动态面部表情识别的新范式

摘要

动态人脸表情识别 (DFER) 是一个快速发展的领域，主要用于识别视频人脸表情序列。先前的研究将非目标帧视为噪声帧，但我们提出应该将其作为弱监督问题来处理。我们还发现了 DFER 中短期和长期时间序列关系的不平衡性。因此，我们引入了多 3D 动态人脸表情学习 (M3DFEL) 框架，该框架利用多示例学习 (MIL) 来处理不精确的标签。M3DFEL 生成 3D 实例来建模强的短期时间关系，并利用 3DCNNs 进行特征提取。然后，利用动态长期实例聚合模块 (DLIAM) 学习长期时序关系并动态聚合实例。我们在 DFEW 和 FERV39K 数据集上的实验表明，M3DFEL 优于现有的采用 R3D18 主干的最新方法。

关键词：动态面部表情识别；多示例学习；动态多实例归一化

1 引言

面部表情在日常交流中至关重要 [13, 14]。通过他人的面部表情理解情绪在对话中至关重要。因此，面部表情的自动识别在诸如人机交互（HCI）[12]、心理健康诊断 [6]、驾驶员疲劳监测 [11] 和数字人 [2] 等各个领域都是一项重大挑战。目前在静态面部表情识别（SFER）方面已经取得了重大进展，越来越多的工作关注对动态面部表情的识别。

随着大规模野外数据集，如 DFEW [5] 和 FERV39K [18] 的出现，人们已经为 DFER 提出了几种方法 [9, 10, 15]。以前的工作 [15, 22] 只是简单地应用一般的视频理解方法来识别动态面部表情。Li 等人 [10] 观察到 DFER 含有大量的噪声帧，并提出了一个动态类标记和基于片段的过滤器来抑制这些帧的影响。Li 等人 [9] 提出了一种强度感知损失（Intensity Aware Loss），以考虑 DFER 中大量的类内和类外损失，使网络对于最不确定的类别提供更多的注意力。然而，我们认为 DEFR 需要特殊的设计，而非被视作视频理解和 SFER 的结合。尽管这些工作 [15, 22] 发现了 DFER 中的一些问题，他们的模型只是用了很粗糙的方式解决它们。

首先，这些工作没有认识到，DFER 中非目标帧的存在实际上是由弱监督造成的。在收集大规模的视频数据集时，注释标签的精确位置是消耗大量人力且具有挑战性的。一个动态的面部表情可能包含非目标情绪和目标情绪之间的变化，如图1所示。如果没有一个可以指导模型的位置标签，忽略不相关的帧而将注意力放在目标上，模型很可能被不准确的标签所误导。因此，将这些非目标帧直接建模为噪声帧是草率的，而背后的弱监督问题仍然没有得到解决。

第二，以前的工作直接复用时间序列模型，而非针对 DFER 进行专门设计。然而，我们发现在 DFER 中存在小范围时序关系和大范围时序关系的不平衡。例如，一些微表情可能在

一个很短的片段中出现，而一些面部运动可能干扰单个帧，如图1所示。与此形成对比的是，一个在视频开始的开心表情和视频结束的开心表情之间几乎没有时序关系。所以，对整个时序关系进行建模或者使用完全不包含时序关系的聚合方法对于 DFER 都不适用。取而代之的是，一个理想的方法应该学习对强的小范围时序关系和弱的长范围时序关系进行建模。

为了解决第一个问题，我们建议使用弱监督方法来训练 DFER 模型，而非将非目标帧当作噪声帧。特别的，我们提出将 DFER 当作多实例学习问题进行建模，其中每一个视频都被认为是一个包含一序列实例的包。在这个多实例学习的框架下，我们不考虑视频中的非目标帧，只关注目标情绪。然而大部分多实例学习的方法是不考虑时间的，因此我们要设计专注于 DFER 的 MIL 框架，以解决不平衡的小范围时序关系和长范围时序关系。

