INVITED PAPER

Toward Label-Efficient Emotion and Sentiment Analysis

This article introduces label-efficient emotion and sentiment analysis from the computational perspective, focusing on state-of-the-art methodologies, promising applications, and potential outlooks.

By Sicheng Zhao, *Senior Member IEEE*, Xiaopeng Hong, *Senior Member IEEE*, Jufeng Yang, *Member IEEE*, Yanyan Zhao, and Guiguang Ding, *Senior Member IEEE*

ABSTRACT | Emotion and sentiment play a central role in various human activities, such as perception, decision-making, social interaction, and logical reasoning. Developing artificial emotional intelligence (AEI) for machines is becoming a bottleneck in human–computer interaction. The first step of AEI is to recognize the emotion and sentiment that are conveyed in different affective signals. Traditional supervised emotion and sentiment analysis (ESA) methods, especially deep learning-based ones, usually require large-scale labeled training data. However, due to the essential subjectivity, complexity, uncertainty and ambiguity, and subtlety, collecting such annotations is expensive, time-consuming, and difficult in practice. In this article, we introduce label-efficient ESA from the computational perspective. First, we present a hierarchical taxonomy for label-efficient learning based on the availability of sample labels, emotion categories, and data domains during training. Second, for each of the seven paradigms, i.e., unsupervised, semisupervised, weakly supervised, low-shot, incremental, domain-adaptive, and domain-generalizable ESA, we give the definition, summarize existing methods, and present our views on the quantitative and qualitative comparison. Finally, we provide several promising real-world applications, followed by unsolved challenges and potential future directions.

KEYWORDS | Affective computing; artificial emotional intelligence (AEI); emotion and sentiment analysis (ESA); label-efficient learning.

I. INTRODUCTION

The experience of emotions significantly influences various aspects of our daily life, ranging from perception and decision-making to social interaction and logical reasoning. Because of the essential role of emotion in human–computer interaction, it is suspicious to regard machines to be intelligent without emotions [1]. Artificial emotional intelligence (AEI) endows machines with the ability to recognize emotions, generate and adapt emotions, and apply emotional information in goal accomplishment and problem-solving [2], [3]. With emotions, intelligent machines can provide humans, especially vulnerable groups, with humanistic interactions and high-quality service in real-world applications, such as companion robots and mental illness monitoring [4].

There are different modalities that humans use to express their emotions. Generally, these affective signals can be classified into two groups: explicit affective cues

Manuscript received 6 September 2022; revised 20 February 2023 and 28 June 2023; accepted 17 August 2023. Date of publication 15 September 2023; date of current version 11 October 2023. This work was supported in part by the CCF-DiDi GAIA Collaborative Research Funds for Young Scholars; in part by the National Natural Science Foundation of China under Grant 62076195, Grant 62176078, and Grant U1936202; in part by the National Key Research and Development Program of China under Grant 2018AAA0100400; in part by the Natural Science Foundation of Tianjin, China, under Grant 20JCJQJC00020; in part by the Fundamental Research Funds for the Central Universities under Grant AUGA5710011522; and in part by the Supercomputing Center of Nankai University (NKSC). *(Corresponding author: Sicheng Zhao.)*

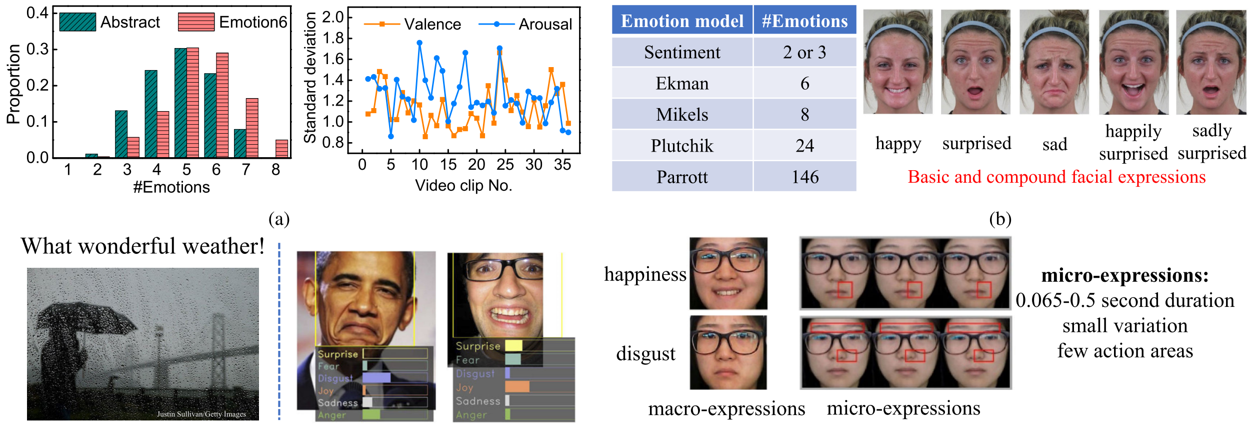
Sicheng Zhao is with the Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China (e-mail: schzhao@tsinghua.edu.cn).

