文献阅读：深度面部表情识别——调查

作者：ReturnYG
原文链接：https://ld246.com/article/1605437547327
来源网站：链滴

原文链接：文献阅读：深度面部表情识别——调查

一、介绍

面部表情是人类传达其情绪状态和意图的最有力，最自然和普遍的信号之一。 由于自动面部表情分析在社交机器人，医学治疗，驾驶员疲劳监视以及许多其他人机交互系统中的实际重要性，因此已经进行了许多研究。 在计算机视觉和机器学习领域，已经探索了各种面部表情识别（FER）系统来对来自面部表示的表情信息进行编码。 早在二十世纪，埃克曼和弗里森就根据跨文化研究定义了六种基本情感，这表明人类无论文化如何都以相同的方式感知某些基本情感。 这些典型的面部表情是愤怒，厌恶，恐惧，幸福，悲伤和惊奇。 轻蔑后来被添加为基本情绪之一。 最近，对神经科学和心理学的高级研究认为，六种基本情绪的模型是特定于文化的，而不是通用的

尽管基于基本情感的情感模型在表达我们日常情感展示和其他情感描述模型（如面部动作编码系统（FACS）和使用情感维度的连续模型）的复杂性和微妙性的能力方面受到限制[11] 被认为代表了更广泛的情感，以离散的基本情感来描述情感的分类模型仍然是FER的最流行观点，这是由于它的开创性研究以及对面部表情的直接和直观定义。 在本次调查中，我们将基于分类模型来限制对FER的讨论。

根据特征表示，**FER系统可分为两大类：静态图像FER和动态序列FER**。 在基于静态的方法中，仅使用来自当前单个图像的空间信息对特征表示进行编码，而基于动态的方法则考虑输入面部表情序列中连续帧之间的时间关系。 基于这两种基于视觉的方法，其他模式（例如音频和生理通道）也已用于多模式系统中，以帮助识别表达。

大多数传统方法都使用手工制作的功能或浅层学习（例如，本地二进制模式（LBP），三个正交平面上的LBP（LBP-TOP），非负矩阵分解（NMF）和稀疏学习）进行FER。 然而，自2013年以来，诸如FER2013和野外情感识别（EmotiW）之类的情感识别竞赛已经从具有挑战性的现实情况中收集了相对足够的培训数据，这隐含地促进了FER从实验室控制向野外过渡设置。 同时，由于芯片处理能力（例如GPU单元）的显着提高和精心设计的网络体系结构，各个领域的研究已开始转移到深度学习方法，从而获得了最新的认可 准确性，大大超过了先前的结果。 同样，由于有了更有效的面部表情训练数据，越来越多的人采用了深度学习技术来应对野外情感识别的挑战性因素。 图1在算法和数据集方面说明了FER的这一演变。

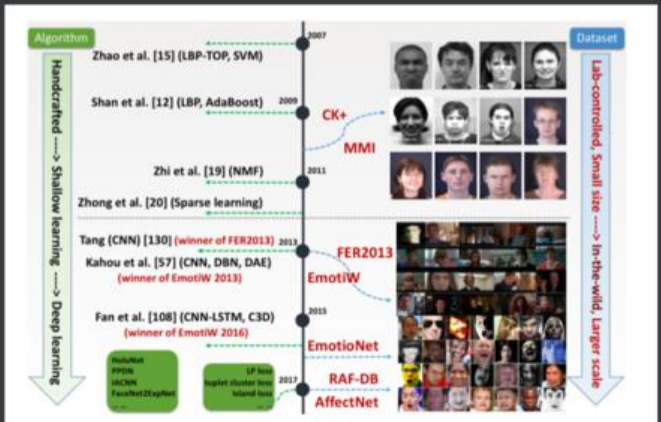


Fig. 1. The evolution of facial expression recognition in terms of datasets and methods.

原文链接：文献阅读：深度面部表情识别——调查

近年来已经发表了有关自动表达分析的详尽调查。 这些调查已经为FER建立了一组标准算法管道。 但是，他们专注于传统方法，深度学习很少受到评论。 最近，在[Facial expression recognition based on deep learning: A survey]中对基于深度学习的FER进行了调查，这是对FER数据集的介绍和关于深度FER的技术细节的简短回顾。 因此，在本文中，我们基于静态图像和视频（图像序列）对FER任务的深度学习进行了系统研究。 我们旨在为新手提供一个有关深度FER的系统框架和主要技能的概述。

尽管深度学习具有强大的功能学习能力，但在应用于FER时仍然存在问题。 首先，**深度神经网络需要大量的训练数据，以避免过度拟合。** 但是，现有的面部表情数据库不足以训练具有深度架构的著名神经网络，该深度网络在对象识别任务中取得了最有希望的结果。 此外，**由于不同的个人属性（例如年龄，性别，种族背景和表达水平），存在较高的主体间差异。** 除了主体身份偏见外，**姿势，照明和遮挡的变化在不受约束的面部表情场景中也很常见。** 这些因素与面部表情非线性耦合，因此加强了对深度网络的要求，以解决较大的类内差异并学习有效的特定表情表示。

在本文中，我们介绍了解决上述FER深度问题的最新研究进展。 我们检查了以前的调查报告中未审查的最新结果。 本文的其余部分安排如下。 第2节介绍了常用的表达数据库。第3节确定了深度FER系统所需的三个主要步骤，并介绍了相关背景。 第4节详细介绍了新颖的神经网络体系结构和基于静态图像和动态图像序列为FER设计的特殊网络训练技巧。 然后，我们将在第5节中介绍其他相关问题和其他实际情况。第6节讨论了该领域中的一些挑战和机遇，并确定了潜在的未来方向。

二、面部表情数据库（FACIAL EXPRESSION DATABASES）

具有足够的标记训练数据，其中包括尽可能多的种群和环境变异，对于深度表达识别系统的设计很重要。 在本节中，我们将讨论包含基本表达式的可公开获得的数据库，这些数据库在我们的综述论文中广泛用于深度学习算法评估。 我们还介绍了新发布的数据库，其中包含从现实世界中收集的大量情感图像，以有益于深度神经网络的训练。 表1概述了这些数据集，包括主要参考资料，主题数量，图像或视频样本数量，收集环境，表达分布和其他信息。

TABLE 1
An overview of the facial expression datasets. P = posed; S = spontaneous; Condit. = Collection condition; Elicit. = Elicitation method.

Database	Samples	Subject	Condit.	Elicit.	Expression distribution	Access
CK+ [33]	593 image sequences	123	Lab	P & S	6 basic expressions plus contempt and neutral	http://www.consortium.ri.cmu.edu/ckagree/
MMI [34], [35]	740 images and 2,900 videos	25	Lab	P	6 basic expressions plus neutral	https://mmifacedb.eu/
JAFFE [36]	213 images	10	Lab	P	6 basic expressions plus neutral	http://www.kasrl.org/jaffe.html
TFD [37]	112,234 images	N/A	Lab	P	6 basic expressions plus neutral	josh@mplab.ucsd.edu
FER-2013 [21]	35,887 images	N/A	Web	P & S	6 basic expressions plus neutral	https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge
AFEW 7.0 [24]	1,809 videos	N/A	Movie	P & S	6 basic expressions plus neutral	https://sites.google.com/site/emotiwchallenge/
SFEW 2.0 [22]	1,766 images	N/A	Movie	P & S	6 basic expressions plus neutral	https://cs.anu.edu.au/few/emotiw2015.html
Multi-PIE [38]	755,370 images	337	Lab	P	Smile, surprised, squint, disgust, scream and neutral	http://www.flintbox.com/public/project/4742/
BU-3DFE [39]	2,500 images	100	Lab	P	6 basic expressions plus neutral	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
Oulu-CASIA [40]	2,880 image sequences	80	Lab	P	6 basic expressions	http://www.cse.oulu.fi/CMV/Downloads/Oulu-CASIA
RaFD [41]	1,608 images	67	Lab	P	6 basic expressions plus contempt and neutral	http://www.socsci.ru.nl:8180/RaFD2/RaFD
KDEF [42]	4,900 images	70	Lab	P	6 basic expressions plus neutral	http://www.emotionlab.se/kdef/
EmotioNet [43]	1,000,000 images	N/A	Web	P & S	23 basic expressions or compound expressions	http://cbcsl.ece.ohio-state.edu/dbform_emotionet.html
RAF-DB [44], [45]	29672 images	N/A	Web	P & S	6 basic expressions plus neutral and 12 compound expressions	http://www.whdeng.cn/RAF/model1.html
AffectNet [46]	450,000 images (labeled)	N/A	Web	P & S	6 basic expressions plus neutral	http://mohammadmahoor.com/databases-codes/
ExpW [47]	91,793 images	N/A	Web	P & S	6 basic expressions plus neutral	http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html

CK+：扩展的CohnKanade（CK +）数据库是**用于评估FER系统的使用最广泛的实验室控制数据库**。CK +包含来自123个主题的593个视频序列。序列的持续时间从10到60帧不等，并且显示了从中性面部表情到峰值表情的转变。在这些视频中，基于面部动作编码系统（FACS），来自118位受试者的327个序列被标记了七个基本表达标签（愤怒，蔑视，厌恶，恐惧，幸福，悲伤和惊奇）。由于CK +不提供指定的训练，验证和测试集，因此在该数据库上评估的算法并不统一。对于基于静态的方法，**最常用的数据选择方法是提取最后一到三个具有峰形成的帧和每个序列的第一个帧（中性面）**。然后，将受试者分为n组用于独立于人的n倍交叉验证实验，其中n的常用选择值为5、8和10。

MMI：MMI数据库是实验室控制的，包括来自32个受试者的326个序列。总共213个序列被六个基本表达式标记（没有“鄙视”），并且在正面视图中捕获了205个序列。**与CK +相反，MMI中的序列在起始-顶点-偏移处标记**，即该序列以中性表达开始，并在返回中性表达之前在中间附近达到峰值。此外，MMI具有更具挑战性的条件，即，人与人之间的差异很大，这是因为受试者的表情不一致，并且他们中的许多人都穿着配饰（例如眼镜，胡须）。对于实验，最常用的方法是在每个额叶序列中选择第一个框架（中性脸）和三个峰值框架，以进行与人无关的10倍交叉验证。

JAFFE：日本女性面部表情（JAFFE）数据库是实验室控制的图像数据库，其中包含来自10位日本女性的213个姿势表情样本。每个人都有3至4张图像，其中有6种基本面部表情（愤怒，厌恶，恐惧，幸福，悲伤和惊奇），每张图像具有中性表情。**该数据库具有挑战性，因为每个主题/表达包含很少的示例**。通常，所有图像都用于遗忘一对象实验。

TFD：多伦多人脸数据库（TFD）是几个面部表情数据集的合并。 TFD包含112,234张图像，其中4,178张图像带有以下七个表达标签之一：愤怒，厌恶，恐惧，幸福，悲伤，惊奇和中立。 面部已经被检测到并规格化为48 * 48，以使所有被摄对象的眼睛相距相同的距离并具有相同的垂直坐标。 TFD中提供了五个官方折叠； 每个折叠包含一个训练集，一个验证集和一个测试集，分别包含70%，10%和20%的图像。

FER2013：在ICML 2013表示学习挑战中引入了FER2013数据库。 FER2013是Google图像搜索API自动收集的大型且不受限制的数据库。 拒绝标记错误的帧并调整裁切区域后，所有图像均已注册并调整为48 * 48像素。 FER2013包含28,709个训练图像，3,589个验证图像和3,589个测试图像，并带有七个表达标签（愤怒，厌恶，恐惧，幸福，悲伤，惊奇和中立）。

AFEW：最早在Acted facial expressions in the wild database中建立并引入了野生面部表情（AFEW）数据库，自2013年以来一直作为年度野外情感识别（EmotiW）的评估平台。AFEW包含从 具有自发表情，各种头部姿势，遮挡和照明的不同电影。 AFEW是一个时间和多模式数据库，可在音频和视频中提供截然不同的环境条件。 样本被标记为七个表达：愤怒，厌恶，恐惧，幸福，悲伤，惊奇和中立。 表达式的注释已不断更新，真人秀节目数据也已不断添加。 EmotiW 2017中的AFEW 7.0在主题和电影/电视来源方面以独立的方式分为三个数据分区：火车（773个样本），瓦尔（383个样本）和测试（653个样本），可确保这三个区域中的数据 场景属于互斥的电影和演员。

SFEW：通过基于面部点聚类计算关键帧从AFEW数据库中选择静态帧来创建“野外静态面部表情”（SFEW）。 最常用的版本SFEW 2.0是EmotiW 2015中SReco子挑战的基准数据[22]。 SFEW 2.0已分为三组：Train（958个样本），Val（436个样本）和Test（372个样本）。 **将每个图像分配给七个表达类别之一，即，愤怒，厌恶，恐惧，中立，幸福，悲伤和惊奇。** 训练和验证集的表达标签是公开可用的，而测试集的表达标签则由挑战组织者保留。

Multi-PIE：CMU多重PIE数据库包含337,370个对象的755,370张图像，这些对象在15个视点和19个光照条件下最多进行四个记录会话。 **每个面部图像都标记有以下六个表情之一：厌恶，中立，尖叫，微笑，斜眼和惊奇。该数据集通常用于多视图面部表情分析。**

BU-3DFE：宾厄姆顿大学3D面部表情（BU-3DFE）数据库包含从100个人中捕获的606个面部表情序列。 对于每个对象，**通过多种方式以多种强度激发出六种普遍的面部表情（愤怒，厌恶，恐惧，幸福，悲伤和惊奇）。** 与Multi-PIE相似，此数据集通常用于多视图3D面部表情分析。

