DFEW:一个大规模的真实场景动态表情数据库*

江星洵　宗源　郑文明　唐传高　夏万闯　路成　刘佳腾
{jiangxingxun,xhzongyuan,wenming_zheng}@seu.edu.cn
东南大学

摘要

近年来，面部表情的研究热点逐渐从实验室场景下的识别，转向真实场景下的识别——这对算法的实际运用是十分重要的。本文中，我们聚焦于这个困难但有趣的问题，研究并作出三点贡献。首先，我们建立了一个大规模的自然场景动态表情数据库，简称*DFEW*数据库。它由从上千部电影中收集到的超过*16000*个视频片段组成。这些视频片段包括了真实生活中各种极具挑战的场景，例如：极端的光照、遮挡及各式各样的姿态变化。其次，我们提出了一个简称*EC-STFL*的框架，来处理真实场景下的动态表情识别问题。第三，我们在*DFEW*数据库上进行了大量实验：各种时空神经网络及我们所提的*EC-STFL*。实验结果表明，*DFEW*是一个精心标注但极具挑战的数据库；我们所提的*EC-STFL*方法可以在自然场景动态表情识别问题中，有效改善时空神经网络的性能。我们的数据库是开源的，可以在*https://dfew-dataset.github.io/*下载。

1. 引言

识别表情，是人们日常生活中情感交流最自然的方式之一 [3]。如果机器能够理解人们的情感，那么人机交互系统将变得更加自然、更加友好。因为这个原因，面部表情识别逐渐成为人机交互与多媒体分析中的研究热点。在过去的十几年中，研究者们提出了许多用于面部表情识别的好方法，并且，这些方法在实验室场景下的面部表情数据库上取得了优越的性能 [44, 42, 43, 25, 31, 8]。然而，当前的面部表情识别技术，与实际应用仍相距甚远。原因之一在于实验室场景与真实场景下的研究是差别较大。真实场景中的人脸往往会受到遮挡、光照与姿态的变化和其他无法预测且困难的干扰条件的影响。这些影响因素也使得目前已有的面部表情识别技术，在真实场景中的性能下降严重。所以，许多研究者们也逐渐将研究重点转到这个困难但意义重大的问题中——自然场景下的面部表情识别。这里的自然场景是指在不受限制且极具挑战条件的真实环境。

自然场景下的面部表情识别可按形式分成两类：一类是静态表情识别，旨在从不受限制的人脸图片中识别表情类别；另一类是动态表情识别，旨在从视频片段或者图片序列中获取情感。深度学习是一种需要大量数据驱动的方法，在很多计算机视觉任务中取得了成功。为实现更加准确的表情识别，受此启发，研究者们开始从互联网中获取面部表情数据，构建自然场景场景下的大规模人脸表情数据库。例例如，Benitezquiroz等人 [1]从互联网上收集面部表情图片，创建了一个叫做EmotioNet的大规模自然场景人脸表情数据库。 EmotioNet包含1,000,000张面部表情图片，其中25,000张图片手工标注了11个人脸运动单元(AUs)。紧接着，Mollahosseini等人 [29]构建了一个更大的数据库，命名为AffectNet。Affect包含450,000张从互联网上获取并仔细标注的图片。最近，Li等人 [22, 21]建立了一个名为RAF-DB的新的静态表情数据库，RAF-DB数据库包

*本文为MM'20论文 [16]的中文翻译版。

表 1. 现有自然场景动态表情数据库总结表

数据库	样本数	来源	标注类别	标注次数	可否获取
Aff-Wild [18]	298	Web	激活度-唤醒度	8	Yes
AFEW 7.0 [4]	1,809	54部电影	7种基本表情	2	Yes
AFEW-VA [19]	600	AFEW数据库	激活度-唤醒度	2	Yes
CAER [20]	13,201	79集电视剧	7种基本表情	3	Yes
DFEW	16,372	1500部电影	7种基本表情	10	Yes

含了接近30,000张从互联网上获得的人脸图片。与EmotioNet数据库和AffectNet数据库相比，RAF-DB的主要优势在于标注：RAF-DB雇佣了315个人作为标注人员，每张图片都被标注了大约40次，以保证标注的可靠性。

