arXiv:2303.00180v4 [cs.CV] 4 Sep 2024

MMA-MRNNet：结合多重情感模型与动态掩码循环神经网络，实现精准的面部表情强度估计

Dimitrios Kollias[1], Andreas Psaroudakis[2], Anastasios Arsenos[2], Paraskevi Theofilou[2], Chunchang Shao[1], Guanyu Hu[1], 和 IoannisPatras[1]

[1] 伦敦玛丽女王大学，英国[2] 雅典国立技术大学，
希腊 d.kollias@qmul.ac.uk

摘要。 本文介绍了 MMA-MRNNet，这是一种新颖的深度学习架构，用于从视频数据中动态进行多输出面部表情强度估计（FEIE）。传统方法通常依赖复杂的三维卷积神经网络，需要大量预训练，并假设面部表情在视频的所有帧中均匀分布。这些方法难以处理长度各异的视频，往往采用临时应对策略，导致有价值的信息被丢弃或引入偏差。MMA-MRNNet 通过两阶段流程解决这些挑战。首先，情感多模型（MMA）提取器组件是一种多任务学习卷积神经网络，可在每帧中同时估计效价-唤醒度、识别基本面部表情并检测动作单元。随后，掩码循环神经网络组件对这些表征进行处理，以捕捉时间依赖关系，并根据输入视频的实际长度动态更新权重，确保最终预测仅使用最相关的特征。我们在 Hume-Reaction 数据集上评估了所提出的单模态非集成学习模型 MMA-MRNNet。结果表明，其性能显著优于其他方法，大幅超越最先进方法，无论这些方法采用单模态、多模态还是集成方法。最后，我们在多个自然场景数据集上展示了所提出方法中 MMA 组件的有效性；在各种指标上，该组件始终优于所有最先进方法。

关键词： MMA-MRNNet · 掩码循环神经网络 · 路由 · 面部表情强度估计 · ABAW · MUSE · Hume-Reaction 数据集5 · 效价-唤醒度估计 · 基本表情识别 · 动作单元检测

1 引言

人类情绪是复杂的、有意识的体验，会深刻影响行为，并可通过各种形式表达。这些情绪在

心理过程中发挥着关键作用，并显著影响人类行为。人工智能（AI）和深度学习（DL）的出现推动了智能系统的发展，使其能够识别和解读人类情绪。心理学家提出了多种描述符，用于量化和分类情绪状态：稀疏描述符，如面部动作单元（AUs），用于捕捉特定的面部肌肉激活情况 [10]；连续描述符，如效价和唤醒度，其中效价表示情绪的积极或消极程度，唤醒度反映激活或被动状态的程度 [50]；以及离散类别描述符，如六种基本表情（愤怒、厌恶、恐惧、快乐、悲伤、惊讶）和中性状态 [9]。本文聚焦于动态多输出面部表情强度估计（FEIE），具体针对崇拜、娱乐、焦虑、厌恶、共情性疼痛、恐惧和惊讶等表情的强度估计。

本文介绍了我们提出的方法 MMA-MRNNet，这是一种新颖的深度学习架构，旨在解决仅提供视频级标注（即整个视频只有一个标注）而非帧级标注时面部表情强度估计（FEIE）所面临的复杂问题。MMA-MRNNet 要应对的关键挑战包括处理视频长度的变化，以及准确汇总各帧的时间信息，从而得出可靠的最终预测。

处理视频数据等三维信号的传统方法通常采用三维卷积神经网络，每个信号仅生成一个预测结果。然而，这些架构本身较为复杂，参数量庞大，而且通常需要在大型三维数据集上进行预训练，才能取得令人满意的性能。另一种常见方法是将视频级标签统一分配给每一帧，然后使用卷积神经网络-循环神经网络在这些已标注的帧上进行训练。这种方法假设所有帧中的面部表情强度都保持一致，但实际情况未必如此，因为可能只有部分帧确实呈现了所标注的强度 [1, 12, 23, 24, 26, 31, 34, 39]。

此外，我们的方法解决了输入视频长度不一所带来的挑战。传统方法通常依赖临时策略来处理帧数变化，例如设定固定的输入长度，然后丢弃多余的帧（这可能导致关键信息丢失），或在较短的视频中复制帧（这可能使模型偏向重复数据）。这些策略不仅效果欠佳，还需要针对每个特定数据集进行经验性调优，从而限制了其泛化能力和有效性。

MMA-MRNNet由两个主要组件构成：情感多模型（MMA）提取器和掩码循环神经网络与路由网络（MRNN）。MMA组件是一个多任务学习（MTL）卷积神经网络，通过同时估计效价-唤醒度（VA）、识别7种基本表情并检测多个动作单元（AUs），从每一帧中提取情感表征。为确保这些表征的可靠性和一致性，我们提出了一种新颖的损失函数，其中融入了不同情感描述符之间关系的先验知识，从而缓解多任务学习场景中常见的噪声梯度和收敛不佳等问题。

随后，将提取的表征传递给MRNN组件。该组件包含一个循环神经网络，用于捕捉帧序列中的时间依赖关系。为处理输入视频长度不一的问题，MRNN中采用了掩码层。该层根据视频的实际帧数动态选择相关的循环神经网络输出，使模型能够适应不同的输入长度，同时不损害时间信息的完整性。随后，将选出的特征传入全连接层，为整个视频生成最终的强度估计。

据我们所知，MMA-MRNNet 是首个将效价-唤醒度、动作单元和基本表情用作面部表情强度估计任务中间表征的架构。这种方法不仅增强了模型捕捉情绪表达细微动态变化的能力，还为处理输入条件各异的真实世界数据提供了稳健的框架。