Oulu-CASIA：Oulu-CASIA数据库包含从80个受试者收集的2880个图像序列，这些受试者带有六个基本情感标签：**愤怒，厌恶，恐惧，幸福，悲伤和惊奇。** 在两个不同的照明条件下，使用两个成像系统之一（即近红外（NIR）或可见光（VIS））捕获每个视频。 **与CK +类似，第一帧为中性，最后一帧为峰值。** 通常，在正常室内照明下，由VIS系统收集的480个视频中，只有最后三个峰值帧和第一个帧（中性脸）用于10倍交叉验证实验。

RaFD：Radboud Faces Database（RaFD）是由实验室控制的，具有来自67个对象的总共1,608张图像，具有**三个不同的注视方向，即前，左和右。** 每个样本都标记有以下八种表达方式之一：**愤怒，鄙视，厌恶，恐惧，幸福，悲伤，惊奇和中立。**

KDEF：实验室控制的Karolinska定向情感面孔（KDEF）数据库最初是为在心理和医学研究中使用而开发的。KDEF由来自70个演员的图像组成，这些图像有五个不同的角度，分别标记有六个基本的面部表情和中性的表情。除了这些用于基本情绪识别的常用数据集之外，最近两年还出现了一些从互联网收集的，建立良好且大规模的，可公开使用的，适合训练深度神经网络的面部表情数据库。

EmotioNet：EmotioNet是一个大型数据库，具有从Internet收集的一百万张面部表情图像。 Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild中的自动动作单元（AU）检测模型对总共950,000张图像进行了注释，其余**25,000张图像由11个AU进行了人工注释**。EmotioNet挑战[51]的第二条轨道提供了六个基本表达式和十个复合表达式，并且提供了带有表达式标签的2,478张图像。

RAF-DB：真实世界的面部表情数据库（RAF-DB）是一个真实世界的数据库，其中包含从互联网下载的29,672张高度多样化的面部图像。 **通过手动众包注释和可靠的估计，为样本提供了七个基本和十一个复合情感标签**。 具体而言，将来自基本情感集的15339张图像分为两组（12271个训练样本和3068个测试样本）进行评估。

AffectNet：AffectNet包含超过一百万张来自Internet的图像，这些图像是通过使用与情感相关的标签查询不同的搜索引擎而获得的。 它是迄今为止最大的数据库，它以两种不同的情感模型（分类模型和维度模型）提供面部表情，其中450,000张图像具有手动注释的用于8种基本表情的标签。

ExpW：表达式野生数据库（ExpW）包含使用Google图像搜索下载的91,793张面孔。 每个面部图像都被手动注释为七个基本表情类别之一。 非面部图像在注释过程中被删除。

三、深度面部表情识别（DEEP FACIAL EXPRESSION RECOGNITION）

在本节中，我们描述了自动深度FER中常见的三个主要步骤，即预处理，深度特征学习和深度特征分类。 我们简要总结了每个步骤中广泛使用的算法，并根据参考文献推荐了现有的最佳实践方案。

3.1 预处理

在不受限制的情况下，与面部表情无关的变化（例如不同的背景，光照和头部姿势）非常普遍。因此，在训练深度神经网络学习有意义的特征之前，需要进行预处理以对齐和规范化由面部传达的视觉语义信息。

3.1.1 人脸对齐(Face alignment)

人脸对齐是许多人脸相关识别任务中的传统预处理步骤。 我们列出了一些著名的方法以及广泛用于深度FER的公开实现。

给定一系列训练数据，第一步是检测面部，然后去除背景和非面部区域。 **Viola-Jones（V&J）面部检测器是用于面部检测的经典且广泛采用的实现方式**，该功能强大且计算简单，可检测近额面部。

尽管面部检测是启用特征学习的唯一必不可少的过程，但使用局部界标的坐标进行进一步的面部对齐可以大大提高FER性能。 此步骤至关重要，因为它可以减少面部比例和面内旋转的变化。 表2研究了在深层FER中广泛使用的面部标志检测算法，并在效率和性能方面进行了比较。 活动外观模型（AAM）是一种经典的生成模型，可从整体面部外观和整体形状模式中优化所需的参数。 在判别模型中，树木（MoT）结构化模型与判别响应图拟合（DRMF）的混合使用基于零件的方法，这些方法通过每个地标周围的局部外观信息来表示人脸。 此外，许多判别

模型直接使用级联回归函数将图像外观映射到地标位置，并显示出更好的结果，例如，在IntraFace中实施的监督下降方法（SDM），面部对齐3000 fps和增量面部对齐。 最近，深层网络已被广泛用于面部对齐。 级联CNN是以级联方式预测地标的早期工作。 基于此，任务约束深度卷积网络（TCDCN）和多任务CNN（MTCNN）进一步利用多任务学习来提高性能。 总的来说，**级联回归（cascaded regression）以其高速度和准确性而成为最流行的人脸对齐方法。**

TABLE 2
Summary of different types of face alignment detectors that are widely used in deep FER models.

type		# points	real-time	speed	performance	used in
Holistic	AAM [53]	68	✗	fair	poor generalization	[54], [55]
Part-based	MoT [56]	39/68	✗	slow/ fast	good	[57], [58]
	DRMF [59]	66	✗			[60], [61]
Cascaded regression	SDM [62]	49	✓	fast/ very fast	good/ very good	[16], [63]
	3000 fps [64]	68	✓			[55]
	Incremental [65]	49	✓			[66]
Deep learning	cascaded CNN [67]	5	✓	fast	good/ very good	[68]
	MTCNN [69]	5	✓			[70], [71]

与仅使用一个检测器进行面部对齐相比，一些方法提出了在复杂的无约束环境中处理面部时组合多个检测器以实现更好的界标估计的方法。 Yu等。 串联了三个不同的面部标志检测器以相互补充。 Kim等。 考虑到不同的输入（原始图像和直方图均等化图像）和不同的面部检测模型（V&J和MoT），因此选择了Intraface提供的具有最高置信度的地标集。

3.1.2 资料扩充(Data augmentation)

深度神经网络需要足够的训练数据，以确保对给定识别任务的通用性。 但是，大多数公开可用的FER数据库没有足够数量的图像用于训练。 因此，**数据增强是深度FER的关键步骤。 数据增强技术可以分为两类：即时数据增强和常规数据增强。**

通常，即时数据增强(on-the-fly data augmentation)被嵌入深度学习工具包中，以缓解过度拟合。 在训练步骤中，从图像的四个角和中心随机裁剪输入样本，然后水平翻转，**这可能导致数据集比原始训练数据大十倍。** 在测试过程中采用两种常见的预测模式：**仅使用面部的中心补丁进行预测，或者对所有十种作物的预测值取平均值。**

除了基本的即时数据扩充外，还设计了各种精细的数据扩充操作以进一步扩展数据的大小和多样性。 **最常用的操作包括随机扰动和变换，例如旋转，移位，偏斜，缩放，噪声，对比度和色彩抖动。** 例如，使用通用噪声模型，盐和胡椒和斑点噪声以及高斯噪声来扩大数据大小。 对于对比度转换，每个像素的饱和度和值（HSV颜色空间的S和V分量）都将更改以进行数据增强。 多种操作的组合可以生成更多看不见的训练样本，并使网络对于偏斜和旋转的脸部更加健壮。 在[A deep-learning approach to facial expression recognition with candid images]中，作者应用了五个图像外观过滤器（磁盘，平均，高斯，不清晰和运动过滤器）和六个微小的变换矩阵，这些矩阵通过向恒等矩阵添加轻微的几何变换来形式化。 在[Image based static facial expression recognition with multiple deep network learning]中，提出了一个更全面的仿射变换矩阵来随机生成随旋转，偏斜和比例变化的图像。 此外，基于深度学习的技术可以应用于数据增强。 例如，在[Using synthetic data to improve facial expression analysis with 3d convolutional networks]中**创建了具有3D卷积神经网络（CNN）的合成数据生成系统**，以保密地创建表情具有不同饱和度的面部。 **生成对抗网络（GAN）还可以通过生成姿势和表情不同的外观来应用于增强数据。** （请参阅第4.1.7节）。

3.1.3 人脸归一化(Face normalization)

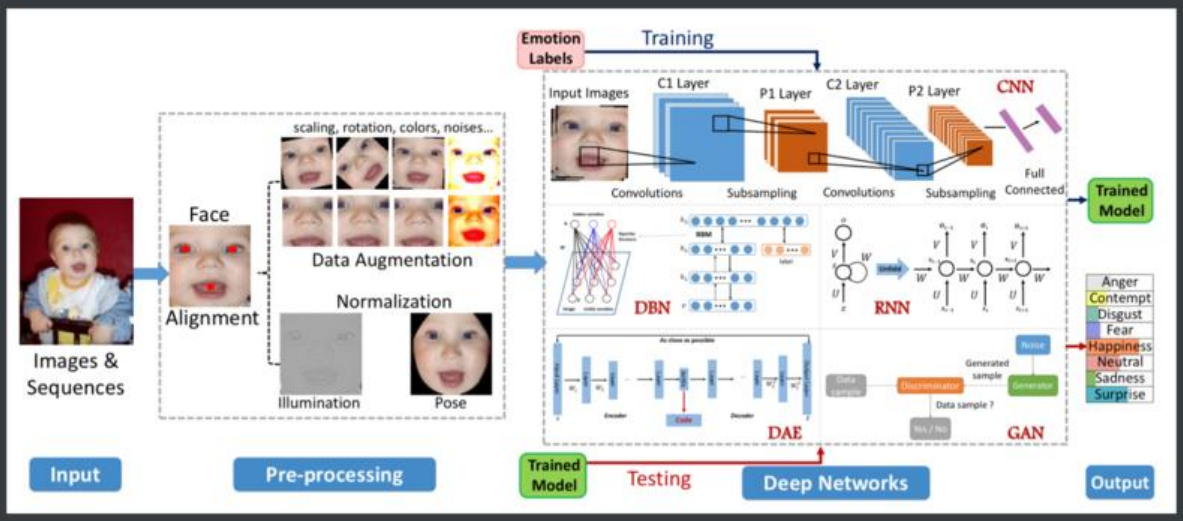
照度和头部姿势的变化会导致图像发生较大变化，从而损害FER性能。 因此，我们**引入了两种典型的人脸归一化方法来改善这些变化：照明归一化和姿势归一化（正面化）**。

照度归一化(Illumination normalization): 即使来自同一个人的相同表情，照明和对比度在不同的图像中也会有所不同，尤其是在不受限制的环境中，这可能会导致较大的组内差异。在[Baseline cnn structure analysis for facial expression recognition]中，针对照明归一化，评估了几种常用的照明归一化算法，**即基于各向同性扩散（IS）的归一化，基于离散余弦变换（DCT）的归一化和高斯差（DoG）**。并且[Facial expression recognition using deep neural networks]中使用**基于同态滤波的归一化技术**，据报道在所有其他技术中可产生最一致的结果，以消除照明归一化。 此外，相关研究表明，与单独使用照明归一化相比，**直方图均衡与照明归一化相结合可获得更好的人脸识别性能。**在深度FER文献中，许多研究都采用直方图均衡化来增加图像的整体对比度，以进行预处理。当背景和前景的亮度相似时，此方法有效。 但是，直接应用直方图均衡化可能会过分强调局部对比度。 为了解决这个问题，[A compact deep learning model for robust facial expression recognition]提出了一种加权求和方法，将直方图均衡和线性映射相结合。 在[Enhancing cnn with preprocessing stage in automatic emotion recognition]中，作者比较了三种不同的方法：**全局对比度归一化（GCN），局部归一化和直方图均衡化。 据报道，GCN和直方图均衡化分别在训练和测试步骤中实现了最佳准确性。**

姿势归一化(Pose normalization): 在不受约束的环境中，相当大的姿势变化是另一个常见且棘手的问题。 一些研究采用姿势归一化技术为FER产生6个正面视图，其中最流行的是Hassner等人提出的。 具体来说，**在定位面部标志之后，将生成一个针对所有面部通用的3D纹理参考模型，以有效地估计可见的面部成分。** 然后，通过将每个输入面部图像反投影到参考坐标系来合成初始的正面面部。 另外，Sagonas等。 [Effective face frontalization in unconstrained images]提出了一个有效的统计模型，**可以同时定位地标和仅使用正面将面部姿势转换**。 最近，提出了一系列**基于GAN的深度模型用于前视图合成**（例如FF-GAN [94]，TP-GAN [95]和DR-GAN [96]），并报告了令人鼓舞的性能。

3.2 用于功能学习的深度网络(Deep networks for feature learning)

深度学习最近已成为热门研究主题，并已在各种应用程序中实现了最先进的性能。 深度学习尝试通过多个非线性变换和表示的层次结构来捕获高级抽象。 在本节中，我们简要介绍一些已应用于FER的深度学习技术。 这些深度神经网络的传统架构如图2所示。



3.2.1卷积神经网络（CNN）

CNN已广泛用于包括FER在内的各种计算机视觉应用中。 在21世纪初，FER文献[Robust face analysis using convolutional neural networks]，[Head-pose invariant facial expression recognition using convolutional neural networks]中的几项研究发现，**CNN能够很好地应对位置变化和尺度变化，并且在以前看不见的情况下，其性能优于多层感知器（MLP）**。 面部姿势变化， [Subject independent facial expression recognition with robust face detection using a convolutional neural network]**使用CNN解决面部表情识别中的主题独立性以及平移，旋转和尺度不变性的问题。**

CNN由三种类型的异构层组成：**卷积层，池化层和完全连接的层。 卷积层具有一组可学习的滤波器**，可以对整个输入图像进行卷积，并生成各种特定类型的激活特征图。 卷积操作与三个主要优点相关联：本地连通性，它学习相邻像素之间的相关性； 同一特征图中的权重共享，大大减少了要学习的参数数量； 以及对对象位置的平移不变性。 **池化层位于卷积层之后，用于减少特征图的空间大小和网络的计算成本**。 平均池化和最大池化是平移不变性最常用的两种非线性下采样策略。 通常将完全连接的层包括在网络的末端，以确保该层中的所有神经元都完全连接到上一层中的激活，并使2D特征图可以转换为1D特征图以进行进一步的特征表示和分类。

我们在表3中列出了一些已经应用于FER的著名CNN模型的配置和特性。除了这些网络之外，还存在一些著名的派生框架。 在[Combining multimodal features within a fusion network for emotion recognition in the wild]，[Facial expression recognition in the wild based on multimodal texture features]中，基于区域的CNN（R-CNN）[Rich feature hierarchies for accurate object detection and semantic segmentation]用于学习FER的功能。 在[Facial expression recognition with faster r-cnn]中，Faster R-CNN通过生成高质量的区域提议来识别面部表情。此外，Ji等。 提出的3D CNN捕获在多个相邻帧中编码的运动信息，以通过3D卷积进行动作识别。

TABLE 3
Comparison of CNN models and their achievements. DA = Data augmentation; BN = Batch normalization.