Xiaopeng Hong and **Yanyan Zhao** are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: hongxiaopeng@hit.edu.cn; yyzhao@ir.hit.edu.cn).

Jufeng Yang is with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: yangjufeng@nankai.edu.cn).

Guiguang Ding is with the Beijing National Research Center for Information Science and Technology (BNRist) and the School of Software, Tsinghua University, Beijing 100084, China (e-mail: dinggg@tsinghua.edu.cn).

Digital Object Identifier 10.1109/JPROC.2023.3309299



Emotion model	#Emotions
Sentiment	2 or 3
Ekman	6
Mikels	8
Plutchik	24
Parrott	146

Fig. 1. ***Illustration of the essential properties of emotions. (a) Subjectivity: the emotional reactions among viewers to the same stimuli might differ significantly. Left: distribution of images on Abstract [9] and Emotion6 [10] datasets that are labeled with different emotion categories. It is clear that almost none of the images are labeled with only one emotion category. Right: standard deviations (STDs) on valence and arousal of the 58 subjects on the 36 video clips in the ASCERTAIN dataset [11]. (b) Complexity: different emotion categories are contained in different psychological models (left) and compound facial expressions also exist (right) [12]. (c) Uncertainty and ambiguity: the emotions the same content expresses might be uncertain and ambiguous [4], [13]. (d) Subtlety: variations of microexpressions are subtle with a short duration.***

and implicit affective stimuli [4]. The former group usually corresponds to specific physical changes in human bodies, such as facial expression, action, and physiological signals, while the latter group refers to the external signals, such as text, image, and video, that are widely used by humans to share their opinions on social networks.

The first step of AEI is to recognize the emotion and sentiment conveyed in the abovementioned affective signals, which is the focus of this article. Two groups of psychological models are often employed to measure emotions [5]. One group is categorical emotion states (CESs), which represent emotions using specific discrete categories, such as binary sentiment (positive versus negative) and Ekman's six emotions (anger, disgust, fear, happiness, sadness, and surprise) [6]. The other group is dimensional emotion space (DES), where emotions are represented with a continuous multidimensional Cartesian space, such as valence–arousal–dominance (VAD) [7]. Corresponding to the psychological models, different emotion and sentiment analysis (ESA) tasks can be performed: classification, regression, detection, retrieval, and distribution learning. The first four tasks can be conducted from the viewpoint of either affective signals or users [5], i.e., dominant emotion versus personalized emotion, while the distribution learning task is usually affective signal centric [8].

A general ESA framework includes three components: affective signal collection and annotation, feature extraction, and classifier learning. To train an effective ESA model, existing methods, especially deep learning-based ones, usually require large-scale labeled training data. However, due to the essential subjectivity, complexity, uncertainty and ambiguity, and subtlety (see Fig. 1), collecting such annotations with high quality is expensive, time-consuming, and difficult in practice.

