文章编号：1003-0530(2023)04-0588-28

情感识别中的迁移学习问题综述

黄兆培[1] 张峰源[1] 赵金明[2] 金 琴[1]
（1. 中国人民大学信息学院，北京 100872；2. 启元实验室，北京 100095）

摘 要： 情感识别是实现自然人机交互的必要过程。然而，情感数据高昂的采集和标注成本成为了限制情感识别研究发展的一大瓶颈。在无标注或有限标注的场景下，利用知识的跨领域或跨任务迁移提升情感识别效果的问题值得探索。本文对情感识别中的迁移学习问题进行了梳理和分析。首先，将迁移学习问题划分为针对领域差异和针对任务差异的两大部分，并进一步将每部分问题细分为多种不同的情况。随后，基于情感识别领域的研究现状，分别总结不同情况下的现有工作。在目标领域训练资源匮乏的情况下，可以利用其他带标注的数据集作为源领域训练模型，并对齐不同领域下的特征分布，或将特征映射到域间共享的空间。考虑到情感标签所提供的监督信息往往较为有限，为了进一步提升模型的识别效果，可以引入其他相关任务进行联合训练，或将预训练模型、外部知识库提供的先验语义知识迁移到情感识别任务中。最后，讨论了情感识别领域中未来需要得到更多关注和探索的迁移学习问题，旨在为研究者带来新的启发。

关键词： 情感识别；迁移学习；深度学习
中图分类号： TP391.4 **文献标识码：** A **DOI：** 10.16798/j.issn.1003-0530.2023.04.002

引用格式： 黄兆培，张峰源，赵金明，等. 情感识别中的迁移学习问题综述[J]. 信号处理，2023，39(4)：588-615. DOI：10.16798/j.issn.1003-0530.2023.04.002.
Reference format：HUANG Zhaopei，ZHANG Fengyuan，ZHAO Jinming，et al. A survey of transfer learning problems in emotion recognition[J]. Journal of Signal Processing，2023，39(4)：588-615. DOI: 10.16798/j.issn.1003-0530.2023.04.002.

A Survey of Transfer Learning Problems in Emotion Recognition

HUANG Zhaopei[1] ZHANG Fengyuan[1] ZHAO Jinming[2] JIN Qin[1]
（1. School of Information，Renmin University of China，Beijing 100872，China；2. Qiyuan Lab，Beijing 100095，China）

Abstract: Emotion recognition is an essential process of natural human-computer interaction. However，the high cost of data collection and labeling has become a bottleneck in the development of emotion recognition research. It is worth exploring how to improve the recognition performance by the cross-domain or cross-task knowledge transfer in scenarios with no or limited annotations. This paper organizes and analyzes transfer learning problems in emotion recognition. Firstly，transfer learning problems are divided into two parts，which are problems for domain discrepancies and for task differences. Each part of them is further subdivided into several different situations. Then，existing works of emotion recognition in different situations are summarized respectively. In the case of scarce training resources in the target domain，other annotated datasets can be used as the source domain to train the model. During this process，the feature distributions from different domains should be aligned，or the features should be mapped to a shared space. Considering that the supervision information provided by emotion annotations is often limited，in order to further improve the recognition performance of the model，other related tasks can be introduced for joint training，or the prior semantic knowledge provided by the pre-training model and external knowledge base can be transferred to the emotion recognition task. Finally，transfer learning problems in the emotion recognition task

收稿日期：2022-11-01；修回日期：2023-01-19
基金项目：国家重点研发计划(2020AAA0108600)；国家自然科学基金(62072462)

which need more attention and exploration in the future are discussed, aiming to bring new inspiration to researchers.

Key words: emotion recognition; transfer learning; deep learning

1　引言

情感识别是情感计算研究当中的重要组成部分，理解人类情感是实现自然人机交互的前提。情感有很多种不同的表达或体现方式。例如，一个人的面部表情、眼部动作、讲话声音、动作姿态以及一些生理上的活动指标都能够直接地反映出他此时此刻的情感状态，我们通常将这些信息称为显式的情感线索。而随着网络和移动设备的发展，人们越来越习惯于在社交媒体上发表自己对于事物的看法或者分享自己当前的状态，一个人发表的言论内容或者他分享出的图片、音视频往往也蕴含了他想要表达的情感，因此我们可以将这些信息称为隐式的情感线索[1-2]。这两种线索中的各模态信息都能被计算机用于人类情感的识别。情感状态的表示方式也并不唯一，通常可以分为离散状态表示和连续状态表示[3]。离散状态表示源于美国心理学家Ekman的“基本情感理论”，即人们的情感可以归为高兴、生气、伤心、厌恶、惊讶、害怕六个类别之一[4]。此后，研究者们又提出了多种不同的分类方式。在当前的情感识别研究中，通常会加入“中性”的情感类别以表示平静的状态。连续状态表示则认为情感状态应被视为多维连续空间中的一个点，例如Mehrabian提出的“情感三维理论”采用“效价”(valence)、“唤醒”(arousal)和“支配”(dominance)三个维度描述情感状态，分别表示情感的愉悦程度、强度和控制程度[5]。在采用计算机进行自动情感识别的任务当中，情感状态的表示方式就对应于数据的标签形式。两种标签形式均广泛存在于当前常用的情感识别数据集中。

当前，已有诸多情感识别模型被提出，它们在一些公开的数据集上展现出了良好的识别效果。然而，情感数据高昂的采集和标注成本成为了限制该领域进一步发展的一大瓶颈。因此，许多研究者开始关注于有限资源情况下的情感识别问题。例如，我们可以利用带标注的情感数据集训练识别模型，并将其迁移至无标注或训练资源匮乏的目标场景中应用。然而，训练数据与目标应用场景中的数据可能存在一定的差异，例如两个集合中的数据属于不同语言、不同模态，或者在分布上存在不一致。如何缓解这种差异带来的负面影响对于跨数据集的迁移问题至关重要。另外，由于情感信息较为抽象，且主观性较强，单纯使用有限规模数据集中的情感标注作为监督信号难以使模型学习到足够的语义知识。如果能够充分利用到现有数据中其他相关任务的标签信息，或是引入在更广泛的数据场景中学习到的通用语义知识，往往有助于增强模型对于目标数据的理解和表示能力，从而提升情感识别效果。因此，这种不同任务之间的知识迁移问题也十分值得探索。

当前已有很多综述文献整理并总结了不同模态上的情感识别工作，包括视觉模态的表情识别[6]、语音模态的情感识别[7-9]、利用生理信号的情感识别[10-12]以及文本模态的情感识别[13-14]等。也有一些综述文献系统地整理并总结了多个模态的数据集、常用特征和识别方法，同时针对情感数据的多模态融合问题做了总结或讨论[15-17]。与上述工作不同的是，本文重点关注于情感识别任务当中的迁移学习问题。Feng等人[18]的综述与本文最为接近，该综述回顾了不同模态的自动情感识别任务中的迁移学习工作，包括语音、人脸和生理信号模态。然而，该工作并未包含利用隐式情感线索信息进行识别的工作(如对于文本评论所蕴含情感的识别)，且涉及的内容仅局限于领域自适应相关的问题和方法。本文则同时涵盖了显示和隐式情感线索的识别工作，并关注于更广泛的迁移学习问题，以便进一步推动情感识别领域的发展。

本文后续内容安排如下：第2章基于过往工作中对于迁移学习相关概念的定义，给出了迁移学习问题的分类以及每类问题在情感识别任务上的具体体现；第3章按照分类分别介绍不同类型的情感识别迁移学习问题以及方法；第4章对情感识别任务上未来值得关注的一些迁移学习问题进行了展望；第5章对本文内容进行了总结。

2 迁移学习问题的概念及分类

迁移学习是机器学习中的前沿研究方向之一。在传统的机器学习问题设定下，训练样本与测试样本所属的领域一般是相同的，训练过程和推理过程所对应的任务一般也相同。即我们认为，训练阶段与测试阶段中样本变量$X(X=\{x_1,\cdots,x_n\})$与其对应的标签变量$Y(Y=\{y_1,\cdots,y_n\})$的联合概率分布$P(X,Y)$是相同的。而迁移学习则放宽了这样的限制，使得在某个领域或任务上学习到的知识或模型可以应用于不同但相关的领域或任务当中[19]。对于“领域”和“任务”的概念，Pan等人[20]给出了规范的定义。“领域”被定义为：$\mathcal{D}=\{\mathcal{X},P(X)\}$。其中，$\mathcal{X}$表示输入样本的集合$X$所处的特征空间，即$X\in\mathcal{X}$；而$P(X)$则表示输入样本集合$X$自身的概率分布，也即数据集中样本和标签共同构成的二维随机变量(X,Y)关于样本X的边缘概率分布[1]。任务的定义为：$\mathcal{T}=\{\mathcal{Y},f(\cdot)\}$。其中，$\mathcal{Y}$表示标签集合所处的空间，即$Y\in\mathcal{Y}$；$f(\cdot)$则表示特征到标签的映射关系，也可表示为条件概率分布$P(Y|X)$。迁移学习问题包含这样的两类情况：训练数据所对应的源领域$\mathcal{D}_S$和测试数据对应的目标领域$\mathcal{D}_T$存在差异，即$\mathcal{D}_S\neq\mathcal{D}_T$；或者源任务$\mathcal{T}_S$和目标任务$\mathcal{T}_T$不同，即$\mathcal{T}_S\neq\mathcal{T}_T$。我们可以将前者称为“针对领域差异的迁移学习问题”，而后者可以被称为“针对任务差异的迁移学习问题”。图1展示了迁移学习问题的示意图，其核心在于从源领域或源任务上学习到的知识向目标领域或目标任务迁移的过程。

从上面的定义中可以看出，源领域和目标领域的差异又可以被分为两种情况。如果两个领域的特征空间不一致，即$\mathcal{X}_S\neq\mathcal{X}_T$，则可以称为“特征空间差异问题”。例如，源领域和目标领域的数据来源于不同的模态，因此两个域的特征分别处于不同模态对应的表示空间中；或者两个域的数据虽然都来自于文本模态，但是属于不同语言，于是两个域的特征分别处于不同语言对应的嵌入空间中。另一种情况是两个领域的特征空间相同而特征的分布存在差异，$\mathcal{X}_S=\mathcal{X}_T$且$P(X_S)\neq P(X_T)$，我们可以称其为“特征分布差异问题”。例如，两个域的数据都是同一种语言的评论文本，但分别指向不同的主题。我们一般通过领域自适应解决该类问题。

在当前情感识别领域的工作中，源任务和目标任务不同的情况大多体现为$\mathcal{Y}_S\neq\mathcal{Y}_T$且$f_S\neq f_T$，因此我们在第3章中主要针对这种情况展开介绍。我们可以进一步将这种情况分为“多任务联合训练中的迁移问题”以及“预训练模型或知识库的迁移问题”两种类型。对于前者的问题类型，源任务的标签信息一般与情感信息存在潜在的关联，将两种相关的任务联合训练往往能够为情感识别任务提供额外的线索。而对于后者，源任务则一般是作为通用语义知识的学习任务，使模型能够先通过源任务的训练学习到一定的语义理解能力，从而为后续情感识别任务的训练和推理提供良好的基础。此外，$\mathcal{Y}_S\neq\mathcal{Y}_T$且$f_S=f_T$的情况意味着两个集合的特征空间

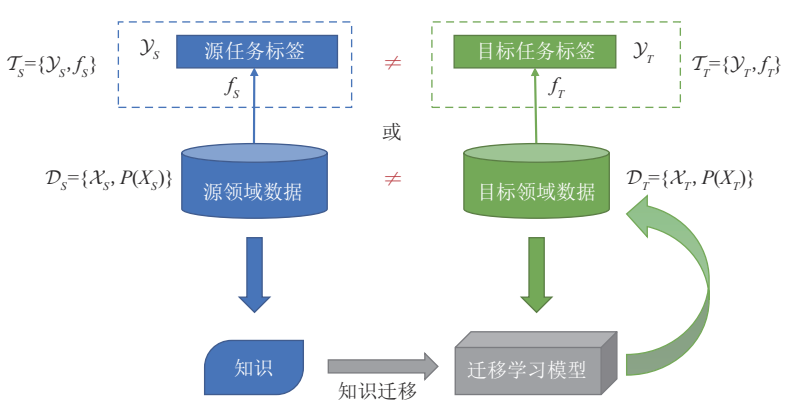

图1 迁移学习问题示意图

Fig. 1 Illustration of the Transfer Learning Problem

[1]另有些迁移学习相关工作将“领域”定义为样本与标签的联合分布$P(X,Y)$，因此将$P(X)$称为“领域的边缘分布”。为避免概念的混淆，本文中将不再采用“边缘分布”的说法，直接称$P(X)$为数据或特征的分布。

到标签空间的映射关系是相同的，差异仅体现在标签空间上。我们可以将这种特殊的情况理解为训练集所包含的标签类别和测试集所包含的标签类别有所差异，例如训练集中的数据只来自于六个情感类别，而测试集中则包含额外的两个情感类别中的数据。我们将在第4.3节中对于这种零样本类别的识别问题进行简单的讨论。而$\mathcal{Y}_S = \mathcal{Y}_T$且$f_S \neq f_T$的情况则意味着两个集合中的标签空间完全相同，但从特征到标签的映射过程存在差异。例如，当我们将一个训练好的情感识别模型应用到实际场景中时，该模型基于原先的映射所输出的特征虽然可以在情感的识别中展现出良好的表现，但也容易被用于一些敏感信息的预测。为了避免隐私的泄露，我们可以在应用时为输入特征加入额外的噪声映射，去除特征中与敏感信息相关的成分，实现隐私的保护。我们将在4.4节的内容中提及该问题。

我们基于Pan等人[20]所给出的定义，将情感识别中的迁移学习问题划分为不同的类型，如图2所示。对于已经在情感识别领域得到较多探索的问题，我们将在第3章中分别介绍；而对于一些尚未在情感方面得到较多关注的问题，我们则将在第4章中进行简单的讨论。

3　情感识别中的迁移学习问题

本章中，我们将主要介绍目前已在情感识别领域得到较多探索的问题。我们根据迁移学习的定义，将这些问题划分为针对领域差异和针对任务差异的两大部分，并进一步将每部分问题细分为多种不同的情况。我们将逐一介绍每种情况对应的情感识别相关工作，并重点关注于不同工作中实现知识跨领域或跨任务迁移的方法。