本文提出的 M3DFEL 被设计来用一种统一的方法解决不平衡的小范围时序关系和大范围时序关系，以及弱监督问题。它使用 3D-Instance 和 R3D18 的结合，来加强小范围时序关系的学习。DLIAM 被专门设计来捕捉实例之间长范围的时序关系，实例的特征被提取后，立即输入到 DLIAM 当中，它把特征聚合成包级别的表达。此外，DMIN 通过进行动态归一化来维持包级别和实例级别的时间一致性。

图 1. 野外的动态面部表情

总体上，我们的贡献可以被概括为如下：

1. 我们提出了一个弱监督方法来对 DFER 作为一个多实例学习问题进行建模。我们认识到 DFER 中小范围时序关系和长范围时序关系的不平衡，这使得建模整个时序关系或者使用时间无关模型都是不合适的。
2. 我们使用了 M3DFEL 框架提供一个统一的解决方案，针对弱监督问题和 DFER 中不平

衡的小范围时序关系和长范围时序关系的建模。

3. 我们在 DFEW 和 FERV39K 上进行了广泛的实验，实验结果表明即使只用了最 vanilla R3D18 作为 backbone，我们的 M3DFEL 模型达到了最好的效果。

2 相关工作

2.1 动态面部表情识别（DFER）

随着 DNNs 在计算机视觉任务中的成功，自动面部表情识别（FER）也通过深度学习得到了改善。DFER 方法与 SFER 方法不同，因为它们除了考虑每张图像的空间特征外，还需要考虑时间信息。一些方法采用 CNN 来提取每一帧的空间特征，然后用 RNN 来分析时间关系 [**?**, 21]。有人提出了 3DCNNs 来为 3D 数据建模，并联合学习空间和时间特征。Fan 等人 [1] 提出了一个混合网络，利用后期融合将循环神经网络（RNN）和三维卷积网络（C3D）结合起来。Lee 等人 [8] 提出了一个场景感知的混合神经网络，该网络以一种新颖的方式结合了 3DCNN、2DCNN 和 RNN。Lee 等人 [7] 提出了 CAER-Net，这是一个用于情境感知情感识别的深度网络，它以联合和提升的方式利用了人类面部表情和情境信息。

最近，基于变换器的网络在提取空间和时间信息方面得到了普及。例如，Zha 等人 [22] 提出了一个动态面部表情识别变换器（Former-DFER），它由一个卷积空间变换器（CS-Former）和一个时间变换器（T-Former）组成。Ma 等人 [15] 提出了空间-时间变换器（STT）来捕捉每一帧内的判别特征，并对各帧间的上下文关系进行建模。

动态-静态融合模块 [9, 10] 用于从静态特征和动态特征中获得更加稳健和具有鉴别力的空间特征，可以有效减少噪声帧对 DFER 任务的干扰。此外，Wang 等人 [19] 提出了双路径多激励协作网络（DPCNet），从较少的关键帧中学习关键信息用于从较少的关键帧中学习面部表情表现的关键信息。

上述方法将 DFER 作为一个一般的视频理解任务来处理，没有考虑到由于不精确的众包注释而导致的问题的弱监督性质。此外，他们忽略了 DFER 中不平衡的小范围和大范围时间关系的问题，而仅仅依赖于一个序列模型。相比之下，M3DFEL 框架通过解决弱监督的问题，和对不平衡的小范围和长范围时间关系统一建模，从根本上解决了这些挑战。

2.2 多实例学习（MIL）

MIL 是一种旨在解决不精确标记问题的技术 [3]。传统上，每个样本都被当作一个实例包，只有当所有的实例都是负面的时候，这个包才被标记为负面。否则，该包被认为是积极的。MIL 通常用于有大量的只有一个标签的样本的情况。在这些情况下，方法必须准确识别和含有大量负面实例的数据集中的正面实例。

MIL 已被应用于各个领域，如 WSOD（弱监督物体检测）[4, 17]、动作定位和 WSI（整个幻灯片图像）分类。虽然没有研究将野外 DFER 制定为 MIL 问题，但我们可以从 WSOD 方法中得到启发，这些方法也解决了基于视频任务的 MIL 问题。例如，Feng 等人提出了一个端到端的弱监督旋转不变量空中物体检测网络，以解决没有相应约束的物体旋转。同时，Tang 等人 [17] 引入了一种新的在线实例分类器改进算法，将 MIL 和实例分类器改进程序整合到