1) *Subjectivity:* For given affective stimuli, the emotional reactions across viewers might be different [5]. Even the same viewer may react differently at different times. Viewers may have different physical and psychological changes when perceiving the same emotion [11].
2) *Complexity:* Emotion is becoming diverse and fine-grained. No consensus on how many emotion categories are necessary has been reached among psychologists [5]. The number of emotion categories in current psychological models ranges from 2 to over 100. Besides Ekman's basic facial expressions, there are also compound ones [12], such as happily surprised as shown in Fig. 1(b), which makes it difficult to collect sufficient data for all these categories.
3) *Uncertainty and Ambiguity:* The emotions that the same content expresses might be uncertain and ambiguous. For example, if one friend posted a tweet, "What wonderful weather!," we may infer that the friend is having fun outside on a sunny day, but if an image about a storm is attached, we can understand that the friend uses sarcasm in the textual content to reflect the bad mood. Simply using the dominant emotion category is insufficient to represent the blended facial expressions [13].
4) *Subtlety:* Variations among emotions might be subtle. For example, the duration of microexpressions is only between 0.065 and 0.5 s and the variation is very small in a few facial action areas [14].

How to perform ESA with noisy, limited, or even no labels? How to deal with incremental emotion categories and training samples? How to increase the transferability and generalizability of the trained models to new and unseen domains?

In this article, we attempt to answer these questions by introducing label-efficient ESA (LeESA) from the computational perspective with proper depth for both experts in this area and nonspecialists. First, we give a hierarchical taxonomy based on the training settings of sample labels, emotion categories, and data domains in Section II. Second, for each of the seven LeESA paradigms, we give a brief definition, summarize some representative and the latest methods, and compare them both quantitatively and qualitatively from Section III to Section IX. Third, we introduce some promising applications based on LeESA in Section X, followed by potential future directions in Secion XI. Finally, we conclude this article in Section XII.

There have been some other recent surveys, reviews, and tutorials on ESA, such as facial expression recognition (FER) [15], [16], [17], [18], microexpression recognition [14], [19], textual sentiment classification [20], [21], [22], [23], speech and music emotion recognition [24], [25], [26], affective image content analysis [5], [27], [28], bodily expressed emotion recognition [29], [30], emotion recognition from physiological signals [31], [32], [33], and multimodal emotion recognition [4], [34], [35], [36]. All these articles mainly cover ESA for a single specific modality from the perspective of supervised learning based on the assumption that sufficient training samples in the target domain are cleanly annotated with a predefined label set. Typically, these articles concentrate on reviewing dataset construction, emotional feature extraction, and classification strategy design.

Some other articles summarize and compare different label-efficient learning paradigms, such as unsupervised learning [37], [38], [39], semisupervised learning (SSL) [40], [41], low-shot learning [42], [43], [44], incremental learning [45], [46], [47], weakly supervised learning [48], [49], transfer learning [50], [51], domain adaptation (DA) [52], [53], [54], [55], and domain generalization (DG) [56], [57]. These articles are mainly presented from the viewpoint of machine learning with applications in traditional computer vision (CV) and natural language processing (NLP) tasks. Directly applying these methods to ESA cannot guarantee to perform well without considering the essential properties of emotions. There exist clear differences between label-efficient methods designed for ESA and other general tasks.

1) For unsupervised ESA, instead of learning general semantic information [58], emotional information conveyed by social data, such as sentiment words and emoticons, is utilized in many LeESA methods [59], [60].
2) For semisupervised ESA, the intrinsic ambiguity of emotions leads to low accuracy of pseudo label [61]. Therefore, it is necessary to combine the prior knowledge (e.g., polarity) and effective techniques (e.g., label smoothing) for training LeESA [62].
3) In terms of weakly supervised ESA, most general tasks solve coarse label problems, such as instance-level annotation for detection and segmentation [48], [49]. However, the LeESA approaches are mainly concerned with the problem of learning with noisy clues [63], [64].
4) For low-shot ESA, the complexity caused by a large number of visual concepts makes it more challenging than the general tasks [5]. To address this issue, prior knowledge, such as adjective–noun pairs (ANPs) for visual concepts is usually introduced [65].
5) For incremental ESA, due to the subjectivity of emotions, the data may have severe bias. Therefore, ensemble learning that integrates multiple models is leveraged in ESA methods [66].
6) For DA and DG, although discrepancy-based methods have good theoretical guarantees, the complexity leads to suboptimal performance when directly minimizing the distance between distributions for ESA. In contrast, the methods that utilize emotion cues to design self-supervised tasks achieve better performance [67].

Differently, this article aims to hierarchically organize different label-efficient paradigms, comprehensively introduce state-of-the-art methodologies as well as promising applications for ESA without explicitly distinguishing different modalities, and rationally provide potential outlooks for future research.