2 相关工作

[16]提出了用于情绪识别的监督评分集成（SSE）。该方法提出了一种新的融合结构，将不同互补特征层的逐类评分激活拼接起来，作为二级监督的输入，从而在单个卷积神经网络架构中实现深度特征集成。 [60] 提出了一种用于视频情绪识别的深度视觉-音频注意力网络（VAANet）；VAANet 将空间注意力、通道注意力和时间注意力融入视觉三维卷积神经网络，并将时间注意力融入音频二维卷积神经网络。该方法还提出了极性一致交叉熵损失，以极性-情绪层级约束为依据引导注意力生成。 [13]构建了一种用于识别人类情绪的视听混合网络。对于视觉数据，该架构采用 VGG-Face 提取逐帧特征，并使用长短期记忆网络（LSTM）根据这些特征的时间依赖关系对其进行关联。

[44] 是第五届ABAW挑战赛 [1, 18–23, 25–27, 29–31, 33– 37,40–42]情绪反应强度（ERI）估计挑战赛的获胜方法。该方法由音频特征编码模块（基于DenseNet121和DeepSpectrum）、视觉特征编码模块（基于PosterV2-Vit）以及视听模态交互模块组成。 [53] 提出了ViPER，这是一种模态无关后期融合网络，利用基于Transformer的模型融合视频帧、音频录音和文本标注，以进行面部表情强度估计（FEIE）。 [56] 提出了一种双分支面部表情强度估计模型：一个分支（由时间卷积神经网络和Transformer 编码器组成）处理视觉模态，另一个分支处理音频模态；此外，还加入了模态丢弃来进行视听特征融合。 [51] 在第五届ABAW挑战赛的ERI挑战赛中获得第三名；其提出的方法使用视觉Transformer、HuBERT和DeBERTa提取视觉、音频和文本模态的特征。该方法采用时间增强和SE模块来提升时序泛化能力 [2–4, 17, 48] 以及上下文理解能力。随后，将各模态的特征

输入上下文层进行处理，并通过后期融合策略进行融合。 [54] 提出了一种方法，使用FAb-Net、EfficientNet和DAN等模型从视频帧中提取视觉特征，以捕捉面部表情和属性。音频特征则使用Wav2Vec2和VGGish模型提取。随后，将提取的特征输入时间卷积网络以捕捉局部时序信息，再通过Transformer 编码器利用动态注意力对长程依赖关系进行建模。 [59]提出了一种方法，使用最先进的模型提取音频和视觉特征，并通过仿射模块将这些特征对齐至共同维度。随后，使用多模态多头注意力模型融合对齐后的特征。

3 方法论

问题表述 我们的方法以由多个实例（即视频帧）组成的视频作为输入，$\boldsymbol{X} = \{\mathbf{x}_1, ..., \mathbf{x}_K\}$，其中 $\mathbf{x}_K \in \Re^{H \times W \times 3}$。 K 表示实例（帧）的数量，不同视频的实例数量各不相同； H 和 W 分别表示 RGB 图像（帧）的高度和宽度。每个视频都有一个视频级标签 Y。我们进一步假设这些实例也有相应的实例级标签$\{\mathbf{y}_1, ..., \mathbf{y}_K\}$，但在训练期间这些标签未知；同一视频中所有实例的实例级标签不一定与视频级标签一致。数据库 $DB = \{\boldsymbol{X}_n, Y_n\}_{n=1}^{N}$由 N个这样的视频-标签对构成。我们的目标是学习一个最优函数，以视频中的实例为输入来预测视频级标签。为此，我们的方法应能够：
1) 有效考虑到输入视频具有可变长度这一事实（换言之，该方法应对不同视频的总帧数各不相同这一情况加以处理）

2）聚合实例 $\{\boldsymbol{x}_k\}_{k=1}^{K}$ 的信息，以作出最终决策。一种广泛采用的聚合方法是基于嵌入的方法，它将$\boldsymbol{X}$ 映射为视频级表示 $\boldsymbol{z} \in \Re^F$，并使用$\boldsymbol{z}$ 来预测 Y。

首先，所有视频 $\{\boldsymbol{X}_n\}_{n=1}^{N}$ 都会填充至统一长度 t，从而得到视频序列 $\boldsymbol{X}_N = \{\mathbf{x}_1, ..., \mathbf{x}_t\}$。随后，每个视频 $\boldsymbol{X}$都会经过情感多模型（MMA）提取器组件处理，该组件对每个二维帧进行局部分析，将 $\boldsymbol{X}$ 映射为多重情感级别表示矩阵 $\boldsymbol{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_t\} \in \Re^{d \times t}$。随后，该矩阵被传入位于 MMA 组件之上的循环神经网络，以捕获所有 $\{\boldsymbol{z}_k\}_{k=1}^{t}$之间的时间依赖关系。循环神经网络将 $\boldsymbol{Z}$ 转换为嵌入矩阵$\boldsymbol{Z}' = \{\mathbf{z}'_1, ..., \mathbf{z}'_t\} \in \Re^{d' \times t}$，对整个视频进行全局分析。后续模块将嵌入集合 $\{\boldsymbol{z}'_k\}_{k=1}^{t}$ 聚合为单个视频级向量嵌入 $\boldsymbol{z}' \in \Re^{d' \cdot t}$，然后将其输入掩码层。掩码层根据视频的“真实”帧数动态选择嵌入，同时考虑填充前的原始帧数。这一步至关重要，因为视频级标注表明，为了准确预测，所有帧作为整体而非单独帧都承载着重要信息。掩码层的输出 $\boldsymbol{z}'' \in \Re^{d' \cdot t}$ 随后被映射为另一个嵌入

$\boldsymbol{z}''' \in \Re^{d''}$，并通过前馈层完成这一映射。最后，$\boldsymbol{z}'''$ 经由由 $\boldsymbol{W}$ 参数化的前馈层转换为对数几率 $\boldsymbol{u}$，从而得到视频级分类结果：$\boldsymbol{u} = \boldsymbol{W}^T \boldsymbol{z}'''$。