一个深度网络中。将 MIL 和实例分类器精炼程序整合到一个单一的深度网络中，并在仅有图像级监督的情况下对网络进行端到端的训练。只用图像级别的监督来训练网络。

在识别情绪方面，MIL 的使用已被探索。Romeo 等人 [16] 探讨了一些现有的基于 MIL 的 SVM 在使用生理信号检测情绪方面的应用。Chen 等人 [2] 主要关注疼痛检测的行动单元编码，并在 MIL 中应用基于聚类的最大运算进行实例融合。Wu 等人 [20] 在 MIL 中采用了可微的 OR 操作以隐马尔科夫模型作为分类器，在实验室控制的 DFER，使用面部地标作为输入特征。所有这些方法都使用手工的特征，并采用传统的机器学习 MIL 方法来完成他们的任务。此外，实验室控制的 DFER 样本更加明确，环境和面部表情的动态是固定的。而野外的样本则更加复杂和具有挑战性，他们应用 MIL 方法的方法是不适用于我们在野外 DFER 的情况。

通过对 DFER 的高级假设和观察，我们通过融合 MIL 管道内的不平衡时间关系的建模，设计了我们的新型 MIL 框架。与使用现有的 MIL 方法来融合手工的特征不同，我们在特征提取过程中对强的小范围时间关系进行建模，并在实例融合过程中学习长范围时间关系。

3 本文方法

3.1 本文方法概述

MIL 流水线主要包含 4 个步骤：实例生成、实例特征提取、实例聚合，以及分类。在 DFER 问题中，我们使用的 M3DFEL 遵循这一流水线，并且优化了 3DCNN，来从生成的 3D 实例中提取特征，并且学习小范围的时序关系。DILIAM 被用来读长范围时序关系进行建模，同时动态地将实例融进一个包。为了维持在包级别和实例级别时间一致性，DMIN 被引入。被使用的 M3DFEL 框架的概览见图2。