3.1　针对领域差异的迁移学习问题

在考虑情感识别的应用问题时，我们经常会遇到目标领域的数据集无情感标注，或训练资源极为匮乏的情况。此时我们不得不利用其他资源更充足的领域训练模型，并将模型迁移至目标领域的数据上进行应用。但源领域和目标领域之间的差异往往会为模型在不同领域之间的迁移带来困难。根据第2章中所介绍的对于“领域”的定义，不同领域之间存在的差异可以分为特征分布的差异以及特征空间的差异两种情况。在本节内容中，我们将分别介绍这两种差异的具体概念，以及每种差异下的多种与情感识别任务相关的具体问题。

3.1.1　特征分布差异问题

当我们想要对一个无标注集合中的数据进行情感识别时，我们往往需要先利用其他带标注的情感数据集训练出识别模型。然而，一般数据集内的数据只是在有限的甚至是单一的场景下收集，特别是对于显式情感线索的数据来说，同一个数据集收集过程中的受试者数量往往也是极为有限的。这

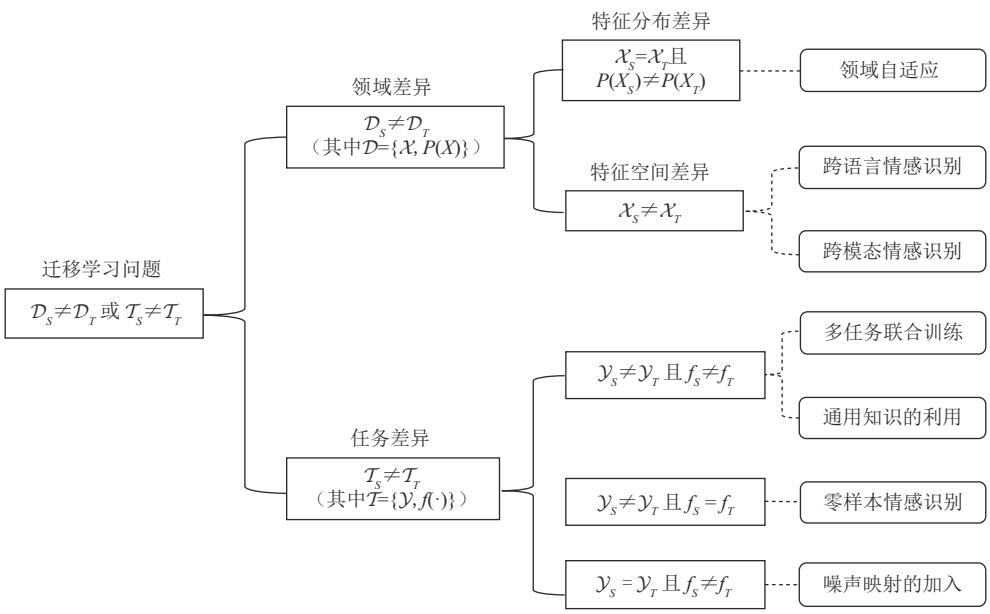

图2　情感识别中的迁移学习问题总结

Fig. 2　Summary of Transfer Learning Problems in Emotion Recognition

就导致不同数据集之间存在着很多由场景的不同、受试者的差异等因素所带来的数据分布差异，限制了模型迁移后的表现。针对该问题，我们往往需要从数据中捕获到更多与领域无关的特征成分，从而对齐两个领域之间情感特征的分布，使得在源领域上训练出的模型可以更有效地应用于目标领域数据的识别任务中。图3为特征分布差异问题的示意图。本节将依次介绍在人脸表情、语音、脑电信号、文本四方面的情感识别任务中解决域间特征分布差异问题的工作。

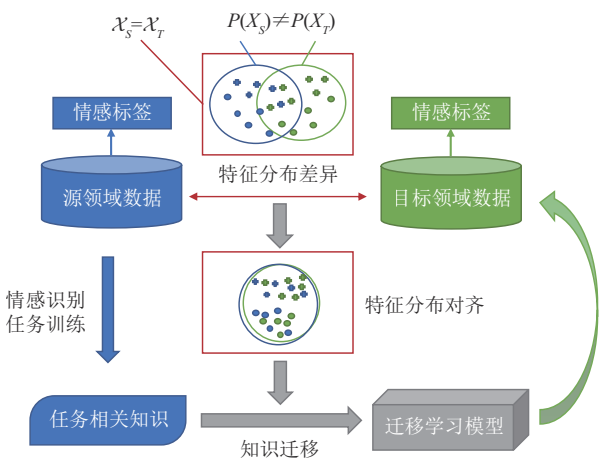

图3 特征分布差异问题示意图

Fig. 3 Illustration of Feature Distribution Differences Problems

3.1.1.1 人脸表情特征分布差异问题

人脸表情识别在人机交互、情感计算等任务中承担着关键角色。在实际应用中，人脸表情识别模型的泛化能力是决定下游任务性能的关键因素之一。然而，各人脸表情数据集的数据分布存在较大的差异，一定程度上阻碍了模型泛化能力的提升。目前的人脸表情识别工作大多假设训练数据与测试数据从属同一分布，然而这在实际应用中并不成立。为了解决该问题，跨领域人脸表情识别问题得到了广泛关注。

现有的表情数据集可根据数据采集场景分为实验室场景和自然场景两类。实验室场景的数据集CK+[21]、JAFFE[22]、MMI[23]等规模较小，数据分布较为一致，且拍摄参数的变化较小。对于这类数据，利用手工设计的特征就能够取得较好的识别效果。然而，对于自然场景下的大型表情数据集RAF-DB[24]、AffectNet[25]、FERPlus[26]，由于采集过程中不同数据在拍摄角度、镜头参数、曝光时间等方面均存在较大差异，使用手工设计的特征进行识别的效果较差。对此，许多工作[27-28]提出利用深度卷积神经网络（Deep Convolutional Neural Networks，DCNN）自动提取表情特征，这些特征在两种场景中都能展现出较好的识别效果，因此具备更强的可迁移性。Li等人[29]基于此探究了将注意力机制与DCNN结合的方法，从与情感更相关的人脸区域中提取深度特征，降低对于背景区域的关注，从而进一步提升情感特征的迁移能力，在跨数据集的设定下取得了不错的识别效果。

增强不同类别特征在空间中的可区分性同样有利于特征的跨领域迁移。Li等人[30]在使用DCNN提取特征的基础上添加了局部性保持损失，在最大化不同类特征之间距离的同时拉近相同类内特征的距离，提高深度特征的可判别性。具体来说，令x_i为源领域中的随机样本，该文试图计算与x_i最近的k个样本的距离之和，并令其最小化。实验表明，相比直接采用DCNN提取的特征，他们的方法能够取得更好的跨领域识别效果。Zhu等人[31]则将应用于源域数据上的区分性损失与同时作用于两个领域上的领域匹配损失相结合，在对齐两个领域整体特征分布的同时增加特征的类间区分性。

然而，仅增强特征的类间区分性无法保证不同域的类别中心能在特征空间中对齐。针对该问题，Ji等人[32]提出了ICID模型，其中ID channel用于增强特征的类间判别性，而IC channel则用于拉近不同域间同类特征之间的距离，从而在增强特征类间区分性的同时实现逐类别的领域特征分布对齐。Li等人[33]提出了语义度量学习（Semantic Metric Learning）的方法实现类似的目的。在目标域无监督的设定下，他们为目标域数据分配伪标签，并分别为两个领域中的每个类别构建了对应的类别中心。在此基础上，他们一方面在每个领域内拉近每个特征样本与对应的类别中心的距离，推远与其他类别中心的距离；另一方面拉近两个领域间每一对同类的类别中心在空间中的距离。此外，他们还为每个特征样本分配了领域标签，通过最小化特征与领域标签的互信息促进域间整体特征分布的对齐。Chen等人[34]则将类别信息显式注入了用于建模特征表示的图结构。具体来说，他们在训练过程中不断利

用K均值聚类算法(K-means)对每个领域中的样本进行聚类，并采用移动平均的方式动态更新每个聚类中心。对于每个源域的输入样本，他们找到空间中与其距离最近的目标域聚类中心，分别使用该源域样本特征以及该目标域聚类中心表示初始化两个域对应的图结点；对于每个目标域的输入样本也找到其对应的源域聚类中心并进行类似的初始化过程。在随后使用图卷积网络建模整体特征和局部特征关联的过程中，不同域间每个类别层面的特征分布也能够被进一步对齐。

表1对人脸表情特征分布差异相关工作及采取的方法进行了总结。

3.1.1.2　语音特征分布差异问题

语音情感识别旨在通过人类说话所产生的音频信号识别人类的情感状态。传统的语音情感识别问题通常假设训练数据与测试数据服从同一分布。然而，在实际应用中，训练的数据和测试的数据可能来自两个完全不同的集合。在实际的数据集采集过程中，采集的设备、采集音频的环境等均可能在一定程度上影响语音信号的分布，且不同数据集的采集场景可能也具有很大的差异，导致在某一数据集上训练得到的语音情感识别模型可能难以适应其他数据集的情况。

在语音情感识别任务当中，多模态情感数据集IEMOCAP[35]和MSP-IMPROV[36]中包含的语音数据得到了较为广泛的使用。IEMOCAP包含10名演员开展的5段会话的视频和音频记录，其中对话的形式分为即兴创作和按照剧本表演两种。MSP-IMPROV则包含12名演员进行的6段会话。该数据集采集的过程中会为每段会话预先定义好一个目标句，要求演员分别用不同的情感演绎出包含目标句的情景。收集到的数据被划分为4个子集，分别为：即兴表演中关于目标句的记录，即兴表演中关于非目标句的记录，在非表演时段所记录的演员间的自然交互，以及演员以指定情感阅读目标句的记录。此外，只包含语音单模态数据的MSP-Podcast数据集[37]也经常在相关任务上被研究者使用。与上述两数据集不同的是，该数据集并未邀请演员表演特定的情感情景，而是直接利用互联网上的播客(podcasts)记录作为候选数据，因此更接近于真实场景下人的情感表达状态。由此可见，不同数据集在采集场景以及设置方面均存在很大差异。因此跨数据集的语音情感识别问题具有较大的挑战，也得到了较多研究者的关注。

现有的许多跨数据集语音情感识别方法基于一些领域分布差异的度量实现分布对齐。例如，Liu等人[38]基于最大均值差异(Maximum Mean Discrepancy，MMD)的领域分布度量构建损失函数，将两个领域特征映射到再生核希尔伯特空间(Reproducing Kernel Hilbert Space，RKHS)，实现特征分布的对齐。Zhang等人[39]则同时考虑到域间整体分布的对齐以及逐类别层面分布的对齐，在对齐时保留了更多类别层面的语义信息，从而达到了更好的迁移效果。

除此以外，受生成对抗网络(Generative Adversarial Network，GAN)启发，也有方法采用深度对抗迁移解决跨数据集语音情感识别问题。Sahu等人[40]利用实验验证对抗训练过程对提升跨数据集语音情感识别的作用。具体来说，该文研究了两种训练过程：对抗训练和虚拟对抗训练，并在IEMOCAP数据集进行k折交叉验证来阐述对抗训练过程的有效性。Fu等人[41]提出了基于自动编码器的对抗分类器，通过对抗训练过程对齐隐空间的特征分布并平滑分类边界。除此以外，该文章还采用了多示例学习(multi-instance learning)方法将语音信号多段划分，以此学到情感表达最显著的时刻。Latif等人[42]认为现有大多数工作只关注了较小语料库差异时

表1　人脸表情特征分布差异问题工作

Tab. 1　Works to Solve Discrepancies of Facial Expression Feature Distribution

方法分类	简介
可迁移特征选取	使用DCNN提取可迁移性更强的深度表情特征[29-30]。
源域类间区分性增强	利用源域的类别信息，在源域数据上计算区分性损失，在空间中拉近同类特征并推远不同类特征[30-31]。
类别层面跨域分布对齐	同时在两个域内增强各自特征的类间区分性，并拉近不同域间相同类别样本或类别中心的距离[32-34]。

的语音情感识别问题，而并未解决语料库差异较大时的语音情感识别问题。因此，该文提出了三方对抗博弈的模型来学到域不变的特征表示，并同时结合自监督方法，利用无监督数据进行自监督预训练来辅助迁移。具体来说，该文首先提出对抗双判别器网络（Adversarial Dual Discriminator network，ADDi），利用编码器对原始样本进行编码，并用生成器作为解码器将中间特征重构成输入样本，并将重构的样本与原始样本一同输入两个判别器中，利用对抗损失函数同时训练生成器与两个判别器。为了给下游任务提供更有效的监督信号，该文还对模型结构进行改进，提出自监督对抗双判别网络，将生成情感语音数据作为前置任务，为下游任务提供可判别的特征表示。作者将该模型与其他对抗域适应模型进行对比，并通过消融实验验证了该文提出的对抗双判别器网络在不同的跨语料库任务上都取得了更好的效果。

表2对语音特征分布差异相关工作及采取的方法进行了总结。

3.1.1.3 脑电信号特征分布差异问题

基于显式情感线索的情感识别大多是利用表情、语音等模态的信息识别人的情感状态。然而，这些识别方法易受地域、文化、语言的影响，对不同文化背景受试者采集的数据可能差别较大。而相比其他模态，脑电信号（Electroencephalogram，EEG）有更强的可靠性，同时蕴含更为丰富的情感信息。因此基于脑电信号的情感识别也越来越多地得到了研究者的关注。

在脑电相关研究中，由Zheng等人[43]公开发布的基于脑电信号的情感计算数据集SEED（SJTU Emotion EEG Dataset）数据集得到了广泛的应用。该数据集共采集了15名被试的数据，包括8名男性和7名女性。在数据采集过程中，采集者通过为被试者提供中文电影片段诱发被试者产生情感。而由Koelstra等人[44]构建的数据集DEAP（Database for Emotion Analysis using Physiological Signals）则采用了音乐视频作为情感诱发材料。在该数据集的构建过程中共有32名被试者受邀参与了实验，包括16名男性和16名女性。由于采集成本高昂，现有的脑电情感数据集普遍规模较小。另外，不同受试者之间的脑电信号往往具有较大的差异。因此，跨领域的脑电情感识别问题具有较大的研究意义。

现有的跨领域脑电情感识别方法大多通过统计方法处理特征以完成迁移。其中，绝大多数方法均利用最大均值差异（Maximum Mean Discrepancy，MMD）实现特征分布度量与对齐。Bethge等人[45]利用MMD为每个领域分别构造一个编码器，以解决脑电情感识别中的多源域迁移问题。Ju等人[46]提出了基于联邦学习结构的脑电分类算法，并在算法中主要采用MMD将不同域数据映射到同一特征空间中。Chen等人[47]通过MMD构造多源域特征分布迁移方法，文中对MMD损失项的消融实验也进一步验证了其迁移性能。其他一些方法采用了主成分分析[48]（Principal Components Analysis，PCA）及递归特征消除[49]（Recursive feature elimination，RFE）等，也取得了较好的迁移效果。