然而，与自然场景下的静态表情相比，当前只有极少数的自然场景动态表情数据库公开。在工作 [5]中，Dhall等人建立了一个不受约束的动态面部表情数据库，即AFEW。到目前为止，AFEW已经更新到第7个版本，即AFEW7.0 [4]。 AFEW收集了来自54部电影的1809个面部表情的视频片段。近来，Lee等人 [20]建立了一个叫做CAER的自然场景动态表情数据集。 CAER收集了来自79集电视剧视频的13,201个视频片段。每个视频片段都由3个标注者分别标注过。据我们所知，CARE是第一个较大的自然场景动态表情数据库。由于缺少大规模数据库，自然场景动态表情识别问题中的深度学习方法研究受到了严重阻碍。比方说，在2019年举办的比赛EmotiW2019(ACM ICMI会议工作坊)中，冠军Li等人 [23]提出了一种加权融合方法，整合来自不同时空模型的预测得分结果。但是，他们在测试集上的准确度仅达到62.8%（7类表情分类任务）。这个识别准确度不高，无法运用于实际。

为了降低由于数据量不足对自然场景动态表情识别产生的影响，本文中，我们建立了一个大规模的精心标注的动态表情数据库，简称DFEW。DFEW可以为研究者们研发、评估动态表情识别算法提供对比基准。

我们在表 1中列出了现有自然场景动态表情数据库和我们的DFEW数据库，以进一步认识DFEW数据库。从表 1中，我们可以看到，DFEW较之现有数据库(Aff-Wild [18], AFEW [4], AFEW-VA [19], CAER [20])拥有三大优点。首先，DFEW数据库是目前最大的自然场景动态表情数据库，拥有超过16000个视频片段。其次，由于DFEW中的样本收集自不同国家的超过1500部电影，他可以很好地模拟真实条件下的变化多样、极具挑战的干扰条件：极端的光照、自遮挡和反复无常的姿态变化。最后，DFEW中的每个样本，都是在专业指导下，被精心地、独立地标注过10次。

除此之外，我们还提出了EC-STFL框架来处理自然场景下的动态表情识别问题。 EC-STFL框架可以加强如C3D [38]、P3D [33]之类的深度神经网络，使网络可以学到更加具有判别性的特征，用于动态表情识别。最后，我们在DFEW上提出了用于对比的实验协议，并给出了多个常用的神经网络的性能结果。实验结果表明，我们提出的EC-STFL框架能在自然场景表情识别问题中显著改善现有时空神经网络的性能。

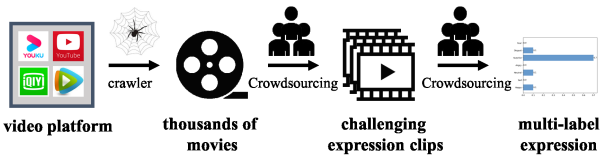

图 1. DFEW数据库数据收集及标注的整体过程

2. DFEW数据库

2.1. 数据收集

电影来源于生活，并模仿真实生活。我们认为，电影中演员们的这种不受约束的表情，与真实世

图 2. DFEW数据库单标签子集部分的样本示例

界中的差别不大。因此，我们可以抽取来自电影中不同表情的视频片段，来建立自然场景下的动态表情数据库。在过去的十几年中，研究者们相继建立了AFEW-Wild [18], AFEW [5, 4]和CAER [20]等动态表情数据库。他们也确实促进了自然场景下动态表情的研究。本文中，我们同样利用电影资源，收集不受限制的动态面部表情，建立我们的DFEW数据库。

DFEW数据库建立流程如图 1所示。首先，我们运用网络爬虫技术，从从互联网上收集了超过1500部高分辨率的、与生活密切相关的电影，将其作为获取面部动态表情视频片段的来源。电影包括各种题材：喜剧、悲剧、战争、爱情等等。然后，我们雇佣了数十名学生，让他们用视频剪辑软件，手动定位为剪辑包含七种基本表情的视频片段。我们为每个学生分配了不同的电影，以保证剪辑出的样本不重复。同时，我们设置了一些规则，以保证剪辑样本的多样性：每部电影只允许剪辑不超过20个样本；同一个人在同一个场景中的剪辑样本，只算一个；按剪辑少样本数量的个数，对剪辑诸如厌恶、恐惧等少样本表情类别片段的同学，给予更高的报酬（表情为预估表情，非真实标签）。

2.2. 数据标注

为收集到的数据打上高质量的标签，是建立高质量数据库的一大挑战：首先，标注如此大规模的数据库是十分耗时的，需要强而有效的人员管理；其次，虽然心理学家P. Ekman认为七种情感是普遍一致、与文化背景不相关的 [6]，但由于文化不同引起的差异，确是切实存在并应尽力去除的。为有效管理标注者和提高标签质量，我们与京东微工 [1]（一家专业的重众包公司）想合作，雇佣了多名标注人员，并保证他们在正式标注前，都接受了专业的情感知识训练。他们被要求为上一步获取到的视频片段打上与七种基本表情（开心、伤心、中性、愤怒、惊讶、厌恶、害怕）最相近的情感标签。由于中途2名标注者离职，我们雇佣了另外2名标注者——前后一共雇佣了12名标注人员，让他们独立地标注样本。这么一来，每个样本都被独立标注了10次。通过这种方式，我们便获取到了16372个视频片段