	AlexNet [25]	VGGNet [26]	GoogleNet [27]	ResNet [28]
Year	2012	2014	2014	2015
# of layers†	5+3	13/16 + 3	21+1	151+1
Kernel size⋆	11, 5, 3	3	7, 1, 3, 5	7, 1, 3, 5
DA	✓	✓	✓	✓
Dropout	✓	✓	✓	✓
Inception	✗	✗	✓	✗
BN	✗	✗	✗	✓
Used in	[110]	[78], [111]	[17], [78]	[91], [112]

† number of convolutional layers + fully connected layers
⋆ size of the convolution kernel

[Learning spatiotemporal features with 3d convolutional networks]提出了精心设计的C3D，它利用大规模监督训练数据集上的3D卷积来学习时空特征。 许多相关研究已将此网络用于涉及图像序列的FER。

3.2.2深度信任网络(Deep belief network (DBN))

由Hinton等人提出的DBN。 [A fast learning algorithm for deep belief nets]是学习提取训练数据的深层次表示的图形模型。 传统的DBN是由一堆受限的Boltzmann机器（RBM）构建的，这些机器是由可见单元层和隐藏单元层组成的两层生成随机模型。 RBM中的这两层必须形成没有横向连接的二部图。 在DBN中，对高层中的单元进行训练以了解相邻较低层中的单元之间的条件相关性，但最顶层的两层除外，它们具有无方向的连接。 DBN的训练包含两个阶段：**预训练和微调**。 首先，采用有效的逐层贪心学习策略以无人监督的方式初始化深度网络，这可以在某种程度上防止较差的局部最优结果，而无需大量标记数据。 在此过程中，使用对比散度来训练DBN中的RBM，以估计对数似然的近似梯度。 然后，在监督下通过简单的梯度下降对网络的参数和所需的输出进行微调。

3.2.3 深度自动编码器(Deep autoencoder (DAE))

DAE是在[Reducing the dimensionality of data with neural networks]中首次引入的，旨在学习有效的降维编码。 与前面提到的经过训练可以预测目标值的网络相反，DAE经过优化，可以通过最小化重构误差来重构其输入。 存在DAE的变体，例如降噪自动编码器，它可以从部分损坏的数据中恢复原始的未失真输入； 稀疏自动编码器网络（DSAE），它对所学习的特征表示实施稀疏性； 压缩自编码器（CAE 1），它添加了一个依赖于活动的正则化以诱导局部不变特征； 卷积自动编码器（CAE 2），对网络中的隐藏层使用卷积（和可选的池化）层； 可变自动编码器（VAE），它是一种具有某些类型的潜在变量的定向图形模型，用于设计复杂的数据生成模型。

3.2.4 循环神经网络(Recurrent neural network (RNN))

RNN是一种捕获时间信息的连接主义者模型，更适合于具有任意长度的顺序数据预测。 除了以单一前馈方式训练深度神经网络外，**RNN还包括跨越相邻时间步长并在所有步长上共享相同参数的循环边。 经典的时间反向传播（BPTT）用于训练RNN。** Hochreiter＆Schmidhuber引入的长期短期记忆（LSTM）是传统RNN的一种特殊形式，用于解决训练RNN时常见的梯度消失和爆炸问题。 LSTM中的细胞状态由三个门控制和控制：一个输入门允许或阻止输入信号改变细胞状态，一个输出门使细胞状态能够影响其他神经元，或阻止该状态影响其他神经元。 调制单元的自循环连接以累积或忘记其先前状态。 通过结合这三个门，LSTM可以按顺序对长期依赖性进行建模，并且已被广泛用于基于视频的表情识别任务。

3.2.5生成对抗网络(Generative Adversarial Network (GAN))

GAN是由Goodfellow等人于2014年首次提出的，它通过生成器G（z）之间的极小极大两人博弈训练模型，生成器G（z）通过将潜势z映射到z～p（z）的数据空间来生成综合输入数据， 分配概率y = Dis（x）∈[0，1]的D（x）是x的实际训练样本，用于区分真实的输入数据和假的输入数据。 生成器和鉴别器是经过交替训练的，并且都可以通过相对于D / G最小化/最大化二进制交叉熵L GAN = log（D（x））+ log（1- D（G（z））来提高自身。 x是训练样本，z～p（z）。 存在GAN的扩展，例如添加条件信息以控制发生器输出的cGAN [126]，采用反卷积和卷积神经网络分别实现G和D的DCGAN，使用学习的特征表示的VAE / GAN 在GAN鉴别器中作为VAE重建目标的基础，而InfoGAN则可以以完全无监督的方式学习解缠结的表示。

3.3面部表情分类(Facial expression classification)

在学习了深刻的特征之后，FER的最后一步是将给定的面孔分类为基本的情感类别之一。

与传统方法不同，传统方法的特征提取步骤和特征分类步骤是独立的，深度网络可以以端到端的方式执行FER。 具体来说，在网络末端增加一个损耗层，以调节反向传播误差。 然后，每个样本的预测概率可以由网络直接输出。 在CNN中，softmax损失是最常用的函数，它可以最大程度地减少估计的类概率和地面真理分布之间的交叉熵。 另外，[Deep learning using linear support vector machines]证明了使用线性支持向量机（SVM）进行端到端训练的好处，该方法可最大程度地减少基于边际的损失而不是交叉熵。 同样，[Investigating deep neural forests for facial expression recognition]研究了深层神经森林（NFs）的适应性，后者用NFs取代了softmax损失层，并获得了竞争性的FER结果。

除了端到端学习方法外，另一种选择是采用深度神经网络（特别是CNN）作为特征提取工具，然后将其他独立分类器（例如支持向量机或随机森林）应用于提取的表示。 此外，[Deep covariance descriptors for facial expression recognition]，[Covariance pooling for facial expression recognition]表明，在DCNN特征上计算的协方差描述符以及在对称正定义（SPD）流形上与高斯核的分类比与softmax层的标准分类更有效。

四、最新技术

在本节中，我们回顾了为FER设计的现有新型深度神经网络以及为解决表达特异问题而提出的相关训练策略。根据数据类型，我们将文献中介绍的工作分为两个主要组：用于静态图像的深度FER网络和用于动态图像序列的深度FER网络。 然后，我们就网络体系结构和性能提供了当前深度FER系统的概述。 由于某些评估的数据集未提供用于训练，验证和测试的明确数据组，并且相关研究可能会在不同的实验条件下使用不同的数据进行实验，因此我们总结了表达识别性能以及有关数据选择和分组方法的信息。

4.1用于静态映像的Deep FER网络（Deep FER networks for static images）

由于数据处理的便利性以及相关培训和测试材料的可用性，大量的现有研究基于静态图像执行表情识别任务，而没有考虑时间信息。 我们首先介绍FER的特定预训练和微调技能，然后回顾该领域中的新型深度神经网络。 对于每个最频繁评估的数据集，表4显示了该领域中最新的方法，这些方法是在独立于人员的协议中明确进行的（训练和测试集中的主题已分离）。

TABLE 4
Performance summary of representative methods for static-based deep facial expression recognition on the most widely evaluated datasets. Network size = depth & number of parameters; Pre-processing = Face Detection & Data Augmentation & Face Normalization; IN = Illumination Normalization; $\mathcal{NE}$ = Network Ensemble; $\mathcal{CN}$ = Cascaded Network; $\mathcal{MN}$ = Multitask Network; LOSO = leave-one-subject-out.

TABLE 4
Performance summary of representative methods for static-based deep facial expression recognition on the most widely evaluated datasets. Network size = depth & number of parameters; Pre-processing = Face Detection & Data Augmentation & Face Normalization; IN = Illumination Normalization; $\mathcal{NE}$ = Network Ensemble; $\mathcal{CN}$ = Cascaded Network; $\mathcal{MN}$ = Multitask Network; LOSO = leave-one-subject-out.

Datasets	Method	Network type		Network size		Pre-processing			Data selection	Data group	Additional classifier	Performance[1](%)
CK+	Ouellet 14 [110]	CNN (AlexNet)		-	-	V&J	-	-	the last frame	LOSO	SVM	7 classes[‡]: (94.4)
	Li et al. 15 [86]	RBM		4	-	V&J	-	IN			✗	6 classes: 96.8
	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-	the last three frames and the first frame	8 folds	AdaBoost	6 classes: 96.7
	Liu et al. 13 [137]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	8 classes: 92.05 (87.67)
	Liu et al. 15 [138]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	7 classes[‡]: 93.70
	Khorrami et al. 15 [139]	zero-bias CNN		4	7m	✓	✓	-		10 folds	✗	6 classes: 95.7; 8 classes: 95.1
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-		10 folds	✗	6 classes: (98.6); 8 classes: (96.8)
	Zeng et al. 18 [54]	DAE (DSAE)		3	-	AAM	-	-	the last four frames and the first frame	LOSO	✗	7 classes[‡]: 95.79 (93.78) 8 classes: 89.84 (86.82)
	Cai et al. 17 [140]	CNN	loss layer	6	-	DRMF	✓	IN	the last three frames	10 folds	✗	7 classes[‡]: 94.39 (90.66)
	Meng et al. 17 [61]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-		8 folds	✗	7 classes[‡]: 95.37 (95.51)
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN		8 folds	✗	7 classes[‡]: 97.1 (96.1)
	Yang et al. 18 [141]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	7 classes[‡]: 97.30 (96.57)
	Zhang et al. 18 [47]	CNN	$\mathcal{MN}$	-	-	✓	✓	-		10 folds	✗	6 classes: 98.9
JAFFE	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-	213 images	LOSO	AdaBoost	7 classes[‡]: 91.8
	Hamester et al. 15 [142]	CNN, CAE	$\mathcal{NE}$	3	-	-	-	IN			✗	7 classes[‡]: (95.8)
MMI	Liu et al. 13 [137]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-	the middle three frames and the first frame	10 folds	SVM	7 classes[‡]: 74.76 (71.73)
	Liu et al. 15 [138]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	7 classes[‡]: 75.85
	Mollahosseini et al. 16 [14]	CNN (Inception)		11	7.3m	IntraFace	✓	-	images from each sequence	5 folds	✗	6 classes: 77.9
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN	the middle three frames	10 folds	✗	6 classes: 78.53 (73.50)
	Li et al. 17 [44]	CNN	loss layer	8	5.8m	IntraFace	✓	-		5 folds	SVM	6 classes: 78.46
	Yang et al. 18 [141]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	6 classes: 73.23 (72.67)
TFD	Reed et al. 14 [143]	RBM	$\mathcal{MN}$	-	-	-	-	-	4,178 emotion labeled 3,874 identity labeled	5 official folds	SVM	Test: 85.43
	Devries et al. 14 [58]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN	4,178 labeled images		✗	Validation: 87.80 Test: 85.13 (48.29)
	Khorrami et al. 15 [139]	zero-bias CNN		4	7m	✓	✓	-			✗	Test: 88.6
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-			✗	Test: 88.9 (87.7)
FER 2013	Tang 13 [130]	CNN	loss layer	4	12.0m	-	✓	IN	Training Set: 28,709 Validation Set: 3,589 Test Set: 3,589		✗	Test: 71.2
	Devries et al. 14 [58]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN			✗	Validation+Test: 67.21
	Zhang et al. 15 [144]	CNN	$\mathcal{MN}$	6	21.3m	SDM	-	-			✗	Test: 75.10
	Guo et al. 16 [145]	CNN	loss layer	10	2.6m	SDM	✓	-			k-NN	Test: 71.33
	Kim et al. 16 [146]	CNN	$\mathcal{NE}$	5	2.4m	IntraFace	✓	IN			✗	Test: 73.73
	pramerdorfer et al. 16 [147]	CNN	$\mathcal{NE}$	10/16/33	1.8/1.2/5.3 (m)	-	✓	IN			✗	Test:75.2
SFEW 2.0	levi et al. 15 [78]	CNN	$\mathcal{NE}$	VGG-S/VGG-M/ GoogleNet		MoT	✓	-	891 training, 431 validation, and 372 test		✗	Validation: 51.75 Test: 54.56
	Ng et al. 15 [63]	CNN	fine-tune	AlexNet		IntraFace	✓	-	921 training, ? validation, and 372 test		✗	Validation: 48.5 (39.63) Test: 55.6 (42.69)
	Li et al. 17 [44]	CNN	loss layer	8	5.8m	IntraFace	✓	-	921 training, 427 validation		SVM	Validation: 51.05
	Ding et al. 17 [111]	CNN	fine-tune	8	11m	IntraFace	✓	-	891 training, 425 validation		✗	Validation: 55.15 (46.6)
	Liu et al. 17 [77]	CNN	loss layer	11	-	IntraFace	✓	IN	958training, 436 validation, and 372 test		✗	Validation: 54.19 (47.97)
	Cai et al. 17 [140]	CNN	loss layer	6	-	DRMF	✓	IN			✗	Validation: 52.52 (43.41) Test: 59.41 (48.29)
	Meng et al. 17 [61]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-			✗	Validation: 50.98 (42.57) Test: 54.30 (44.77)
	Kim et al. 15 [76]	CNN	$\mathcal{NE}$	5	-	multiple	✓	IN			✗	Validation: 53.9 Test: 61.6
	Yu et al. 15 [75]	CNN	$\mathcal{NE}$	8	6.2m	multiple	✓	IN			✗	Validation: 55.96 (47.31) Test: 61.29 (51.27)

[1] The value in parentheses is the mean accuracy, which is calculated with the confusion matrix given by the authors.
[†] 7 Classes: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise.
[‡] 7 Classes: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise.