与跨数据集语音情感识别类似，另一些方法也采用生成对抗网络的思想解决跨领域脑电情感识别的问题。大多数已有的脑电情感识别模型均基于Ganin等人[50]提出的DANN对抗迁移方法，该方法利用对抗网络对齐源域和目标域特征分布以实现迁移。Özdenizci等人[51]直接运用DANN提取脑电数据的域不变特征。He等人[52]将对抗迁移模型ADDA[53]直接运用到跨领域脑电情感识别问题中，验证了深度对抗迁移方法的有效性。Rayatdoost等人[54]结合受试人无关特征表示及DANN，同时最小化情感识别损失及最大化受试者混淆损失，有效提升了迁移效果。

表2 语音特征分布差异问题相关工作

Tab. 2 Works to Solve Discrepancies of Speech Feature Distribution

方法分类	简介
MMD及其变体	通过设计类似MMD的结构将源域和目标域样本特征映射到RKHS高维空间，在高维空间中度量并减小源域和目标域间距离[38-39]。
GAN及其变体	在GAN的基础上进一步提出新的对抗网络结构，利用生成器提取源域和目标域的中间表示，通过逐渐欺骗域判别器完成源域和目标域整体分布的对齐[40-42]。

表3对脑电信号特征分布差异相关工作及采取的方法进行了总结。

3.1.1.4　文本特征分布差异问题

文本情感分类任务是自然语言处理领域的一大经典任务，其目的是分析出一段文本内容所传达的情感倾向，从而识别出用户在发表该言论时的情感状态。我们在互联网上能够获取到大量表达观点或态度的文本内容，例如微博、推特等社交媒体上用户发表的博文以及各类商品评价网站上的评论文本。然而，对于来源于不同领域的文本来说，最能体现出情感的词语或表述方式往往也存在着一定的区别。例如，我们常用“真实”、“动人”等词汇表达我们对一部电影的肯定，但是在对于餐馆的评价当中则几乎不会出现这类用词，即不同领域的文本中词汇的分布情况可能存在着明显的差异。因此，很多研究者关注到了跨域文本情感分类任务，希望能够将在带情感标注的源域上训练出的模型有效地迁移到无标注的目标域上。

在跨域文本情感分类任务中，最为常用的基准数据集是Blitzer等人[55]构建的亚马逊产品评论数据集，其中共包含对于4种产品类型的评论：书本、数字化视频光盘(DVD)、电子设备以及厨房用具，每种产品类型下共有1000条标注为正向情感的评论和1000条标注为负向情感的评论。在对迁移学习模型进行评测的实验中，通常将其中的某一种类型产品的评论数据作为源领域，将另一种类型产品的评论数据作为目标领域，由此可构建出12组跨域情感分类任务。与之类似地，He等人[56]也利用亚马逊评论数据构建了一个新的跨域文本情感分类数据集，共包含对于书本、电子设备、化妆品以及音乐4种类型产品的评论数据，每类产品中共含有正向、负向、中性情感标注的数据各2000条。除此之外，电影评论数据集IMDb[2]、SST[57]，商户评论数据集Yelp[3]以及亚马逊网站中对于其他类型产品的评论数据也经常被用于跨域文本情感分类任务的训练和评测。

Blitzer等人[55]首次将结构对应学习(Structural Correspondence Learning, SCL)应用到情感分类任务中，提出了基于枢轴词(pivot words)的迁移方法。他们找出一些同时在源域和目标域中具有较高出现频率的单词，并在源域数据中计算每个高频出现单词与情感标签的互信息，从而挑选出与源域情感标签高度相关的高频词，并将这样的单词定义为枢轴词。随后，他们以无监督的形式同时利用两个域的数据，在文本输入中移除枢轴词特征，对于每个枢轴词，都进行二分类任务预测其是否原本在当前输入语句中出现，从而为每个枢轴词训练一个对应的线性分类器。接下来，对每个线性分类器的权重做奇异值分解(Singular Value Decomposition, SVD)，处理后的权重矩阵就可以将两个领域的文本特征映射到共享的分布上，从而可以直接利用源域的带标注数据训练用于目标域数据的情感分类器。随着深度学习的发展，也有一些研究者尝试将经典的结构对应学习思想与神经网络模型相结合，在跨域文本分类任务上取得了较好的效果。Yu等人[58]在获取到两个域的高频词后，计算每个词与源域正向和负向标签的加权对数似然比(Weighted Logarithm Likelihood Ratio, WLLR)，分别筛选出对应于正向情感和负向情感的枢轴词表。接下来，他们设计了两个辅助任务，使用基于遮蔽枢轴词后的输入语句抽取出的特征表示，分别预测原始语句中是否包含至少一个正向枢轴词以及至少一个负向枢轴词。同时，他们将基于原始输入语句抽出的特征表示与用于辅助任务的特征表示拼接，用于预测情感类别，并将该情感分类任务与上述两个辅助任务联合训练。

表3　脑电信号特征分布差异问题相关工作

Tab. 3　Works to Solve Discrepancies of EEG Feature Distribution

方法分类	简介
统计学习方法	大多数工作通过设计类似MMD的结构将源域和目标域样本特征映射到RKHS高维空间进行对齐，另有一些工作采用主成分分析及递归特征消除等方法也取得了不错的效果[45-49]。
对抗学习方法	在GAN的基础上进一步提出新的对抗网络结构，利用生成器提取源域和目标域的中间表示，通过逐渐欺骗域判别器完成源域和目标域整体分布的对齐[50-52, 54]。

[2]https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[3]https://www.kaggle.com/yelp-dataset/yelp-dataset

Li 等人[59]则基于 Transformer 模型学习迁移性较强的枢轴词，他们认为选出的枢轴词对应的特征表示不仅要满足与源域情感标签高度相关的条件，还应该与领域本身无关，以至于可以混淆领域判别器对其所属领域的预测。于是他们设计了针对枢轴词的遮蔽单词预测任务，以较大概率遮蔽其中的枢轴词，以较小的概率遮蔽其他单词，将遮蔽后的语句表示送入 Transformer 编码，并对每个遮蔽的枢轴词对应的隐藏状态应用了领域判别器。他们利用 Kendall[60]等人提出的贝叶斯不确定性估计方法计算判别器预测的不确定程度，并将其与互信息相结合，作为枢轴词筛选过程的优化目标，通过模型的训练不断更新枢轴词表。在词表大小小于设定的阈值后，他们认为模型已具备学习到可迁移特征的能力，于是接入分类器，采用源域带标注数据进行情感分类任务的训练。实验表明，不确定性估计方法的引入可以帮助去除一些和领域信息相关性较高的枢轴词，从而帮助提升跨域文本分类的效果。

基于枢轴词的方法高度依赖于枢轴词的选取质量，然而基于确定规则或度量的选取方法往往较为繁琐，且准确性难以得到保证。因此也有许多研究者引入常用的领域自适应方法解决文本模态的特征分布差异问题。Li 等人[61]引入了 DANN[50]中提出的领域判别器和梯度反转层结构，构建了基于对抗训练的记忆网络 AMN。他们将一段文本输入中的所有单词堆叠进一个外部存储块中，并构建了一个查询向量，在模型的编码器部分基于该查询向量对存储块中的所有单词计算注意力权重，并进行加权求和。将注意力加权后的结果与查询向量经过线性映射后的结果相结合，作为下一次注意力加权的查询向量，重复多次后将最后一次注意力加权的结果作为模型编码的文本特征表示。随后，将带标签的源域数据抽取出的特征表示送入情感分类器中，同时将两个域所有输入数据中抽取出的特征表示送入领域判别器中，将领域判别器回传的梯度经过梯度反转层后与情感分类器回传的梯度结合进行联合优化。经过这样的训练过程，查询向量能够为输入语句中与情感相关度较高而与特定领域相关性较低的单词赋予更大的权重，为其他单词赋予较小的权重，实质上是以自动的方式实现了枢轴词的选取及利用，从而学习到了适用于情感分类任务的领域无关语句级别特征表示。He 等人[56]则基于 Zhuang 等人[62]提出的对称 KL 散度显式地最小化源域与目标域的实例在特征嵌入空间中的距离。在对齐两个域间的整体的特征分布后，为了使基于源域带标注数据训练出的分类器对目标域数据也能具有较强的区分能力，他们基于半监督学习中的方法，最小化模型对于所有目标域数据预测结果的熵值；同时还在两个域的数据上加入了自集成的优化过程，以进一步提升分类器预测的可靠程度。

随着预训练语言模型的发展，一些研究者开始关注于利用预训练模型在海量语料上获取到的通用语言知识解决下游任务上不同领域之间的特征分布差异问题。Ye 等人[63]利用预训练语言模型编码下游源域和目标域数据的输入，以得到具备域间可迁移性的文本特征。他们认为，预训练模型的中间层特征往往比高层特征具备更强的可迁移性，因此他们同时提取出预训练语言模型的最后 N 层输出送入一个由前馈神经网络和注意力机制层构成的特征迁移模块中，对 N 层的输出做注意力加权聚合，作为用于下游跨域情感分类任务的特征。为了进一步加强模型对目标域数据的区分能力，他们为目标域数据生成了伪标注，并将筛选出的高质量伪标目标域数据加入到分类任务的训练中。为了避免伪标签的噪声给训练过程带来负面影响，他们还将预训练模型的输出与特征迁移模块的输出之间的互信息作为监督信号，在目标域数据集中进行互信息最大化的蒸馏训练，使得最终用于目标域分类任务的特征能够最大程度地保留预训练模型中包含的语义信息。Du 等人[64]使用 BERT 预训练模型[65]在下游的情感分类数据上进行了后训练（post-training）过程。他们设计了领域区分任务和目标域遮蔽语言建模任务，前者同时使用两个域的数据，以 50% 的概率从两个域中各取一句话拼接在一起，另外 50% 的概率下从目标域取出两句话拼在一起，要求模型判断拼接后的两句话是否来源于同一个域；后者则是只利用目标域数据，延续 BERT 中的 MLM 任务。经过后训练的阶段，模型更加适应于下游任务中评论类型文本的分布，且学习到了区分下游数据领域的能力，有助于随后的微调阶段采用对抗训练的方式进一步对齐域间特征分布。而 Wu 等人[66]则将最近提出的软提示词（soft prompt）方

法[67]与对抗训练的过程结合，迫使领域相关的信息更多体现在软提示词向量中，从而减小领域差异对于负责输出预测结果的遮蔽token位置特征表示的影响。

表4对文本特征分布差异相关工作及采取的方法进行了总结。

3.1.2　特征空间差异问题

在3.1.1节所述的工作中，两个领域上的原始输入虽然存在分布差异，但是总体上都是处于同一特征空间中，因此直接对齐空间中两个领域对应的特征分布便可实现高质量的迁移。然而在有些情况下，不同领域中的原始输入会处于不同的特征空间中，例如属于不同语言的两段文本或来自不同模态的两段信息。由于在源域上训练的模型学习到的是源域的特征空间到标签空间的映射关系，在源域和目标域原始特征空间不同的情况下，模型无法直接应用于目标域数据的预测任务。因此我们必须首先将两个域的原始特征映射到一个共同的特征空间中。随后，利用源领域资源训练出的模型才能够被有效地迁移到目标领域数据上。图4为特征空间差异问题的示意图。

我们依据具体的应用场景将特征空间差异问题划分为跨语言及跨模态的情感识别两类问题。针对两类问题的一些解决方法有一定相似之处：例如可以利用现有的技术将一个领域的原始特征直接转换到另一个领域的特征空间中，或者可以利用两个领域上的平行数据学习出跨领域的特征空间。特别地，得益于自然语言处理领域研究的发展，我们还可以直接利用跨语言或多语言预训练资源为共享特征空间的学习提供良好的基础；但在跨模态情感识别问题上尚缺乏可利用的预训练资源。以下将分别介绍跨语言以及跨模态情感识别问题下的具体方法。

3.1.2.1　不同语言的特征空间差异问题

对于文本模态的信息来说，即使是含义相同的一句话，如果采用不同的语言表述，原始的输入信息也会处于不同的特征空间之中。如果我们希望构建出一个能够在目标语料上表现优异的模型，最好的方式是直接利用与目标语料相同语言的数据训练该模型。然而，不同语言对应的训练资源规模存在着巨大的差异，对于一些小众的语言来说，可能很难找到对应的大规模语料，甚至可能难以找到任何带标注的文本数据。因此在实际场景下，我们经常会利用在资源丰富的语言上训练出的模型帮助解决资源匮乏的语言上的各类问题。而如何解决跨语言输入信息的特征空间差异就成为了影响低资源语言场景下模型效果的关键因素。图5展示了跨语言情感识别方法的发展脉络。

早期的跨语言工作一般是直接利用机器翻译模型将其中一种语言的输入转换成另一种语言，通过这样的转换，源域和目标域的语料在送入分类器时便可统一成一种语言，使得利用源域数据训练好的分类器能够直接应用于目标域的数据上。Balahur等人[68]提出的跨语言情感识别模型将源语言中带情感标注的数据翻译成目标语言，并使用翻译后的结果以及对应的原标签训练适用于目标语言的情感分类器。而Hajmohammadi等人[69]提出的方法则与之相反，将无标注的目标语言翻译成源语言，再利用在源语言语料上训练好的分类器预测情感。这种直接对输入数据进行翻译的方式虽然简单直接，但会极大地受到机器翻译系统本身翻译错误的限制，且翻译的过程往往并不能解决两种语言之间

表4　文本特征分布差异问题相关工作

Tab. 4　Works to Solve Discrepancies of Text Feature Distribution

方法分类	简介
基于枢轴词	基于规则或度量选取出两个域中出现频率较高，且与源域情感标签高度相关的枢轴词，利用枢轴词得到具备可迁移性的文本情感表示[55, 58-60]。
基于对抗训练	利用领域判别器，自动提取出句子内与领域无关的成分[61, 64, 66]。
基于分布差异度量	基于对称KL散度显式最小化不同域间分布的距离[56]。
基于预训练模型	利用预训练模型的通用语言知识。利用预训练模型中间层特征较强的可迁移性[63]；利用后训练过程使模型更适应下游目标领域文本[64]；利用可学习的软提示词向量削弱要预测的情感词位置特征表示的域间差异[66]。