[1]http://weigong.jd.com/

样本的七维情感向量，或称为情感分布标注。我们用$L_j = \{l_1, \ldots, l_k, \ldots, l_7\}$来表示第$j$个视频片段的七维情感标签，这里，$l_k$表示第k种情感的标注次数。$k \in \{1, 2, 3, 4, 5, 6, 7\}$，分别对应开心、伤心、中性、愤怒、惊讶、厌恶和害怕这七种情感。

表 2. DFEW数据库单标签子集部分的基本信息

表情	视频片段数				百分比(%)
	0-2秒	2-5秒	5秒+	小计	
开心	852	1252	384	2488	20.63
悲伤	440	915	653	2008	16.65
中性	832	1335	542	2709	22.46
愤怒	762	1091	376	2229	18.48
惊讶	691	648	159	1498	12.42
厌恶	71	58	17	146	1.22
害怕	408	435	138	981	8.14
小计	4056	5734	2269	12059	100.00

然而，按照得到的情感分布，并非所有片段都可以被明确地划分为七种基本之一。因此，为了得到单标签情感更加准确的标注，我们挑选满足$l_k > r$的第k种表情，作为视频片段样本的标签。此处我们设定$r = 6$，即超过一半的标注者认为第k种情感应为样本的标签。通过这种方式，我们在DFEW数据库中挑选出12059个片段作为单标签子集（可被分为七种基本情感之一的样本）。我们将DFEW数据库的基本信息总结如表 2所示，并在图 2中展示了单标签子集样本。注意，我们将同时提供DFEW数据库的七维情感标注信息，和单标签标注信息。

2.3. 一致性检测

本章中，我们基于Fleiss's Kappa Test [10]讨论标签质量。Fleiss' s Kappa Test可以检测标注中的一致部分是否受偶然因素影响，能够很好地评估标签的可靠性和质量。我们令n_{ij}表示标注者们认为第i个样本为第j种情感的投票数。因此，整个标注任务中第j种情感的标注比例p_j为

$$\begin{cases} p_j = \frac{1}{N \times n} \sum_{i=1}^{N} n_{ij} \\ \sum_{j=1}^{K} p_j = 1 \end{cases} \tag{1}$$

我们用N表示视频片段样本数。每个样本的标注次数$n = 10$, 标注的情感总类别数$K = 7$。我们可以计算标注者们对第i个视频片段标注认同的程度P_i。

$$P_i = \frac{1}{n \times (n-1)} \left[(\sum_{j=1}^{K} n_{ij}^2) - n \right] \tag{2}$$

进而，我们可以计算$\bar{P}$(P_i的平均值)和$\bar{P}_e$

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{3}$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2 \tag{4}$$

从而计算指标κ

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{5}$$

通过Fleiss's Kappa Test，我们计算出DFEW数据库及其单标签部分的一致性指标分别为$\kappa = 0.70$和$\kappa = 0.63$。结合表 3的解释,我们可以认为，标注者们认为标注结果的一致性好，也即标注质量高。

3. 表情聚集的特征学习

自然场景下的面部动态表情，受到多方面环境干扰的影响：光照、姿态、遮挡和尺度上的各种变化。自然场景下的动态表情识别所面临的挑战是：如何学习到用以描述面部动态表情的鲁棒的判别性较高的特征。时空神经网络在视频理解中占据重要地位，它能够较好地从视频样本的时空域中捕捉动态人脸的运动信息。较之传统方法，当前时空神经

表 3. Fleiss's Kappa Test中指标κ的解释

κ	解释
<0	较差的一致
0.01-0.20	轻微的一致
0.21-0.40	公平的一致
0.41-0.60	比较一致
0.61-0.80	实质性一致
0.81-1.00	近乎完美的一致

网络对干扰条件，也表现出更强的鲁棒性。然而，由于自然场景下极具挑战的干扰条件，不同表情对应特征的边界是模糊的。为了使识别自然场景的动态面部表情时，不同表情对应的特征更加清晰，我们提出了“表情聚集的时空特征学习”框架，简称EC-STFL。 EC-STFL借鉴于LDA的思想，可以提高同类表情特征的相关性，降低不同表情特征的相关性，并可方便地嵌入常用的时空神经网络中。

$$\min_{W} \sum_{i,j} \frac{P_{ij}\phi(x_i, x_j)}{Q_{ij}\phi(x_i, x_j)} \tag{6}$$

