文结果说明，限制原生ViT在面部表情识别中表现的关键并不只是网络结构本身，更在于预训练目标与后训练方式是否契合表情识别的细粒度判别需求。未来工作将进一步探索该训练范式在视频表情识别、多模态情感理解以及更大规模视觉模型中的适用性。

## 参考文献(References)

Clavel C, Vasilescu I, Devillers L, Richard G and Ehrette T. 2008. Fear-type emotion recognition for future audio-based surveillance systems. Speech Communication, 50(6): 487-503

Yang D, Alsadoon P C, Prasad P C, Singh A K and Elchouemi A. 2018. An emotion recognition model based on facial recognition in virtual learning environment. Procedia Computer Science, 125: 2-10

Pioggia G, Igliozzi R, Ferro M, Ahluwalia A, Muratori F and De Rossi D. 2005. An android for enhancing social skills and emotion recognition in people with autism. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 13(4): 507-515

Ren F and Quan C. 2012. Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing. Information Technology and Management, 13(4): 321-332

Kopalidis T, Solachidis V, Vretos N, Daras P. Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. Information, 2024, 15(3): 135

Li S and Deng W. 2020. Deep facial expression recognition: A survey. IEEE Transactions on Affective Computing

Li S, Deng W and Du J. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 2852-2861

Mao J, Xu R, Yin X, et al. Poster++: A simpler and stronger facial expression recognition network [J]. Pattern Recognition, 2025, 157: 110951.

Shen Z. A comparative study of hybrid CNN and vision transformer models for facial emotion recognition[C]//2024 11th International Conference on Dependable Systems and Their Applications (DSA).

IEEE, 2024: 401-408.

Steiner A., Kolesnikov A., Zhai X., Wightman R., Jakobsen J., Beyer L., Ros H., Resnick C., Gelly S., & Moulin J. 2022. How to train your ViT? Data, augmentation, and regularization in Vision Transformers // Transactions on Machine Learning Research.

Li H, Wang N, Ding X, Yang X and Gao X. 2021. Adaptively learning facial expression representation via cf labels and distillation. IEEE Transactions on Image Processing, 30: 2016-2028

She J, Hu Y, Shi H, Wang J, Shen Q and Mei T. 2021. Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE: 6248-6257

Jin H, Jian M, Ding D and Yu H. 2026. HMamba-3DFT: A hierarchical mamba framework for emotion-driven semantic 3D facial tracking. Pattern Recognition, 178: 113415.

Jian M, Wang R, Yu X, Xu F, Yu H and Lam K M. 2024. UniFRD: A Unified Method for Facial Image Restoration Based on Diffusion Probabilistic Model. IEEE Transactions on Circuits and Systems for Video Technology, 34(12): 13494-13506.

Cui X Y, He C, Zhao H K and Wang M L. 2024. Combining ViT with contrastive learning for facial expression recognition. Journal of Image and Graphics, 29(01): 0123-0133 (崔鑫宇,何翀,赵宏珂,王美丽. 2024. 融合ViT与对比学习的面部表情识别. 中国图象图形学报,29(01): 0123-0133)[DOI:10.11834/jig.230043]

Lu L D, Xia H Y, Tan Y M and Song S X. 2024. Attention-guided local feature joint learning for facial expression recognition. Journal of Image and Graphics, 29(08): 2377-2387 (卢莉丹,夏海英,谭玉枚,宋树祥. 2024. 注意力引导局部特征联合学习的人脸表情识别. 中国图象图形学报,29(08): 2377-2387)[DOI:10.11834/jig.230410]

Wang Y J, He J, Zhang J X, Sun R H and Liu X L. 2025. Semi-supervised facial expression recognition robust to head pose empowered by dual consistency constraints. Journal of Image and Graphics, 30(02): 0435-0450 (王宇键,何军,张建勋,孙仁浩,刘学亮. 2025. 头姿鲁棒的双一致性约束半监督表情识别. 中国图象图形学报,30(02): 0435-0450)[DOI:10.11834/jig.240205]

Xue F, Wang Q and Guo G. 2021. Transfer: Learning relation-aware facial expression representations with transformers // Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 3601-3610

Tian Y, Zhu J, Yao H, Chen D. 2024. Facial expression recognition based on Vision Transformer with hybrid local attention. Applied Sciences, 14(15): 6471.