下面，我们将进一步详细说明所提出方法的各个组成部分。图 1 概述了我们提出的 MMA-MRNNet 框架。

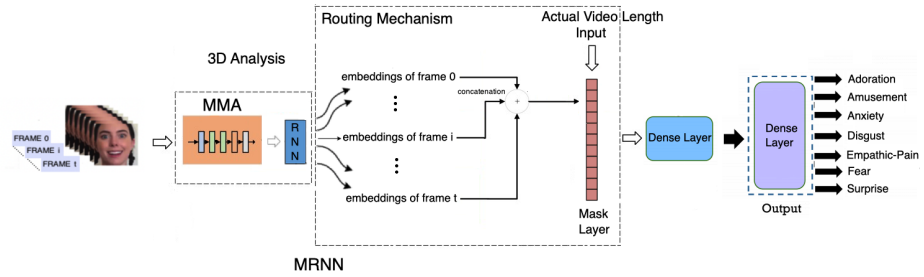

图 1： 所提出的用于动态多输出面部表情强度估计的 MMA-MRNNet 概述。MMA-MRNNet 包含两个主要组成部分：情感多模型（MMA）提取器，用于为每个视频帧生成情感表征（效价-唤醒度、基本表情和动作单元）；以及掩码循环神经网络与路由（MRNN），用于捕获时间依赖关系，并根据输入视频的可变长度动态选择关键特征（并更新权重）。

3.1 MMA：多重情感模型提取器组件

情感多模型（MMA）提取器组件通过从输入视频 $\boldsymbol{X}$ 的每一帧中提取情感表征来处理该视频，所用方法是采用三个不同的情感模型。具体而言，MMA 是一个多任务学习（MTL）卷积神经网络模型，可同时执行：(i) 以效价和唤醒度（VA）表示的连续情感估计；(ii) 7种基本面部表情识别；以及 (iii) 17个动作单元（AU）检测。如图 2 所示，MMA 的架构以残差单元为基础，其中“bn”表示批归一化层。该模型在由共享前馈层生成的同一嵌入空间中，整合了效价-唤醒度估计、7种基本表情识别和17个AU检测任务。因此，MMA 处理 $\boldsymbol{X}$ 时的输出是一个多重情感层级表征矩阵 $\boldsymbol{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_t\} \in \Re^{26 \times t}$。

在训练 MMA 时，我们使用了多个自然场景数据集，包括 Aff-Wild2 [18, 20, 22, 28, 32–34,36–38,40–42, 57], AffectNet [49], 和 EmotioNet [5],。这些数据集标注了效价-唤醒度、7种基本表情和17个动作单元（这些动作单元是所有数据集的汇总）。

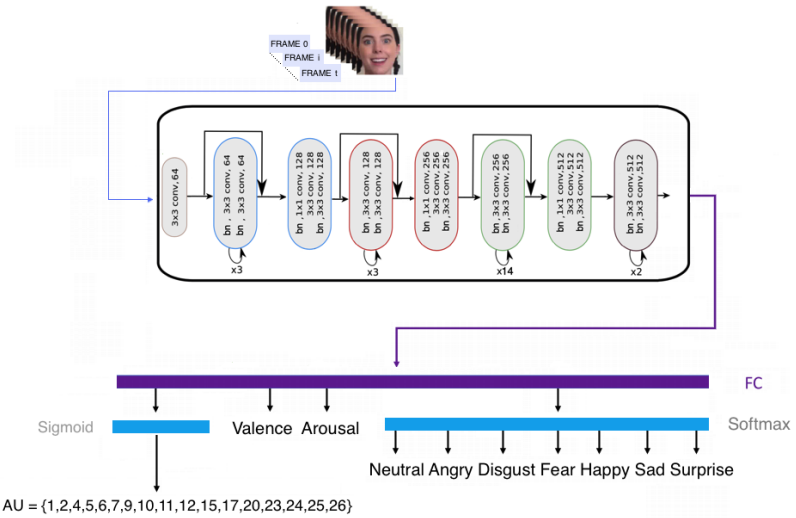

图 2： 多模型情感提取组件（MMA），针对每一帧输出以下情感描述符：效价和唤醒度、17 个动作单元以及 7 种基本表情

我们近期的研究 [14, 15] 指出，当前情感分析方法的评估面临诸多挑战，并指出数据库划分和评估实践存在不一致，导致比较结果有偏且不公平。为解决这些问题，研究者提出了一种数据库划分统一协议，以确保公平性和可比性，同时考虑受试者人口统计信息。研究表明，在原始划分上曾被视为最先进的方法，在这一新协议下未必能保持其性能。因此，在本文中，我们采用了这一更新后的数据划分协议。

利用这些数据集的一项主要挑战在于，它们的特定任务标注彼此不重叠。例如，EmotioNet 仅包含 AU 标注，不包含效价-唤醒度和 7 种基本表情标签。若直接使用这些数据集训练 MMA，并对所有任务采用组合损失函数，就会产生噪声梯度并导致收敛不佳，因为并非所有损失项都会持续为整体目标函数作出贡献。这可能引发多任务学习（MTL）中常见的问题，例如任务不平衡，即某项任务可能主导训练；或负迁移，即多任务学习模型的表现不如单任务模型 [32–34]。

为解决这一问题，我们根据 MMA 为每一帧生成的 7 种基本表情表征（r_{expr}），生成 AU 伪表征（r'_{AU}）。具体而言，我们利用表情与动作单元之间的关系，该关系定义见 [8] 中的表1。 [8] 对面部表情与动作单元激活之间的关联进行了认知与心理学分析，

并总结了研究结果，制成一张表，详细说明表情与对应动作单元之间的关联。该表作为参考列于表1。原型动作单元是指所有标注者均一致认为已激活的动作单元，而观察动作单元则是指只有部分标注者标记为已激活的动作单元。

表1： 根据 [8] 推断的表情与动作单元的相关性

表情	原型动作单元	观察动作单元
快乐	12, 25	6
悲伤	4, 15	1,6 , 11, 17
fear	1, 4, 20, 25	2, 5, 26
愤怒	4, 7, 24	10, 17, 23
惊讶	1, 2, 25, 26	5
厌恶	9, 10, 17	4, 24