The value in parentheses is the mean accuracy, which is calculated with the confusion matrix given by the authors.7 Classes: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise. 7 Classes: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise.

4.1.1预训练和微调（Pre-training and fine-tuning）

如前所述，在相对较小的面部表情数据集上进行深度网络的直接训练很容易过度拟合。 为了缓解这个问题，许多研究使用了额外的面向任务的数据来从头对自己的自建网络进行预训练，或者在众所周知的预训练模型（例如AlexNet，VGG，VGG-face和GoogleNet）上进行精细调整。 Kahou等。 [Combining modality specific deep neural networks for emotion recognition in video]，[Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks]指出，使用附加数据可以帮助获得具有高容量的模型而不会过度拟合，从而增强了FER性能。

为了选择适当的辅助数据，适合使用大规模人脸识别（FR）数据集（例如CASIA WebFace，野外名人脸（CFW），FaceScrub数据集）或相对较大的FER数据集（FER2013和TFD）。 Kaya等。 [Video-based emotion recognition in the wild using deep transfer learning and score fusion]建议为FR训练的VGG-Face压倒了为目标识别而开发的ImageNet。 Knyazev等人观察到的另一个有趣的结果。 [Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video]是在较大的FR数据上进行预训练会积极地影响情绪识别的准确性，并且使用其他FER数据集进行进一步的微调可以帮助改善性能。

代替直接使用预先训练或精细调整的模型来提取目标数据集上的特征，多阶段精细调整策略（请参见图3中的“提交3”）可以实现更好的性能：在第一阶段精细调整之后， 在预训练的模型上使用FER2013进行调整，基于目标数据集（EmotiW）训练部分的第二阶段细调用于调整模型以适应更具体的数据集（即目标数据集）。

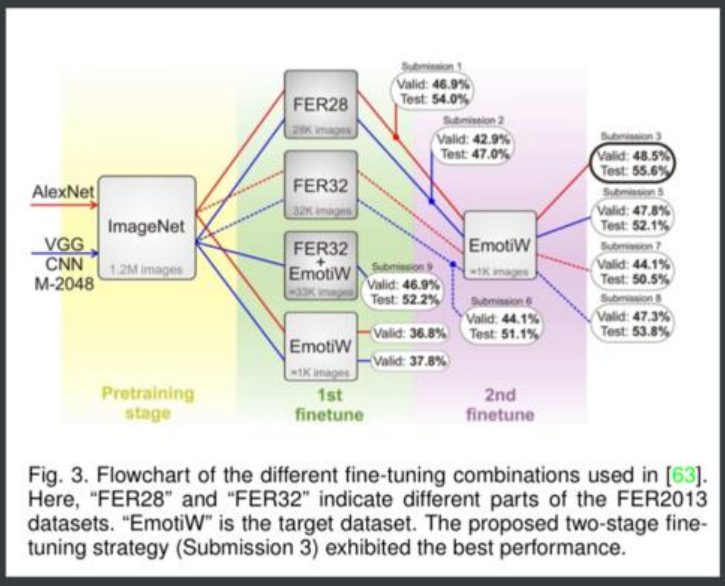


Fig. 3. Flowchart of the different fine-tuning combinations used in [63]. Here, “FER28” and “FER32” indicate different parts of the FER2013 datasets. “EmotiW” is the target dataset. The proposed two-stage fine-tuning strategy (Submission 3) exhibited the best performance.

尽管对外部FR数据进行预训练和微调可以间接避免训练数据较小的问题，但网络是与FER分开进行训练的，并且面部主导的信息仍保留在学习的特征中，这可能会削弱网络表示表情的能力 。 为了消除这种影响，提出了一个两阶段的训练算法FaceNet2ExpNet（见图4） 。 精细调整的面部网络可以很好地初始化表情网络，并且仅用于指导卷积层的学习。 并且使用表达式信息从头开始训练完全连接的层，以规范目标FER网络的训练。

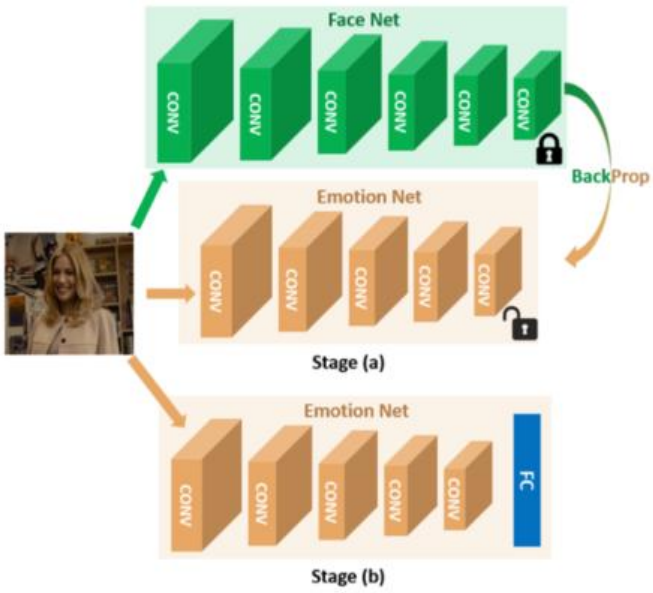


Fig. 4. Two-stage training flowchart in [111]. In stage (a), the deeper face net is frozen and provides the feature-level regularization that pushes the convolutional features of the expression net to be close to the face net by using the proposed distribution function. Then, in stage (b), to further improve the discriminativeness of the learned features, randomly initialized fully connected layers are added and jointly trained with the whole expression net using the expression label information.

4.1.2多样化的网络输入(Diverse network input)

传统做法通常使用RGB图像的整个对齐面作为网络的输入，以学习FER的功能。 但是，这些原始数据缺少重要的信息，例如均匀或规则的纹理以及图像缩放，旋转，遮挡和照明方面的不变性，这可能代表FER的混淆因素。 一些方法采用了各种手工制作的功能及其扩展作为缓解网络问题的网络输入。

低级表示对给定RGB图像中小区域的特征进行编码，然后将这些特征与局部直方图聚类并合并在一起，这些局部直方图对于照明变化和较小的配准误差具有鲁棒性。 提出了一种新颖的映射LBP特征（见图5）用于照度不变的FER。 针对图像缩放和旋转具有鲁棒性的Scaleinvariant特征变换（SIFT）特征用于多视图FER任务。 将轮廓，纹理，角度和颜色的不同描述符组合为输入数据也可以帮助增强深度网络性能。

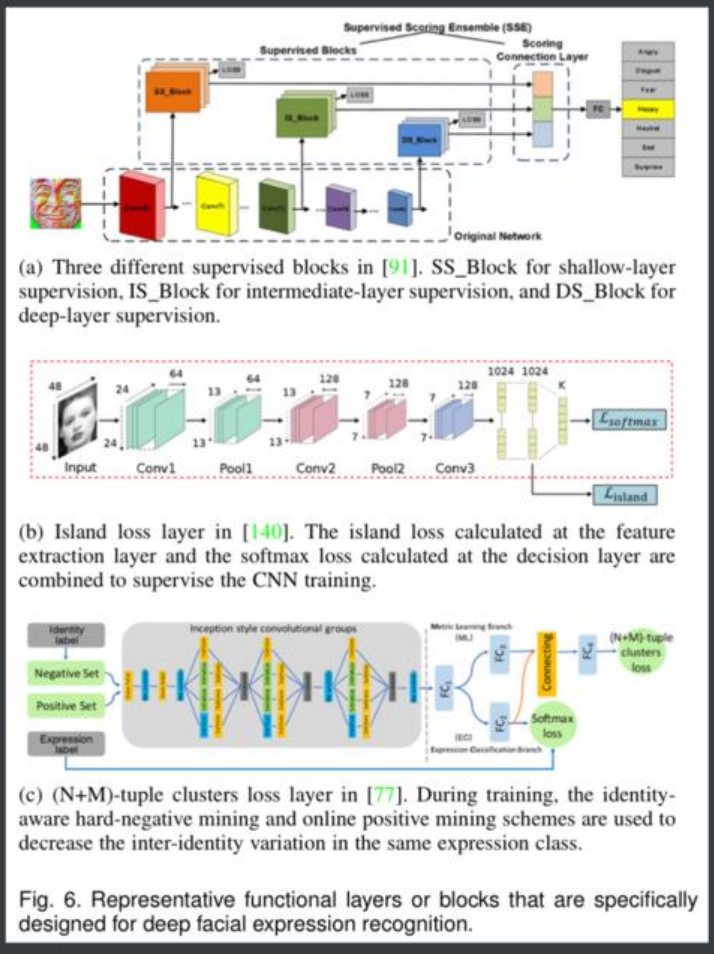
Fig. 5. Image intensities (left) and LBP codes (middle). [78] proposed mapping these values to a 3D metric space (right) as the input of CNNs.

基于零件的表示会根据目标任务提取特征，这些特征会从整个图像中删除非关键部分并利用对任务敏感的关键部分。 [Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction]指出，**三个感兴趣的区域（ROI），即眉毛，眼睛和嘴巴，与面部表情的变化密切相关**，并裁剪了这些区域作为DSAE的输入。 **其他研究建议自动学习面部表情的关键部分**。 例如，[Facial expression recognition using visual saliency and deep learning]采用了一个深层的多层网络来检测显着性图，该显着性图将强度施加在需要视觉注意力的零件上。 文献[Adaptive feature mapping for customizing deep learning based facial expression recognition model]应用邻居中心差向量（NCDV）来获得具有更多固有信息的特征。

4.1.3 辅助块和层(Auxiliary blocks & layers)

基于CNN的基础架构，一些研究提出了添加设计良好的辅助块或层以增强学习特征的与表达相关的表示能力。

一种新颖的CNN体系结构HoloNet，是为FER设计的，其中CReLU与强大的残差结构相结合，可在不降低效率的情况下增加网络深度，并具有初始残差块，[Inception-v4, inception-resnet and the impact of residual connections on learning]专为FER设计以学习多尺度特征 捕获表达式中的变化。 为了提高FER的监督程度，引入了另外一个CNN模型Supervised Scoring Ensemble（SSE），在主流CNN的早期隐藏层中分别嵌入了三种类型的受监管块，分别用于浅层，中层和深层监管(见图2).6(a)。



(a) Three different supervised blocks in [91]. SS_Block for shallow-layer supervision, IS_Block for intermediate-layer supervision, and DS_Block for deep-layer supervision.

(b) Island loss layer in [140]. The island loss calculated at the feature extraction layer and the softmax loss calculated at the decision layer are combined to supervise the CNN training.

(c) (N+M)-tuple clusters loss layer in [77]. During training, the identity-aware hard-negative mining and online positive mining schemes are used to decrease the inter-identity variation in the same expression class.

Fig. 6. Representative functional layers or blocks that are specifically designed for deep facial expression recognition.