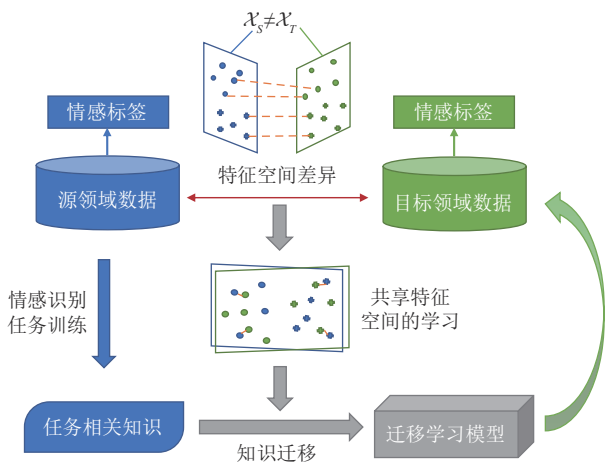

图4 特征空间差异问题示意图

Fig. 4 Illustration of Feature Space Differences Problems

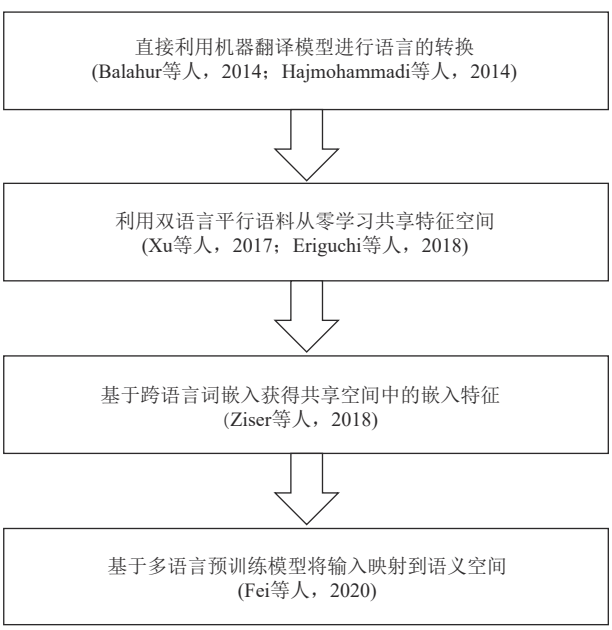

图5 文本模态跨语言情感识别方法发展脉络图

Fig. 5 The Development Diagram of Cross-lingual Emotion Recognition Method in Text Modality

句法结构或语义概念分布上的差异。

也有一些研究者关注于利用双语言的平行语料数据训练模型，使模型能够将源语言和目标语言分别映射到共享的特征空间当中。Eriguchi 等人[70]使用平行语料，训练了一个编码器-解码器结构的多语言机器翻译模型，其中编码器是对于不同语言共享的，而解码器部分则是分别针对不同语言采用不同的分支。在训练好之后，去掉该模型的解码器部分，只保留多语言的编码器并将其冻结，用于将不同语言的输入映射到共享空间中。同时，在编码器后接入情感分类器，采用源语言中的带标注数据训练分类器，训练好的分类器便可应用于共享空间中的目标语言特征上。Xu 等人[71]则借鉴了模型压缩领域常用的知识蒸馏方法学习跨语言的共享特征空间。他们首先在含情感类别标签的源语言数据集中训练了一个基于 TextCNN 结构[72]的源分类器。随后，在包含源语言和目标语言的无监督平行语料数据上，通过蒸馏的方式将源分类器的知识迁移到另外一个相同结构的目标语言数据分类器上。具体的做法是将一对平行的双语言数据分别送入到各自语言对应的分类器中，并分别计算出两个分类器的情感类别预测“软分布”$p:p_i = \mathrm{softmax}(x_i/t)$，其中 x_i 表示模型对于数据 i 的 logits 输出，t 是 Hinton 等人[73]提出的温度超参，能够让模型的预测分布变得更平缓，有利于提升蒸馏效果。该阶段训练过程的优化目标是最小化两个模型“软分布”之间的差异，本文采用的是交叉熵损失约束两分布的差异。通过蒸馏过程训练好目标分类器之后，便可直接将目标分类器应用于测试集中的目标语言数据上。该作者还通过理论分析证明，在源域和目标域输入的特征分布相同的前提下，本文提出的两阶段跨语言蒸馏方式等价于直接利用目标语言数据和标签训练目标分类器的过程。而为了进一步实现特征分布的对齐，作者在两个训练阶段均额外引入了基于对抗训练的 DANN[50]方法。

随着词向量表示模型在自然语言处理领域的成功应用，很多研究者对于跨语言词嵌入的构建展开了探索，构造的方法包括有监督、半监督和无监督的形式。有监督的方法同样需要利用到大规模的双语平行文本，将词级别的对齐关系作为跨语言的监督信号；半监督的跨语言词嵌入方法则是只依赖于少量平行语料学习不同语言特征空间之间的映射关系，通过迭代优化确定出有限数据下的最佳映射，并将这种映射推广到源语言和目标语言的全部空间上；而无监督的词嵌入方法则常常是通过对抗训练的方式从非平行的大规模语料资源中挖掘出两种语言之间的关联[74]。构建好的跨语言词嵌入模型就可以在单词级别上将两种不同语言的输入映射到共享的嵌入空间中，解决跨语言特征空间差异的问题。例如，Ziser 等人[75]将 Smith 等人[76]构

建的多语言词嵌入与基于枢轴词的迁移方法相结合，以解决跨语言特征空间差异和跨主题特征分布差异同时存在的问题。他们在模型的输入层中直接使用了多语言词嵌入获得共享空间中的嵌入特征，同时还将该多语言词嵌入用于构成枢轴词分类器的输出层权重，从而进一步实现跨语言的共享空间学习和跨领域分布的对齐。

近期，大规模预训练模型在自然语言处理的各类任务上得到了广泛的应用，许多研究者利用多种语言的大规模语料，训练出了 mBERT[77]，XLM[78]，XLM-R[79]等等多种多语言预训练模型。mBERT 采用了多种语言文本构成的大规模语料库进行预训练，采用统一的模型学习所有语言的上下文语义表示。预训练任务与 BERT[65]中的遮蔽语言建模（Masked Language Modeling，MLM）任务相同，每次均送入一个遮蔽部分 token 的单语言句子，并输出对于遮蔽部分的预测。通过 MLM 任务训练的跨语言共享权重的编码器可以将文本输入映射到对应语言的语义空间，而不同语言的高层语义信息是具有一定的一致性的，因此这样的预训练过程可以在一定程度上解决跨语言的特征空间差异问题。XLM 则加入了部分双语平行语料进行预训练，并在 MLM 任务的基础上加入了翻译语言建模（Translation Language Modeling，TLM）的新任务，该任务是将成对的双语平行语句拼接作为一个新的输入语句，并随机遮蔽该语句中的 token，使模型根据跨语言的上下文预测遮蔽部分的内容，从而得以学习到更强的跨语言迁移能力。Fei 等人[80]利用 XLM 的预训练权重初始化一个编码器-解码器（encoder-decoder）结构作为无监督机器翻译模型。在此基础之上，他们同时利用了源语言和目标语言上的非平行单语言语料，设计了输入文本重建的自监督训练任务，包括同语言内加噪文本的恢复以及跨语言的还原翻译（back-translation），并通过最小化 KL 散度的方式拉近编码器对于翻译前与翻译后的句子的表示，从而实现跨语言特征空间的进一步对齐。他们还将编码器输出的特征表示送入一个额外的语言判别器，通过对抗训练的方式对齐跨领域的特征分布差异。实验表明，他们提出的方法优于直接利用 XLM 微调的方式，且在多个情感分类任务上都达到了当时的最佳水平。

3.1.2.2　不同模态的特征空间差异问题

我们所处的世界是包含多个模态信息的，例如对于一个人说话过程的视频记录，我们可以提取出他的脸部表情信息、语音信息以及话语内容。这些不同模态的信息往往都蕴含着该说话人的情感状态，只不过体现的角度不同。因此我们在对一个人进行情感识别时也可以根据实际的情况，从不同的模态入手。例如对于通过电话与我们进行交流或访谈的人，我们如果想通过计算机自动评估其情感状态，最直接的办法就是利用基于语音数据训练的情感识别模型。然而，当前的机器学习研究在各个模态上的发展程度并不均衡。在视觉和文本模态上，由于数据更易获取，研究者们已经进行了较多的探索，并且利用大规模数据构建了很多预训练模型，能够有效提升对应模态下游任务的表现；而在语音模态上，数据的规模较为有限，且存在噪声较大、信息冗余程度较大的问题，虽然近期已有一些语音预训练模型的研究工作，但相比于另两个模态的训练资源来说仍存在较大差距，也在一定程度上限制了语音情感识别模型的发展。另外，在某些特定的情感类别或维度上，语音信息也天然地存在识别能力的限制。例如，生气和高兴情绪状态下的语音可能都表现出音高较高、能量强度较强、词间间隙更短等特性，但通过面部表情则很容易区分出这两种情绪[81]；文本模态信息对于效价度的识别效果显著优于语音信息[82-83]。因此，有些研究者认为可以将资源较丰富的视觉或文本模态信息迁移到语音模态，以帮助提升语音情感识别模型的效果。而使用这种方法所面临的一大关键问题即为：如何应对不同模态的输入信息在特征空间上的差异。

与跨语言的工作类似的是，知识蒸馏的方法也可以用于解决跨模态的特征空间差异问题。我们可以先在源模态上利用丰富的有监督资源预训练好一个情感识别模型作为教师模型。接下来，我们需要构建一个目标模态的情感识别模型作为学生模型，并在多模态的平行数据上将教师模型与学生模型分别抽取出的对应模态特征表示在空间上对齐，以实现知识从源模态到目标模态的迁移。最后，我们可以在目标模态的有监督数据集上微调学生模型，或者直接用该模型在目标模态识别任务上评测。由于不同模态特征的粒度可能会有所差异，

源模型和目标模型的结构通常也会因对应模态的特点而异，且在学习两模态共享特征空间之前需要先解决特征粒度的对齐问题。

Albanie 等人[84]最早将跨模态蒸馏方法引入情感识别任务中。他们将 SENet 模型[85]在 VGG-Face2 数据集[86]上做身份验证的预训练，并在 FERPlus 数据集[26]上进行表情识别的微调，得到视觉模态的教师模型。对于语音模态上的学生模型则使用基于 VGG-M 的结构[87]。他们利用大规模无情感标注的 VoxCeleb 数据集[88]进行跨模态蒸馏，该数据集包含说话人脸部的视频记录以及与此同步的音频记录，采用自动化的方法构建而成。蒸馏的具体做法与 Eriguchi 等人[70]类似，同样是计算出教师模型与学生模型的“软分布”，并最小化两者之间的交叉熵。需要说明的是，由于教师模型在预训练阶段做的是帧级别的表情识别任务，在蒸馏前需要先将该模型对于一个片段内所有帧的预测结果进行最大池化，以得到片段级别的预测，这样才能与语音模态特征的粒度对齐。作者最后将训练好的学生模型直接用于语音情感识别的评测数据上，效果显著优于随机权重的分类器，表明学生模型将从视觉模态上学到的知识有效地迁移到了语音情感识别任务上。

Li 等人[89]则考虑到文本模态与语音的关联性更强，且能够避免表情识别中说话人面部肌肉运动给识别效果带来的负面影响，因而提出了从文本模态向语音模态迁移知识的多层次跨模态情感蒸馏方法 MCED，如图 6 所示。他们首先通过三阶段的训练过程，利用现有的预训练语言模型、领域相关无标注文本语料以及带标注的文本情感数据集，训练出文本模态的教师模型。接下来，同时采用特征层面和输出层面的蒸馏方法迁移知识。其中，输出层面的蒸馏方法是采用 KL 散度约束两个模型“软分布”之间的差异，特征层面的蒸馏则是通过对两个模型对应的中间层表示进行基于注意力机制的“软对齐”实现。具体来说，他们将教师模型的中间层特征表示通过线性映射后作为查询，将学生模型对应的中间层表示通过线性映射后作为键和值，计算注意力加权后的语音特征表示，并最小化其与对应的文本模态中间层特征表示的均方误差。将两个层面的损失值求和作为蒸馏过程的总体损失值。最后，使用训练好的学生模型在语音情感识别的基准数据集上微调并评测，效果显著优于直接在该集合上训练的方法，且能够胜过先前提出的许多其他语音情感识别方法，充分说明了从文本模态到语音模态进行跨模态蒸馏的有效性。

还有一些工作引入了自动语音识别（Automatic Speech Recognition，ASR）技术，将语音信息显式地转换到文本模态的特征空间中，帮助缓解文本模态

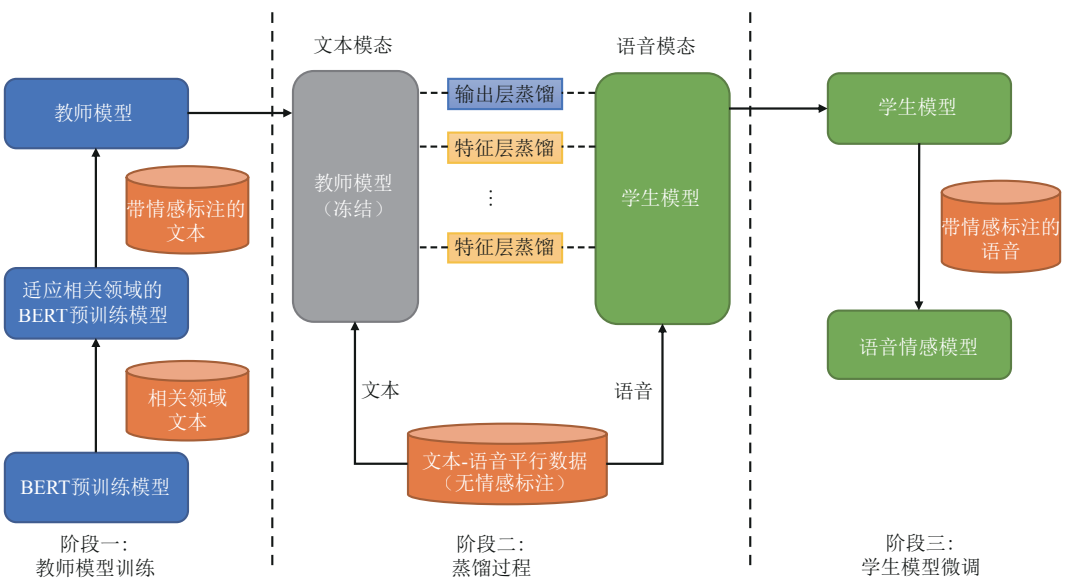

图 6 Li 等人[89]提出的多层次跨模态情感蒸馏方法示意图

Fig. 6 The Diagram of Multi-level Cross-modal Emotion Distillation Proposed by Li et al [89]