II. LeESA TAXONOMY

In this section, we define a hierarchical taxonomy of LeESA and compare different paradigms with the traditional supervised learning setting.

Let $\mathbf{x}$ and y that are drawn from a given data distribution $P(\mathbf{x}, y)$, respectively, represent the affective signal and emotion label.[1] For a given target domain, suppose that the distribution is $P_T(\mathbf{x}, y)$, and the target dataset drawn from P_T is $D_T = \{(\mathbf{X}_T, Y_T)\} = \{(\mathbf{x}_T^j, y_T^j)\}_{j=1}^{N_T}$. Similarly, if we have another source domain that is related to the target domain, we can have source distribution $P_S(\mathbf{x}, y)$ and source dataset $D_S = \{(\mathbf{X}_S, Y_S)\} = \{(\mathbf{x}_S^i, y_S^i)\}_{i=1}^{N_S}$. Here, $\mathbf{x}_S^i$ and $\mathbf{x}_T^j$ are the observed affective signals in the source and target domains, respectively, N_S and N_T are the numbers of source and target samples, respectively, and y_S^i and y_T^j are the corresponding emotion labels. Suppose that the label sets in the source and target domain are $\mathcal{C}_S$ and $\mathcal{C}_T$, respectively. Let N_T^L and Y_T^L denote the number of labeled training samples in the target domain and corresponding labels, respectively. Please note that our final task is to perform ESA on the target test set. The traditional supervised ESA is usually performed under such settings: the training set of the target domain is fully labeled (i.e., $N_T^L = N_T$), emotion categories are fixed ($\mathcal{C}_T$ is defined in advance) and provided once, the training samples are provided once,

[1] For example, $\mathbf{x}$ can be facial expressions, text, images, and any other modalities that are used to express emotions; y can be discrete emotion categories for emotion classification, continuous VAD values for emotion regression, and probability distributions for emotion distribution learning.

