多模态迁移学习在复杂环境中的动态面部情绪识别

Ezra Engel
eengel9@gatech.edu
Georgia Institute of Technology

Chris Hudy
chudy3@gatech.edu
Georgia Institute of Technology

Lishan Li
lli680@gatech.edu
Georgia Institute of Technology

Robert Schleusner
rschleusner@gatech.edu
Georgia Institute of Technology

Abstract

面部表情识别（FER）是计算机视觉的一个子集，对人机交互、医疗保健和客户服务具有重要的应用价值。FER 代表了一个具有挑战性的问题领域，因为准确分类需要模型区分面部特征的细微变化。在本文中，我们研究了多模态迁移学习在提高基于视频的挑战性 FER 数据集——野外动态面部表情（DFEW）上的性能的应用。通过结合预训练的 ResNets、OpenPose 和 OmniVec 网络，我们探讨了跨时间、多模态特征对分类准确率的影响。最终，我们发现这些精细调整后的多模态特征生成器适度提高了基于变压器的分类模型的准确性。

1. 介绍/背景/动机

面部情感识别（FER）是计算机视觉的一个具有挑战性的子领域，对于人机交互有着重要的应用。人脸是人们传递信息的重要机制，当前文献表明，人类沟通中有很大一部分由肢体语言、面部表情和语调组成 [7][9]。随着计算机成为我们日常生活和工作中越来越重要的一部分，提高它们识别并响应这些非言语交流信号的能力至关重要。

自动计算机识别人类情绪在广泛的人机交互系统中具有广泛应用，包括客户服务、驾驶员疲劳检测、计算治疗、视频游戏设计等 [3][5]。然而，在实验室环境之外的面部表情识别仍然是一个极其困难的任务 [8]。许多当前模型仅关注单一通信模式（文本、面部表情和音频情感）[3]。我们旨在使用一系列模型来整合这些模态中的信息，并提高人类情绪分类在实际应用中的性能。

许多当前和历史模型专注于静态面部表情识别。在这些模型中，一系列静止的、实验室控制下的面部图像被分类为七种情感类别：快乐、悲伤、中性、愤怒、惊讶、厌恶和恐惧 [3]。实验室任务使用标准化光照条件下的正面面部图像，这极大地简化了任务。在非受控的真实世界设置（野外）中进行准确的 FER 仍然极其困难。

强调野外应用的方法仍然主要集中在使用单一模态来分类面部表情。这些包括静态（图像）和动态（视频）分类任务。先前的研究在将 3D 卷积网络应用于序列图像数据 [15][10] 方面取得了一些成功，但单模模型的整体性能对于许多实际应用来说仍显不足 [3]。

多模态方法近年来因其深度学习促进从手工特征向学习表示的转变而获得了显著的关注。许多当前的研究展示了多模态情感数据在提高野外分类任务模型准确性方面的实用性。

基于视频的 FER 数据集，包括音频和其他情境线索，促进了超越静态图像的发展。在这项工作中，我们关注野外动态面部表情（DFEW）数据集——一个包含

来自超过 1,500 部电影的 16,372 个视频片段的大规模 FER 数据集，每个视频片段都标注了七种主要情感类别的强度。DFEW 提供了简短、不受限制的视频片段，使其非常适合时空和多模态方法。

我们旨在将这些发现应用于一个相对较新的数据集，该数据集专注于野外动态 FER，DFEW[13]。与其他目前可用的野外 FER 数据集主要由静态图像组成不同，DFEW 数据集包含 16,372 个不受限制的表情视频片段，并配有稳健的注释。

在本文中，我们提出了一种多模态 FER 模型，该模型利用了三个模态（视觉、音频和姿态）中的 SOTA 预训练模型的迁移学习。我们的架构集成了一个 ResNet-18 CNN（用于面部帧）、一个 OpenPose 网络（用于身体姿势）和一个 Wav2Vec2 模型（用于语音音频），以从每个片段中提取互补特征。然后使用基于变压器的序列模型将这些特征组合起来，捕捉时间依赖性，最后在浅层密集网络中进行分类。最终，这种方法在动态野外数据上的表现得到了提升，超越了原始 DFEW 基准，并接近近期 SOTA 架构的准确性。