与语音模态之间的特征空间差异问题。在Ghriss等人[90]的工作中，他们先利用RoBERTa预训练模型[91]在推特情感数据集上做文本情感分析任务，得到一个较为可靠的文本情感模型，用于在后续使用的无情感标注数据集上生成伪标签。接下来，他们构建出了一个多任务预训练框架，利用一个ASR语音数据集学习注入文本情感知识的ASR特征。具体来说，该框架接收ASR数据集中的语音特征作为输入，并通过一个ASR编码器将语音特征转换到文本特征空间。转换后的文本特征序列一方面被送入token级别的预测层，计算ASR的损失；另一方面又被送入一个额外的情感分类层，将得到的预测结果与上述预训练的文本情感模型生成的伪标签计算交叉熵损失。在联合优化上述两种损失的过程当中，ASR编码器得到了有效的训练，使得语音模态输入能够初步转换到文本模态空间；同时情感分类的分支又以一种类似蒸馏的方式将预训练的文本情感模型中的知识迁移到了编码器上，进一步促进ASR编码器的输出与文本情感模型编码的情感特征之间的对齐。最后，将训练好的ASR编码器后面加入情感回归层，并在带有连续维度情感标注的语音数据上微调，实现融合文本模态知识的语音情感识别。该工作在效价维度上的表现很好，胜过了Li等人[92]利用自监督语音预训练做情感识别工作的效果，说明文本模态的知识被有效地迁移到了语音模态上。

3.2　针对任务差异的迁移学习问题

情感是人基于对客观事物的认识所产生出的一种主观反应，因此情感信息是一种较为抽象的高层次语义信息，且表现形式往往因人而异。这就导致在规模有限的情感数据集中，模型很难准确地捕获到数据的原始输入特征到其对应的情感标签之间的映射关系。尤其是对于一般的粗粒度标注情感数据集来说，仅仅利用句子级或片段级的情感标签作为监督信号并不足以使模型学习到对于句子或视频片段中细粒度语义内容的理解能力，导致模型缺乏预测更高层次情感信息所需的语义基础。因此，我们可以考虑引入其他任务，利用额外的数据或当前数据集上其他方面的监督信号学习到额外的知识，并将知识迁移到作为目标的情感识别任务中，加强模型对于语义内容的理解能力以及对于特定被识别人在特定情景下情感倾向的认识，从而帮助提升情感识别效果。我们可以将与情感关联度较高的信息的预测任务作为源任务，与情感识别任务联合训练；也可以将输入信息的上下文语义建模任务或细粒度成分的理解任务作为源任务，为模型提供更强的语义表示知识作为后续进行情感识别任务的基础。本节将分别介绍在这两种情况下实现知识跨任务迁移的具体工作。

3.2.1　多任务联合训练中的迁移问题

在很多情感数据集的数据中，除了情感的标注信息之外，我们还能获取到很多其他信息，例如说话人的身份、说话人的意图等等。在这些额外的信息当中，可能会存在一些与说话人的情感状态或情感倾向较为相关的部分。我们可以通过多任务学习的方式，将其他相关信息所对应的预测任务作为源任务，并与作为目标任务的情感识别任务进行联合训练，使模型同时学习到两种具体任务对应的特征。此外，有些数据集可能同时含有不同情感表示类型的标注，例如对于同一数据既有属于离散状态表示的情感类别标注，又有连续状态表示下的各个维度的取值标注。若目标任务为其中的某一种情感表示的预测，则可以将另一种表示形式作为源任务，同时采用两种任务对应的特征进行联合训练。而如果能够在训练或推理阶段对两种特征之间的潜在关联进行显式的建模，就可以进一步促进模型将从源任务中学到的知识迁移到目标任务上，从而提升在目标任务上的预测效果。图7为多任务联合训练的示意图。

在自然语言处理领域，研究者们普遍认为，在一段对话交互过程当中，交互参与者的意图（也被称为“对话行为”）与情感具有一定的关联性[93-94]。典型的意图类别包括提问、回答、同意、否定等等[94]。举例来说，在一个双人对话交互的场景下，对话者A先对某个产品给出了正面的评价，但接下来对话者B针对A的言论做出了反驳的对话行为，即B此时的意图是属于“否定”的，那么我们就可以推断出，B此时所表达的情感也很可能是负向的情感。因此，如果一个模型对于对话参与者的意图有着较强的识别能力，它就可以利用识别到的意图信息帮助提升其对于情感的理解能力，反之亦然。Cerisara等人[94]首次探索了将意图识别和情感识别联合训练的效果。他们首先构建了一个对话交互数据

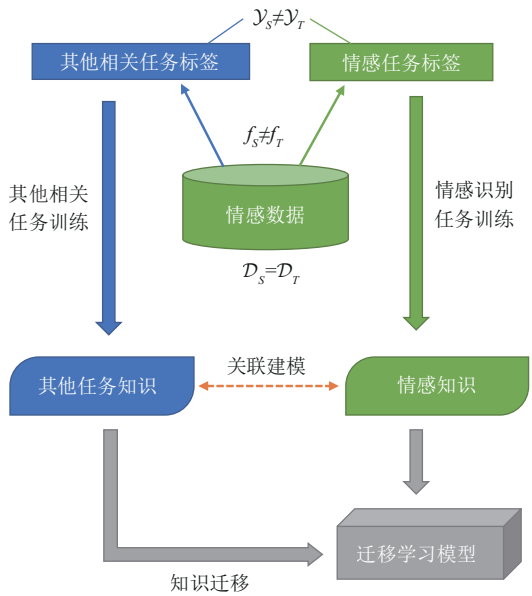

图7 多任务联合训练示意图

Fig. 7 Illustration of Multi-task Joint Training

集 Mastodon，其中的每句话都类似于推特的形式。接下来，他们提出了一个多任务模型在 Mastodon 上同时做意图识别和情感识别任务。该模型采用双向长短期记忆网络（Long Short-Term Memory，LSTM）对词嵌入向量编码句特征，接下来将交互中所有句子对应的句特征送入一个循环神经网络（Recurrent Neural Network，RNN）以得到整段对话的特征。最后，将得到的对话级别特征分别送入两个不同的2层多层感知机网络（Multilayer Perceptron，MLP）中，分别用于预测意图和情感信息。作者先在数据集的全部训练数据上进行了评测，发现将两个任务联合训练的效果和只训练目标任务的效果无明显差异。但如果只采用少量目标任务的带标注数据，并加入另一个任务的全部带标注数据进行联合训练，效果要明显好于只采用目标任务的少量数据进行单任务训练的效果。这说明模型在意图识别或情感识别这两个任务之一上学到的知识能够有效地迁移到另一个任务上，不过在该工作提出的多任务框架下，这种迁移效果体现得还不够显著。

Cerisara 等人[94]工作的最大局限在于对两个任务采用了完全相同的对话级别特征进行表示，在后续相似问题的工作当中，研究者们往往采用了具有非共享层的编码器分别编码对应于不同任务的特征，并设计了交互与融合模块显式对对话的时序信息以及不同任务之间的关联进行建模。Kim 等人[95]提出了同时进行言语行为识别、谓词识别和情感识别三个任务的框架。其中“谓词”代表与句子主要含义相关的语义焦点。该模型采用三个非共享权重的卷积神经网络（Convolutional Neural Network，CNN）作为编码器，分别提取出对应于三个任务的初始单任务特征，并将初始特征通过一些全连接层得到多任务的融合特征。该工作的作者认为，言语行为具有较强的时序依赖性，而谓词与情感则会主要受到当前时刻的对话行为的影响。因此，他们先将上一时刻的言语行为预测结果与当前时刻该任务相关的融合特征结合，进行当前时刻的言语行为预测；而对于谓词识别或情感识别任务，他们则将该任务相关的融合特征与刚刚得到的当前时刻的言语行为预测结果结合，分别进行这两个任务的预测。Qin 等人[96]提出的 DCR-Net 模型利用注意力机制建模了意图和情感之间的关联。他们先采用一个共享权重的双向 LSTM 得到包含时序信息的句子表示，如果用于意图任务则将该表示通过另一个双向 LSTM 网络编码，如果用于情感任务则用一个 MLP 网络编码，得到特定于任务的表示。接下来，采用共注意力机制显式建模两个任务特征之间的交互。例如对于情感识别任务，将编码的情感特征作为查询，意图特征作为键和值，并将利用注意力机制加权后的结果与初始的情感特征相加，得到交互后的情感特征表示。Qin 等人[97]在之后又提出了 Co-GAT 模型，该模型则是利用图网络实现对话上下文的建模和两个任务间的交互。具体来说，该模型的编码器部分由一个共享权重的双向 LSTM、一个注入说话人信息的图注意力网络和两个对应于不同任务的非共享权重的 LSTM 构成。交互模块则包含多层堆叠的图注意力网络，编码器对于每个任务的每句话编码的特征表示都作为图中的一个结点。在同一个任务的所有结点中，将每个结点与其上下文的所有结点连边，以建模对话上下文的信息；同时将每个结点与另一个任务的所有结点连边，以建模不同任务间的关联关系。在前向传播过程中，相邻结点之间的信息能够被传递到当前结点上，因此更新后的结点特征表示就包含了上下文信息和不同任务之间的关联信息。

在与人脸表情相关的研究当中，面部动作单元（Action Units，AU）的概念受到了广泛的关注。这一

概念来源于Ekman等人[98]提出的面部动作编码系统(Facial Action Coding System，FACS)，他们在人的脸部划分出多个局部肌肉区域，将某一个或某几个肌肉区域的运动定义为一组AU。事实上，人所表现出的各种面部表情均可表示为多个AU的组合。因此，宏观的面部表情与局部的AU之间存在天然的关联。将AU检测和表情识别两种任务联合训练有助于充分利用这种关联性。Chu等人[99]提出了适应性权重共享网络(Adaptively Weights Sharing Network，AWS-Net)学习AU检测与表情识别两种任务之间的信息交互。他们分别采用两个VGG19网络[100]作为两个任务对应的网络分支，同时在两个分支相对应的每一层之后加入一个共享的AWS单元网络层。该AWS单元网络层接收两个分支在对应层的特征图作为输入，并采用可学习的权重对两张特征图进行两种不同的线性组合，将得到的结果分别作为两个任务的网络分支在下一层的输入。实验结果表明AWS单元网络层的引入可以同时提升模型在两个任务上的预测效果。在最近举办的第4届自然场景情感行为分析竞赛(Affective Behavior Analysis in-the-wild，ABAW)[101]中，主办方也设置了多任务学习赛道，希望研究者们在包含人脸的视频数据中探索将效价度-激活度预测、情感类别的识别以及AU检测这几种不同的任务进行联合训练的效果。该赛道的冠军Zhang等人[102]分别尝试了三种不同的多任务学习框架。第一种框架采用统一的时序编码器抽取出共享的时序特征，再分别送入各自任务对应的预测器当中。第二种框架将时序编码器分为两部分，底层为任务间共享，用于捕获人脸图像中的一些基本信息；而上层则为任务特定的网络层，用于抽取对应于每种任务的时序特征并送入各自任务对应的预测器中。第三种框架同样将时序编码器分为共享和非共享的两部分，除此之外，考虑到有些任务的高层语义信息之间也可能具有一定的依赖关系，他们将其中一种或几种任务的非共享编码器部分抽取出的高层任务特征与共享编码器部分抽取出的特征拼接，并送入到另外一些任务的非共享编码器中，再分别将每个非共享编码器的输出送入任务预测器。实验发现，在不同的任务组合以及目标任务上，取得最好效果的多任务框架可能有所不同，但三种多任务框架下的预测效果均普遍优于单任务训练的效果。

也有一些研究者探索了在语音模态上将情感识别任务与其他相关任务进行联合训练的效果。Sharma[103]将多个不同语言背景下的语音情感数据集结合，构建了一个较大规模的多语言情感数据集。在该集合上，他采用多语言语音预训练模型XLSR-53[104]进行情感识别任务的微调，同时探索了在微调过程中加入不同辅助任务进行联合训练对情感识别效果带来的影响。他尝试了5种辅助任务：性别预测、语言预测、基频(F0)回归、能量(energy)回归和声音比率(voice ratio)回归，并对比了不同辅助任务组合的效果，发现将性别预测、语言预测和基频均值的回归这三种任务与情感识别任务联合训练时效果最好。事实上，性别的分类有助于模型捕获到与情感特性相关的音高(pitch)和梅尔频谱(Mel Frequency Cepstrum Coefficient，MFCC)信息[105]；语言的分类任务有助于模型学到不同语言的声调；而基频信息更是与情感高度相关：高基频往往代表兴奋的状态，低基频往往意味着忧伤的状态。因此，加入这些辅助任务共同训练后，一些较为重要的特征得到了强化或补充，从而提升了情感识别的效果。Zhang等人[106]则将一个由演员表演的情感数据集和一个自发表现的情感数据集的训练数据合并，得到复合场景的训练数据集，并在其上探索加入数据类型的判别任务(即判断数据属于表演类型还是表现类型)对于情感识别任务的影响。他们认为，中性的情感是更具有普适性的，与数据类型或数据集无关；而其他情感类别则可能与数据的类型更为相关。因此他们对数据类型的判别任务进行了改进，即只对除中性类别数据之外的情感类别数据加入类型判别任务。实验结果表明，这种选择性的类型判别任务能够在一定程度上帮助提升模型的情感识别效果。

3.2.2　预训练模型或知识库的迁移问题

我们人类在解决现实问题时总是会潜在地利用到很多先验知识。例如当我们在接收到一些信息时，我们可能基于信息的某些成分产生联想，获取到其潜在的内涵或可能相关的内容，同时利用已有的经验得到对于这些信息的整体性理解。这些联想到的内容以及对信息进行处理加工的经验往往适用于多种具体的任务。在使用计算机进行情感识别任务时，我们同样可以为模型注入类似的先

验知识，帮助提升模型的识别效果。此时的先验知识可能是对于输入信息中细粒度成分的语义补充，或是对于各局部成分的上下文语义理解。我们可以利用模型的预训练或者外部知识库的构建作为源任务获取到这样的先验知识，随后将知识迁移到作为目标任务的情感识别任务中。

基于预训练模型的知识迁移通常属于隐式的迁移形式，一般我们会先利用一些预训练任务使模型学习到较强的上下文语义表示能力，随后基于预训练阶段得到的模型权重，在目标任务上继续训练，如图8所示。在自然语言处理领域，大规模预训练模型在各种下游任务上得到了广泛的应用。对于文本情感识别这种文本分类任务，通常采用BERT[65]，RoBERTa[91]等自编码式的预训练模型。BERT采用MLM和下一句预测（Next Sentence Prediction，NSP）的预训练任务学习融合上下文信息的文本特征表示，而RoBERTa则去掉了NSP任务，并调整了预训练过程中的一些设定。在利用大规模无监督语料库学到通用的自然语言知识后，我们可以在带标注的小规模情感数据集上微调模型，使得模型可以在已有的先验语言知识基础之上进一步学习与目标任务相关的知识，效果也往往显著优于在目标任务上从零开始训练的方法。