通过在AlexNet中嵌入特征选择机制来设计特征选择网络（FSN），该机制会根据学习的面部表情特征图自动过滤不相关的特征并强调相关的特征。 有趣的是，Zeng等。 [Facial expression recognition with inconsistently annotated datasets]指出不同FER数据库之间的注释不一致是不可避免的，当通过合并多个数据集来扩大训练集时，这将损害性能。 为了解决这个问题，作者提出了对潜在真相的不一致伪注释（IPA2LT）框架。 在IPA2LT中，端到端可训练的LTNet旨在通过最大化这些不一致的注释的对数似然性，从人类注释和从不同数据集中训练的机器注释中发现潜在的真相。

CNN中的传统softmax损失层只是简单地迫使不同类别的特征保持分离，但现实世界中的FER不仅遭受类别间相似性高而且类别间变异高的困扰。 因此，一些工作提出了用于FER的新颖的损耗层。 受中心损失的启发，该损失惩罚了深层特征与其对应的类中心之间的距离，提出了两种变体来协助对softmax损失进行监督，以对FER进行更多区分：（1）正式确定了岛损失以进一步增加成对 不同类别中心之间的距离（见图6（b））和（2）保留位置的损失（LP损失）被形式化为将同一类别的局部邻近特征拉在一起，从而使每个类别的类别内局部聚类 类很紧凑。 此外，基于三重态损失，这需要一个正例比一个具有固定间隙的负例更接近锚点，提出了两种变体来替代或帮助监督softmax损失：（1）指数三重态 在更新网络时，将丢失形式化以赋予困难的样本更大的权重，并且将（2）（N + M）-元组聚类损失形式化以减轻同一性FER的三重态损失中锚点选择和阈值验证的困难（见图。 有关详细信息，请参见图6（c））。 此外，提出了一种特征损失，以在早期训练阶段为深度特征提供补充信息。

	definition
Majority Voting	determine the class with the most votes using the predicted label yielded from each individual
Simple Average	determine the class with the highest mean score using the posterior class probabilities yielded from each individual with the same weight
Weighted Average	determine the class with the highest weighted mean score using the posterior class probabilities yielded from each individual with different weights

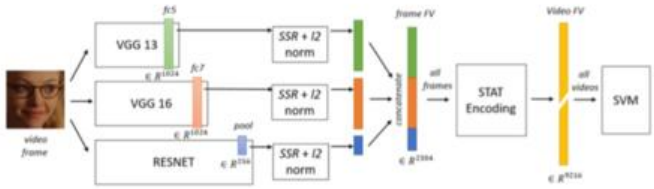
4.1.4 网络合奏(Network ensemble)

先前的研究表明，多个网络的组合可以胜过单个网络。 在实现网络集成时，应考虑两个关键因素：（1）网络的足够多样性以确保互补性；（2）一种可以有效聚合委员会网络的适当集成方法。

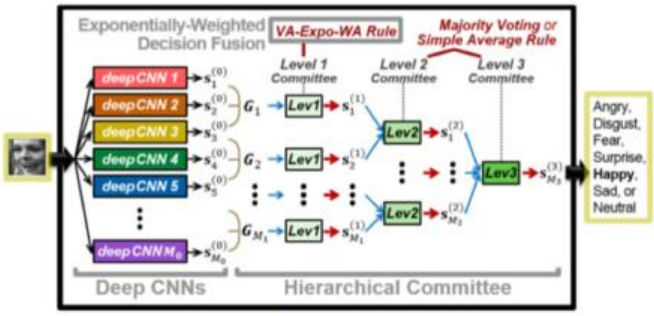
在第一个因素方面，考虑使用不同种类的训练数据和各种网络参数或体系结构来产生不同的委员会。 几种预处理方法（例如变形和规范化）以及第4.1.2节中描述的方法可以生成不同的数据来训练各种网络。 通过更改过滤器的大小，神经元的数量和网络的层数，并应用多个随机种子进行权重初始化，还可以增强网络的多样性。 此外，可以使用不同的网络架构来增强多样性。 例如，将以监督方式训练的CNN和以非监督方式训练的卷积自动编码器（CAE）结合在一起用于网络集成。

对于第二个因素，委员会网络的每个成员都可以在两个不同的级别进行组装：要素级别和决策级别。 对于特征级别的集成，最常用的策略是连接从不同网络学习到的特征。 例如，[Emotion recognition in the wild from videos using images]将从不同网络学习到的特征连接起来，以获得用于描述输入图像的单个特征向量（见图7（a））。 对于决策级集合，将应用三个广泛使用的规则：多数表决，简单平均和加权平均。 表5提供了这三种方法的摘要。 由于加权平均规则考虑了每个人的重要性和信度，因此提出了许多加权平均方法以找到网络集成的最佳权重集。 [Combining modality specific deep neural networks for emotion recognition in video,]提出了一种随机搜索方法来加权每种情绪类型的模型预测。 [Image based static facial expression recognition with multiple deep network

learning]使用对数似然损失和铰链损失为每个网络自适应地分配不同的权重。 [Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition]提出了基于验证准确性的指数加权平均值，以强调合格的个人（见图7（b））。 [Supervised committee of convolutional neural networks in automated facial expression analysis]使用CNN来学习每个模型的权重。



(a) Feature-level ensemble in [88]. Three different features (*fc5* of VGG13 + *fc7* of VGG16 + *pool* of Resnet) after normalization are concatenated to create a single feature vector (FV) that describes the input frame.



(b) Decision-level ensemble in [76]. A 3-level hierarchical committee architecture with hybrid decision-level fusions was proposed to obtain sufficient decision diversity.

Fig. 7. Representative network ensemble systems at the feature level and decision level.

4.1.5 多任务网络(Multitask networks)

许多现有的FER网络都专注于单个任务，并学习对表达式敏感的功能，而无需考虑其他潜在因素之间的相互作用。 但是，在现实世界中，FER与各种因素交织在一起，例如头部姿势，照明和受试者身份（面部形态）。 为了解决此问题，引入了多任务学习以从其他相关任务中转移知识并消除令人讨厌的因素。

里德等。 [Learning to disentangle factors of variation with manifold interaction]构造了一个高阶玻尔兹曼机（disBM）来学习表达相关因素的流形坐标，并提出了解开纠缠的训练策略，使得与表达相关的隐藏单元对于面部形态不变。 其他工作[Multi-task learning of facial landmarks and expression]，[Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition]建议将FER与其他任务同时进行，例如面部界标定位和面部AUs检测，可以共同提高FER性能。

此外，一些工作采用多任务学习进行身份不变的FER。 在[Identity-aware convolutional neural network for facial expression recognition]中，提出了具有两个相同子CNN的身份识别CNN（IACNN）。 一个流使用表达敏感的对比损失来学习表达区分特征，另一流使用身份敏感的对比损失来学习关于身份不变FER的身份相关特征。 在[Facial expression recognition based on deep evolutional spatial-temporal networks]中，提出了一种多信号CNN（MSCNN），它在FER和人脸验证任务的监督下进行了训练，以迫使模型将注意力集中在表情信息上（见图8）。

此外，提出了一种多功能的CNN模型，以同时解决包括笑容检测在内的多种面部分析任务。 首先使用预先在面部识别上训练的权重来初始化网络，然后通过对多个数据集进行训练，使用基于域的正则化从特定的层次中分支特定于任务的子网。 具体来说，由于微笑检测是一项与主题无关的任务，它更多地依赖于可从较低层获得的本地信息，因此作者建议融合较低的卷积层以形成用于微笑检测的通用表示形式。 常规的监督多任务学习需要为所有任务标记训练样本。 为了放松这一点，[From facial expression recognition to interpersonal relation prediction]提出了一种新颖的属性传播方法，该方法可以利用面部表情与其他异构属性之间的固有对应关系，尽管不同数据集的分布不同。

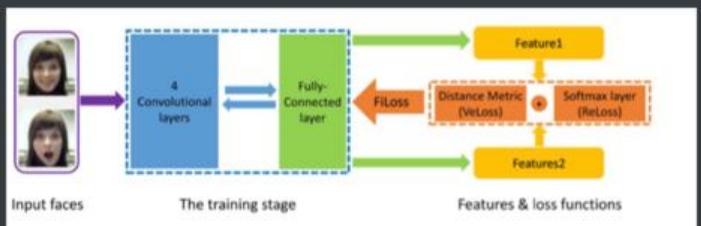


Fig. 8. Representative multitask network for FER. In the proposed MSCNN [68], a pair of images is sent into the MSCNN during training. The expression recognition task with cross-entropy loss, which learns features with large between-expression variation, and the face verification task with contrastive loss, which reduces the variation in within-expression features, are combined to train the MSCNN.

4.1.6 级联网络(Cascaded networks)

在级联网络中，将用于不同任务的各种模块顺序组合以构建更深的网络，其中前一个模块的输出将被后者的模块利用。 相关研究提出了不同结构的组合，以学习特征的层次结构，通过该层次结构，可以逐步筛选出与表达无关的变异因素。

最常见的是，不同的网络或学习方法是顺序地和单独地组合在一起的，它们中的每一个都有不同的层次作用。在[Facial expression recognition via deep learning]中，对DBN进行了训练以首先检测面部并检测与表达相关的区域。 然后，通过堆叠的自动编码器对这些解析的面部分量进行分类。 在[Disentangling factors of variation for facial expression recognition]中，提出了一种多尺度压缩卷积网络（CCNET）来获得局部翻译不变（LTI）表示。然后，设计收缩自动编码器，以从主体身份和姿势中分层分离出与情感相关的因素。 在[Au-aware deep networks for facial expression recognition]，[Au-inspired deep networks for facial expression feature learning]中，首先使用CNN架构学习了过度完整的表示，然后利用多层RBM来学习FER的更高级别的功能（见图9）。 Liu等人，而不是简单地连接不同的网络。 [Facial expression recognition via a boosted deep belief network]提出了一种增强的DBN（BDBN），它可以在单循环状态下迭代执行特征表示，特征选择和分类器构造。 与没有反馈的串联相比，该循环框架向后传播分类错误，以交替启动特征选择过程，直到收敛为止。 因此，在此迭代过程中，可以大大提高FER的判别能力。

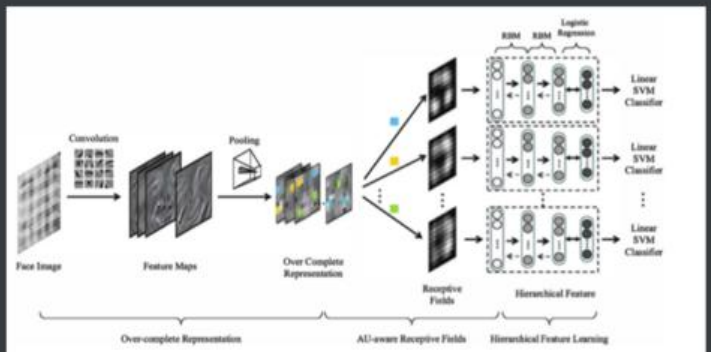


Fig. 9. Representative cascaded network for FER. The proposed AU-aware deep network (AUDN) [137] is composed of three sequential modules: in the first module, a 2-layer CNN is trained to generate an over-complete representation encoding all expression-specific appearance variations over all possible locations; in the second module, an AU-aware receptive field layer is designed to search subsets of the over-complete representation; in the last module, a multilayer RBM is exploited to learn hierarchical features.

4.1.7 生成对抗网络（GAN）

最近，基于GAN的方法已成功地用于图像合成中，以产生令人印象深刻的逼真的人脸，数字和各种其他图像类型，这对训练数据增强和相应的识别任务是有益的。 一些工作提出了基于GAN的姿势不变FER和身份不变FER的新颖模型。

对于姿态不变的FER，Lai等人。 [Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition]提出了一种基于GAN的人脸正面化框架，其中生成器将输入的人脸图像正面化，同时保留身份和表达特征，鉴别器将真实图像与所产生的人脸正面图像区分开。 张等。 [Joint pose and expression modeling for facial expression recognition]提出了一种基于GAN的模型，该模型可以针对多视图FER在任意姿势下生成具有不同表情的图像。 对于身份不变的FER，Yang等。 [Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks]提出了一个由两个部分组成的身份自适应生成（IA-gen）模型。 上部分别使用cGAN生成具有不同表情的同一主题的图像。 然后，下部在不涉及其他个体的情况下针对每个单个身份子空间进行FER，因此可以很好地缓解身份变化。 Chen等。 [Vgan-based image representation learning for privacy-preserving facial expression recognition]提出了一种隐私保护表示学习变体GAN（PPRL-VGAN），该算法结合了VAE和GAN来学习一种身份不变的表示形式，该表示形式与身份信息明确分离，并生成了用于保留表情的面部图像合成。 杨等。 [141]提出了一种去表达残基学习（DeRL）过程，以探索表达信息，该信息在去表达过程中被过滤掉，但仍嵌入生成器中。 然后，该模型直接从生成器中提取此信息，以减轻主体变化的影响并提高FER性能。

4.1.8讨论(Discussion)

现有的结构良好的深度FER系统着重于两个关键问题：缺乏足够的多样化训练数据以及表达无关的变化，例如照明，头部姿势和身份。 表6显示了这些不同类型的方法相对于两个未解决的问题（数据大小要求和与表达式无关的变化）以及其他重点（计算效率，性能和网络培训难度）的相对优缺点。

TABLE 6
Comparison of different types of methods for static images in terms of data size requirement, variations* (head pose, illumination, occlusion and other environment factors), identity bias, computational efficiency, accuracy, and difficulty on network training.