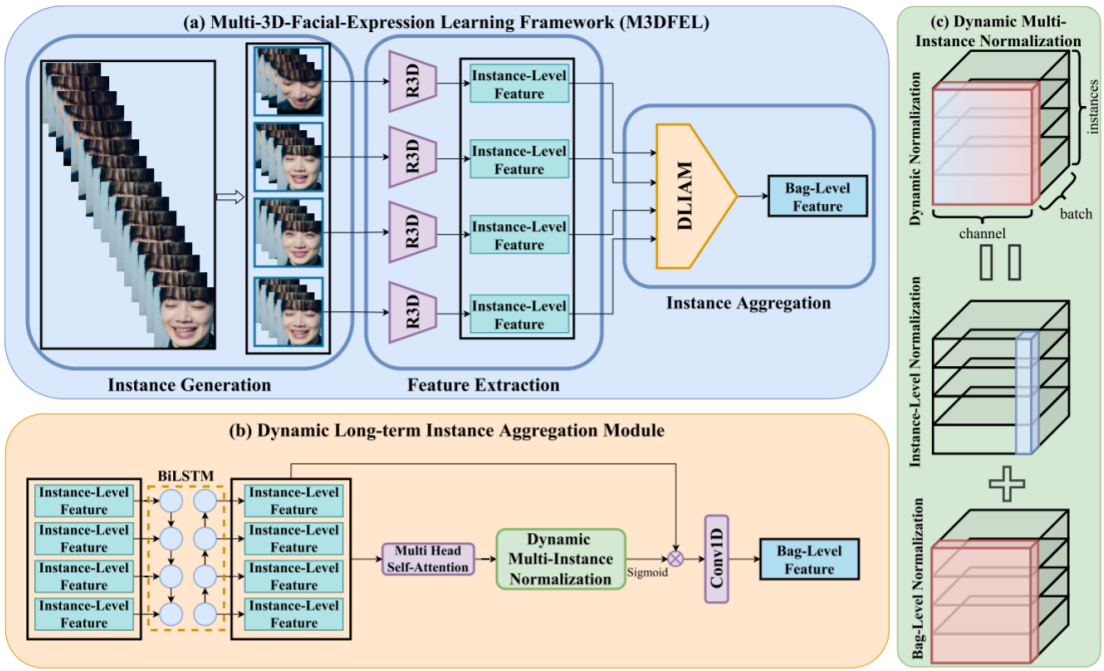


图 2. 本文提出的 M3DFEL 框架的概述

3.2 3D 实例生成

通过将视频裁剪成帧来生成实例是 MIL 任务的常见方法，因为它们通常是基于帧的任务，例如弱监督对象检测或动作定位。然而，在 DFER 中，一些帧可能在对象说话时无法捕获典型的面部表情。虽然这些帧本身看起来不正常，但它们实际上代表了面部运动的运动。此外，与其他 MIL 任务相比，类之间面部运动的差异是微妙的，这意味着即使是很小的运动也会导致预测的情绪和特征的变化。

给定包含 T 帧图像的视频 V，我们将视频裁剪成维度 T 上的 N 个部分。然后，袋可以被定义为实例序列 $I = [I_1, I_2, \cdots, I_N]$，其中 I_n $R^{C\times T\times H\times W}$ 表示第 n 个 3D 实例。这种设计使得特征提取器能够通过捕获跨实例的面部运动的运动以及当对象说话时的一致情感来对强短期时间关系进行建模。

3.3 实例特征提取

利用 R3D18 用于提取包内每个实例 I_n 的特征 F_n。R3D18 模型提取压缩帧表示，并结合每个实例的相邻帧的时间信息。每个包中的实例特征表示为 F $R^{N\times C}$，其中 C 表示通道的数量。基于 3D-Instance-based MIL 设置加强了短期时间学习。

3.4 动态长期实例聚合

动态长期实例聚合模块（DLIAM）提出了动态聚合的实例，同时建模的长期时间关系。

第一步是使用 BiLSTM 捕获实例之间的长期时间关系

为了动态聚合实例的表示，我们首先应用多头自注意（MHSA）学习实例间的关系，得到一个注意权重 A $R^{N\times C}$

经过上面两步之后表情在短期内不太稳定，为了解决这个问题，设计了一个动态多实例归一化（DMZ）方法，以保持包和实例级别的时间一致性。我们定义一组归一化器 $K = bn\ in$，并动态调整重要性权重，其中 bn 表示袋级归一化器，in 表示实例级归一化器。令 A_{nc} 和 $\hat{A}_{nc}$ 是归一化之前和之后的第 n 个实例的第 c 个通道值，并且归一化过程可以如下所示：

$$\hat{A}_{nc} = \frac{A_{nc} - \sum_{k\in\mathcal{K}} w_k \mu_k}{\sqrt{\sum_{k\in\mathcal{K}} w'_k \sigma_k^2 + \epsilon}} * \gamma + \beta \tag{1}$$

其中 μ_k 和 σ_k 分别是使用实例的特定通道值的归一化器 k 估计的均值和方差值。ϵ 是为了数值稳定性而添加的一个小数字。可学习的仿射变换参数由 γ 和 β 表示。归一化子 k 的重要性权重由 w_k 和 w'_k 表示，并且是动态调整的。

两个正则化器之间的区别在于用于估计统计值的数值设置，包级别的正则化器沿着每个包的 N 个实例 C 个通道计算统计量。

$$\mu_{bn} = \frac{1}{NC}\sum_{n,c}^{N,C} A_{nc}, \quad \sigma_{bn} = \frac{1}{NC}\sum_{n,c}^{N,C} (A_{nc} - \mu_{bm})^2 \tag{2}$$

其中 $\mu_{bn}, \sigma_{bn} \in \mathbb{R}^1$，表示单个包的值共享同一个包级别统计量。

实例级别的正则化沿着 N 个维度计算统计量。

$$\mu_{\text{in}} = \frac{1}{N}\sum_{n}^{N} A_{nc}, \quad \sigma_{\text{in}} = \frac{1}{N}\sum_{n}^{N} (A_{nc} - \mu_{\text{in}})^2 \tag{3}$$