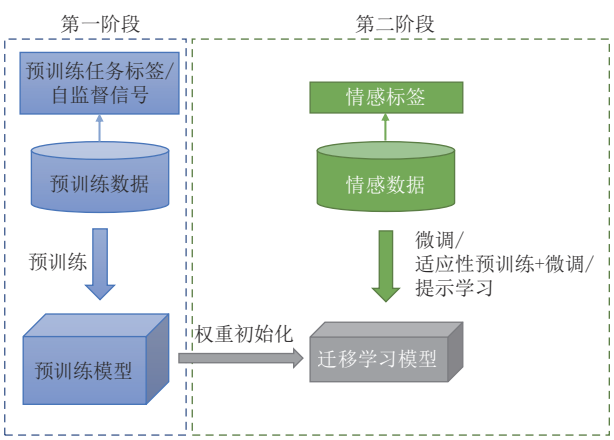

图8 预训练模型的迁移问题

Fig. 8 Transfer of Pre-trained Models

在经典的“预训练-微调”迁移范式中，预训练语料的数据领域与下游任务的数据领域并不一致，模型对于预训练任务与下游任务的输出形式也会存在较大的差异。这就导致从源任务迁移到目标任务的过程当中，模型需要同时适应数据领域和输出形式的两大差异，限制了知识迁移的效果。一些研究者提出可以加入适应性预训练（adaptive pre-training）阶段缓解该问题，即在完成一般意义上的预训练之后，继续在与目标任务相关领域的数据上做预训练任务，强化模型对于目标领域数据的理解能力，促进先验语义知识向目标领域的迁移。Gururangan等人[107]总结了适应性预训练的具体方法并进行了对比分析。一种方法是找到与目标任务中的数据相近的大规模无监督语料，继续在这样的语料上做预训练任务。例如对于评论文本情感识别的任务，搜集大量的各类评论文本作为相近领域的语料用于适应性预训练。另一种方法是仅在目标任务的训练数据集上继续做无监督的预训练任务，虽然预训练数据较为有限，但由于此时训练数据的领域分布与最终推理阶段测试集的领域分布更为贴近，往往能够达到和第一种方法可比的效果。Sun等人[108]探究了使用这两种适应性预训练方法对于文本情感分类任务的影响，发现相比于经典的“预训练-微调”策略，这两种方法的加入都能为模型的分类效果带来有效的提升。还有一些研究者提出了提示学习（prompt learning）[109]的方法，将下游任务的输出形式修改为与预训练任务相似的形式，以减小先验知识迁移到下游任务的难度。例如对于文本情感分类任务，可以在原文本序列后加入一段提示词和一个遮蔽token，引导模型在遮蔽token位置预测出一个与情感类别有关的词语，再根据该词得到最终的情感预测。Mao等人[110]探索了在不同的文本预训练模型上基于提示学习进行情感识别任务的效果，并对该任务上提示学习的具体实现方式展开了分析。

在语音领域，近期也涌现出了许多通用的自监督预训练模型，如Wav2vec 2.0[111]，TERA[112]，Mockingjay[113]，HuBERT[114]等。Morais等人[115]探索了将Wav2vec 2.0和HuBERT学到的语音通用知识迁移到情感识别任务上的效果。他们先将通用的语音预训练模型在带标注的情感识别数据集上微调，接下来将微调后的模型固定作为特征抽取器，并在其后接入ECAPA-TDNN网络[116]将特征在时序维度上聚合，得到句子级别的特征用于最终的情感识别任务。在IEMOCAP数据集[35]上，使用基于Wav2vec 2.0和HuBERT融合特征的模型取得的识别效果达到了当前语音单模态方法的最佳水平，甚至接近于

采用文本+语音双模态信息的方法[117]的效果，表明语音预训练模型中学到的先验知识也能为下游的情感识别任务提供有效的帮助。

此外，也有研究者专门针对情感识别的目标任务，设计了多种自监督任务作为上游预训练任务以学习更有效的特征表示。Li 等人[118]认为先前的脑电情感识别研究大多依赖于有监督的单任务学习，但因脑电数据集的规模较小且标注噪声较大，这种较为简单的训练方式容易产生过拟合现象，对于新数据集的泛化能力也较差。因此他们根据脑电数据的特点，提出了三种自监督任务：空间拼图任务（spatial jigsaw puzzle task）、频段拼图任务（frequency jigsaw puzzle task）和对比学习任务。在空间拼图任务中，他们首先根据大脑的空间区域将收集到的脑电数据划分为 10 块，随后将每个原始输入重排列为预定义好的 128 种空间排列之一，该任务的目标即为预测该排列的序号。通过该任务的训练，模型能够捕获到大脑不同区域之间的内在空间关系。在频段拼图任务中，原始的输入数据则会被转换为与情感相关的 5 个频段之间的重排列，通过预测排列序号的目标，模型能够更好地捕获到可用于情感识别的关键频段。而对比学习任务则将原始数据经过空间或频率上的变换得到扩增样本作为正样本，将其他数据及其扩增样本作为负样本，通过对于正负样本的对比提升特征表示的鲁棒性。在作者进行的无监督模式实验中，他们利用上述任务训练出的模型抽取特征用于下游情感识别任务的训练，效果优于其他通用的自监督学习方法，表明本文针对脑电信息特点进行的任务设计能够为特征表示的学习带来更大的帮助。而 Zhao 等人[119]则提出了包含语音、文本、人脸三模态信息的多模态情感预训练模型 MEmoBERT。他们从电影和电视剧中采集了大量富含情感内容的片段，在此基础上采用了 4 种任务进行预训练。其中，针对三个不同的模态，分别采用了对应的遮蔽内容预测的任务。具体来说，针对文本模态，他们随机遮蔽掉句子中的一些单词，迫使模型利用剩余的单词以及另两个模态的信息预测出遮蔽部分的词；而对于人脸或语音模态，他们则随机遮蔽掉连续的多帧，使模型利用对应模态剩余帧的信息以及另两个模态的信息对遮蔽内容做回归。这三个预训练任务有助于模型对于各个模态内的局部信息学到跨模态的上下文语义表示，可以被看作多模态场景下的先验语义知识。而另外的一个任务则是将一个预训练的表情识别模型对人脸的预测结果作为监督信号，以类似蒸馏的方式将该人脸模态模型的情感相关知识迁移到多模态的 MEmoBERT 中。在下游的 IEMOCAP[35]和 MSP-IMPROV[36]数据集上，使用 MEmoBERT 微调或做提示学习的方法效果均胜过从零初始化的效果，且使用提示学习的效果更佳，说明该预训练模型学到的多模态先验语义知识也被有效地迁移到了下游的多模态情感识别任务上。

基于外部知识库的知识迁移则属于显式的迁移过程，如图 9 所示。较为简单的做法是直接从已有的外部知识库中检索出与输入信息的细粒度成分相关联的外部知识，并将其与根据输入信息本身得到的语义表示相结合，作为模型的识别过程所使用的融合特征。也有一些研究者先利用外部知识库中的数据训练得到知识生成模型，随后将该模型应用于目标任务的输入数据上，以生成的方式获取到与输入信息相关联的外部知识。在对话情感识别任务中，许多工作就采用了基于外部知识库的知识迁移方式加强模型对于对话交互过程以及交互

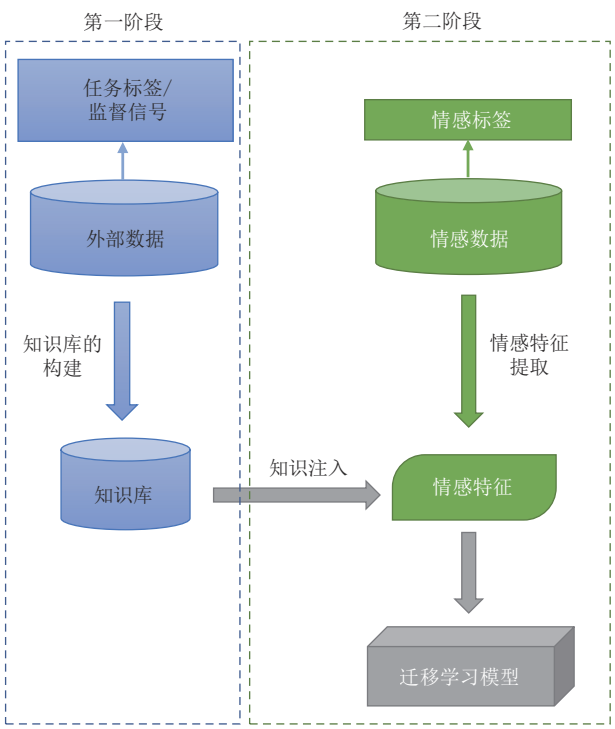

图 9　外部知识库的迁移问题

Fig. 9　Transfer of External Knowledge Bases

参与者自身状态的理解能力。Zhong 等人[120]提出的 KET 模型利用了 ConceptNet 知识库[121]和 NRC_VAD 情感词典[122]。ConceptNet 中的知识以<概念 1,关系,概念 2>的三元组形式存在,每个三元组附有一个置信度分数。NRC_VAD 中则包含很多单词及其对应的词级别效价、唤醒、支配度值(以下简称为“VAD 值”)的映射关系。在 KET 中,对于输入语句中的每个非停用词,先在 ConceptNet 中检索与其关联到的所有概念,筛选出有效的关联概念后再从 NRC_VAD 中检索其对应的 VAD 值。于是对于每个词都可以得到多个元组,每个元组的形式为:(关联概念,置信度,VAD 值)。随后,KET 利用这样的知识信息计算每个关联概念与主题的相关性以及情感强度,并根据这两点因素对每个词的所有关联概念对应的知识进行注意力加权求和,得到词级别知识表示,并与该词的语义表示拼接作为融入知识的特征。Xie 等人[123]提出的 CKE-Net 则仅采用 ConceptNet 作为词级别知识库,通过检索的方式得到每个词对应的所有关联概念知识元组:(关联概念,置信度)。他们采用图注意力网络建模词级别的知识表示,并将其与经过对话建模的词级别语义表示以及语音的词级别表示结合,作为融合特征。

也有些知识库的知识是以句子级别存在的,如 Sap 等人[124]构建的 ATOMIC 知识图谱,每条知识以<事件 1,关系,事件 2>的三元组形式存在,其中“关系”共有 9 种类型。Ghosal 等人[125]提出的 COSMIC 模型是以基于 ATOMIC 训练的知识生成模型 COMET[126]作为知识库,他们将输入语句分别与 5 种关系(主体意图、主体影响、主体反应、客体影响、客体反应)拼接,送入 COMET 的编码器中,将编码的特征表示直接作为 5 种关联事件的知识表示。随后,他们又利用 5 种不同的门控循环单元网络(Gated Recurrent Unit, GRU)建模对话上下文特征与不同关联事件特征之间的关系。而 Zhu 等人[127]提出的 TODKAT 则同时利用 ATOMIC 和 COMET 作为知识库。对于某个输入语句,他们先利用 SBERT[128]在 ATOMIC 检索出几条与输入最相近的事件,并取出其对应的 3 种关系(主体意图、主体反应、客体反应)下的关联事件,再利用 COMET 直接生成上述 3 种关系对应的关联事件,随后利用指针网络[129]从上述检索和生成两种方式得到的知识中选取出较好的一种保留。最后,采用预训练的文本模型抽取保留下的关联事件语句的语义和主题特征,并利用注意力机制融合输入语句的语义、关联事件的语义和主题信息,得到知识表示。上述注入外部知识的对话情感识别工作的方法总结见表 5。

4 展望

尽管当前许多情感识别领域的研究者已经关注到了任务中的一些迁移学习问题,并开展了一定的探索。但我们认为,仍存在一些尚未在情感领域被充分探究的问题,包括脑电情感模型的泛化性问题、多模态的跨领域迁移问题、零样本情感识别问题以及模型的安全迁移问题。我们在此对这些问题进行简单探讨,希望能够为情感识别领域后续的研究工作带来启发。

4.1 脑电情感模型的泛化性问题

生理信号中的脑电信息可以客观地反映情感状态变化,因此在情感识别任务上具有较高的研究价值。但当前的脑电情感识别面临着诸多挑战,例如现有的脑电数据集规模较小,不同被试者之间的脑电信号差异一般较大,情感标签噪声较大等等。这些因素都会导致训练出的情感模型缺乏泛化能力,难以应用于其他受试者或其他集合中的脑电数

表 5 注入外部知识的对话情感识别工作

Tab. 5 Dialogue Emotion Recognition with External Knowledge

工作	外部知识库	知识层级	知识获取方式	知识形式	知识迁移建模方式
KET[120]	ConceptNet[121]; NRC_VAD[122]	词级别	检索	所有关联概念对应的元组:(关联概念,置信度,VAD 值)	融合相关度和情感强度信息的图注意力网络
CKE-Net[123]	ConceptNet	词级别	检索	所有关联概念对应的元组:(关联概念,置信度)	图注意力网络
COSMIC[125]	基于 ATOMIC[124] 训练的 COMET[126]	句子级别	生成	5 种关系的关联事件特征表示	5 种不同的 GRU
TODKAT[127]	ATOMIC;COMET	句子级别	检索/生成	3 种关系的关联事件语句	注意力机制

据。因此，增强模型的泛化性对于脑电情感识别问题尤为重要。

在本文的3.1.1.3节中介绍了关于脑电特征分布差异问题的有关工作。该类工作致力于将在源领域数据上训练得到的模型迁移至指定的目标被试者或目标集合的数据上进行应用，采取的做法一般是对齐源领域和目标领域的特征分布，学习到同时适用于两个领域的特征表示。但是这种针对于指定目标领域的迁移过程难以满足更广泛场景下的应用需求。在本文的3.2.2节中则介绍了利用自监督任务学习脑电特征的工作，更为充分地挖掘了脑电信息本身的特性，使得学到的特征表示在应用于作为目标的情感识别任务时具有更强的鲁棒性。

然而，上述方法均局限于对脑电特征进行处理。事实上，不同个体由于自身先天特质不同，其情绪在生理层面的表达必然也会存在较大的个体差异。基于此，Hu等人[130]提出应在脑电情感模型当中融入个体特点的建模。具体来说，可以将个体应对特定刺激所产生的情绪反应拆分为共性反应和个性反应两部分，对于共性反应部分可借鉴现有工作提取出较为鲁棒的脑电特征，对于个性反应则可采用一些人格特征元素进行建模。个性化元素的引入有望增强脑电情感模型对于更广泛人群的适用性，在未来的研究当中值得探索。