Network type	data	variations*	identity bias	efficiency	accuracy	difficulty
Pre-train & Fine-tune	low	fair	vulnerable	high	fair	easy
Diverse input	low	good	vulnerable	low	fair	easy
Auxiliary layers	varies	good	varies	varies	good	varies
Network ensemble	low	good	fair	low	good	medium
Multitask network	high	varies	good	fair	varies	hard
Cascaded network	fair	good	fair	fair	fair	medium
GAN	fair	good	good	fair	good	hard

预训练和精细调整已成为深度FER的主流，以解决训练数据不足和过度拟合的问题。 事实证明，特别有用的实用技术是使用辅助数据（从大型异议或面部识别数据集到小型FER数据集，即从大到小以及从常规到通用）在多个阶段对网络进行预训练和微调。 但是，与端到端的训练框架相比，与表达式无关的表示结构仍保留在现成的预训练模型中，例如与异议网的较大领域差距以及面部网络中的对象识别干扰。 因此，提取的特征通常容易受到身份变化的影响，并且性能将下降。 值得注意的是，随着大规模野生FER数据集（例如AffectNet和RAF-DB）的出现，使用具有中等规模的深层网络进行的端到端训练也可以实现竞争表现。

除了直接使用原始图像数据来训练深层网络外，建议使用各种预先设计的功能，以增强网络对常见干扰（例如照明，头部姿势和遮挡）的鲁棒性，并迫使网络将更多注意力集中在面部 具有表达性信息的区域。 此外，使用多个异构输入数据可以间接扩大数据大小。 但是，这种方法通常会忽略身份偏差问题。 此外，生成各种数据会占用更多时间，并且将这些多个数据组合会导致高维度，这可能会影响网络的计算效率。

训练具有大量隐藏层和灵活过滤器的深度和广泛的网络是学习区分目标任务的深层高级功能的有效方法。 但是，此过程易受训练数据大小的影响，如果没有足够的训练数据来学习新参数，则该过程可能会表现不佳。 解决此问题的自然研究方向是将多个相对较小的网络并联或串联集成。 网络集成在功能或决策级别集成了各种网络，以结合其优势，通常在情感比赛中应用以帮助提高性能。 然而，设计不同种类的网络以相互补偿会明显增加计算成本和存储需求。 此外，通常根据原始训练数据的性能来学习每个子网的权重，从而导致对新看不见的测试数据的过度拟合。 考虑到目标FER任务与其他次要任务（例如面部界标定位，面部AU识别和面部验证）之间的交互作用，多任务网络共同训练了多个网络，因此可以很好地消除与表达无关的因素，包括身份偏差。 这种方法的缺点是，它需要所有任务的标记数据，并且随着涉及更多任务，训练变得越来越麻烦。 可替代地，级联网络以分级方法顺序地训练多个网络，在这种情况下，所学习特征的辨别能力被不断增强。 通常，此方法可以缓解过度拟合的问题，同时可以逐步消除与面部表情无关的因素。 值得考虑的不足之处是，大多数现有级联系统中的子网都是在没有反馈的情况下进行单独训练的，而端到端的训练策略对于提高训练效果和性能是可取的。

理想情况下，深层网络（尤其是CNN）具有处理头部变化的良好功能，但是大多数当前的FER网络并未明确解决头部变化，并且未在自然场景中进行测试。 可以利用生成对抗网络（GAN）来解决这个问题，方法是正面化人脸图像，同时保留表情特征或合成任意姿势以帮助训练姿势不变网络。 GAN的另一个优点是，可以通过生成相应的中性人脸图像或合成不同的表情来明确区分身份变化，同时为身份不变的FER保存身份信息。 此外，GAN可以帮助增加规模和多样性方面的培训数据。 GAN的主要缺点是训练不稳定以及视觉质量和图像多样性之间的权衡。

4.2 用于动态图像序列的深度FER网络(Deep FER networks for dynamic image sequences)

尽管以前的大多数模型都集中在静态图像上，但是面部表情识别可以从序列中连续帧的时间相关性中受益。 我们首先介绍现有的帧聚合技术，这些技术可以策略性地结合从基于静态FER网络中学习到的深层功能。 然后，考虑到在视频流中人们通常以不同的强度显示相同的表情，因此，我们进一步回顾了使用不同表情强度状态下的图像进行不变性FER的方法。 最后，我们介绍了深层FER网络，该网络考虑了视频帧中的时空运动模式以及从时间结构中得出的学习特征。 对于每个最频繁评估的数据集，表7显示了在独立于人员的协议中执行的当前最新方法。

TABLE 7
Performances of representative methods for dynamic-based deep facial expression recognition on the most widely evaluated datasets. Network size = depth & number of parameters; Pre-processing = Face Detection & Data Augmentation & Face Normalization; IN = Illumination Normalization; $\mathcal{FA}$ = Frame Aggregation; $\mathcal{EIN}$ = Expression Intensity-invariant Network; $\mathcal{FLT}$ = Facial Landmark Trajectory; $\mathcal{CN}$ = Cascaded Network; $\mathcal{NE}$ = Network Ensemble; S = Spatial Network; T = Temporal Network; LOSO = leave-one-subject-out.

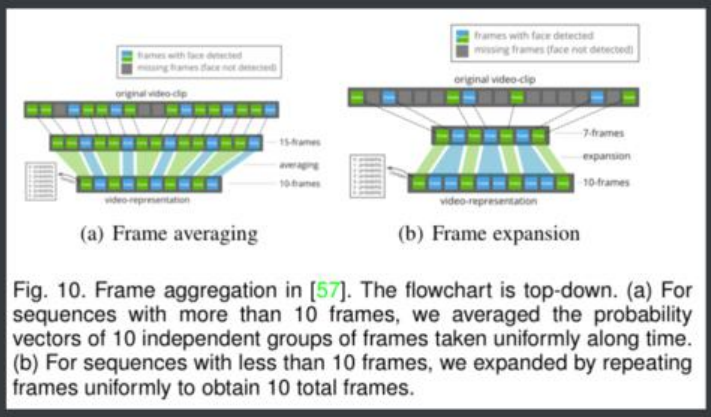
Datasets	Methods	Network type	Network size		Pre-processing			Training data Selection in each sequence	Testing data selection in each sequence	Data group	Performance[1](%)
CK+	Zhao et al. 16 [17]	$\mathcal{EIN}$	22	6.8m	✓	-	-	from the 7th to the last[2]	the last frame	10 folds	6 classes: 99.3
	Yu et al. 17 [70]	$\mathcal{EIN}$	42	-	MTCNN	✓	-	from the 7th to the last[2]	the peak expression	10 folds	6 classes: 99.6
	kim et al. 17 [184]	$\mathcal{EIN}$	14	-	✓	✓	-	all frames	the same as the training data	10 folds	7 classes: 97.93
	Sun et al. 17 [185]	$\mathcal{NE}$	3 * GoogLeNetv2		✓	-	-	S: emotional T: neutral+emotional		10 folds	6 classes: 97.28
	Jung et al. 15 [16]	$\mathcal{FLT}$	2	177.6k	IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 92.35
	Jung et al. 15 [16]	C3D	4	-	IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 91.44
	Jung et al. 15 [16]	$\mathcal{NE}$	$\mathcal{FLT}$/C3D		IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 97.25 (95.22)
	kuo et al. 18 [89]	$\mathcal{FA}$	6	2.7m	IntraFace	✓	IN	fixed length 9		10 folds	7 classes: 98.47
	Zhang et al. 17 [68]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the last frame T: all frames		10 folds	7 classes: 98.50 (97.78)
MMI	Kim et al. 17 [66]	$\mathcal{EIN}, \mathcal{CN}$	7	1.5m	Incremental	✓	-	5 intensities frames	the same as the training data	LOSO	6 classes: 78.61 (78.00)
	kim et al. 17 [184]	$\mathcal{EIN}$	14	-	✓	✓	-	all frames		10 folds	6 classes: 81.53
	Hasani et al. 17 [112]	$\mathcal{FLT}, \mathcal{CN}$	22	-	3000 fps	-	-	ten frames		5 folds	6 classes: 77.50 (74.50)
	Hasani et al. 17 [55]	$\mathcal{CN}$	29	-	AAM	-	-	static frames		5 folds	6 classes: 78.68
	Zhang et al. 17 [68]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the middle frame T: all frames		10 folds	6 classes: 81.18 (79.30)
	Sun et al. 17 [185]	$\mathcal{NE}$	3 * GoogLeNetv2		✓	-	-	S: emotional T: neutral+emotional		10 folds	6 classes: 91.46
Oulu-CASIA	Zhao et al. 16 [17]	$\mathcal{EIN}$	22	6.8m	✓	-	-	from the 7th to the last[2]	the last frame	10 folds	6 classes: 84.59
	Yu et al. 17 [70]	$\mathcal{EIN}$	42	-	MTCNN	✓	-	from the 7th to the last[2]	the peak expression	10 folds	6 classes: 86.23
	Jung et al. 15 [16]	$\mathcal{FLT}$	2	177.6k	IntraFace	✓	-	fixed number of frames	the same as the training data	10 folds	6 classes: 74.17
	Jung et al. 15 [16]	C3D	4	-	IntraFace	✓	-	fixed number of frames		10 folds	6 classes: 74.38
	Jung et al. 15 [16]	$\mathcal{NE}$	$\mathcal{FLT}$/C3D		IntraFace	✓	-	fixed number of frames		10 folds	6 classes: 81.46 (81.49)
	Zhang et al. 17 [68]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the last frame T: all frames		10 folds	6 classes: 86.25 (86.25)
	kuo et al. 18 [89]	$\mathcal{NE}$	6	2.7m	IntraFace	✓	IN	fixed length 9		10 folds	6 classes: 91.67
AFEW* 6.0	Ding et al. 16 [186]	$\mathcal{FA}$	AlexNet		✓	-	-	Training: 773; Validation: 373; Test: 593			Validation: 44.47
	Yan et al. 16 [187]	$\mathcal{CN}$	VGG16-LSTM		✓	✓	-	40 frames		3 folds	7 classes: 44.46
	Yan et al. 16 [187]	$\mathcal{FLT}$	4	-	[188]	-	-	30 frames		3 folds	7 classes: 37.37
	Fan et al. 16 [108]	$\mathcal{CN}$	VGG16-LSTM		✓	-	-	16 features for LSTM			Validation: 45.43 (38.96)
	Fan et al. [108]	C3D	10	-	✓	-	-	several windows of 16 consecutive frames			Validation: 39.69 (38.55)
	Yan et al. 16 [187]	fusion	/					Training: 773; Validation: 383; Test: 593			Test: 56.66 (40.81)
	Fan et al. 16 [108]	fusion	/					Training: 774; Validation: 383; Test: 593			Test: 59.02 (44.94)
AFEW* 7.0	Ouyang et al. 17 [189]	$\mathcal{CN}$	VGG-LSTM		MTCNN	✓	-	16 frames			Validation: 47.4
	Ouyang et al. 17 [189]	C3D	10	-	MTCNN	✓	-	16 frames			Validation: 35.2
	Vielzeuf et al. [190]	$\mathcal{CN}$	C3D-LSTM		✓	✓	-	detected face frames			Validation: 43.2
	Vielzeuf et al. [190]	$\mathcal{CN}$	VGG16-LSTM		✓	✓	-	several windows of 16 consecutive frames			Validation: 48.6
	Vielzeuf et al. [190]	fusion	/					Training: 773; Validation: 383; Test: 653			Test: 58.81 (43.23)

[1] The value in parentheses is the mean accuracy calculated from the confusion matrix given by authors.
[2] A pair of images (peak and non-peak expression) is chosen for training each time.
* We have included the result of a single spatio-temporal network and also the best result after fusion with both video and audio modalities.
† 7 Classes in CK+: Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise.
‡ 7 Classes in AFEW: Anger, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise.

4.2.1 帧聚合(Frame aggregation)

由于给定视频剪辑中的帧表达强度可能会发生变化，因此直接测量每帧错误不会产生令人满意的性能。 已经提出了各种方法来聚合每个序列中的帧的网络输出以提高性能。 我们将这些方法分为两组：决策级帧聚合和特征级帧聚合。

对于决策级帧聚合，将序列中每个帧的n类概率向量进行积分。 最方便的方法是直接连接这些帧的输出。 但是，每个序列中的帧数可能不同。已经考虑了两种聚合方法来为每个序列生成一个固定长度的特征向量：帧平均和帧扩展（有关详细信息，请参见图10）。 一种不需要固定帧数的替代方法是应用统计编码。 平均值，最大值，平方平均值，最大抑制向量的平均值等可以用于总结每个序列中每帧的概率。



(a) Frame averaging (b) Frame expansion

Fig. 10. Frame aggregation in [57]. The flowchart is top-down. (a) For sequences with more than 10 frames, we averaged the probability vectors of 10 independent groups of frames taken uniformly along time. (b) For sequences with less than 10 frames, we expanded by repeating frames uniformly to obtain 10 total frames.