Li N, Huang Y, Wang Z, Fan Z, Li X, Xiao Z. 2024. Enhanced hybrid Vision Transformer with multi-scale feature integration and patch dropping for facial expression recognition. Sensors, 24(13): 4153.

Zhang Y, Zheng X, Liang C, Hu J, Deng W. 2024. Generalizable facial expression recognition // Proceedings of the European Conference on Computer Vision. Cham: Springer: 231-248.

Tao H and Duan Q. 2023. Hierarchical attention network with progressive feature fusion for facial expression recognition. Neural Networks

Gao H, Wu M, Chen Z, Li Y, Wang X and An S, et al. 2023. Ssa-icl: Multi-domain adaptive attention with intra-dataset continual learning for facial expression recognition. Neural Networks, 158: 228-238

Guo Y, Zhang L, Hu Y, He X and Gao J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition // Computer Vision-ECCV 2016: 14th European Conference. Amsterdam: Springer: 87-102

He K, Chen X, Xie S, Li Y, Dollar P and Girshick R. 2022. Masked autoencoders are scalable vision learners // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 16000-16009

Xie Z, Zhang Z, Cao Y, Lin Y, Bao J and Yao Z, et al. 2022. Simmim: A simple framework for masked image modeling // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 9653-9663

Gunel B, Du J, Conneau A and Stoyanov V. 2021. Supervised contrastive learning for pre-trained language model fine-tuning[EB/OL].[2024-03-29]. https://openreview.net/forum?id=cu7IUiOhujH

Moukafih Y, Ghogho M and Smaili K. 2023. Supervised contrastive learning as multi-objective optimization for fine-tuning large pre-trained language models[EB/OL].[2024-03-29]. https://doi.org/10.1109/ICASSP49357.2023.10095108

He K, Fan H, Wu Y, Xie S and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 9729-9738

Chen X, Fan H, Girshick R and He K. 2020. Improved baselines with momentum contrastive learning [EB/OL].[2024-03-29]. https://arxiv.org/pdf/2003.04297.pdf

Khosla P, Teterwak P, Wang C, Sarna A, Tian Y and Isola P, et al. 2020. Supervised contrastive learning. Advances in Neural Information Processing Systems, 33: 18661-18673

Zhang K, Zhang Z, Li Z and Qiao Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10): 1499-1503

Amos B, Ludwiczuk B and Satyanarayanan M. 2016. Openface: A general-purpose face recognition library with mobile applications. Pittsburgh: CMU School of Computer Science

Hu Y, Zeng Z, Yin L, Wei X, Zhou X and Huang T S. 2008. Multi-view facial expression recognition // Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. Amsterdam: IEEE: 1-6

Luo Y, Wu C M and Zhang Y. 2013. Facial expression recognition

based on fusion feature of pca and lbp with svm. Optik, 124(17): 2767-2770

Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z and Matthews I. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression // Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. San Francisco: IEEE: 94-101

Zhao G, Huang X, Taini M, Li S Z and Pietikainen M. 2011. Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): 607-619

Barsoum E, Zhang C, Ferrer C C and Zhang Z. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution // Proceedings of the 18th ACM International Conference on Multimodal Interaction. Tokyo: ACM: 279-283

Mollahosseini A, Hasani B and Mahoor M H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1): 18-31

Wang K, Peng X, Yang J, Meng D and Qiao Y. 2020. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29: 4057-4069