2. 方法

与许多先前的 FER 方法在架构早期执行特征融合不同，我们的设计利用了每个模态的模块化预训练特征提取器，并将它们的集成推迟到更深的层。通过为视觉、听觉和姿态线索使用单独的预训练模型，我们可以利用从大规模单模态数据集（如 ImageNet）中学到的表示。决策级融合还提供了训练灵活性，因为每个模态可以在平行且干扰有限的情况下进行训练。我们假设这种架构有助于跨模态关系的学习，因为每个输入流首先被投影到一个流水线潜在特征空间。

我们将 FER 分类任务分解为四个步骤：(1) 图像/音频预处理，(2) 特征提取，(3) 跨模态和时序整合，以及 (4) 决策网络。具体来说，在训练一个多层 Transformer 编码器以跨模态及时序空间整合特征的同时，我们以端到端的方式微调预训练的特征提取器。我们的架构完整图示见图 1。

2.1. 图像预处理

DFEW 数据集的创建者包括原始片段和预处理帧。原作者使用 OpenCV 从原始片段中提取图像帧，然后应用 face++ API 提取面部区域图像和面部特征点 [1]。然后，他们使用 SeetaFace[16] 对提取的面部进行归一化。最后，作者通过在处理后的序列图像之间插值来标准化序列长度，为每个片段创建 16 个预处理帧 [24][23]。

2.2. 音频预处理

对于音频数据，我们利用预训练的 Wave2Vec2 模型从原始音频片段中提取高维嵌入。具体而言，我们将原始音频通过 Wave2Vec2 框架进行处理，该框架生成一个 1024 维特征向量 [2]。提取的嵌入经过归一化以确保一致的比例。随后，我们通过对归一化的嵌入应用一系列线性层和批归一化层来降低维度，首先降至 512 维，然后降至 128 维 [19, 12]。最终的 128 维特征向量代表了处理后的音频，并为与视频和 OpenPose 特征在后续阶段进行集成做好准备。

2.3. 特征提取

我们使用三个预训练模型从原始剪辑、预处理的面部帧和音频数据中提取时空特征。其中两个模型（ResNet [11] 和 Wave2Vec2 [2]）在训练变换器和决策层时进行了微调。

我们使用在 ImageNet 上预训练的 Torch Vision 实现的 ResNet-18 作为面部特征提取的基础模型 [17]。此流程图显示在图 1 的顶部分支中。首先，我们将预训练模型的最后一层全连接层剥离（512x1000）。然后，对于每个经过预处理的 16 帧，我们生成一个长度为 512 的向量，该向量捕捉来自帧的有意义的抽象特征。这些向量被拼接成一个长度为 16 的序列，然后与可学习的位置嵌入相加，并输入到多层变压器编码器中。

我们使用预训练的 PyTorch 版本的 OpenPose[4] 提取肢体语言信息。该模型处理从原始视频片段中提取的三个全身帧，生成每个帧形状为 (17, 3) 的张量的各种关节关键点数据。我们丢弃第三个值，这个值代表一个始终等于 1 的置信度得分，将张量减少到 (17, 2)。这些关键点被展平并通过全连接层传递，将其转换成形状为 (512, 3) 的更高维度表示，其中每一列对应一个帧。最后，这些特征与基于视频的嵌入特征结合，形成长度为 19 的组合特征序列。