4.2　多模态的跨领域迁移问题

当前，多模态情感识别已经得到了广泛研究，然而对于多模态输入数据的跨领域迁移问题还未在情感识别领域得到充分探索。在第3.1.1节中，我们已经总结了应对各模态内跨领域特征分布差异问题的方法，这些方法可用于解决各自模态内的跨领域迁移问题。在多模态的场景下，应对跨领域迁移问题的一种简单思路是直接对齐多模态融合特征的跨领域分布差异。例如，Yin等人[131]提出的SIDANN模型采用直接拼接的方式融合视觉、语音、文本的三模态输入特征，并采用基于对抗训练的方法学习与说话人无关的多模态特征表示用于跨领域迁移。然而，该方法的模态融合过程较为简单，且直接对齐融合后的特征往往效果有限。

而在动作识别领域，已经有一些工作将视觉信息中的RGB信息（红、绿、蓝三个颜色通道的信息）和光流信息视为两个不同的模态，并基于对比学习的方法同时解决多模态表示学习和跨领域特征迁移的问题。例如，Kim等人[132]提出了跨模态和跨领域的对比损失函数。他们在两个域中分别采样出一些视频，根据跨模态损失，在每个域内拉近来自同一段视频的两个模态特征，推远来自不同视频的两个模态特征。由于不同模态的特征处于不同的特征空间，因此计算该损失时不同模态的特征会通过一个额外的映射层。通过该过程，模型能够更好地从不同模态输入中学习到具有一致性的高层语义信息。与此同时，他们将目标域某个带伪标注的视频作为锚点，采样源域标注为相同类别的视频作为正样例，再采样目标域伪标为不同类别的视频作为负样例，并分别在两个模态上计算跨领域对比损失函数，分别对每个模态的特征实现了类别层面的特征分布对齐。考虑到情感信息也属于较高层次的语义信息，而语义信息相比各个模态上的私有成分往往更加具有跨领域的一致性，因此我们也可以在跨领域情感识别任务中考虑采用类似的对比学习方法，促进各个模态内部特征分布的对齐以及跨模态情感表示的学习。

4.3　零样本情感识别问题

本文3.1.1节中介绍的应对特征分布差异的方法虽然可以用于解决跨数据集或跨领域情感识别的问题，但前提是要求两个集合对应的标签空间完全一致。然而在现实生活中，我们可能希望模型能够识别出训练阶段未曾见过的情感类别。例如在人机交互系统中，系统需要较为准确地判断说话人是否可信、友好，但现有的公开数据集中极少存在这些情感类别的标注数据。因此，为了将情感识别技术推广至更广泛的场景，零样本情感识别问题需要得到更多关注。

Xu等人[133]基于原型学习（prototype learning）的方法，对语音情感识别任务中的零样本识别问题进行了初步探索。他们首先利用文本模态的表示模型将训练和测试集中可能出现的所有情感类别对应的描述词映射到文本特征空间，作为语义嵌入原型。接下来，他们分别尝试了逐样本学习（sample-wise learning）和逐情感学习（emotion-wise learning）的训练策略，利用可见类别的数据训练用于识别不可见类别样本的模型。其中前者将语义嵌入原型作为初始化，得到动态更新的样本级别原型，并在

训练过程中拉近每个样本与其对应的样本级别原型在特征空间中的距离；而后者则是直接基于语义嵌入原型学习语音特征和情感类别标签之间的对应关系。通过这两种策略的训练，测试集中新类别样本的语音特征能够较好地匹配到该类别情感描述词所对应的类别原型，从而实现零样本的情感识别。不过，该工作采用的类别原型表示方式仍较为简单。未来，可以参考 Wang 等人[134]在图像分类领域的零样本识别工作，采用图神经网络（Graph Neural Network，GNN）为情感类别原型的特征表示提供补充。此外，我们还可以借鉴 Li 等人[135]的思路，采用生成对抗网络（Generative Adversarial Network，GAN）直接在训练阶段生成不可见情感类别的样本，解决训练和测试阶段标签空间的差异问题。

4.4 模型的安全迁移问题

当前，机器学习模型已经在现实世界的各类场景中得到了广泛的应用，为人们的工作和生活提供了巨大的帮助。然而，研究者们发现，当前的许多模型其实很容易受到黑客的攻击，且模型所承受的攻击风险贯穿于其训练和推理过程。例如，在训练阶段进行的“投毒攻击”可以利用向训练集中加入的特定样本使训练出的模型失去应有的预测能力；而在推理阶段进行的“隐私攻击”则可能使模型泄露用户的敏感信息。

在 3.2.2 节中我们提到，目前在文本和语音模态上已有许多基于大规模通用预训练模型的迁移进行情感识别的工作。然而 Zhang 等人[136]发现，针对通用预训练模型的投毒攻击所带来的负面影响将很可能被继承到下游任务微调后的模型中。因此他们提出了相关模型切片（Relevant Model Slicing，ReMoS）的方法，保留预训练模型的权重中与下游目标任务密切相关的部分，去除掉易受攻击的与源任务较为相关的部分，随后再在目标任务数据上微调，实现了模型从源任务到目标任务的安全迁移。

此外，显式的情感线索中往往蕴含大量和被识别人的其他个人特征相关的成分。例如我们根据一个用户的人脸相貌或者语音信息往往就能推断出该用户的性别及年龄，甚至可能推测出该用户来自的地区，而这些信息往往是不应被轻易泄露的。因此，在将训练好的情感识别模型迁移至实际的应用场景时，我们也需要极力避免模型遭受隐私攻击，防止模型输出的特征被轻易用于这些敏感属性的预测。Feng 等人[137]对此进行了初步的探索，他们将 Mireshghallah 等人[138]提出的 Cloak 方法与针对性别属性的对抗训练相结合，学习到一种针对输入数据的噪声映射，再将映射后的特征送入原来的情感识别模型中。经过噪声映射处理后，模型在情感识别任务上依然能表现出较好的效果，但将模型的输出特征用于性别预测的效果相比处理前大幅下降，实现了对于用户性别属性的保护。不过该工作未对其他敏感属性的隐私攻击防御问题展开探究，对于将情感识别模型迁移至应用场景时所面临的该类问题，在未来还有很大的探索空间。

5 结论

本文对情感识别中的迁移学习问题进行了综述。依据迁移学习的定义，本文将迁移学习问题分成了针对领域差异和针对任务差异的两大部分。对于领域差异的问题，本文先总结了人脸表情、语音、脑电信号、文本四方面的情感识别工作中对齐不同领域中特征分布差异的方法，随后又介绍了跨语言以及跨模态场景下解决不同领域之间的特征空间差异的方式。对于任务差异的问题，我们则重点关注于知识从源任务迁移到目标任务的过程，并将现有工作分为多任务联合训练中的迁移问题、预训练模型或知识库的迁移问题两类。最后，本文还提出了一些情感识别领域未来需要得到更多关注的迁移学习问题，并结合当前已有的初步探索以及其他相关领域的工作进行了简单的讨论，为后续的研究提供新的方向和思路。

参考文献

[1] ZHAO Sicheng, JIA Guoli, YANG Jufeng, et al. Emotion recognition from multiple modalities: Fundamentals and methodologies [J]. IEEE Signal Processing Magazine, 2021, 38(6): 59-73.

[2] SOLEYMANI M, GARCIA D, JOU B, et al. A survey of multimodal sentiment analysis [J]. Image and Vision Computing, 2017, 65: 3-14.

[3] SCHERER K R. Psychological models of emotion[J]. The Neuropsychology of Emotion, 2000, 137(3): 137-162.

[4] ORTONY A, TURNER T J. What's basic about basic emotions?[J]. Psychological Review, 1990, 97(3): 315.

[5] MEHRABIAN A. Pleasure-arousal-dominance: A gen-

eral framework for describing and measuring individual differences in temperament[J]. Current Psychology, 1996, 14(4): 261-292.

[6] BARRETT L F, ADOLPHS R, MARSELLA S, et al. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements[J]. Psychological Science in the Public Interest: A Journal of the American Psychological Society, 2019, 20(1): 1-68.

[7] KHALIL R A, JONES E, BABAR M I, et al. Speech emotion recognition using deep learning techniques: A review[J]. IEEE Access, 2019, 7: 117327-117345.

[8] AKÇAY M B, OĞUZ K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication, 2020, 116: 56-76.

[9] SHAH FAHAD M, RANJAN A, YADAV J, et al. A survey of speech emotion recognition in natural environment [J]. Digital Signal Processing, 2021, 110: 102951.

[10] EGGER M, LEY M, HANKE S. Emotion recognition from physiological signal analysis: A review[J]. Electronic Notes in Theoretical Computer Science, 2019, 343: 35-55.

[11] TORRES P E P, TORRES E A, HERNÁNDEZ-ÁLVAREZ M, et al. EEG-based BCI emotion recognition: A survey [J]. Sensors, 2020, 20(18): 5083.

[12] SUHAIMI N S, MOUNTSTEPHENS J, TEO J. EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities[J]. Computational Intelligence and Neuroscience, 2020, 2020.

[13] ALSWAIDAN N, MENAI M E B. A survey of state-of-the-art approaches for emotion recognition in text[J]. Knowledge and Information Systems, 2020, 62(8): 2937-2987.

[14] DENG Jiawen, REN Fuji. A survey of textual emotion recognition and its challenges[J]. IEEE Transactions on Affective Computing, 2021.

[15] SAXENA A, KHANNA A, GUPTA D. Emotion recognition and detection methods: A comprehensive survey[J]. Journal of Artificial Intelligence and Systems, 2020, 2(1): 53-79.

[16] ZHANG Jianhua, YIN Zhong, CHEN Peng, et al. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review[J]. Information Fusion, 2020, 59: 103-126.

[17] SHARMA G, DHALL A. A survey on automatic multimodal emotion recognition in the wild[J]. Advances in Data Science: Methodologies and Applications, 2021, 189: 35-64.

[18] FENG Kexin, CHASPARI T. A review of generalizable transfer learning in automatic emotion recognition [J]. Frontiers in Computer Science, 2020, 2: 9.

[19] 龙明盛. 迁移学习问题与方法研究[D]. 北京：清华大学, 2014.
LONG Mingsheng. Transfer learning: Problems and methods[D]. Beijing: Tsinghua University, 2014. (in Chinese)

[20] WEISS K, KHOSHGOFTAAR T M, WANG Dingding. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1): 1-40.

[21] LUCEY P, COHN J F, KANADE T, et al. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression [C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco, CA, USA: IEEE, 2010: 94-101.

[22] LYONS M, AKAMATSU S, KAMACHI M, et al. Coding facial expressions with Gabor wavelets[C]//Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. Nara, Japan: IEEE, 1998: 200-205.

[23] PANTIC M, VALSTAR M, RADEMAKER R, et al. Web-based database for facial expression analysis[C]// 2005 IEEE International Conference on Multimedia and Expo. IEEE, 2005.

[24] LI Shan, DENG Weihong. A deeper look at facial expression dataset bias[J]. IEEE Transactions on Affective Computing, 2022, 13(2): 881-893.

[25] MOLLAHOSSEINI A, HASANI B, MAHOOR M H. AffectNet: A database for facial expression, valence, and arousal computing in the wild[J]. IEEE Transactions on Affective Computing, 2019, 10(1): 18-31.

[26] BARSOUM E, ZHANG Cha, FERRER C C, et al. Training deep networks for facial expression recognition with crowd-sourced label distribution[C]//ICMI'16: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016: 279-283.

[27] LI Shan, DENG Weihong. Deep facial expression recognition: A survey[J]. IEEE Transactions on Affective Computing, 2022, 13(3): 1195-1215.

[28] WANG Kai, PENG Xiaojiang, YANG Jianfei, et al. Suppressing uncertainties for large-scale facial expression recognition [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA. IEEE, 2020: 6896-6905.

[29] LI Yong, ZENG Jiabei, SHAN Shiguang, et al. Occlusion aware facial expression recognition using CNN with attention mechanism [J]. IEEE Transactions on Image

Processing: A Publication of the IEEE Signal Processing Society, 2018, 28(5): 2439-2450.

[30] LI Shan, DENG Weihong, DU Junping. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 2852-2861.

[31] ZHU Ronghang, SANG Gaoli, ZHAO Qijun. Discriminative feature adaptation for cross-domain facial expression recognition [C]//2016 International Conference on Biometrics (ICB). Halmstad, Sweden: IEEE, 2016: 1-7.

[32] JI Yanli, HU Yuhan, YANG Yang, et al. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network[J]. Neurocomputing, 2019, 333: 231-239.

[33] LI Yingjian, GAO Yingnan, CHEN Bingzhi, et al. JDMAN: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition[C]// MM '21: Proceedings of the 29th ACM International Conference on Multimedia, 2021: 3312-3320.

[34] CHEN Tianshui, PU Tao, WU Hefeng, et al. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44 (12): 9887-9903.

[35] BUSSO C, BULUT M, LEE C C, et al. IEMOCAP: Interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): 335-359.

[36] BUSSO C, PARTHASARATHY S, BURMANIA A, et al. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception[J]. IEEE Transactions on Affective Computing, 2017, 8(1): 67-80.

[37] LOTFIAN R, BUSSO C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings[J]. IEEE Transactions on Affective Computing, 2017, 10(4): 471-483.

[38] LIU Na, ZONG Yuan, ZHANG Baofeng, et al. Unsupervised cross-corpus speech emotion recognition using domain-adaptive subspace learning [C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5144-5148.

[39] ZHANG Jiacheng, JIANG Lin, ZONG Yuan, et al. Cross-corpus speech emotion recognition using joint distribution adaptive regression[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 3790-3794.

[40] SAHU S, GUPTA R, SIVARAMAN G, et al. Smoothing model predictions using adversarial training procedures for speech based emotion recognition[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 4934-4938.

[41] FU Changzeng, LIU Chaoran, ISHI C T, et al. MAEC: Multi-Instance learning with an adversarial auto-encoder-based classifier for speech emotion recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6299-6303.

[42] LATIF S, RANA R, KHALIFA S, et al. Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition[J]. IEEE Transactions on Affective Computing, 2022.

[43] ZHENG Weilong, LU Baoliang. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks[J]. IEEE Transactions on Autonomous Mental Development, 2015, 7(3): 162-175.

[44] KOELSTRA S, MUHL C, SOLEYMANI M, et al. DEAP: A database for emotion analysis; using physiological signals [J]. IEEE Transactions on Affective Computing, 2012, 3(1): 18-31.

[45] BETHGE D, HALLGARTEN P, GROSSE-PUPPENDAHL T, et al. Domain-invariant representation learning from EEG with private encoders[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 1236-1240.

[46] JU Ce, GAO Dashan, MANE R, et al. Federated transfer learning for EEG signal classification[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2020: 3040-3045.