其中 $\mu_{in}, \sigma_{in} \in \mathbb{R}^C$ ，表明实例级别的统计量在每个包的单个通道上是共享的。对于重要性权重，ω_k 和 ω'_k ，我们用 softmax 操作来保证 $\sum_{k\in\mathcal{K}} w_k = 1, \sum_{k\in\mathcal{K}} w'_k$ ，且所有标量都被限制在 $0-1$ ，

$$w_k = \frac{e^{\lambda_k}}{\sum_{j\in\mathcal{K}} e^{\lambda_j}} \tag{4}$$

其中 λ 是可学习参数，用于调整不同归一化方法的权重。对于实例的最终汇总，权重首先与经过一个 sigmoid 函数后的实例相乘。然后，利用 Conv1D 层将实例级特征 X 汇总为包级别特征 $\mathrm{Z} \in \mathbb{R}^{N\times C}$ ，

$$\mathrm{Z} = \mathrm{Conv1D}(\mathrm{X} * \mathrm{Sigmoid}(\hat{\mathrm{A}})) \tag{5}$$

然后包级别的特征被输进全连接层，以得到预测结果，交叉熵损失函数被用来监督结果。

4 复现细节

4.1 与已有开源代码对比

本复现工作基于作者发布于 GitHub：https://github.com/faceeyes/M3DFEL。
在已有开源代码的基础上，我们做出了如下改进：

- 在原先交叉熵损失的基础上添加了情感强度感知损失，使网络额外关注每个样本中最容易混淆的类别。
- 在 DFER 数据集上的实验表明我们的模型相比于原文能够更好的识别类别标签较少的动态面部表情。

4.2 情感感知损失

表情强度是人类面部表情的一个重要属性。当我们定义强度 $\in (0,1)$ 时，显然当强度收敛到零时，所有的非中性表情都趋向于中性表情。

$$\lim_{intensity\to 0} NNE = NE \tag{6}$$

其中 NNE 和 NE 分别是非中性和中性表情，看来把面部表情作为一个回归任务可能更合适。然而，由于具有连续标签的 DFER 数据集的注释成本很高，因此此类数据集的规模通常是有限的。此外，注释的强度标准难以统一。为了解决这个问题，我们提出了强度感知损失，以减少低强度样本的 DFER 任务带来的影响。

基于低强度样本可能与其他类别的低强度样本混淆的假设，网络应该格外注意每个样本中最容易混淆的类别。因此，所提出的强度感知损失可以公式化为：

$$\mathcal{L}_{IA} = -\log(P_{IA}) \tag{7}$$

$$P_{IA} = \frac{e^{x_t}}{e^{x_t} + e^{x_{max}}} \tag{8}$$

最终的损失函数重构为：

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\mathcal{L}_{IA} \tag{9}$$

4.3 实验设计

我们的整个框架是用 PyTorch-GPU 实现的，并在 Tesla P100 GPU 上进行训练。对于特征提取，我们采用了 vanilla R3D18 模型并利用 Torchvision 提供的预训练权重。这些模型使用 AdamW 优化器和余弦调度器训练了 300 个 epoch，其中有 20 个 warm-up epoch。学习率被设置为 5e-4，最小学习率被设置为 5e-6，权重衰减被设置为 0.05。我们使用 256 的 batch size，并应用 0.1 值的标签平滑。我们的数据增强方法包括随机裁剪、水平线翻转和 0.4 的颜色抖动。对于每个视频，我们总共提取 16 帧作为样本。在所有的实验中，我们使用加权平均召回率（WAR）和非加权平均召回率（UAR）作为评价指标。在接下来的实验中，我们主要使用 DFEW 来进行进一步分析和讨论。

5 实验结果分析

加入情感感知损失模块之后，WAR 虽然有所下降，但是 UAR 提升较为明显，这是因为情感感知损失让模型更加关注最容易混淆的类别，从而让小样本的类别也能正确分类。实验结果如表1所示，本文的方法在 WAR 上下降了 0.47%，但是在 UAR 上上升了 0.53%，这和我们先前的猜想一致，证明了情感感知损失模块的有效性。