最后，我们使用一个预训练的音频模型（例如，Wave2Vec2[19, 12]）从音频数据中提取一个 1024 维的嵌入。在归一化这个嵌入之后，我们应用一系列线性

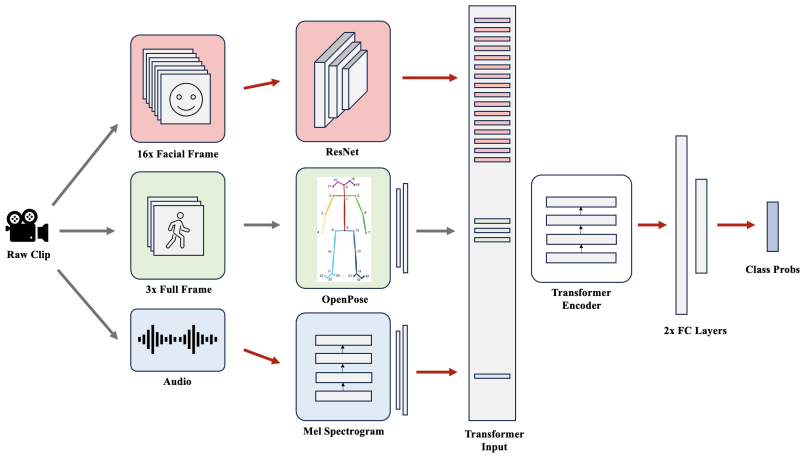

图 1. **原始多模态 FER 模型架构**。我们结合三个预训练模型来提取面部特征、身体姿态数据和音频情感信息。然后，我们利用多层变压器编码器学习时空依赖性和关系以提高 FER 分类的准确性。最终，由于初始结果不佳，我们选择直接将 Mel 谱图和 OpenPose 特征连接到全连接层。

和批标准化层，首先将其维度降低到 512，然后降低到 128。得到的 128 维音频特征向量随后与视频和 OpenPose 特征拼接在一起，然后再传递给序列变换器。

2.4. 变压器层

我们的 20 长度顺序特征向量是直接输入到使用内置 PyTorch 实现的多层变压器编码器 [17][18] 。我们使用一个具有 6 层、8 个头和前馈大小为 2048 的编码器。变压器编码器的目标是从顺序多模态输入中提取信息，并利用它们生成一组新的特征，这些特征能够有效捕捉我们提取的特征之间的跨时间与跨模态关系。编码器的输出然后被展平并通过我们的决策网络。

2.5. 决策网络

决策网络由两个全连接层组成，最终输出大小为 7。最终仿射变换的输出传递到 log-softmax 函数以计算归一化类别分数。DFEW 数据集包括一类属性的一热标签和每个片段的个别类分数 [13]。我们从知识蒸馏中获得一些灵感，并选择直接在类分数（而不是一热标签）上进行训练。为此，我们在预测类别概率和真实类别分数之间使用 Kullback-Leibler 损失函数。我们的希望是通过包含模糊和细微的得分，模型能够更好地学习提取特征之间的关系和相关性。我们预计这将提高整体模型性能。

2.6. 评估指标

为了便于与 DFEW 作者的基准进行直接比较，并处理类别不平衡问题，我们使用加权平均召回率（WAR）作为分类任务 [13] 的评估指标。WAR 仅仅是分类器的准确率。原始作者还使用了方程 1 中描述的非加权平均召回率，其中 n 是类别数量，tp_k 是真正例，tn_k 是真负例，而 T 是总预测数。（下标 k 表示每类的预测）。

$$UAR = \sum_{k=1}^{n} \frac{1}{n} \frac{tp_k + tn_k}{T} \tag{1}$$

我们将主要分析限制在 WAR 上，但在第 3.3 节中包括精确召回曲线。在计算指标时，我们首先选择忽略所有类别得分均不大于 6 的样本（这是原 DFEW 论文中使用的相同阈值）。该阈值用于消除足够模糊的样本以进行 WAR 计算，理解这些示例无法干净地映射到单个类别标签。

2.7. 方法讨论

许多之前的多模态 FER 研究主要集中在在网络特征提取部分应用模态融合。虽然这种方法已经实现了相对较高的性能，但特征融合使得利用更大语料库的预训练变得困难，并且也增加了使用自监督或半监督技术进行性能提升的复杂性。

通过将相对模块化的特征提取器与稳健的变压器

架构相结合，我们希望利用多模态信息和在更大数据集和半监督任务上的预训练优势。通过有效地将音频和手势信息整合到时空决策架构（变压器和全连接层）中，我们希望能观察到相比于仅使用 16 个预处理面部帧的基线 ResNet/Transformer 模型有显著改进。

在野外进行 FER 任务是一个极其困难的任务（DFEW 作者建立了一个大约 45.35%的 UAR 基准）。鉴于该任务的难度以及一个相对高容量模型，我们预计在训练和过拟合方面会遇到困难。出于这些原因，我们预期需要执行大量的超参数调整（ResNet 深度、变换器深度、学习率、调度程序、dropout）以在评估集上实现合理的表现。

我们的初始 ResNet 模型（单模式）的表现明显低于 DFEW 基准。然而，在超参数调整过程中，我们发现通过降低学习率可以显著提高性能。我们还发现，将多模态特征作为输入传递给变压器会严重损害整体模型的性能。绕过变压器层并将这些多模态输入直接作为全连接决策层的输入对整体性能有一定的影响，尽管这种影响较小但可察觉。这些实验在下一节中进一步讨论。