此处定义：神经网络的参数为W；第i个样本和第j个样本的时空特征的距离差为$\phi(x_i, x_j) = \|x_i - x_j\|$；$x \in \mathbb{R}^d$为神经网络分类层之前的最后特征向量。相似度矩阵$P$和$Q$定义如下：

$$P_{ij} = \begin{cases} 0, & \text{当 } x_i \text{ 和 } x_j \text{ 标签相同} \\ 1, & \text{其他} \end{cases} \tag{7}$$

$$Q_{ij} = \begin{cases} 0, & \text{当 } x_i \text{ 和 } x_j \text{ 标签不同} \\ 1, & \text{其他} \end{cases} \tag{8}$$

可以看到，EC-STFL是通过最小化同类表情的特征间距和最大化不同类表情的特征间距，来扩大表情特征边距的。受限于显存大小，为了更加有效地使用EC-STFL，我们将EC-STFL的计算放在mini-batch中完成。另外，我们注意到自然场景表情识别中较为严重的类别不平衡问题 [23, 24, 15, 9]。类别不平衡的数据，使模型的学习过程产生偏颇——模型更倾向于将样本识别为数量较多的情感类别。同样的问题，在DFEW数据库中也存在。所以，我们对EC-STFL损失函数动态加权，以此在批损失函数更新过程中平衡不同表情类别的损失函数。我们将EC-STFL扩展为：

$$L_{EC-STFL} = \frac{\sum_{1\le i,j\le n, x_j \in \mathcal{N}\{x_i\}} \frac{\|x_i - x_j\|}{N_{x_i}}}{\sum_{1\le i,j\le n, x_j \notin \mathcal{N}\{x_i\}} \frac{\|x_i - x_j\|}{N_{x_j}}} \tag{9}$$

我们定义：$\mathcal{N}\{x_i\}$表示批更新时mini-batch中与第i个样本的特征x_i具有相同标签的样本特征的集合；N_{x_i}表示集合$\mathcal{N}\{x_i\}$中的样本数量；n表示mini-batch的大小。 EC-STFL通过N_{x_i}和N_{x_j}的动态加权，平衡不同表情的损失函数，从而在一定程度上减缓面部表情识别中的类别不平衡问题。

我们采用联合训练的方式，同时训练softmax损失函数和EC-STFL损失函数。我们定义：L_s 表示softmax 损失函数，超参数λ表示损失函数softmax loss和EC-STFL loss之间权衡系数。整体损失函数表示为：$L = L_s + \lambda L_{EC-STFL}$。注意，当$L_{EC-STFL}$无意义即mini-batch中只包含一类表情时，我们将不进行反向传播。

4. 实验

本章内容包括：数据预处理、实验协议和评价指标。我们在DFEW上进行了大量时空神经网络及EC-STFL的相关实验。并运用迁移实验，验证了DFEW数据库具备为真实场景的面部表情识别提供足够的迁移知识的能力：将DFEW数据库上经过预训练模型迁移到AFEW数据库，其性能优于将动作数据库上预训练模型迁移到AFEW数据库上的性能。

4.1. 实验步骤

数据和协议. DFEW数据库的单标签子集部分共有12059个视频片段。为了更好地评估这部分数据，我们采用五折交叉验证的协议：我们将所有的样本分成不重叠的五份，对每折数据（第一折 ~第五折）

而言，其中一折作为测试集，其他四折作为训练集；将四折预测的结果串起来，与真实标签相对比，得到识别率。

预处理. 首先，我们用OpenCV将视频离散成图片帧，用face++ [34]提供的API，来获取人脸区域和人脸关键点。对未检测出人脸的图片，我们将其移除，并统计每个样本的有效图片比例。然后，我们移除掉有效图片比例小于50%的样本——共有362个视频片段被移除。再来，我们利用获取的人脸关键点信息和SeetaFace [26]进行了人脸矫正。最后，我们用时间插值算法 [46, 45]，将任意长度大小的图片序列变成16帧。

评价指标. 我们挑选了非加权的平均召回率（各类准确度之和除以类别，简称UAR）和加权的平均召回率（各类准确度按样本数据加权，简称WAR）这两个指标 [35]作为评价指标。用这两个指标来评价自然场景下的动态表情研究是十分合适的。UAR指标反应了不同类别表情的平均准确度，适合于样本不平衡问题的研究。 WAR指标反应了表情的整体识别率。我们希望学习到的模型，同时在WAR和UAR上有所改善。