Li Y, Zeng J, Shan S and Chen X. 2018. Occlusion aware facial expression recognition using cnn with attention mechanism. IEEE Transactions on Image Processing, 28(5): 2439-2450

Wang K, Peng X, Yang J, Lu S and Qiao Y. 2020. Suppressing uncertainties for large-scale facial expression recognition // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 6897-6906

Zeng J, Shan S and Chen X. 2018. Facial expression recognition with inconsistently annotated datasets // Proceedings of the European Conference on Computer Vision. Munich: Springer: 222-237

Hadsell R, Chopra S and LeCun Y. 2006. Dimensionality reduction by learning an invariant mapping // Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York: IEEE: 1735-1742

Chen T, Kornblith S, Norouzi M and Hinton G. 2020. A simple framework for contrastive learning of visual representations // Proceedings of the 37th International Conference on Machine Learning. Vienna: PMLR: 1597-1607

Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X and Unterthiner T, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale [EB/OL]. [2024-03-29]. https://arxiv.org/pdf/2010.11929.pdf

Liu Z, Lin Y, Cao Y, Hu H, Wei Y and Zhang Z, et al. 2021. Swin transformer: Hierarchical vision transformer using shifted windows // Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 10012-10022

Zhang H, Cissé M, Dauphin Y N and Lopez-Paz D. 2017. Mixup: Beyond empirical risk minimization [EB/OL]. [2024-03-29]. https://arxiv.org/pdf/1710.09412.pdf

Yun S, Han D, Oh S J, Chun S, Choe J and Yoo Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features // Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 6023-6032

Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L. 2009. Imagenet: A large-scale hierarchical image database // Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE: 248-255

Farzaneh A H and Qi X. 2020. Discriminant distribution-agnostic loss for facial expression recognition in the wild // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE: 406-407

Gera D and Balasubramanian S. 2021. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. Pattern Recognition Letters, 145: 58-66

Zhang Y, Wang C and Deng W. 2021. Relative uncertainty learning for facial expression recognition. Advances in Neural Information Processing Systems, 34: 17616-17627

Farzaneh A H and Qi X. 2021. Facial expression recognition in the wild via deep attentive center loss // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE: 2402-2411

Zhao Z, Liu Q and Wang S. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Transactions on Image Processing, 30: 6544-6556

Li H, Sui M, Zhao F, Zha Z and Wu F. 2021. Mvt: mask vision transformer for facial expression recognition in the wild [EB/OL]. [2026-03-29]. https://arxiv.org/pdf/2106.04520.pdf

Ma F, Sun B and Li S. 2021. Facial expression recognition with visual transformers and attentional selective fusion. IEEE Transactions on Affective Computing

Zhang Y, Wang C, Ling X and Deng W. 2022. Learn from all: Erasing attention consistency for noisy label facial expression recognition // Computer Vision-ECCV 2022: 17th European Conference. Tel Aviv: Springer: 418-434

Feng L, Lv J, Han B, Xu M, Niu G and Geng X, et al. 2020. Provably consistent partial-label learning. Advances in Neural Information Processing Systems, 33: 10948-10960

Wen H, Cui J, Hang H, Liu J, Wang Y and Lin Z. 2021. Leveraged weighted loss for partial label learning // Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 11091-11100

Wang H, Xiao R, Li Y, Feng L, Niu G and Chen G, et al. 2022. Pico: Contrastive label disambiguation for partial label learning [EB/

OL]. [2026-03-29]. https://arxiv.org/pdf/2201.08984.pdf

Xue F, Wang Q, Tan Z, Ma Z and Guo G. 2022. Vision transformer with attentive pooling for robust facial expression recognition. IEEE Transactions on Affective Computing

Chen D, Wen G, Li H, Chen R and Li C. 2023. Multi-relations aware network for in-the-wild facial expression recognition. IEEE Transactions on Circuits and Systems for Video Technology, 33: 3848-3859