3. 结果与讨论

我们的最终模型是通过实证驱动的迭代设计过程得出的结果。在开发和测试过程中，我们发现（1）该模型对我们超参数和学习率的选择非常敏感；（2）将跨模态特征向量作为变换器输入会降低模型的性能。

3.1. 超参数实验

骨干模型由预训练的 ResNet18 层作为特征提取器组成，配以变压器模型来学习如何关注连续面部帧（16 帧）之间的变化，以便识别每个片段中表达的核心情感。第一个实验涉及端到端地训练 ResNet-Transformer 混合模型。调整模型的最终目标是在保持训练和测试指标差异较小的同时实现相对较高的训练准确性，以减轻高方差和过拟合。

我们使用两种不同的损失函数训练了基线模型：均方误差和 Kullback-Leibler 散度。由于类标签的模糊性，我们决定探索这两种方法。7 长度注释向量的一种可能解释是每个单独类别的归一化分数列表。对于这些标签代表归一化分数的情况，均方误差将是一个适当的指标。对标签的另一种有效解释是在 7 个情感类别上的离散概率分布。在这种情况下，KL-散度是一个更合适的指标。因此，我们探索了这两种损失函数的表现，并最终得出它们表现相似的结论。这些实验在以下各节中进行了详细说明。

3.1.1 均方误差

最初，两个模型都显示出有限的进展，训练准确率在 10 个周期内停滞在 26%到 35%之间，表明模型难以从训练数据中学习。进一步分析后，我们发现了一个类别不平衡问题，最大的类别包含 4,209 个样本，而最小的类别仅有 145 个样本。在基于类别分布进行分层数据分割并对训练数据进行归一化处理后，模型开始逐步学习。

图 2 显示了 MSE 模型前 10 个周期的结果，在此期间我们观察到训练和验证损失之间的差距逐渐增大，而训练和验证准确率稳步上升，表明有效学习。为进一步提升性能，我们在额外的 5 个周期中使用 L2 正则化来防止过拟合对模型进行了训练。然而，在这些额外周期之后，验证准确率并未显著提高。

参数调整过程主要集中在调整 Adam 优化器中的学习率和权重衰减。基于验证损失的最佳 MSE 模型实现了 80.13%的训练准确率和 70.37%的验证准确率，在保持过拟合控制的同时提供了稳定的结果。

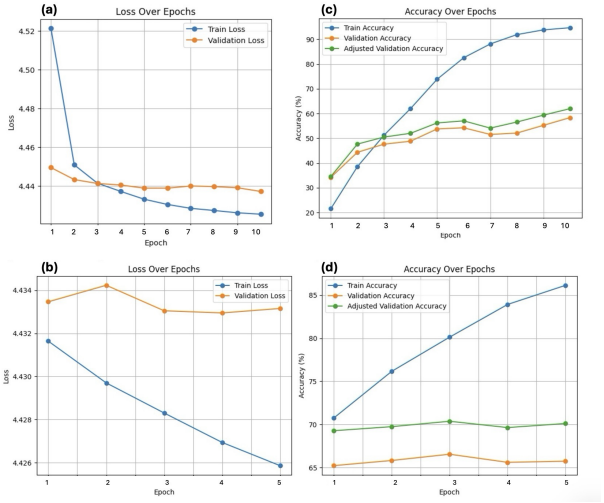

图 2. 情感分类器使用均方误差（MSE）损失训练的模型性能。图 (a) 和 (c) 显示了前 10 个周期无正则化的结果，而图 (b) 和 (d) 表示将正则化项（权重衰减）设置为 0.0001 时的性能。

3.1.2 Kullback-Leibler 散度

KLD 模型的训练从默认参数值开始：学习率为 0.0001，使用 Adam 优化器，并将 KL 散度作为损失函数。在最初的 10 个纪元中，训练损失持续下降，而验证损失开始增加。尽管存在这种分歧，图 3 中显示了验证准确率有所提高的迹象。

为了解决类别不平衡问题，我们切换到了带有类别权重的交叉熵损失函数，但在额外的 5 个周期后并未观察到显著改进。为了缩小训练和验证性能之间的差距，应用了权重衰减为 0.0001 的正则化方法，但这同样未能产生可测量的改进。