实施细节. 我们选取12G的Titan Xp作为硬件平台，使用PyTorch框架 [32]来实现所提方法。通过网格搜索策略，我们用合适的初始化学习率初始化模型，并设置学习策略：当损失函数饱和时，学习率下调10倍。首先，我们给出DFEW数据库上基准神经网络方法的结果。我们给出的基准神经网络是从头（零）开始训练的，其batch的数量大小定为24（这是C3D模型下Titan Xp能容纳的最大batch数量大小）。对于EC-STFL和center loss的参数设置：我们将Softmax损失函数与EC-STFL的权衡系数λ设置为10；并遵从 [41]，将softmax损失函数与center loss的权衡系数设置为 1×10^{-4}。其次，我们选取C3D模型和3D Resnet18模型，进一步对EC-STFL的batch大小及权衡系数λ进行了超参讨论。第三，我们进行了跨数据库迁移实验——我们运用其他研究者提供的预训练模型，选择合适的初始化学习率，微调神经网络模型。这里的训练方式与实现基准神经网络的训练相同。

4.2. 实验结果

基线结果. 我们可以将输入为视频RGB帧的时空神经网络分为两类：3D卷积神经网络和CNN-RNN网络。我们挑选了五种3D卷积神经网络(C3D [38]、I3D-RGB [2]、R3D18 [39]、3D Resnet18 [12]、P3D [33])和两种CNN-RNN神经网络(VGG11+LSTM和Resnet18+LSTM)，作为DFEW的基准模型结果。其中，VGG11 [36]和Resnet18 [13]是经轻微修改，以适应112×112的图片输入尺寸。实验结果如表4所示。

由表4可得，P3D [33]在WAR指标上达到了最好的结果,为54.47%；3D Resnet18在UAR上达到了最好的结果，为44.73%。在七类表情中，3D卷积神经网络在开心、伤心、愤怒、惊讶和害怕的表情中得到了最好的性能；CNN-RNN模型在中性和厌恶的表情中得到了最好的结果。另外，从表4可以看出，模型对特征学习各有偏向。但整体而言，我们还发现开心的表情是最容易识别的，厌恶的表情是最难识别的。我们认为这可能是因为开心的表情类内方差较低，而厌恶的表情类内方差较大；不过这也可能是因为厌恶表情的样本数量较少导致的。厌恶等表情数量远少于其他表情，使识别任务中产生了较为严重的类别不平衡问题。据我们所知，厌恶表情识别问题确实是自然场景表情识别中的一大难题。

EC-STFL. 为了使学习到的特征判别性更强，我们设计了EC-STFL。 EC-STFL模块的消融实验如表5所示。我们发现，所有带有EC-STFL的模型，其性别性能皆优于不带有EC-STFL的模型。结果表明，所提出的EC-STFL在UAR和WAR指标上平衡能有1.61个百分点和02.08个百分点的涨幅。另外，从表5中我们发现，带有EC-STFL模块的3D Resnet18模型在UAR和WAR指标上能获取到最好的结果。

表 4. 多种神经网络模型在DFEW数据库的单标签子集上七种基本情感上的分类结果。模型包括:C3D, P3D, R3D18, 3D Resnet18, I3D-RGB, VGG11+LSTM, Resnet18+LSTM。评价指标包括:UAR（不加权的平均召回率）和WAR（加权的平均召回率）。

模型	情感							评价指标	
	开心	伤心	中性	愤怒	惊讶	厌恶	害怕	UAR	WAR
C3D [38]	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
P3D [33]	74.85	43.40	54.18	60.42	**50.99**	0.69	23.28	43.97	54.47
R3D18 [39]	**79.67**	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22
3D Resnet18 [12]	73.13	**48.26**	50.51	**64.75**	50.10	0.00	**26.39**	**44.73**	**54.98**
I3D-RGB [2]	78.61	44.19	56.69	55.87	45.88	2.07	20.51	43.40	54.27
VGG11+LSTM [36, 14, 11]	76.89	37.65	**58.04**	60.70	43.70	0.00	19.73	42.39	53.70
Resnet18+LSTM [13, 14, 11]	78.00	40.65	53.77	56.83	45.00	**4.14**	21.62	42.86	53.08

表 5. 有无EC-STFL模块时各神经网络模型在DFEW数据库上的表现性能。

模型	评价指标	
	UAR	WAR
C3D	42.74	53.54
C3D,EC-STFL	**45.10**	**55.50**
P3D	43.97	54.47
P3D,EC-STFL	**45.22**	**56.48**
R3D18	42.79	53.22
R3D18,EC-STFL	**45.05**	**56.19**
3D Resnet18	44.73	54.98
3D Resnet18,EC-STFL	**45.35**	**56.51**
I3D-RGB	43.40	54.27
I3D-RGB,EC-STFL	**45.05**	**56.19**
VGG11+LSTM	42.39	53.70
VGG11+LSTM,EC-STFL	**44.78**	**56.25**
Resnet18+LSTM	42.86	53.08
Resnet18+LSTM,EC-STFL	**43.60**	**54.72**