AU 伪表征被建模为基本表情类别的混合：

$$r'_{AU} = \sum\nolimits_{expr} r_{expr} \cdot r_{AU|expr} \tag{1}$$

其中，$r_{AU|expr}$ 根据表1确定性地定义；对于原型或观察动作单元，其值为1，否则为0。

然后，我们通过最小化软目标二元交叉熵损失，使 MMA 的动作单元表示（r_{AU}）与 AU 伪表征相匹配：

$$\mathcal{L}_{DM} = \mathbb{E}\left[\sum_{i=1}^{17}[-r'^{\,i}_{AU} \cdot \log\, r^{i}_{AU}]\right] \tag{2}$$

通过这一损失，我们旨在将关于任务关系的先验知识（依据表1）注入网络，从而引导生成更优且一致的表情表征和动作单元表示。例如，如果网络以概率1预测为快乐，同时又预测动作单元4、15和1被激活，那么这就是错误的，因为这些动作单元与悲伤表情相关。在这种情况下，动作单元表示与表情表征相互冲突。因此，MMA训练期间最小化的总体目标函数（$\mathcal{L}_{MMA}$）为：

$$\mathcal{L}_{MMA} = \mathcal{L}_{CCC} + \mathcal{L}_{CCE} + \mathcal{L}_{BCE} + \mathcal{L}_{DM} \tag{3}$$

其中：
$\mathcal{L}_{CCC}$ 是与效价-唤醒度估计任务相关的损失项，且$\mathcal{L}_{CCC} = 1 - 0.5 \cdot (\rho_a + \rho_v)$，其中$\rho_{a/v}$ 为唤醒度/效价的协调相关系数（CCC）；

$\mathcal{L}_{CCE}$ 是与7类基本表情识别任务相关的分类交叉熵损失；而$\mathcal{L}_{BCE}$ 是与17个AU检测任务相关的二元交叉熵损失。

3.2 MRNN：掩码循环神经网络与路由组件

如上一节所述，MMA组件通过使用三种不同的情感模型，从输入视频$\boldsymbol{X}$ 的每一帧（$\boldsymbol{x}_i$）中提取情感表征。由此得到情感层级表征矩阵$\boldsymbol{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_t\} \in \Re^{26\times t}$。随后，该矩阵被输入置于MMA组件之上的循环神经网络，以捕捉视频连续帧之间的时间依赖关系和序列信息。循环神经网络从第0帧到第 t帧依次处理提取出的向量表征$\mathbf{z}_k\}_{k=1}^{K}$，并将其映射为嵌入$\mathbf{z}'_k\}_{k=1}^{K}$，其中每个$\mathbf{z}'_k \in \Re^{d'}$。

这些嵌入（对应所有视频帧）被拼接成单个向量嵌入$\boldsymbol{z}' \in \Re^{d'\cdot t}$，以符合估计整个视频序列中各种面部表情强度的目标，并与给定标注保持一致。接着，该嵌入$\boldsymbol{z}'$被输入掩码层，生成新的嵌入$\boldsymbol{z}'' \in \Re^{d'\cdot t}$。输入视频未经填充的原始长度 l会传递至掩码层，以引导路由机制。在训练过程中，路由机制根据视频长度 l，从$\boldsymbol{z}'$ 中的不同位置动态选取元素，保留所选元素的值，并将其余元素置零。此过程仅将相关元素路由至后续层，从而使模型能够更加关注关键时间特征。

嵌入$\boldsymbol{z}'' \in \Re^{d'\cdot t}$ 随后通过前馈层转换为另一个嵌入$\boldsymbol{z}''' \in \Re^{d''}$ 。该层经过训练，旨在从“掩码后”的嵌入$\boldsymbol{z}''$中提取高层信息。在训练过程中，只有将前馈层神经元连接到掩码层路由至$\boldsymbol{z}'$ 中元素的权重会更新。当对应的前馈层神经元连接到掩码层在另一视频输入中选取的$\boldsymbol{z}'$中元素时，其余权重才会更新。损失函数最小化的方式与采用动态路由的网络类似：未参与路由过程的权重保持不变，且忽略$\boldsymbol{z}'$ 中与未路由元素对应的连接。最后，嵌入$\boldsymbol{z}'''$ 通过由$\boldsymbol{W}$参数化的前馈层映射为逻辑值$\boldsymbol{u}$，从而得到视频级分类，表示为：$\boldsymbol{u} = \boldsymbol{W}^T \boldsymbol{z}'''$。

我们用于训练 MMA-MRNNet 的损失函数并非典型的均方误差（MSE），而是基于皮尔逊相关系数的损失，因为该相关性指标是所用数据库的评估标准：

$$\mathcal{L}_{total} = 1 - \sum_{i=1}^{7} \frac{\rho_i}{7} = 1 - \frac{1}{7}\sum_{i=1}^{7} \frac{s_{i,xy}}{\sqrt{s_{i,x} \cdot s_{i,y}}} \tag{4}$$

其中： i 表示面部表情； ρ_i 是皮尔逊相关系数；$s_{i,\ x}$ 和 $s_{i,\ y}$ 分别是表情标签和预测值的方差；$s_{i,\ xy}$是它们的协方差。

3.3 数据集、预处理与实现细节

Hume-Reaction 数据集同时用于 MuSe 2022 [6]的情绪反应子挑战赛，以及 2023 [1,18–23,25–27,29–31,33– 37,40–42]第5届 ABAW 竞赛中的情绪反应强度估计挑战赛。该子挑战赛的参赛者将探索一项多输出回归任务，使用七种由参与者自行标注、细致区分的情绪类别：“崇拜”“娱乐”“焦虑”“厌恶”“共情性疼痛”“恐惧”和“惊讶”。该数据集包含来自 2,222 名受试者的 25,067 段视频，其中 15,806 段构成训练集，4,657 段构成验证集，另有 4,604 段构成测试集。