我们通过使用初始学习率为 0.0001 并结合调度器在平台期降低学习率的方法，进一步提高了模型的准确性。这个最终的 KLD 模型达到了 98.67%的训练准确率（由于权重衰减）和 72.40%的验证准确率。

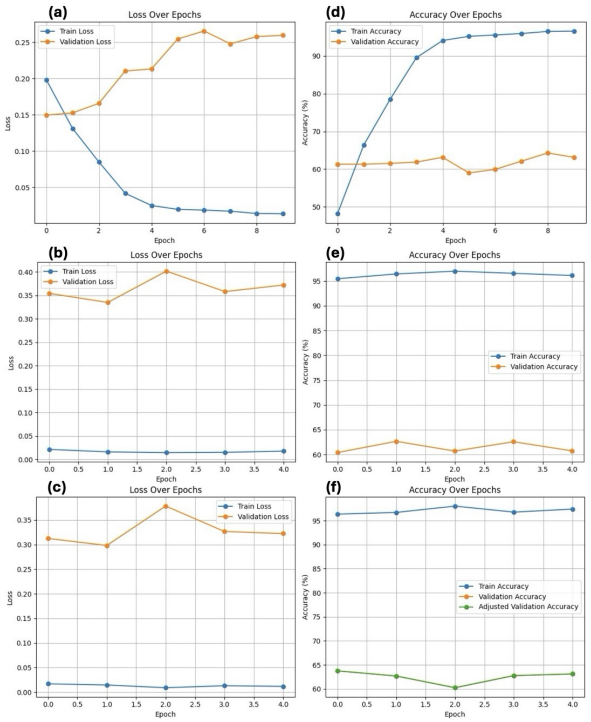

图 3. 情感分类器使用 KL 散度（KL）损失训练的模型性能如图所示。图 (a) 和 (d) 显示了前 10 个周期使用 KL 损失的结果。图 (b) 和 (e) 描绘了从 KL 损失切换到加权类分布的交叉熵损失后的性能。最后，图 (c) 和 (f) 展示了当应用正则化项（权重衰减）值为 0.0001 时的性能。

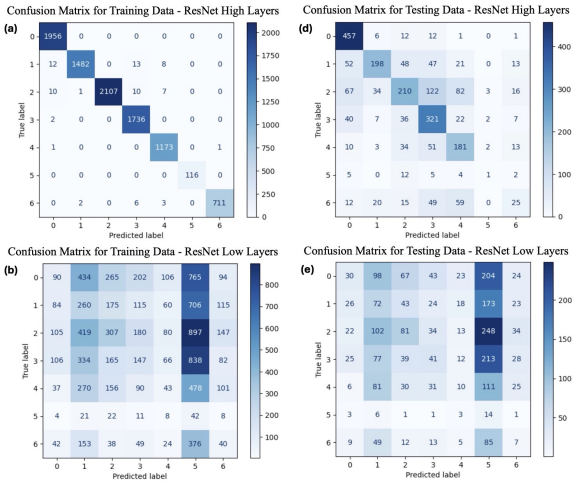

图 4. 混淆矩阵用于训练具有较低或较高 ResNet 层的分类模型。

3.1.3 带有冻结层的 ResNet KLD 模型

初始方法是通过预训练的 ResNet 模型传递图像以提取面部特征，旨在提高训练效率。骨干模型进行了端到端的训练，结合了 ResNet 和变压器架构。如先前结果所示，这种方法在区分人类情感方面表现出强大的能力。ResNet 在包含 1,000 个对象类别的 ImageNet 数据集上的数百万张图像上进行了预训练，作为一个有效的特征提取器。然而，尚不清楚是否需要重新训练整个 ResNet 模型，或者仅微调部分层就足够了。为了解决这个问题，我们进行了一项比较分析。

关于 CNN 模型如何学习视觉特征 [22] 的研究论文表明，ResNet 的较低层侧重于学习基本特征，如颜色、边缘和线条，而较高层则捕捉更抽象的任务特定特征。低级特征是通用且可迁移的，而高级特征则是针对特定任务定制的。图 4 说明了这一点：冻结所有 ResNet 层（除了最后三层高级层）的结果与端到端模型的性能相当。相比之下，冻结除最初的三个低级层之外的所有层（低级 ResNet），导致表现不佳，预测偏向于八个情感类别中的单一类别。