表 1. 与原文方法的对比

Method	WAR	UAR
M3DFEL (Theirs)	68.69%	57.17%
M3DFEL (Ours)	68.22%	57.70%

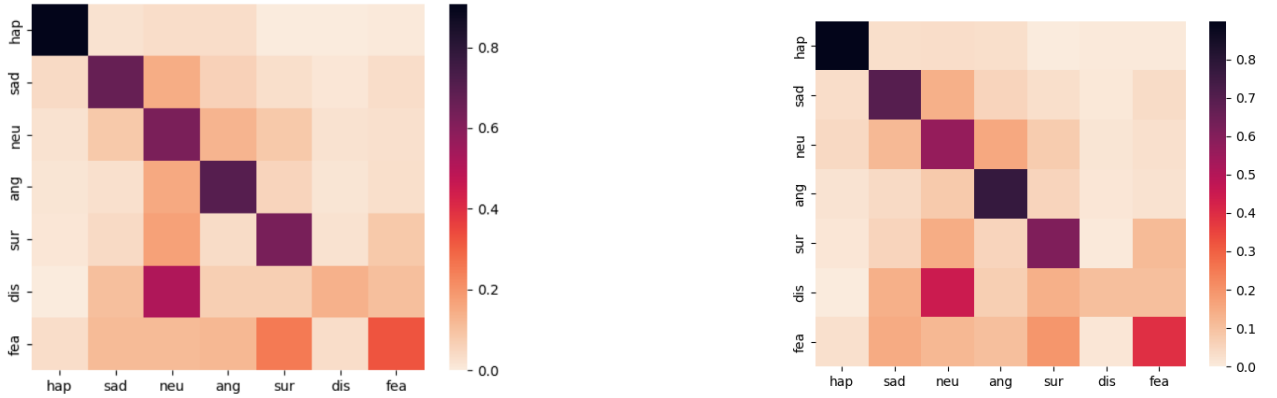
根据 bag size 的不同，我们对 DFEW 进行了消融研究，以展示 MIL 设置中包大小的影响。当包大小设置为 16 时，与采样帧数相同，3DMIL 设置退化成 2D 的形式。将包大小设置为 1 表示一次性将所有帧送入特征提取器从而导致聚合模块失效。结果在表2中展示。我们

表 2. Bag Size 的影响

Bag Size	WAR	UAR
1	67.53%	57.03%
2	67.81%	57.21%
4	68.22%	57.70%
8	67.35%	57.42%
16	66.24%	56.23%

将本文提出的方法和原文提出的方法在 DFEW Fold 1 上评估的混淆矩阵可视化以分析结果。从图3中，我们观察到原文模型很难预测被标记为厌恶的视频的情绪。这是由于 DFEW 中标签严重不平衡，其中厌恶视频的比例仅为 1.22 %。因此，模型在训练过程中更容易忽略带有”厌恶” 标签的视频，导致模型对这种情绪的表现较差。恐惧标签也出现了类似的情况，比例为 8.14 %。该模型倾向于预测一些将恐惧标注为其他情绪的视频，因为缺乏足够的针对该情

绪的训练实例。此外，我们观察到模型倾向于更频繁地预测中性标签，这是因为将这些样本预测为任何其他情绪比预测为中性更有风险，但是本文的方法中预测为中性的样本明显减少，这也证明了本文方法的有效性。



(a) 原文方法　　(b) 本文方法

图 3. 混淆矩阵可视化展示

6 总结与展望

我们利用多实例学习（MIL）并采用 M3DFEL 框架来以统一的方式解决弱监督问题以及不平衡的短期和长期时间关系。复现了 M3DFEL 框架包括 3D 实例生成模块（可学习强短期时间关系）和动态长期实例聚合模块（DLIAM），它模拟了弱的长期时间关系。实现了动态多实例规范化，以保持包级和实例级的时间一致性。基于一个低强度样本很可能与其他类的低强度样本混淆的假设，对原文中的 L2 损失进行改进，加入了强度感知损失。