Aff-Wild2 数据库 [18, 20, 22, 28, 32–34,36–38, 40–43, 57] 是规模最大的自然场景数据库，也是唯一一个按帧对七种基本表情（即快乐、惊讶、愤怒、厌恶、恐惧、悲伤和中性状态）、十二个动作单元（AUs 1,2,4,6,7,10,12,15,23,24,25, 26）以及效价和唤醒度进行标注的数据库。该数据库共包含 564 个视频，约 280 万帧，涉及 554 名受试者。Aff-Wild2 在受试者的年龄、族裔和国籍方面呈现出很大的多样性，环境也存在显著差异。

AffectNet 数据集 [49] 包含约 100 万张面部图像，其中 30 万张经过人工标注，标注内容包括 7 种离散表情（另加轻蔑）以及效价-唤醒度。该数据库原始训练集约包含 29 万张图像，原始验证集包含 4,000 张图像。根据我们先前的工作 [14, 15] （如第 3.1 节所述），我们采用该数据库更新后的数据划分协议来评估我们的方法。这一新的划分包括约 16 万张图像的训练集、约 4.5 万张图像的验证集以及约 9 万张图像的测试集。

EmotioNet 数据库 [5] 包含约 100 万张图像，并于 2017 年为 EmotioNet 挑战赛发布。其中 95 万张图像经过自动标注，其余 5 万张图像则由人工标注了 11 个动作单元（1、2、4、5、6、9、12、17、20、25、26）；后者约有一半构成挑战赛的验证集，另一半构成测试集。我们根据先前的工作 [14, 15] （如第 3.1 节所述），使用该数据库更新后的数据划分协议评估我们的方法。新的数据划分包括约 2.5 万张图像的训练集、约 7 千张图像的验证集，以及约 1.4 万张图像的测试集。

我们使用 RetinaFace 检测器 [7] 从所有图像中提取人脸边界框和 5 个人脸关键点；后者用于人脸对齐。所有裁剪并对齐后的图像均调整至 $112 \times 112 \times 3$ 像素分辨率，并将其强度值归一化至 $[-1, 1]$。

我们选择批量大小为 4，长度 t 为 480；从头开始训练时使用 Adam优化器，学习率为 10^{-4}；初始化每个子网络后，采用端到端方式训练时的学习率为 10^{-5}。对于循环神经网络，我们使用

1 层门控循环单元，包含 128 个单元；前馈层包含 32 个单元。训练在 Tesla V100 32GB GPU 上进行，耗时 3 天。我们使用了 TensorFlow 平台。

4 实验结果

4.1 与最先进水平的比较

首先，我们将 MMA-MRNNet 的性能与多种基线 [6] 方法和最先进方法进行比较：ViPER 和网易伏羲虚拟人方法（这两种多模态方法利用音频、视觉和文本信息）；表现最佳的 HFUT-CVers 方法（在相关工作一节中介绍，是一种利用音频和视觉信息的集成多模态方法）；USTC-IAT-United 方法（在相关工作一节中介绍，是一种利用音频和视觉信息的多模态方法）；以及 USTC-AC 和 NISL-2023 方法（均在相关工作一节中介绍，都是利用音频和视觉信息的集成多模态方法）。表 2 显示，我们的单模态、非集成学习方法 MMA-MRNNet（仅利用视觉信息，不采用任何集成学习）以显著优势胜过所有其他方法（尽管其中一些方法采用多模态，甚至集成方法）。还需指出，所有基线方法和最先进方法都采用了一种特设策略：通过删除或复制每段视频序列中的图像，选取固定的输入长度。

表 2： MMA-MRNNet、基线方法与最先进水平在 Hume-Reaction 数据集测试集上的比较；皮尔逊相关系数结果以 % 表示

方法	皮尔逊相关系数（ρ）
HFUT-CVers [44]	47.3
USTC-IAT-United [56]	43.8
网易伏羲虚拟人 [51]	40.5
USTC-AC [54]	37.3
NISL-2023 [59]	36.7
ViPER [53]	29.7
FAU-Baseline [6]	28.0
VGGface 2-Baseline [6]	18.3
融合基线 [6]	20.30
MMA-MRNNet	**53.1**

4.2 消融研究

我们开展了一系列消融实验，以评估不同元素和组件对模型性能的影响。

最初，我们仅将单任务情感表征（从 MMA 中提取）作为循环神经网络的输入。随后，我们测试了两个任务的组合（例如，效价-唤醒度和动作单元），最后同时使用了来自全部三个任务的情感表征。结果汇总于表3中。为避免结果过于冗杂，我们仅列出每项实验的最佳表现。值得注意的是，即使仅使用效价和唤醒度表征，我们的网络表现也优于除 HFUT-CVers 之外的所有其他方法。加入额外的逐帧特征（如 7种基本表情或17个动作单元）后，模型性能显著提升。在双任务实验中，皮尔逊相关系数进一步提高了 1% 至 1.5%。最终，同时使用全部三个任务时，我们的方法取得了最高性能。

表3： MMA-MRNNet 在 Hume-Reaction 数据集验证集上的消融结果；皮尔逊相关系数结果以 % 表示

情感表征 来自 MMA 组件	**皮尔逊相关系数（ρ）**
VA	51.1
7种基本表情	52.6
17个动作单元	53.0
效价-唤醒度与7种基本表情	53.5
效价-唤醒度 & 17个动作单元	54.1
17个动作单元 & 7种基本表情	54.1
效价-唤醒度 & 7种基本表情 & 17个动作单元	**54.4**

为确定我们网络的最优架构，我们进行了多种配置的实验，包括使用不同的卷积神经网络（例如，用 ResNet50 替代 MMA）和循环神经网络（例如，用长短期记忆网络替代门控循环单元），以及改变层数和单元数量，具体如表4所示。在评估了广泛的组合后，我们确定最有效的配置由一个包含128个单元的单层门控循环单元组成，后接一个包含32个单元的前馈层。此外，我们还评估了引入掩码层、动态路由以及我们提出的损失函数（作为传统均方误差的替代方案）的影响。表3所示结果表明，这些组件显著提升了 MMA-MRNNet 的性能。