这些结果与预期一致，并得出一个重要结论：当使用预训练模型进行下游任务时，专注于高级层的训练通常会带来更好的性能。此外，端到端模型、高级 ResNet 和低级 ResNet 的可训练参数数量分别为 21,691,463、21,008,391 和 10,524,487。这表明在保持类似模型性能的同时，计算资源大约减少了 1%。另外，高级 ResNet 和低级 ResNet 之间可训练参数数量的显著差异有助于

解释为什么低级 ResNet 失败。

3.2. 多模态实验

我们最初尝试将 OpenPose 和 Wave2Vec2 特征用作变压器输入的方法失败了。这些特征的加入实际上恶化了模型性能。虽然一个调校良好的仅 ResNet 模型可能达到接近 70% 的验证准确率，但多模态模型在训练集上始终无法超越 30% 的验证准确率，并且表现同样糟糕。

我们假设这些多模态特征之间的初始不匹配如此之大，以至于变换器架构中的自注意力头无法识别输入之间的有意义关系。我们最初认为通过反向传播，模型可能能够将 ResNet 特征、OpenPose 特征和 Wave2Vec2 映射到同一个抽象特征空间中。然而，在多次实验之后，很明显试图通过变换器层将这些模型微调到一个连贯的特征空间是不太可能成功的。

相反，我们选择完全绕过带有多模态输入的变换器。我们将这些特征连接到最终的全连接网络，而不是连接到变换器。这避免了许多尝试通过自注意力机制反向传播不同特征空间的问题。一旦我们转向这种新架构，模型性能有了小幅提升。完整的总结结果如表 1 所示。

3.3. 结果

图 5 总结了仅使用 ResNet 的 KL 散度模型的整体性能。该模型实现了相对较高的整体分类准确率，但在较少见的情绪如厌恶和恐惧上表现不佳。这是一个在文献中常见的问题，并且有两个主要原因。首先，在野外数据集倾向于大量偏好快乐、悲伤、愤怒和中性面孔，因为这些阶段通常是非实验室图像中最常见的情绪。

其次，厌恶和恐惧对面部特征的影响往往比其他主要情绪（例如微笑）更为微妙。图 6 显示了大多数情绪类别分布的显著偏斜。具体来说，“快乐”、“悲伤”和“愤怒”类别的评分较高，这表明这些情绪通常与更强的表情相关联。这也使它们更容易被识别，如图 5 所示，与其他四种情绪类型相比。这一点与先前文献中的发现一致 [14]。

表 1 显示了我们选择的 ResNet-Transformer 架构和文献中的模型的 WAR。正如讨论过的，当 OpenPose 和 Wave2Vec2 特征连接到 transformer 层时，它们会通过混淆自我注意机制来降低模型性能。相比之下，当

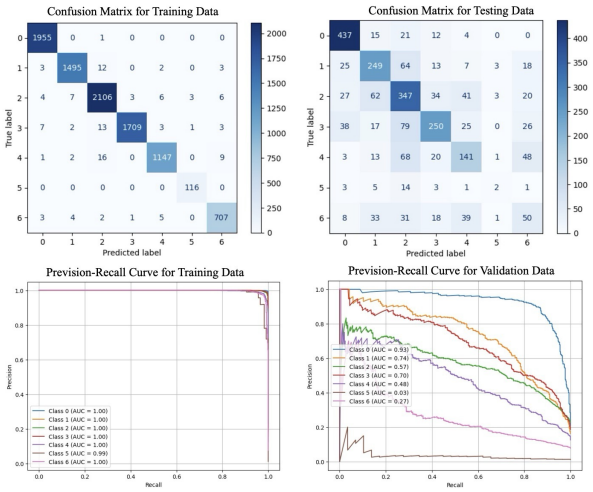

图 5. 混淆矩阵和精确率-召回率曲线对于使用 KL 损失训练的情感分类器的训练集和测试集。

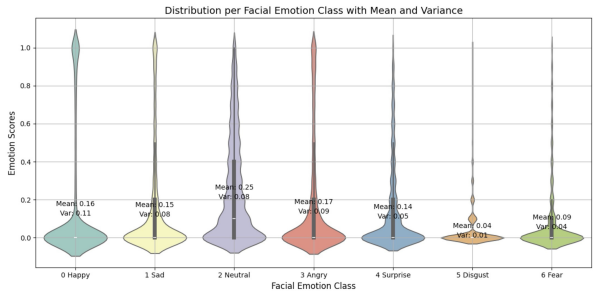

图 6. 每种情绪类别下的数据分布箱线图。

模型	战争
ResNet-Only	72.40%
ResNet+OpenPose^T	25.74%
ResNet+Wav2Vec2^T	26.87%
Full Multimodal^{FC}	72.97%
EC-STFL[13]	56.51%
M3-DFEL[20]	69.25%
MMA-DFER[6] (SOTA)	77.51%