我们对失败案例的分析表明，它们大多发生在 MIL 中的分类阶段，而不是实例融合阶段。例如，如果大部分视频帧是中性的，那么整个包的融合结果就是预期的非中性情感。然而，该模型经常将非中性情绪误分类，例如将恐惧情绪归类为惊讶情绪。这表明当前的性能在很大程度上受限于模型的分类能力。其中一个主要问题是标签不平衡问题，即由于数据集中缺少带有这些标签的样本，从而牺牲了在厌恶和恐惧上的准确率。DFER 中的一些表情的强度远低于静态表情，这与微表情识别（MER）中的关键问题类似。利用光流等 MER 技术可能有助于解决这一问题。此外，一些先验知识，如地标或动作单元，可能会对模型提供有用的提示。噪声标签问题、不确定性问题和难样本问题也都会对 DFER 产生较大的影响。

参考文献

[1] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 445–450, 2016.

[2] Zhixin Fang, Libai Cai, and Gang Wang. Metahuman creator the starting point of the metaverse. In *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*, pages 154–157. IEEE, 2021.

[3] Michael Gadermayr and Maximilian Tschuchnig. Multiple instance learning for digital pathology: A review on the state-of-the-art, limitations & future potential. *arXiv preprint arXiv:2206.04425*, 2022.

[4] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 670–678, 2022.

[5] 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.

[6] Ziv Lautman and Shahar Lev-Ari. The use of smart devices for mental health diagnosis and care, 2022.

[7] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10143–10152, 2019.

[8] Min Kyu Lee, Dong Yoon Choi, Dae Ha Kim, and Byung Cheol Song. Visual scene-aware hybrid neural network architecture for video-based facial expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[9] Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 67–75, 2023.

[10] Hanting Li, Mingzhe Sui, Zhaoqing Zhu, et al. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*, 2022.

[11] Zuojin Li, Liukui Chen, Ling Nie, and Simon X Yang. A novel learning model of driver fatigue features representation for steering wheel angle. *IEEE Transactions on Vehicular Technology*, 71(1):269–281, 2021.

[12] Feng Liu, Si-Yuan Shen, Zi-Wang Fu, Han-Yang Wang, Ai-Min Zhou, and Jia-Yin Qi. Lgcct: A light gated and crossed complementation transformer for multimodal speech emotion recognition. *Entropy*, 24(7):1010, 2022.

[13] Feng Liu, Han-Yang Wang, Si-Yuan Shen, Xun Jia, Jing-Yi Hu, Jia-Hao Zhang, Xi-Yi Wang, Ying Lei, Ai-Min Zhou, Jia-Yin Qi, et al. Opo-fcm: A computational affection based occ-pad-ocean federation cognitive modeling approach. *IEEE Transactions on Computational Social Systems*, 2022.

[14] Feng Liu, Hanyang Wang, Jiahao Zhang, Ziwang Fu, Aimin Zhou, Jiayin Qi, and Zhibin Li. Evogan: An evolutionary computation assisted gan. *Neurocomputing*, 469:81–90, 2022.

[15] Fuyan Ma, Bin Sun, and Shutao Li. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*, 2022.

[16] Luca Romeo, Andrea Cavallo, Lucia Pepa, Nadia Bianchi-Berthouze, and Massimiliano Pontil. Multiple instance learning for emotion recognition using physiological signals. *IEEE Transactions on Affective Computing*, 13(1):389–407, 2019.

[17] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2843–2851, 2017.

[18] Yan Wang, Yixuan Sun, Yiwen Huang, Zhongying Liu, Shuyong Gao, Wei Zhang, Weifeng Ge, and Wenqiang Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20922–20931, 2022.

[19] Yan Wang, Yixuan Sun, Wei Song, Shuyong Gao, Yiwen Huang, Zhaoyu Chen, Weifeng Ge, and Wenqiang Zhang. Dpcnet: Dual path multi-excitation collaborative network for facial expression representation learning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 101–110, 2022.

[20] Chongliang Wu, Shangfei Wang, and Qiang Ji. Multi-instance hidden markov model for facial expression recognition. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–6. IEEE, 2015.

[21] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial–temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, 49(3):839–847, 2018.

[22] Zengqun Zhao and Qingshan Liu. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1553–1561, 2021.