Li C, Li X, Wang X, et al. 2023. Fine-grained associative graph representation for facial expression recognition in the wild. IEEE Transactions on Circuits and Systems for Video Technology, 34(2): 882-896 [DOI: 10.1109/TCSVT.2023.3237006]

Liu Y, Zhang X, Kauttonen J, et al. 2022. Uncertain facial expression recognition via multi-task assisted correction [EB/OL]. [2024-03-29]. https://arxiv.org/abs/2212.07144

Li J, Nie J, Guo D, et al. 2025. Emotion separation and recognition from a facial expression by generating the poker face with vision transformers. IEEE Transactions on Computational Social Systems, 12(4): 1548-1562 [DOI: 10.1109/TCSS.2024.3478839]

Chen X and Huang L. 2024. A lightweight model enhancing facial expression recognition with spatial bias and cosine-harmony loss. Computation, 12 (10): 201 [DOI: 10.3390/computation12100201]

Liu S, Xu Y, Wan T and Zheng Z. 2023. Ada-DF: an adaptive label distribution fusion network for facial expression recognition. IEEE Transactions on Multimedia, 25: 6414-6426 [DOI: 10.1109/TMM.2022.3210960]

Song J, He M, Feng J, et al. 2024. Bridging the gaps: utilizing unlabeled face recognition datasets to boost semi-supervised facial expression recognition [EB/OL]. [2024-03-29]. https://arxiv.org/abs/2410.17622

Liu T, Li J, Wu J, Du B, Zhan Y, Tao D and Wan J. 2025. Facial expression recognition with heatmap neighbor contrastive learning. IEEE Transactions on Multimedia, 27: 4795-4807 [DOI: 10.1109/TMM.2025.3543029]

El-Khashab O, Hamdy A and Mahmoud A. 2023. FerNeXt: facial expression recognition using ConvNeXt with channel attention // Proceedings of the 2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC). Cairo, Egypt: IEEE: 1-8 [DOI: 10.1109/MIUCC58832.2023.10278345]

Yu G. 2025. A multi-granularity feature fusion approach with attention for facial expression recognition. Scientific Reports, 15(1): 42507 [DOI: 10.1038/s41598-025-26533-9]

Pazandeh A M and Fatemizadeh E. 2025. Enhancing vision transformers for facial expression recognition. IEEE Access, 13: 144689-144698 [DOI: 10.1109/ACCESS.2025.3598917]

Song D and Liu C. 2025. A facial expression recognition network using hybrid feature extraction. PLOS ONE, 20(1): e0312359 [DOI: 10.1371/journal.pone.0312359]

Devasena G and Vidhya V. 2025. Twinned attention network for occlusion-aware facial expression recognition. Machine Vision and Applications, 36(1): 23 [DOI: 10.1007/s00138-024-01641-0]

Chen X, Xie S and He K. 2021. An empirical study of training self-supervised vision transformers [EB/OL]. [2026-03-29]. https://arxiv.org/pdf/2104.02057.pdf

Wang H, Tang Y, Wang Y, Guo J, Deng Z H and Han K. 2023. Masked image modeling with local multi-scale reconstruction // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 2122-2131

Van der Maaten L and Hinton G. 2008. Visualizing data using t-sne. Journal of Machine Learning Research, 9(11)

## 作者简介

黎凯耀,男,硕士研究生,研究方向为情感计算和多模态大模型。E-mail: kyaoli@163.com

章帆,男,博士研究生,主要研究方向为多模态大模型和智能体系统。E-mail:fzhang@link.cuhk.edu.hk

李青,女,助理教授,主要研究方向为计算机视觉与多模态大模型。E-mail:liqing@sztu.edu.cn

彭保,男,教授,主要研究方向为人工智能及行业应用。E-mail:Pengb@sziit.edu.cn

彭小江,通信作者,男,教授,主要研究方向为视觉与情感计算、具身智能和多模态大模型。E-mail:pengxiaojiang@sztu.edu.cn