4.3 MMA 评估结果

在此，我们开展了一项广泛的实验研究，采用来自ABAW竞赛的顶尖方法（FUXI [58], SITU [45],CTC

表 4： MMA-MRNNet 在 Hume- Reaction 数据集验证集上的进一步消融结果；皮尔逊相关系数结果以 % 表示

模型	皮尔逊相关系数（ρ）
MMA +门控循环单元 + 全连接层 (64)	53.0
MMA +门控循环单元 + 全连接层 (16)	53.5
MMA + 门控循环单元 +全连接层 (8)	52.8
MMA + 2 × 门控循环单元 +全连接层 (32)	54.2
ResNet50 + 门控循环单元 +全连接层 (32)	51.5
MMA + 长短期记忆网络 +全连接层 (32)	53.2
MMA +门控循环单元 (256) + 全连接层 (32)	53.4
MMA +门控循环单元 (64) + 全连接层 (32)	53.8
MMA-MRNNet（无掩码和路由）	51.9
MMA-MRNNet（采用均方误差）	53.1
MMA-MRNNet	**54.4**

[61]）以及最先进方法（DACL [11], DAN [55], POSTER++ [47]; ME-GraphAU [46]& AUNets [52]），用于 7 种基本表情识别、AU 检测和效价-唤醒度估计，并将其性能与我们提出的MMA组件进行了比较。

如表 5 所示，我们提出的 MMA 组件在所有任务（7种基本表情识别、AU检测和效价-唤醒度估计）以及所有使用的数据库（Aff-Wild2、AffectNet 和 EmotioNet）上，均以显著优势超越了所有这些方法。

表 5： MMA 组件与各种最先进方法的性能比较（单位：%）。“CCC-VA”表示效价和唤醒度的一致性相关系数（CCC）均值。“F1 - Expr”表示 7种基本表情的 F1 分数均值（即宏平均），而“F1 - AUs”表示所用各数据库中所有动作单元的 F1 分数均值。

数据库	Aff-Wild2			AffectNet		EmotioNet
方法	CCC-效价-唤醒度	F1 - 表情	F1 - 动作单元	CCC-效价-唤醒度	F1 - 表情	F1 - 动作单元
SITU	64.14	38.24	54.22	71.1	59.0	77.2
CTC	56.66	33.11	48.87	71.0	57.7	74.6
FUXI	63.72	39.21	55.49	74.0	63.1	77.9
ME-GraphAU	-	-	-	-	-	72.9
AUNets	-	-	-	-	-	82.8
DAN	-	-	-	-	60.0	-
DACL	-	-	-	-	60.3	-
海报++	-	-	-	-	63.2	-
MMA	**67.38**	**43.21**	**58.87**	**78.2**	**65.4**	**85.4**

在多个数据集（Aff-Wild2、AffectNet

和 EmotioNet）上，将所提出的 MMA 方法与多种最先进的方法进行了比较评估，详情见表 5。MMA组件在所有数据集和所有评估指标上，均在所有任务（7类基本表情识别、AU检测和效价-唤醒度估计）中大幅超越了其他方法。具体而言，在 Aff-Wild2 数据集上，MMA 在效价-唤醒度估计方面比最接近的竞争方法 SITU 高出 3.24%。在 7 类基本表情识别方面，MMA 也比最接近的竞争方法 FUXI 高出 4 %，在 AU检测方面则高出 3.38%。在 AffectNet 数据集上，MMA 再次展现出卓越性能，在效价-唤醒度估计方面至少比最先进的方法高出 4.2%，在 7 类基本表情识别方面至少高出 2.2%。在 EmotioNet 数据集上，MMA 至少比最先进的方法高出 2.6%。这些结果凸显了 MMA 在提供精准、可靠的情感表征方面的鲁棒性和优越性。

5 结论

本文介绍了 MMA-MRNNet，这是一种用于从视频数据中进行动态多输出面部表情强度估计（FEIE）的新型深度学习架构。我们的方法利用多任务学习（MTL）框架提取丰富的情感表征，包括效价-唤醒度、基本面部表情和动作单元（AUs），从而克服了传统方法的局限性。随后，通过掩码路由循环神经网络（MRNN）进一步优化这些表征，使模型能够动态适应输入视频的可变长度序列，确保预测结果稳健且准确。

我们在 Hume-Reaction 数据集上证明了 MMA-MRNNet 的有效性，其性能始终大幅超越所有最先进方法。我们还在多个自然场景数据集上验证了 MMA组件的有效性，其在各种指标上始终优于所有最先进方法。我们的方法不仅能够处理复杂的视频级标注，还能缓解处理可变长度序列所带来的挑战，为情感计算领域的实际应用提供了一种灵活且强大的解决方案。

参考文献

1. Arsenos, A., Davidhi, A., Kollias, D., Prassopoulos, P., Kollias, S.：通过医学影像进行数据驱动的新冠肺炎检测。载于：2023年 IEEE 声学、语音与信号处理研讨会国际会议（ICASSPW）。第1–5页。IEEE（2023） 2. Arsenos, A., Karampinis, V., Petrongonas, E., Skliros, C., Kollias, D., Kollias, S., Voulodimos, A.：用于评估和提升空对空视觉目标检测鲁棒性的常见损坏。IEEE 机器人与自动化快报（2024） 3. Arsenos, A., Kollias, D., Petrongonas, E., Skliros, C., Kollias, S.：用于单源域泛化的、不确定性引导的对比学习。载于：ICASSP 2024—2024年 IEEE 声学、语音与信号处理国际会议（ICASSP）。第6935–6939页。IEEE（2024）