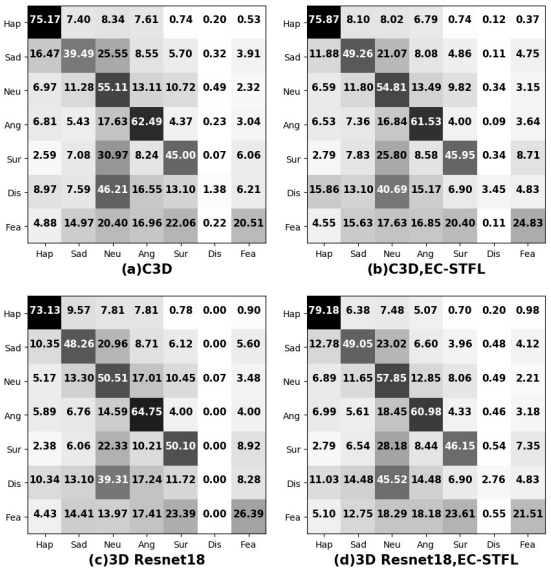

图 3. 有无EC-STFL模块时各神经网络性能的混淆矩阵。(a)C3D，(b)带有EC-STFL模块的C3D，(c)3DResnet18，(d)带有EC-STFL模块的3D Resnet18。

图 3中展示的是C3D模型和3D Resnet18模型有无EC-STFL模块时的混淆矩阵。从混淆矩阵中，我们可以看到各类表情的识别性能。如图 3所示，EC-STFL能有效提高C3D和3D Resnet18模型在开心、伤心、厌恶表情上的识别率；EC-STFL开心、伤心、惊讶、厌恶和害怕的表情识别率上分别提升了0.7%、9.77%、0.95%、2.07%和4.32%；EC-STFL使3D Resnet18在开心、伤心、中性和厌恶表情的识别率上分别提升了6.05%、0.79%、7.34%和2.76%。

为进一步观察由EC-STFL学到的特征，我们用tSNE [28, 40](一种非线性映射方法)，将学习到的高维特征映射到二维平面上，以便可视化。图 4展示的在添加EC-STFL模块前后，学习到不同特征。我们发现添加EC-STFL模块后学到的特征，其类间距离较未添加EC-STFL模块时更加明显，这表明我

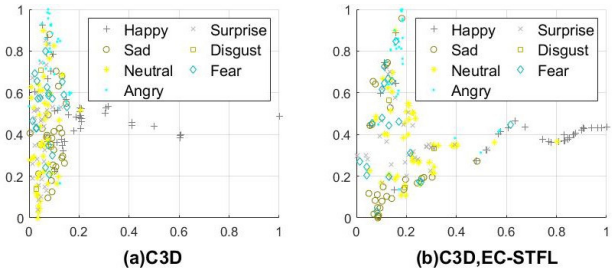

图 4. 各神经网络有无EC-STFL模块时由tSNE方法降维的高维特征。 (a) C3D, (b)带有EC-STFL模块的C3D。如图所示，STFL有助于学习到更具判别性的特征。

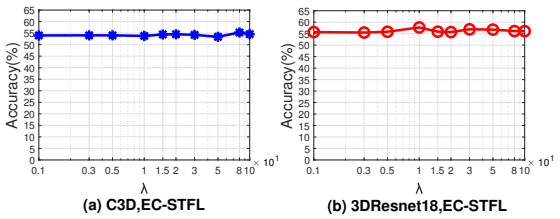

图 5. EC-STFL模块的超参数λ(权衡系数)讨论实验。(a) 带有EC-STFL模块的C3D模型 (b) 带有EC-STFL模块的3D Resnet18模型。权衡系数超参的范围为 $\lambda \in \{1, 3, 5, 10, 15, 20, 30, 50, 80, 100\}$。

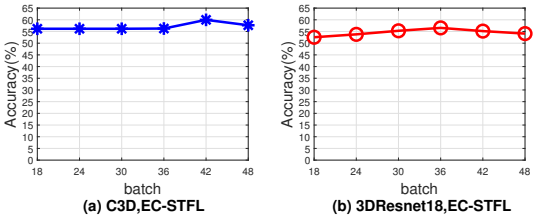

图 6. EC-STFL模块的超参数m(batch大小)讨论实验。(a) 带有EC-STFL模块的C3D模型 (b) 带有EC-STFL模块的3D Resnet18模型。 batch大小的超参范围为 $m \in \{18, 24, 30, 36, 42, 48\}$。