对于特征级别的帧聚合，将序列中帧的学习特征进行聚合。 许多基于统计的编码模块可以应用在该方案中。一种简单有效的方法是将所有框架上的特征的均值，方差，最小值和最大值连接起来。 或者，也可以采用基于矩阵的模型（例如特征向量，协方差矩阵和多维高斯分布）进行汇总。 此外，已经探索了多实例学习用于视频级表示，其中从辅助图像数据计算聚类中心，然后为每个包视频帧获得词袋表示。

4.2.2 表达强度网络(Expression Intensity network)

大多数方法（在第4.1节中介绍）专注于识别峰值高强度表达，而忽略细微的低强度表达。 在本节中，我们介绍了表达强度不变网络，该网络以不同强度的训练样本作为输入，以利用强度变化的序列表达之间的内在联系。

在表达强度不变网络中，带有强度标签的图像帧用于训练。 在测试过程中，表达强度变化的数据用于验证网络的强度不变能力。 赵等。 [Peak-piloted deep network for facial expression recognition]提出了一种峰值导引的深层网络（PPDN），该网络将一对具有相同表达且来自同一对象的峰值和非峰值图像作为输入，并利用L2范数损失来最小化两个图像之间的距离。 在反向传播过程中，提出了一种峰梯度抑制（PGS）来将非峰表达的学习特征趋向于峰表达的特征，同时避免反演。 因此，可以提高对低强度表达的网络判别能力。 Yu等基于PPDN。 [Deeper cascaded peak-piloted network for weak expression recognition]提出了一种更深的级联峰值导频网络（DCPN），该网络使用更深和更大的体系结构来增强学习特征的判别能力，并采用了一种称为级联细调的集成训练方法来避免过度拟合。 在[Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition]中，利用了更多的强度状态（发作，发作到顶点过渡，顶点，顶点到偏移过渡和偏移），并采用了五种损失函数来通过最小化表情分类错误，类内表情变化， 强度分类误差和内部强度变化，以及分别编码中间强度。

考虑到针对个人身份具有不同表达强度的图像并不总是在野外可用，因此提出了几项旨在自动获取强度标签或生成具有目标强度的新图像的作品。 例如，在[Deep peak-neutral difference feature for facial expression recognition]中，峰序列和中性帧自动从序列中分两个阶段进行：聚类阶段，使用K-means算法将所有帧分为峰状组和中性状组，以及 分类阶段，使用半监督SVM检测峰值和中性帧。 在[Deep generativecontrastive networks for facial expression recognition]中，提出了一个深层的生成对比模型，该模型分为两个步骤：一个生成器，通过卷积编码器/解码器为每个样本生成参考（表达较少的）脸；一个对比网络，共同过滤掉不相关的信息 通过对比度量损失和有监督的重建损失来表达。

4.2.3 深度时空FER网络(Deep spatio-temporal FER network)

尽管帧聚合可以将帧整合到视频序列中，但关键的时间依存性并未得到明确利用。 相比之下，时空FER网络将时间窗口中的一系列帧作为单个输入，而无需事先了解表达强度，并利用纹理和时间信息对更细微的表达进行编码。

RNN和C3D:RNN可以通过利用以下事实来从序列中可靠地获取信息：连续数据的特征向量在语义上是相互联系的，因此是相互依赖的。 改进的版本LSTM灵活地以较低的计算成本处理长度可变的顺序数据。 源自RNN的RNN由ReLU组成，并用单位矩阵（IRNN）初始化，用于提供一种更简单的机制来解决消失和爆炸的梯度问题。 并且使用双向RNN（BRNN）来学习原始方向和反向方向上的时间关系。 最近，在[Spatio-temporal convolutional features with nested lstm for facial expression recognition]中提出了带有两个子LSTM的嵌套LSTM。 即，T-LSTM对学习到的特征的时间动态进行建模，而C-LSTM将所有T-LSTM的输出整合在一起，以便对在网络中间层中编码的多层特征进行编码。

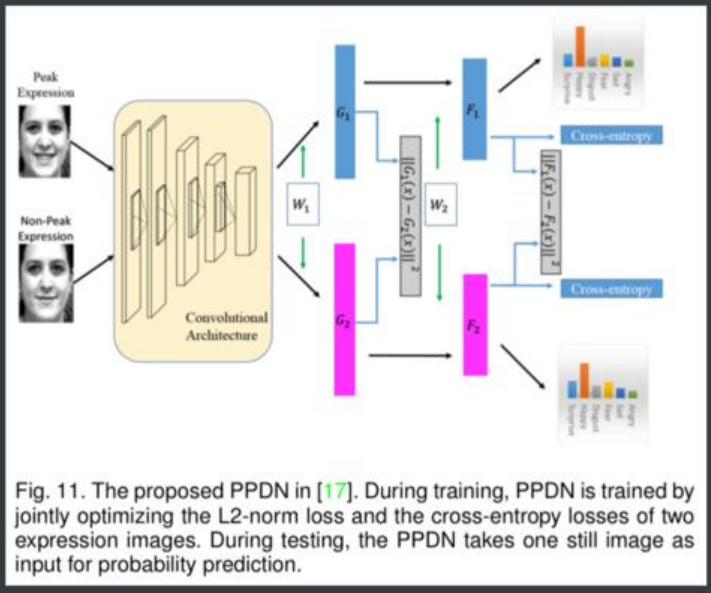


Fig. 11. The proposed PPDN in [17]. During training, PPDN is trained by jointly optimizing the L2-norm loss and the cross-entropy losses of two expression images. During testing, the PPDN takes one still image as input for probability prediction.

与RNN相比，CNN更适合于计算机视觉应用； 因此，它的派生C3D使用沿时间轴具有权重共享的3D卷积内核而不是传统的2D内核，已被广泛用于基于动态的FER捕获时空特征。 基于C3D，已经为FER设计了许多派生结构。在[Deeply learning deformable facial action parts model for dynamic expression analysis]中，将3D CNN与DPM启发的可变形面部动作约束结合在一起，以同时编码动态运动和区分部分的表示形式（有关详细信息，请参见图12）。 在[Joint fine-tuning in deep neural networks for facial expression recognition]中，提出了一种深时空出现网络（DTAN），该网络采用了3D滤波器，并且没有沿时间轴分配重量； 因此，每个过滤器的重要性会随着时间而

变化。 同样，提出了加权C3D，其中从每个序列中提取几个连续帧的窗口，并根据它们的预测分数进行加权。 代替直接使用C3D进行分类，[Deep spatio-temporal features for multimodal emotion recognition]使用C3D进行时空特征提取，然后与DBN级联进行预测。 在[Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild]中，C3D还被用作特征提取器，随后是NetVLAD层，以通过学习聚类中心来聚合运动特征的时间信息。

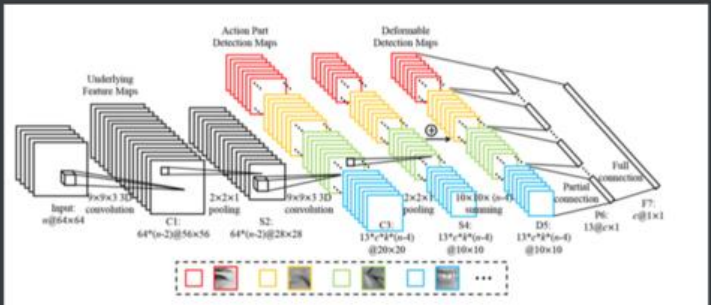


Fig. 12. The proposed 3DCNN-DAP [199]. The input n-frame sequence is convolved with 3D filters; then, $13 * c * k$ part filters corresponding to 13 manually defined facial parts are used to convolve k feature maps for the facial action part detection maps of c expression classes.

面部标志轨迹: 相关的心理学研究表明，表情是通过某些面部部分（例如，眼睛，鼻子和嘴巴）的动态运动来调用的，这些动作包含最能描述表情的信息。 为了获得用于FER的更准确的面部动作，已经提出了面部界标轨迹模型来捕获来自连续帧的面部分量的动态变化。

要提取界标轨迹表示，最直接的方法是通过标准化将随时间变化的帧中的面部界标点的坐标连接起来，以生成每个序列的一维轨迹信号或形成像图像的地图作为CNN的输入。 此外，连续帧中每个界标的相对距离变化也可以用于捕获时间信息。 此外，事实证明，基于局部的模型可以根据面部的物理结构将面部地标分为几个部分，然后分别将它们分别馈送到网络中，这对于局部低级和全局高级特征编码都是有效的（请参阅“ PHRNN” 在图13中）。 Hasani等人（而不是分别提取轨迹特征，然后将它们输入到网络中）。 [Facial expression recognition using enhanced deep 3d convolutional neural networks]通过用面部标志的元素逐次乘法和残差单元的输入张量替换原始3D Inception-ResNet残差单元中的快捷方式，从而合并了轨迹特征。 因此，可以对基于地标的网络进行端到端训练。

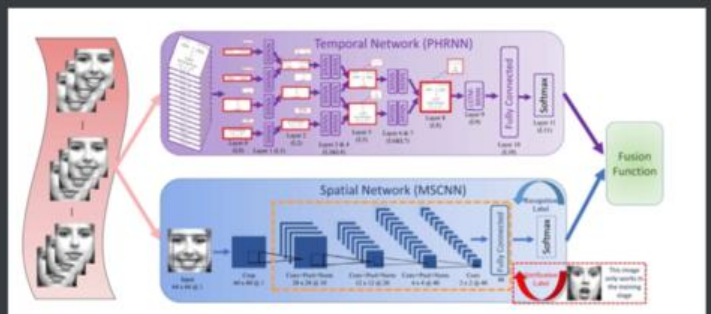


Fig. 13. The spatio-temporal network proposed in [68]. The temporal network PHRNN for landmark trajectory and the spatial network MSCNN for identity-invariant features are trained separately. Then, the predicted probabilities from the two networks are fused together for spatio-temporal FER.

级联网络：通过结合从CNN中学习到的强大的感知视觉表示与LSTM在可变长度输入和输出中的优势，Donahue等人。 提出了一个在空间和时间上都较深的模型，该模型将CNN的输出与LSTM进行级联，以完成涉及时变输入和输出的各种视觉任务。 类似于该混合网络，已经提出了许多用于FER的级联网络。

[Spatiotemporal convolutional sparse auto-encoder for sequence classification]代替卷积神经网络，使用了卷积稀疏自动编码器来处理稀疏和移位不变特征。 然后，对LSTM分类器进行了时间演化训练。 [Audio-visual emotion recognition using deep transfer learning and multiple temporal models]采用了一种更灵活的网络，称为ResNet-LSTM，该网络允许较低CNN层中的节点直接与LSTM接触以捕获时空信息。 除了将LSTM与CNN的完全连接层连接起来之外，基于超列的系统还提取了最后的卷积层特征作为LSTM的输入，以获得更长的范围依赖性，而不会丢失全局一致性。 代替LSTM，在[Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields]中使用有效识别人类活动的条件随机域（CRF）模型来区分输入序列的时间关系。

网络集成：用于视频中动作识别的两流CNN，它在多帧密集光学流上训练CNN的一个流以获取时间信息，而在静止图像上训练CNN的另一个流以获取外观特征，然后将输出融合 Simonyan等人介绍了两个流中的一个。 受此体系结构的启发，针对FER提出了几种网络集成模型。

Sun等。 [Deep spatial-temporal feature fusion for facial expression recognition in static images]提出了一种多通道网络，该网络从表情表情的面孔中提取空间信息，并从情绪面孔和中性面孔之间的变化中提取时间信息（光学流），并研究了三种特征融合策略：分数平均融合，基于SVM的融合 和基于神经网络的融合。 张等。 [Facial expression recognition based on deep evolutional spatial-temporal networks]融合了时间网络PHRNN（**在“地标轨迹”中讨论**）和空间网络MSCNN（在第4.1.5节中讨论），以提取FER的部分整体，几何外观和静态-动态信息（见图）. 13）。 荣格等人没有融合不同权重的网络输出。 [Joint fine-tuning in deep neural networks for facial expression recognition]提出了一种联合精细调整方法，该方法联合训练了DTAN（在“ RNN和C3D”中讨论），DTGN（**在“地标轨迹”中讨论**）和集成网络（有关详细信息，请参见图14）， 胜过加权和策略。

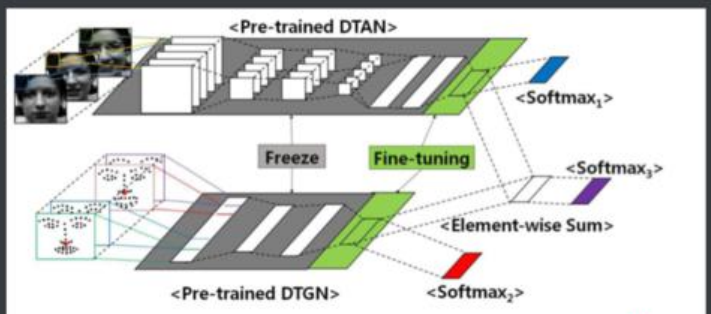


Fig. 14. The joint fine-tuning method for DTAGN proposed in [16]. To integrate DTGA and DTAN, we freeze the weight values in the gray boxes and retrain the top layer in the green boxes. The logit values of the green boxes are used by Softmax3 to supervise the integrated network. During training, we combine three softmax loss functions, and for prediction, we use only Softmax3.

4.2.4 讨论

在现实世界中，人们以动态过程显示面部表情，例如从微妙到明显，并且已成为对序列/视频数据进行FER的趋势。 表8总结了关于动态数据的不同类型方法的相对优点，这些方法在表示空间和时间信息的能力，对训练数据大小和帧长（可变或固定）的要求，计算效率和性能方面具有优势。帧聚合用于组合序列级结果的每个帧的学习特征或预测概率。 可以简单地将每个帧的输出连接起来（每个序列中需要固定长度的帧），也可以统计汇总以获取视频级表示（可变长度的帧可处理）。 该方法计算简单，如果目标数据集的时间变化不复杂，则可以实现中等性能。根据视频序列中表达强度随时间变化的事实，表达强度不变网络考虑具有非峰值表达的图像，并进一步利用峰值和非峰值表达之间的动态相关性来提高性能。 通常，强度不变的FER需要具有特定强度状态的图像帧。

TABLE 8
Comparison of different types of methods for dynamic image sequences in terms of data size requirement, representability of spatial and temporal information, requirement on frame length, performance, and computational efficiency. $\mathcal{FLT}$ = Facial Landmark Trajectory; $\mathcal{CN}$ = Cascaded Network; $\mathcal{NE}$ = Network Ensemble.