表 1. WAR 对我们模型架构和其他文献中的基准进行了多种评估。上标 T 和 FC 分别表示与变压器层和全连接层相连的多模态特征。EC-STFL 是 DFEW 作者的原始基准。

多模态特征直接连接到 FC 层时，我们看到分类准确率有适度的提升。

我们的最终模型优于原始 DFEW 作者实现的表情聚类时空特征学习（EC-STFL），并且与其他基准相比表现相对较好。

本文中提出的 ResNet-Transformer 模型的性能可能通过一些额外的改进进一步提升。多模态特征空间的对齐是多模态学习问题中的一个已知问题 [21]。一种潜在的解决方案（以及未来研究的方向）可能是首先训练一个浅层全连接网络，将这些不同的特征映射到同一特征空间中。这种逐步训练可能避免由于自注意力机制输入不同而导致的问题。

参考文献

[1] M. inc face++ research. [n.d.]. toolkit. www.faceplus.com. 2

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33:12449–12460, 2020. 2

[3] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. Information Sciences, 582:593–617, 2022. 1

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 2

[5] M Kalpana Chowdary, Tu N Nguyen, and D Jude Hemanth. Deep learning-based facial emotion recognition for human–computer interaction applications. Neural Computing and Applications, 35(32):23311–23328, 2023. 1

[6] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4673–4682, 2024. 6

[7] Starkey Duncan Jr. Nonverbal communication. Psychological bulletin, 72(2):118, 1969. 1

[8] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2402–2411, 2021. 1

[9] Judith A Hall, Terrence G Horgan, and Nora A Murphy. Nonverbal communication. Annual review of psychology, 70(1):271–294, 2019. 1

[10] Behzad Hasani and Mohammad H Mahoor. Facial expression recognition using enhanced deep 3d convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 30–40, 2017. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[12] Pourya Jafarzadeh, Amir Mohammad Rostami, and Padideh Choobdar. Speaker emotion recognition: Leveraging self-supervised models for feature extraction using wav2vec2 and hubert. arXiv preprint arXiv:2411.02964, 2024. Accessed: 2024-12-08. 2

[13] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In Proceedings of the 28th ACM international conference on multimedia, pages 2881–2889, 2020. 2, 3, 6

[14] Jeniffer Xin-Ying Lek and Jason Teo. Academic emotion classification using fer: A systematic review. Human Behavior and Emerging Technologies, 2023(1):9790005, 2023. 6

[15] Huibin Li, Jian Sun, Zongben Xu, and Liming Chen. Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network. IEEE Transactions on Multimedia, 19(12):2816–2831, 2017. 1

[16] Xin Liu, Meina Kan, Wanglong Wu, Shiguang Shan, and Xilin Chen. Viplfacenet: an open source deep face recognition sdk. Frontiers of Computer Science, 11:208–218, 2017. 2

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. 2, 3

[18] A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017. 3

[19] Peter Vieting, Ralf Schlüter, and Hermann Ney. Comparative analysis of the wav2vec 2.0 feature extractor. 2023. Accessed: 2024-12-08. 2

[20] Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. Rethinking the learning paradigm for dynamic facial expression recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 17958–17968, 2023. 6

[21] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. arXiv preprint arXiv:1909.05645, 2019. 7

[22] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In European Conference on Computer Vision (ECCV), pages 818–833, 2014. 5

[23] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, and Matti Pietikäinen. A compact representation of visual speech data using latent variables. IEEE transactions on pattern analysis and machine intelligence, 36(1):1–1, 2013. 2

[24] Ziheng Zhou, Guoying Zhao, and Matti Pietikäinen. Towards a practical lipreading system. In CVPR 2011, pages 137–144. IEEE, 2011. 2