们所提的EC-STFL模块能够促使模型学习到更好的特征。

为进一步验证EC-STFL的效果，我们挑选C3D和3D Resnet18两个时空神经网络作为主干网络，将EC-STFL与其他聚类相关的损失函数相比较，如center loss [41]，结果如表 6所示。我们发现，EC-STFL和center loss都能改善模型，以达到更好的分类性能。并且，EC-STFL能表现出比center loss更好的性能，能使模型在UAR和WAR指标上达到最好的性能。

超参讨论. 我们知道，权衡系数λ和batch的数量大小对EC-STFL是较为重要的。所以，我们挑选C3D和3D Resnet18两个模型，在第一折数据上进行了超参敏感实验。超参讨论实验中，我们选取WAR作为评价标准。首先，我们将batch的数量大小固定为$m = 24$，变化超参λ（其取值范围为$\lambda \in \{1, 3, 5, 10, 15, 20, 30, 50, 80, 100\}$）。可以看到，挑选适当的λ有助于EC-STFL性能的提升。然后，我们将超参数λ 固定为$\lambda = 10$，变化超参m（其取值范围为$m \in \{18, 24, 30, 36, 42, 48\}$）。不同超参对应的结果如图 5和图 6所示。可以看到，带有EC-STFL模块的模型，在较大的batch数量范围内，其性能较为稳定。

4.3. 迁移实验

我们认为DFEW数据库是有助于真实场景中表情识别的迁移任务的。为验证我们的猜想，我们挑选了两个时空神经网络及其带有EC-STFL的模型来验证我们的猜想：C3D、带有EC-STFL模块的C3D、3D Resnet18、带有EC-STFL模块的3D Resnet18。我们分别将动作数据库和DFEW数据库中预训练的模型迁移到AFEW数据库 [5]上，并进行讨论。其中，动作数据库包括：UCF101 [37]、Sports1M [17]、Kinect700 [2]和Moments In Time [30]。

预训练模型方面，我们挑选DFEW数据库第二折数据和第五折数据的预训练模型，另外为保证公平，我们使用其他研究人员提供的动作数据库预训练模型。我们用预训练好的C3D模型权重来初始化C3D模型和带有EC-STFL模块的C3D模型，并采用网格搜索法挑出最适学习率，微调神经网络的所有层，以实现迁移学习。我们选取WAR作为评价指标，迁移的结果如表 7所示。我们发现，用DFEW预训练模型初始化迁移到AFEW数据库上的模型性能，优于用动作数据库初始化的模型。另外，我们还将迁移学习的结果与其他目前最优的方法进行了比较。如表 8所示，经由DFEW预训练的3D Resnet18模型，较之其他方法，在WAR性能指标上高出2个百分点。因此，我们验证了我们的猜想：我们所公布

表 6. EC-STFL和center loss在DFEW数据库上的对比结果。

模型	情感							评价指标	
	开心	伤心	中性	愤怒	惊讶	厌恶	害怕	UAR	WAR
C3D	75.17	39.49	55.11	62.49	45.00	1.38	20.51	42.74	53.54
C3D, center loss	75.62	44.67	54.18	63.14	42.21	2.07	22.17	43.44	54.17
C3D,EC-STFL	75.87	49.26	54.81	61.53	45.95	3.45	24.83	**45.10**	**55.50**
3D Renset18	73.13	48.26	50.51	64.75	50.10	0.00	26.39	44.73	54.98
3D Resnet18, center loss	78.49	44.30	54.89	58.40	52.35	0.69	25.28	44.91	55.48
3D Resnet18,EC-STFL	79.18	49.05	57.85	60.98	46.15	2.76	21.51	**45.35**	**56.51**

表 7. AFEW7.0数据库上的迁移学习的表现

预训练	微调的模型			
	C3D	C3D, EC-STFL	3D Resnet18	3D Resnet18, EC-STFL
Sports 1M	41.78	44.91	-	-
UCF101	41.25	42.34	-	-
Kinect700	-	-	49.35	49.61
Kinect700+Moments In Time	-	-	49.35	49.35
DFEW, fd2	**44.91**	**45.56**	**53.00**	**53.26**
DFEW, fd5	**49.87**	**49.87**	**49.61**	**49.66**

表 8. 3D Resnet18的迁移结果与目前其他最优方法在AFEW7.0数据库上的比较。

模型	WAR
Lu et al. [27]	45.31
Fan et al. [9]	45.43
Hu et al. [15]	46.48
Fan et al. [7]	48.04
Liu et al. [24]	51.44
3D Resnet18,DFEW fd2	**53.00**
3D Resnet18,EC-STFL,DFEW fd2	**53.26**