Network type		data	spatial	temporal	frame length	accuracy	efficiency
Frame aggregation		low	good	no	depends	fair	high
Expression intensity		fair	good	low	fixed	fair	varies
Spatio-temporal network	RNN	low	low	good	variable	low	fair
	C3D	high	good	fair	fixed	low	fair
	$\mathcal{FLT}$	fair	fair	fair	fixed	low	high
	$\mathcal{CN}$	high	good	good	variable	good	fair
	$\mathcal{NE}$	low	good	good	fixed	good	low

尽管这些方法具有优势，但帧聚合处理帧时无需考虑时间信息和细微的外观变化，并且表达强度不变的网络需要表达强度的先验知识，而这在现实世界中是不可用的。 相比之下，深度时空网络被设计为对连续帧中的时间相关性进行编码，并且已被证明可以从学习空间特征和时间特征中受益。 RNN及其变体（例如LSTM，IRNN和BRNN）和C3D是学习时空特征的基础网络。 但是，这些网络的性能几乎不能令人满意。 RNN无法捕获强大的卷积功能。 C3D中的3D文件应用于非常短的视频剪辑，而忽略了远程动态。 同样，训练如此庞大的网络在计算上也是一个问题，尤其是对于视频数据不足的动态FER。 或者，面部界标轨迹方法基于面部形态变化的物理结构提取形状特征，以捕获动态面部组件活动，然后应用深层网络进行分类。 该方法计算简单，可以消除光照变化的问题。但是，它对套准错误很敏感，需要精确的面部标志检测，很难在不受限制的条件下进行访问。 因此，此方法执行效果较差，并且更适合于补充外观表示形式。 网络集成用于训练多个网络的时空信息，然后在最后阶段融合网络输出。 血流和面部界标轨迹可以用作时间表示，以协作空间表示。 该框架的缺点之一是光学流或界标轨迹矢量的预先计算和存储消耗。 而且大多数相关研究都随机选择了固定长度的视频帧作为输入，从而导致有用的时间信息丢失。 提出了级联网络，以首先提取面部表情图像的判别表示，然后将这些特征输入到顺序网络中，以增强时间信息编码。 然而，该模型引入了附加参数来捕获序列信息，并且当前作品中的特征学习网络（例如，CNN）和时间信息编码网络（例如，LSTM）没有被共同训练，这可能导致参数设置不理想。 以端到端的方式进行培训仍然很漫长。

与基于静态数据的深度网络相比，表4和表7展示了深度时空网络的强大功能和流行趋势。 例如，在广泛评估的基准（例如CK +和MMI）上的比较结果表明，基于序列数据的训练网络并分析帧之间的时间依赖性可以进一步提高性能。 同样，在2015年EmotiW挑战赛中，只有一个系统为FER使用了深空域网络，而2017年EmotiW挑战赛的7个审查系统中有5个依赖于此类网络。

五、其他相关问题

除了上述最流行的基本表达分类任务外，我们还介绍了一些依赖于深度神经网络和原型表达相关知识的相关问题。

5.1 咬合和非额头姿势(Occlusion and non-frontal head pose)

遮挡和非正面姿势可能会改变原始面部表情的视觉外观，这是自动FER的两个主要障碍，尤其是在实际情况下。

对于面部阻塞，Ranzato等人，提出了一个深层的生成模型，该模型使用mPoT作为DBN的第一层来建模像素级表示，然后训练DBN为其输入拟合合适的分布。 因此，通过使用条件分布序列重建顶层表示，可以填充图像中被遮挡的像素。 程等，采用多层RBMs，对Gabor特征进行预训练和微调，以压缩被遮挡的面部部位的特征。 徐等，将从两个具有相同结构但已针对不同数据进行预训练的CNN转移来的高级学习特征串联在一起：原始MSRA-CFW数据库和MSRA-CFW数据库（带有累加样本）。

对于多视图FER，Zhang等人， 在CNN中引入了一个投影层，该投影层通过在2D SIFT特征矩阵内加权不同的面部界标点来学习判别性面部特征，而无需进行面部姿势估计。 刘等，提出了一种多通道姿势感知CNN（MPCNN），其中包含三个级联部分（多通道特征提取，联合多尺度特征融合和姿势感知识别），以通过最大程度地减少条件关节损失来预测表达标签 姿势和表情识别。 此外，在[Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition]，[Joint pose and expression modeling for facial expression recognition]中已经采用了生成对抗网络技术（GAN），以针对多视角FER在任意姿势下生成具有不同表情的面部图像。

5.2 红外数据的FER（FER on infrared data）

尽管RBG或灰度数据是FER中的最新标准，但这些数据容易受到环境光照条件的影响。 同时，记录由情绪产生的皮肤时间分布的红外图像对照明变化不敏感，这可能是研究面部表情的有希望的替代方法。 例如，He等，采用了DBM模型，该模型由高斯二进制RBM和FER的二进制RBM组成。 通过分层预训练和联合训练对模型进行训练，然后在长波长热红外图像上进行微调以学习热特征。 Wu等，提出了一种三流3D CNN，用于融合FER照明不变的近红外图像上的局部和全局时空特征。

5.3 在3D静态和动态数据上进行FER（FER on 3D static and dynamic data）

尽管2D FER取得了显着进步，但它无法解决两个主要问题：照度变化和姿势变化。 使用具有深度信息的3D面部形状模型的3D FER可以捕获微妙的面部变形，这些变形自然对姿势和光照变化具有鲁棒性。

深度图像和视频根据距深度相机的距离记录面部像素的强度，其中包含面部几何关系的关键信息。 例如，[Facial expression recognition using kinect depth sensor and convolutional neural networks]使用kinect深度传感器获得梯度方向信息，然后在未注册的面部深度图像上使用CNN进行FER。 [A facial expression recognition system using robust face features from depth videos and deep learning]，[Facial expression recognition using salient features and convolutional neural network]从深度视频中提取了一系列显着特征，并将其与用于FER的深度网络（即CNN和DBN）结合在一起。 为了强调面部表情运动的动态变形模式，李等人。 [Automatic 4d facial

expression recognition using dynamic geometrical image network]使用动态几何图像网络探索4D FER（使用动态数据的3D FER）。此外，Chang等。[Expnet: Landmark-free, deep, 3d facial expressions]建议使用CNN从图像强度估计3D表达系数，而无需面部标志检测。因此，该模型对于极端的外观变化（包括平面外头部旋转，缩放比例变化和遮挡）具有很高的鲁棒性。

近来，越来越多的作品趋向于将2D和3D数据结合起来以进一步提高性能。Oyedotun等，[Facial expression recognition via joint deep learning of rgb-depth map latent representations]使用CNN从RGB和深度图潜在模式中共同学习面部表情特征。和李等，[Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network]提出了一种深度融合CNN（DF-CNN），以探索多模式2D + 3D FER。具体来说，首先从带纹理的3D面部扫描中提取六种2D面部属性图（即几何，纹理，曲率，法线分量x，y和z），然后将其联合输入特征提取和特征融合子网中 了解2D和3D人脸表示的最佳组合权重。为了改进这项工作，[Accurate facial parts localization and deep learning for 3d facial expression recognition]提出从纹理和深度图像中提取的不同面部部位提取深度特征，然后将这些特征融合在一起以使其与反馈互连。Wei等，[Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition]进一步探讨了使用无监督域自适应技术的2D + 3D FER中的数据偏差问题。

5.4 面部表情合成

现实的面部表情合成可以为交互式界面生成各种面部表情，是一个热门话题。Susskind等。[Generating facial expressions with deep belief nets]证明了DBN能够捕获表现形式上的大范围变化，并且可以在大型但稀疏标记的数据集上进行训练。鉴于这项工作，[On deep generative models with applications to recognition]，[Modeling natural images using gated mrfs]，[A fast and accurate facial expression synthesis system for color face images using face graph and deep belief network]将DBN与无监督学习结合使用来构建面部表情合成系统。Kaneko等，[Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks]提出了一种具有状态识别和关键点定位的多任务深度网络，以自适应地生成视觉反馈以改善面部表情识别。随着深度生成模型的最新成功，例如变分自动编码器（VAE），对抗性自动编码器（AAE）和生成性对抗网络（GAN），基于这些模型开发了一系列面部表情合成系统。面部表情合成也可以应用于数据扩充，而无需手动收集和标记巨大的数据集。Masi等。[Do we really need to collect millions of faces for effective face recognition?]利用CNN通过增加特定于面部的外观变化（例如3D纹理面部模型中的表情）来合成新的面部图像。

5.5 可视化技术（Visualization techniques）

除了将CNN用于FER外，几项工作在学习的CNN特征上采用可视化技术来定性分析CNN如何促进基于外观的学习过程 FER并定性地解密脸部哪一部分产生最有区别的信息。反卷积的结果都表明，在学习到的特征上某些特定滤波器的激活与对应于面部AU的面部区域有很强的相关性。

5.6 其他特殊问题

在原型表达类别的基础上，已经提出了几种新颖的问题：显性和互补性情感识别挑战和“真实与假冒”表达情感挑战。此外，这两个挑战的参与者已经充分应用了深度学习技术。其他相关的现实世界应用程序，例如用于智能手机的实时FER App，Eyemotion（使用眼动相机的FER），保护隐私的移动分析[246]，不舒服的情绪和抑郁症识别，也得到了发展。

六、挑战与机遇

6.1 面部表情数据集

随着FER文献将主要重点转移到充满挑战的野外环境条件下，许多研究人员已致力于采用深度学习技术来处理困难，例如照明变化，遮挡，非额头姿势，身份偏见和识别。 低强度表达。 鉴于FER是数据驱动的任务，并且训练足够深的网络以捕获与表达相关的细微变形需要大量训练数据，因此，深FER系统面临的主要挑战是缺乏训练数据和质量。

由于年龄范围，文化和性别不同的人以不同的方式显示和解释面部表情，因此理想的面部表情数据集应包括具有精确面部属性标签的丰富样本图像，不仅包括表情，还包括年龄，性别和种族等其他属性，这将有助于使用深度学习技术（例如多任务深度网络和转移学习）进行跨年龄范围，跨性别和跨文化FER的相关研究。 此外，尽管遮挡和多姿势问题在深层人脸识别领域引起了相对广泛的关注，但在深度FER中，遮挡鲁棒和姿势不变的问题受到的关注较少。 主要原因之一是缺少具有遮挡类型和头部姿势注释的大规模面部表情数据集。

另一方面，以自然场景的巨大变化和复杂性来准确注释大量图像数据，显然是构建表达式数据集的明显障碍。 合理的方法是在专家注释者的指导下采用众包模型。 另外，由专家改进的全自动标签工具可以替代，提供近似但有效的注释。 在这两种情况下，都必须进行后续的可靠估计或标记学习过程才能过滤出嘈杂的注释。 特别是，考虑到现实情况并包含多种面部表情的比较大规模的数据集最近已公开可用，例如EmotioNet，RAFDB和AffectNet，我们预计，随着技术的进步和互联网的广泛普及，将构建更多互补的面部表情数据集以促进深度FER的发展。

6.2 整合其他情感模型（Incorporating other affective models）

需要考虑的另一个主要问题是，尽管分类模型中的FER得到了广泛的认可和研究，但原型表达的定义仅涵盖了特定类别的一小部分，无法捕获现实交互的全部表现行为。 还开发了另外两个模型来描述更大范围的情感景观：FACS模型和各种尺寸模型，其中各种面部肌肉AU组合在一起来描述面部表情的可见外观变化。 其中两个连续值变量，即价和唤醒，被提出来连续编码情绪强度的小变化。 Du等人提出了另一种新颖的定义，即复合表达。 有些认为某些面部表情实际上是一种以上基本情绪的组合。 这些作品改善了面部表情的特征，并在一定程度上可以补充分类模型。 例如，如上所述，CNN的可视化结果已证明学习的表示与AU定义的面部区域之间存在一定的一致性。 因此，我们可以设计深层神经网络的过滤器，以根据不同的面部肌肉动作部位的重要性分配不同的权重。

6.3 数据集偏差和分布不平衡（Dataset bias and imbalanced distribution）

由于不同的采集条件和注释的主观性，数据偏倚和注释不一致在不同的面部表情数据集中非常普遍。 研究人员通常会在特定数据集中评估他们的算法，并且可以获得令人满意的性能。 但是，早期的跨数据库实验表明，由于收集环境和构建指标的不同，数据库之间存在差异。 因此，通过数据库内协议评估的算法在看不见的测试数据上缺乏通用性，并且跨数据集设置的性能大大降低。 深层适应和知识提炼是解决这一偏见的替代方法。 此外，由于表达注释不一致，当通过直接合并多个数据集来扩大训练数据时，FER性能无法持续提高。

面部表情中的另一个常见问题是班级失衡，这是数据获取的实用性的结果：诱发和注释微笑很容易，但是，捕获有关厌恶，愤怒和其他较不常见表情的信息可能非常具有挑战性。如表4和表7所示，与平均标准相比，按平均准确度评估的性能（对所有类别均分配了相等的权重）下降，并且这种下降在现实数据集中（例如SFEW 2.0和AFEW）尤其明显）。 一种解决方案是使用数据扩充和合成在预处理阶段平衡类的分布。 另一种选择是在培训期

间为深度网络开发一个成本敏感的损失层。

6.4 多峰影响识别（Multimodal affect recognition）

最后但并非最不重要的一点，现实应用中的人类表达行为涉及从不同角度进行编码，而面部表情只是一种形式。 尽管基于可见的面部图像的纯表情识别可以实现令人鼓舞的结果，但是将其与其他模型结合到高级框架中可以提供补充信息，并进一步增强鲁棒性。 例如，EmotiW挑战和音频视频情感挑战（AVEC）的参与者。[Avec 2017: Real-life depression, and affect recognition workshop and challenge]认为音频模型是第二重要的元素，并采用了多种融合技术进行多模态影响识别。 另外，由于面部表情的巨大互补性，诸如红外图像，3D面部模型的深度信息和生理数据等其他模式的融合正成为有前途的研究方向。