4. Arsenos, A., Petrongonas, E., Filippopoulos, O., Skliros, C., Kollias, D., Kollias, S.：Nefeli：用于提升先进空中交通自主性的深度学习检测与跟踪流程。可见 SSRN 4674579 5. Benitez-Quiroz, C., Srinivasan, R., Martinez, A.：EmotioNet：一种精准的实时算法，可在自然场景中自动标注百万个面部表情。载于：IEEE计算机视觉与模式识别国际会议（CVPR'16）论文集。美国内华达州拉斯维加斯（2016年6月） 6. Christ, L., Amiriparian, S., Baird, A., Tzirakis, P., Kathan, A., Müller, N., Stappen, L., Meßner, E.M., König, A., Cowen, A., Cambria, E., Schuller, B.W.：MuSe 2022多模态情感分析挑战赛：幽默、情绪反应与压力。载于：第三届多模态情感分析挑战赛论文集。计算机协会，葡萄牙里斯本（2022），研讨会于 ACM Multimedia 2022 举行，即将发表 7. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.：RetinaFace：自然场景中的单次多层级人脸定位。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第5203–5212页（2020） 8. Du, S., Tao, Y., Martinez, A.M.：复合情绪面部表情。《美国国家科学院院刊》111（15），E1454–E1462（2014） 9. Ekman, P.：面部动作编码系统（FACS）。A Human Face（2002） 10. Ekman, P., Friesen, W.V.：面部动作编码系统。《环境心理学与非言语行为》（1978） 11. Farzaneh, A.H., Qi, X.：通过深度注意力中心损失实现自然场景中的面部表情识别。载于：IEEE/CVF冬季计算机视觉应用会议论文集。第2402–2411页（2021） 12. Gerogiannis, D., Arsenos, A., Kollias, D., Nikitopoulos, D., Kollias, S.：通过人工智能辅助CT影像分析进行新冠肺炎计算机辅助诊断：部署医疗人工智能系统。arXiv预印本 arXiv:2403.06242（2024） 13. Guo, X., Polanía, L.F., Barner, K.E.：利用深度混合网络进行自然场景中的音视频情绪识别（2020）。https://doi.org/10.48550/ARXIV.2002.09023，09023 14 2002 https://arxiv.org/abs/ . . Hu, G., Kollias, D., Papadopoulou, E., Tzouveli, P., Wei, J., Yang, X.：重新思考情感分析：确保公平性与一致性的协议。arXiv预印本 arXiv:2408.02164（2024） 15. Hu, G., Papadopoulou, E., Kollias, D., Tzouveli, P., Wei, J., Yang, X.：弥合差距：实现公平且一致的情感分析协议。arXiv预印本 arXiv:2405.06841（2024） 16. Hu, P., Cai, D., Wang, S., Yao, A., Chen, Y.：学习用于自然场景情绪识别的监督式评分集成模型。载于：第19届 ACM 多模态交互国际会议论文集。第553–560页。ICMI '17，计算机协会，美国纽约州纽约（2017）。https://doi.org/10.1145/3143009 17 3136755 3143009，10 1145 3136755 / . https://doi.org/ . / . . Karampinis, V., Arsenos, A., Filippopoulos, O., Petrongonas, E., Skliros, C., Kollias, D., Kollias, S., Voulodimos, A.：确保无人机安全：仅依赖视觉的实时框架，通过目标检测、跟踪和距离估计实现碰撞规避。arXiv预印本 arXiv:2405.06749（2024） 18. Kollias, D., Schulc, A., Hajiyev, E., Zafeiriou, S.：分析首届 ABAW 2020竞赛中的情感行为。载于：2020年第15届IEEE自动人脸与手势识别国际会议（FG 2020 800 19 794）（FG）。第–页。Kollias, D.：ABAW：从合成数据中学习与多任务学习挑战。载于：欧洲计算机视觉会议。第157–172页。Springer（2022）

20. Kollias, D.：ABAW：效价-唤醒度估计、表情识别、动作单元检测与多任务学习挑战。arXiv预印本 arXiv:2202.10659（2022） 21. Kollias, D.：ABAW：合成数据学习与多任务学习挑战。载于：欧洲计算机视觉会议。第157–172页。Springer（2023） 22. Kollias, D.：多标签复合表情识别：C-expr 数据库与网络。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第5589–5598页（2023） 23. Kollias, D., Arsenos, A., Kollias, S.：利用人工智能辅助分析三维CT扫描，以诊断新冠肺炎及其严重程度。载于：2023 年 IEEE 国际声学、语音与信号处理会议研讨会（ICASSPW）。第1–5页。IEEE（2023） 24. Kollias, D., Arsenos, A., Kollias, S.：用于协调医学影像中三维输入数据分析与决策的深度神经网络架构。Neurocomputing**542**，126244（2023） 25. Kollias, D., Arsenos, A., Kollias, S.：用于协调医学影像中三维输入数据分析与决策的深度神经网络架构。Neurocomputing**542**，126244（2023） 26. Kollias, D., Arsenos, A., Kollias, S.：医学图像分析人工智能中的领域自适应、可解释性与公平性：基于三维胸部CT扫描的新冠肺炎诊断。arXiv预印本 arXiv:2403.02192（2024） 27. Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.：深度神经网络增强：生成人脸以进行情感分析。《国际计算机视觉杂志》**128**(5)，1455–1484（2020） 28. Kollias, D., Cheng, S., Ververas, E., Kotsia, I., Zafeiriou, S.：深度神经网络增强：生成人脸以进行情感分析。《国际计算机视觉杂志》第1–30页（2020） 29. Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.：使用深度神经网络识别自然场景中的情感。载于：2017年IEEE计算机视觉与模式识别研讨会（CVPRW）。第1972–1979页。IEEE（2017） 30. Kollias, D., Nicolaou, M.A., Kotsia, I., Zhao, G., Zafeiriou, S.：使用深度神经网络识别自然场景中的情感。载于：2017年IEEE计算机视觉与模式识别研讨会（CVPRW）。第1972–1979页。IEEE（2017） 31. Kollias, D., Psaroudakis, A., Arsenos, A., Theofilou, P.：Facernet：面部表情强度估计网络。arXiv预印本 arXiv:2303.00180（2023） 32. Kollias, D., Sharmanska, V., Zafeiriou, S.：定制化面部行为：在单一网络中实现表情、情感与动作单元。arXiv预印本 arXiv:1910.11111（2019） 33. Kollias, D., Sharmanska, V., Zafeiriou, S.：异构多任务学习的分布匹配：一项大规模人脸研究。arXiv预印本 arXiv:2105.03790（2021） 34. Kollias, D., Sharmanska, V., Zafeiriou, S.：分类任务多任务学习的分布匹配：一项关于人脸及更多领域的大规模研究。载于：AAAI人工智能会议论文集。第38卷，第2813–2821页（2024） 35. Kollias, D., Tzirakis, P., Baird, A., Cowen, A., Zafeiriou, S.：ABAW：效价-唤醒度估计、表情识别、动作单元检测与情绪反应强度估计挑战。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第5888–5897页（2023） 36. Kollias, D., Tzirakis, P., Baird, A., Cowen, A., Zafeiriou, S.：ABAW：效价-唤醒度估计、表情识别、动作单元检测与情绪反应强度估计挑战。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第5888–5897页（2023）

37. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Shao, C., Hu, G.：第六届野外情感行为分析竞赛（ABAW）。arXiv预印本 arXiv:2402.19344（2024） 38. Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.：自然场景中的深度情感预测：Aff-Wild数据库与挑战、深度架构及其他。《国际计算机视觉杂志》第1–23页（2019） 39. Kollias, D., Vendal, K., Gadhavi, P., Russom, S.：BTDNet：一种用于脑肿瘤放射基因组学分类的多模态方法。《应用科学》**13**(21)，11984（2023） 40. Kollias, D., Zafeiriou, S.：表情、情感、动作单元识别：Aff-Wild2、多任务学习与ArcFace。arXiv预印本 arXiv:1910.04855（2019） 41. Kollias, D., Zafeiriou, S.：自然场景中的情感分析：效价-唤醒度、表情、动作单元与统一框架。arXiv预印本 arXiv:2103.15792（2021） 42. Kollias, D., Zafeiriou, S.：第二届ABAW2竞赛中的情感行为分析。载于：IEEE/CVF国际计算机视觉大会论文集。第3652–3660页（2021） 43. Kollias, D., Zafeiriou, S., Kotsia, I., Dhall, A., Ghosh, S., Shao, C., Hu, G.：第七届ABAW竞赛：多任务学习与复合表情识别。arXiv预印本 arXiv:2407.03835（2024） 44. Li, J., Chen, Y., Zhang, X., Nie, J., Yu, Y., Li, Z., Wang, M., Hong, R.：利用Transformer进行多模态特征提取与融合，以估计视频中的情绪反应强度并分类表情。arXiv预印本 arXiv:2303.09164（2023） 45. Liu, C., Zhang, X., Liu, X., Zhang, T., Meng, L., Liu, Y., Deng, Y., Jiang, W.：基于多模态特征的野外视频面部表情识别。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第5871–5878页（2023） 46. Luo, C., Song, S., Xie, W., Shen, L., Gunes, H.：学习用于面部动作单元识别的基于多维边缘特征的AU关系图。载于：第三十一届国际人工智能联合会议论文集。第1239–1246页（2022） 47. Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A.：POSTER v2：一种更简单、更强大的面部表情识别网络。arXiv预印本 arXiv:2301.12149（2023） 48. Miah, H., Kollias, D., Pedone, G.L., Provan, D., Chen, F.：机器学习能否辅助诊断原发性免疫性血小板减少症？一项可行性研究。arXiv预印本 arXiv:2405.20562（2024） 49. Mollahosseini, A., Hasani, B., Mahoor, M.H.：AffectNet：一个用于自然场景下面部表情、效价与唤醒度计算的数据库。arXiv预印本 arXiv:1708.03985（2017） 50. Plutchik, R.：情绪的心理进化理论（1982） 51. Qiu, F., Ma, B., Zhang, W., Ding, Y.：利用时间增强进行多模态情绪反应强度估计。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第5777–5784页（2023） 52. Romero, A., León, J., Arbeláez, P.：多视角动态面部动作单元检测。《图像与视觉计算》（2018） 53. Vaiani, L., La Quatra, M., Cagliero, L., Garza, P.：ViPER：用于情绪识别的基于视频的感知器。载于：第三届多模态情感分析研讨会与挑战赛论文集。第67–73页（2022）

54. Wang, S., Wu, J., Zheng, F., Li, X., Li, X., Wang, S., Wu, Y., Chang, Y., Miao, X.：基于多模态数据的情绪反应强度估计。arXiv 预印本 arXiv:2303.09167（2023）55. Wen, Z., Lin, W., Wang, T., Xu, G.：Distract your attention：用于面部表情识别的多头交叉注意力网络。仿生学 8(2), 199（2023）56. Yu, J., Cai, Z., Li, R., Zhao, G., Xie, G., Zhu, J., Zhu, W.：探索大规模未标注人脸以增强面部表情识别。arXiv 预印本 arXiv:2303.08617（2023）57. Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.：Aff-wild：野外效价与唤醒度挑战赛。载于：计算机视觉与模式识别研讨会（CVPRW），2017年IEEE会议。第1980–1987页。IEEE（2017）58. Zhang, W., Ma, B., Qiu, F., Ding, Y.：基于掩码自编码器的多模态面部情感分析。载于：IEEE/CVF计算机视觉与模式识别大会论文集。第5792–5801页（2023）59. Zhang, Z., An, L., Cui, Z., Dong, T., 等：基于Transformer编码器和视听融合的面部情感识别，面向ABAW5挑战赛。arXiv 预印本 arXiv:2303.09158（2023）60. Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H., Keutzer, K.：用于用户生成视频情绪识别的端到端视觉-音频注意力网络（2020）。https://doi.org/10.48550/ARXIV.2003.00832，https: //arxiv.org/abs/2003 00832 61 . . Zhou, W., Lu, J., Xiong, Z., Wang, W.：结合时序卷积网络和 Transformer，实现有效的连续情绪识别视听融合。载于：IEEE/CVF计算机视觉与模式识别大会论文集。第5756–5763页（2023）