的DFEW数据库是有利于开发出适用于真实生活的、更优秀的动态表情识别模型。

5. 总结与展望

本文中，我们公布了一个非受限条件下的动态面部表情数据库DFEW，并且提出了一种新的深度时空神经网络学习框架EC-STFL，来处理自然场景下的面部动态表情识别问题。据我们所知，DFEW数据库是目前最大的自然场景动态表情数据库，它包含了从超过1500部电影里抽取的16372个视频片段。 DFEW还提供了精心标注的情感分布标签——每个样本都被独立地标注过10次。我们为DFEW数据库制定了对比协议，并提供了大量的基线方法结果以便研究者对比。我们也在DFEW数据库上进行了EC-STFL的测试。实验结果表明，所公布的DFEW数据库是一个优秀的、非受限条件下的动态表情数据库；我们所提出的EC-STFL框架能够有效改善动态表情识别领域中时空网络的性能。在未来的工作里，我们会为DFEW数据库继续收集更多的样本、更丰富的标签，以促进面部表情识别领域的发展。

致 谢. 本项目受到了国家重点研发计划(2018YFB1305200)， 国家自然科学基金(61921004, 61902064, 81971282), 中央高校基本科研业务费专项资金(2242018K3DN01)的资助。

参考文献

[1] C. F. Benitezquiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. pages 5562–5570, 2016. 1

[2] J. Carreira, E. Noland, C. Hillier, and A. Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 6, 7, 8

[3] C. Darwin and P. Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998. 1

[4] A. Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pages 546–550, 2019. 2, 3

[5] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE multimedia*, (3):34–41, 2012. 2, 3, 8

[6] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 3

[7] Y. Fan, J. C. Lam, and V. O. Li. Video-based emotion recognition using deeply-supervised neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 584–588, 2018. 9

[8] Y. Fan, V. Li, and J. C. Lam. Facial expression recognition with deeply-supervised attention network. *IEEE Transactions on Affective Computing*, 2020. 1

[9] Y. Fan, X. Lu, D. Li, and Y. 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. 5, 9

[10] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971. 4

[11] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999. 7

[12] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017. 6, 7

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 7

[15] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 553–560, 2017. 5, 9

[16] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. 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. 1

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 8

[18] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6-7):907–929, 2019. 2, 3

[19] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 2

[20] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn. Context-aware emotion recognition networks. 2019. 2, 3

[21] S. Li and W. Deng. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing*, 28(1):356–370, 2018. 1

[22] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 1

[23] S. Li, W. Zheng, Y. Zong, C. Lu, C. Tang, X. Jiang, J. Liu, and W. Xia. Bi-modality fusion for emotion recognition in the wild. In *2019 International Conference on Multimodal Interaction*, pages 589–594, 2019. 2, 5

[24] C. Liu, T. Tang, K. Lv, and M. Wang. Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 630–634, 2018. 5, 9

[25] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014. 1

[26] X. Liu, M. Kan, W. Wu, S. Shan, and X. Chen. VIPLFaceNet: An open source deep face recognition sdk. *Frontiers of Computer Science (FCS)*, 2016. 6

[27] C. Lu, W. Zheng, C. Li, C. Tang, S. Liu, S. Yan, and Y. Zong. Multiple spatio-temporal feature learning for video-based emotion recognition in the wild. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 646–652, 2018. 9

[28] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[29] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 1

[30] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 8

[31] B. Pan, S. Wang, and B. Xia. Occluded facial expression recognition enhanced through privileged information. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 566–573, 2019. 1

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 6

[33] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 2, 6, 7

[34] M. I. F. research. toolkit. www.faceplusplus.com. 6

[35] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010. 6

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6, 7

[37] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2, 6, 7

[39] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6, 7

[40] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014. 7

[41] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 6, 8

[42] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007. 1

[43] W. Zheng, H. Tang, Z. Lin, and T. S. Huang. Emotion recognition from arbitrary view facial images. pages 490–503, 2010. 1

[44] W. Zheng, X. Zhou, C. Zou, and L. Zhao. Facial expression recognition using kernel canonical correlation analysis (kcca). *IEEE Transactions on Neural Networks*, 17(1):233–238, 2006. 1

[45] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen. A compact representation of visual speech data using latent variables. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):1–1, 2013. 6

[46] Z. Zhou, G. Zhao, and M. Pietikäinen. Towards a practical lipreading system. In *CVPR 2011*, pages 137–144. IEEE, 2011. 6