[47] CHEN Hao, JIN Ming, LI Zhunan, et al. MS-MDA: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition[J]. Frontiers in Neuroscience, 2021, 15: 778488.

[48] ZHENG Weilong, LU Baoliang. Personalizing EEG-based affective models with transfer learning[C]//Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 2016: 2732-2738.

[49] YIN Zhong, WANG Yongxiong, LIU Li, et al. Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination[J]. Frontiers in Neurorobotics, 2017, 11: 19.

[50] GANIN Y, LEMPITSKY V. Unsupervised domain adap-

tation by backpropagation [C]//International Conference on Machine Learning. PMLR, 2015: 1180-1189.

[51] ÖZDENIZCI O, WANG Ye, KOIKE-AKINO T, et al. Learning invariant representations from EEG via adversarial inference [J]. IEEE Access, 2020, 8: 27074-27085.

[52] HE Zhipeng, ZHONG Yongshi, PAN Jiahui. Joint temporal convolutional networks and adversarial discriminative domain adaptation for EEG-based cross-subject emotion recognition [C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 3214-3218.

[53] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 2962-2971.

[54] RAYATDOOST S, YIN Yufeng, RUDRAUF D, et al. Subject-invariant EEG representation learning for emotion recognition [C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 3955-3959.

[55] BLITZER J, DREDZE M, PEREIRA F. Biographies, Bollywood, boom-boxes and blenders: domain adaptation for sentiment classification [C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, 2007: 440-447.

[56] HE Ruidan, LEE W S, NG H T, et al. Adaptive semi-supervised learning for cross-domain sentiment classification [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 3467-3476.

[57] SOCHER R, PERELYGIN A, WU J Y, et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1631-1642.

[58] YU Jianfei, JIANG Jing. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 236-246.

[59] LI Liang, YE Weirui, LONG Mingsheng, et al. Simultaneous learning of Pivots and representations for cross-domain sentiment classification [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34 (5): 8220-8227.

[60] KENDALL A, GAL Y. What uncertainties do we need in Bayesian deep learning for computer vision?[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5580-5590.

[61] LI Zheng, ZHANG Yu, WEI Ying, et al. End-to-end adversarial memory network for cross-domain sentiment classification [C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia. California: International Joint Conferences on Artificial Intelligence Organization, 2017: 2237-2243.

[62] ZHUANG Fuzhen, CHENG Xiaohu, LUO Ping, et al. Supervised representation learning: Transfer learning with deep autoencoders [C]//IJCAI'15: Proceedings of the 24th International Conference on Artificial Intelligence, 2015: 4119-4125.

[63] YE Hai, TAN Qingyu, HE Ruidan, et al. Feature adaptation of pre-trained language models across languages and domains with robust self-training[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 7386-7399.

[64] DU Chunning, SUN Haifeng, WANG Jingyu, et al. Adversarial and domain-aware BERT for cross-domain sentiment analysis[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 4019-4028.

[65] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL-HLT, 2019: 4171-4186.

[66] WU Hui, SHI Xiaodong. Adversarial soft prompt tuning for cross-domain sentiment analysis [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 2438-2447.

[67] LESTER B, AL-RFOU R, CONSTANT N. The power of scale for parameter-efficient prompt tuning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 3045-3059.

[68] BALAHUR A, TURCHI M. Comparative experiments using supervised learning and machine translation for multi-

lingual sentiment analysis[J]. Computer Speech & Language, 2014, 28(1): 56-75.

[69] HAJMOHAMMADI M S, IBRAHIM R, SELAMAT A. Density based active self-training for cross-lingual sentiment classification[C]//Advances in Computer Science and its Applications. Berlin, Heidelberg: Springer, 2014: 1053-1059.

[70] ERIGUCHI A, JOHNSON M, FIRAT O, et al. Zero-shot cross-lingual classification using multilingual neural machine translation[EB/OL]. https://arxiv.org/pdf/1809.04686.pdf, 2018.

[71] XU Ruochen, YANG Yiming. Cross-lingual distillation for text classification[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017: 1415-1425.

[72] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1746-1751.

[73] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL]. http://www.csri.utoronto.ca/~hinton/absps/distillation.pdf, 2015.

[74] XU Yuemei, CAO Han, DU Wanze, et al. A survey of cross-lingual sentiment analysis: Methodologies, models and evaluations[J]. Data Science and Engineering, 2022, 7(3): 279-299.

[75] ZISER Y, REICHART R. Deep pivot-based modeling for cross-language cross-domain transfer with minimal guidance[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 238-249.

[76] SMITH S L, TURBAN D H P, HAMBLIN S, et al. Offline bilingual word vectors, orthogonal transformations and the inverted softmax[EB/OL]. https://arxiv.org/pdf/1702.03859.pdf, 2017.

[77] PIRES T, SCHLINGER E, GARRETTE D. How multilingual is multilingual BERT?[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4996-5001.

[78] CONNEAU A, LAMPLE G. Cross-lingual language model pretraining[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 7059-7069.

[79] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 8440-8451.

[80] FEI Hongliang, LI Ping. Cross-lingual unsupervised sentiment classification with multi-view transfer learning[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 5759-5771.

[81] BUSSO C, DENG Zhigang, YILDIRIM S, et al. Analysis of emotion recognition using facial expressions, speech and multimodal information[C]//ICMI '04: Proceedings of the 6th International Conference on Multimodal Interfaces, 2004: 205-211.

[82] GUNES H, SCHULLER B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions[J]. Image and Vision Computing, 2013, 31(2): 120-136.

[83] SRINIVASAN S, HUANG Zhaocheng, KIRCHHOFF K. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition [C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 6442-6446.

[84] ALBANIE S, NAGRANI A, VEDALDI A, et al. Emotion recognition in speech using cross-modal transfer in the wild[C]//Proceedings of the 26th ACM International Conference on Multimedia. Seoul, Republic of Korea. New York: ACM, 2018: 292-301.

[85] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 7132-7141.

[86] CAO Qiong, SHEN Li, XIE Weidi, et al. VGGFace2: A dataset for recognising faces across pose and age[C]// 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). Xi'an, China: IEEE, 2018: 67-74.

[87] CHATFIELD K, SIMONYAN K, VEDALDI A, et al. Return of the devil in the details: Delving deep into convolutional nets[C]// Proceedings of the British Machine Vision Conference 2014. Nottingham: British Machine Vision Association, 2014.

[88] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A large-scale speaker identification dataset[C]//Interspeech 2017. Stockholm, Sweden: ISCA, 2017: 2616-2620.

[89] LI Ruichen, ZHAO Jinming, JIN Qin. Speech emotion recognition via multi-level cross-modal distillation[C]// Interspeech 2021. Brno, Czechia: ISCA, 2021: 4488-4492.

[90] GHRISS A, YANG Bo, ROZGIC V, et al. Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 7347-7351.

[91] LIU Yinhan, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. https://arxiv.org/pdf/1907.11692.pdf%5C, 2019.

[92] LI Mao, YANG Bo, LEVY J, et al. Contrastive unsupervised learning for speech emotion recognition[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6329-6333.

[93] PENG Wei, HU Yue, XING Luxi, et al. Modeling intention, emotion and external world in dialogue systems[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 7042-7046.

[94] CERISARA C, JAFARITAZEHJANI S, OLUOKUN A, et al. Multi-task dialog act and sentiment recognition on Mastodon[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018: 745-754.

[95] KIM M, KIM H. Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances[J]. Pattern Recognition Letters, 2018, 101: 1-5.

[96] QIN Libo, CHE Wanxiang, LI Yangming, et al. DCR-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8665-8672.

[97] QIN Libo, LI Zhouyang, CHE Wanxiang, et al. CO-GAT: A co-interactive graph attention network for joint dialog act recognition and sentiment classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13709-13717.

[98] EKMAN P, FRIESEN W V. Facial action coding system: a technique for the measurement of facial movement [J]. Palo Alto, 1978, 3(2): 5.

[99] WANG Chu, ZENG Jiabei, SHAN Shiguang, et al. Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network[C]// 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019: 56-60.

[100] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. https://arxiv.org/pdf/1409.1556.pdf, 2014.

[101] KOLLIAS D. Abaw: Learning from synthetic data & multi-task learning challenges[EB/OL]. https://arxiv.org/pdf/2207.01138.pdf, 2022.

[102] ZHANG Tenggan, LIU Chuanhe, LIU Xiaolong, et al. Multi-task learning framework for emotion recognition in-the-wild[EB/OL]. https://arxiv.org/pdf/2207.09373.pdf, 2022.

[103] SHARMA M. Multi-lingual multi-task speech emotion recognition using wav2vec 2.0[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 6907-6911.

[104] CONNEAU A, BAEVSKI A, COLLOBERT R, et al. Unsupervised cross-lingual representation learning for speech recognition[C]//Interspeech 2021. Brno, Czechia: ISCA, 2021: 2426-2430.

[105] LI Yuanchao, ZHAO Tianyu, KAWAHARA T. Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning[C]// Interspeech 2019. Graz Austria: ISCA, 2019: 2803-2807.

[106] ZHANG Heran, MIMURA M, KAWAHARA T, et al. Selective multi-task learning for speech emotion recognition using corpora of different styles[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 7707-7711.

[107] GURURANGAN S, MARASOVIĆ A, SWAYAMDIPTA S, et al. Don't stop pretraining: Adapt language models to domains and tasks[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 8342-8360.

[108] SUN Chi, QIU Xipeng, XU Yige, et al. How to fine-tune BERT for text classification?[C]//China National Conference on Chinese Computational Linguistics. Springer, Cham, 2019: 194-206.

[109] LIU Pengfei, YUAN Weizhe, FU Jinlan, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2022.

[110] MAO Rui, LIU Qian, HE Kai, et al. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection[J]. IEEE Transactions on Affective Computing, 2022, PP(99):

1-11.

[111] BAEVSKI A, ZHOU Yuhao, MOHAMED A, et al. Wav2vec 2.0: A framework for self-supervised learning of speech representations[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 12449-12460.

[112] LIU A T, LI Shangwen, LEE H Y. TERA: self-supervised learning of transformer encoder representation for speech [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 2351-2366.

[113] LIU A T, YANG Shuwen, CHI Pohan, et al. Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6419-6423.

[114] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[115] MORAIS E, HOORY R, ZHU Weizhong, et al. Speech emotion recognition using self-supervised features[C]// ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 6922-6926.

[116] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//Interspeech 2020. Shanghai, China: International Speech Communication Association (ISCA), 2020: 3830-3834.

[117] WU Wen, ZHANG Chao, WOODLAND P C. Emotion recognition by fusing time synchronous and time asynchronous representations[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, ON, Canada: IEEE, 2021: 6269-6273.

[118] LI Yang, CHEN Ji, LI Fu, et al. GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition[J]. IEEE Transactions on Affective Computing, 2022, PP(99): 1.

[119] ZHAO Jinming, LI Ruichen, JIN Qin, et al. Memobert: pre-training model with prompt-based learning for multimodal emotion recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore. IEEE, 2022: 4703-4707.

[120] ZHONG Peixiang, WANG Di, MIAO Chunyan. Knowledge-enriched transformer for emotion detection in textual conversations[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 165-176.

[121] SPEER R, CHIN J, HAVASI C. ConceptNet 5.5: An open multilingual graph of general knowledge[C]//Thirty-first AAAI Conference on Artificial Intelligence, 2017: 4444-4451.

[122] MOHAMMAD S. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 174-184.

[123] XIE Yunhe, SUN Chengjie, JI Zhenzhou. A commonsense knowledge enhanced network with retrospective loss for emotion recognition in spoken dialog[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 7027-7031.

[124] SAP M, LE BRAS R, ALLAWAY E, et al. ATOMIC: an atlas of machine commonsense for if-then reasoning [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 3027-3035.

[125] GHOSAL D, MAJUMDER N, GELBUKH A, et al. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 2470-2481.

[126] BOSSELUT A, RASHKIN H, SAP M, et al. COMET: commonsense transformers for automatic knowledge graph construction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4762-4779.

[127] ZHU Lixing, PERGOLA G, GUI Lin, et al. Topic-driven and knowledge-aware transformer for dialogue emotion detection[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 1571-1582.

[128] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 3982-3992.

[129] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. 2015: 2692-2700.

[130] HU Xin, CHEN Jingjing, WANG Fei, et al. Ten challenges for EEG-based affective computing[J]. Brain Science Advances, 2019, 5(1): 1-20.

[131] YIN Yufeng, HUANG Baiyu, WU Yizhen, et al. Speaker-invariant adversarial domain adaptation for emotion recognition[C]// ICMI '20: Proceedings of the 2020 International Conference on Multimodal Interaction, 2020: 481-490.

[132] KIM D, TSAI Y H, ZHUANG Bingbing, et al. Learning cross-modal contrastive features for video domain adaptation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 13598-13607.

[133] XU Xinzhou, DENG Jun, CUMMINS N, et al. Exploring zero-shot emotion recognition in speech using semantic-embedding prototypes[J]. IEEE Transactions on Multimedia, 2022, 24: 2752-2765.

[134] WANG Xiaolong, YE Yufei, GUPTA A. Zero-shot recognition via semantic embeddings and knowledge graphs[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 6857-6866.

[135] LI Jingjing, JING Mengmeng, LU Ke, et al. Leveraging the invariant side of generative zero-shot learning[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 7394-7403.

[136] ZHANG Ziqi, LI Yuanchun, WANG Jindong, et al. ReMoS: reducing defect inheritance in transfer learning via relevant model slicing[C]//2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE). Pittsburgh, PA, USA: IEEE, 2022: 1856-1868.

[137] FENG Tiantian, HASHEMI H, ANNAVARAM M, et al. Enhancing privacy through domain adaptive noise injection for speech emotion recognition[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE, 2022: 7702-7706.

[138] MIRESHGHALLAH F, TARAM M, JALALI A, et al. Not all features are equal: Discovering essential features for preserving prediction privacy[C]//WWW '21: Proceedings of the Web Conference 2021, 2021: 669-680.

作者简介

黄兆培　男，1998年生，北京顺义人。中国人民大学信息学院博士研究生，主要研究方向为多模态情感识别、迁移学习等。

E-mail：huangzhaopei@ruc.edu.cn

张峰源　男，2000年生，江苏淮安人。中国人民大学信息学院硕士研究生，主要研究方向为多模态情感识别、迁移学习等。

E-mail：fy.zhang@ruc.edu.cn

赵金明　男，1992年生，山东青州人。启元实验室研究员，主要研究方向为情感计算、多模态人机交互等。

E-mail：zhaojinming@qiyuanlab.com

金　琴　女，中国人民大学信息学院教授，工学博士，博士生导师，主要研究方向为多媒体智能计算、人机交互等。

E-mail：qjin@ruc.edu.cn