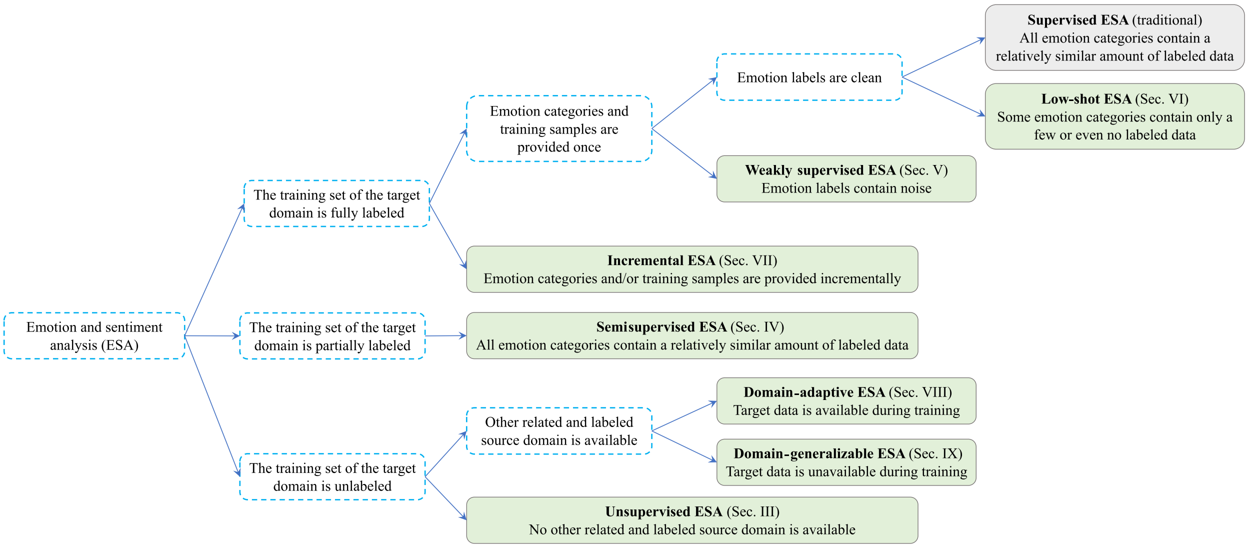

Emotion and sentiment analysis (ESA)
The training set of the target domain is fully labeled
The training set of the target domain is partially labeled
The training set of the target domain is unlabeled
Emotion categories and training samples are provided once
Emotion labels are clean
Supervised ESA (traditional) All emotion categories contain a relatively similar amount of labeled data
Low-shot ESA (Sec. VI) Some emotion categories contain only a few or even no labeled data
Weakly supervised ESA (Sec. V) Emotion labels contain noise
Incremental ESA (Sec. VII) Emotion categories and/or training samples are provided incrementally
Semisupervised ESA (Sec. IV) All emotion categories contain a relatively similar amount of labeled data
Other related and labeled source domain is available
Domain-adaptive ESA (Sec. VIII) Target data is available during training
Domain-generalizable ESA (Sec. IX) Target data is unavailable during training
Unsupervised ESA (Sec. III) No other related and labeled source domain is available

Fig. 2. ***Hierarchical taxonomy of LeESA based on the availability of sample labels, emotion categories, and data domains during training.***

the emotion labels are clean, and all emotion categories contain a relatively similar amount of labeled data.

In real-world applications, such settings are difficult to meet. Thus, it is imperative to study LeESA. Here, we define seven LeESA paradigms.[2]

1) *Unsupervised ESA:* The training set of the target domain is unlabeled (i.e., $N_T^L = 0$), and no other related and labeled source domain is available ($\mathbf{X}_T$ is available during training).

2) *Semisupervised ESA:* The training set of the target domain is partially labeled (i.e., $N_T^L < N_T$), and all emotion categories contain a specific amount of labeled data ($\mathbf{X}_T$ and Y_T^L are available during training, and different categories in $\mathcal{C}_T$ are labeled with a relatively similar amount).

3) *Weakly Supervised ESA:* The training set of the target domain is fully labeled, emotion categories and training samples are provided once, and emotion labels contain noise ($\mathbf{X}_T$ and unclean Y_T are available during training).

4) *Low-Shot ESA:* The training set of the target domain is fully labeled, emotion categories and training samples are provided once, and some emotion categories contain only a few or even no labels ($\mathbf{X}_T$ and Y_T^L are available during training, and some categories in $\mathcal{C}_T$ contain few or no labels).

5) *Incremental ESA:* The training set of the target domain is fully labeled, and emotion categories and/or training samples are provided incrementally ($\mathbf{X}_T$ and Y_T are available during training, and $\mathcal{C}_T$ and/or $\mathbf{X}_T$ are incremental).

6) *Domain-Adaptive ESA:* The training set of the target domain is unlabeled, another related and labeled source domain is available, and the target data are available during training ($\mathbf{X}_S, Y_S$, and $\mathbf{X}_T$ are available during training).

7) *Domain-Generalizable ESA:* The training set of the target domain is unlabeled, another related and labeled source domain is available, and the target data are unavailable during training ($\mathbf{X}_S$ and Y_S are available during training).

A brief hierarchical taxonomy of LeESA is shown in Fig. 2. For the seven paradigms, the methods for a specific modality are beneficial for other modalities. First, some methods can be directly transferred to another modality to solve similar problems. In particular, in the emotion and sentiment clue-based methods in Section III, these clues (e.g., sentiment words and emoticons) are leveraged as auxiliary knowledge for training. These methods can be easily transferred to other modalities to address the problem of lacking annotation [63], [68]. Second, although some techniques based on the data distribution cannot be transferred, e.g., pixel- and word-level preprocessing, the strategy designed for a specific problem is insightful for other modalities. For instance, the regularization-based semisupervised strategies in speech [69] are utilized to address the nonstationary and multirhythm of EEG [70], and sample reweighting designed for DA of NLP tasks [71] is also leveraged for FER [72].

III. UNSUPERVISED ESA

Although there has been significant progress toward ESA, most existing methods mainly rely on large-scale manually labeled corpora. Unfortunately, emotion and sentiment labels are not prelabeled, or even nonexistent, in most practical situations, and are scarce, especially for complex fine-grained ESA tasks, such as the aspect-based textual sentiment analysis.

[2] For simplicity, we focus on the most popular situations when defining these LeESA paradigms. They can be modified and combined in real-world applications. More situations for these seven paradigms can be found in Sections III–IX.

Table 1 Categorization and Representative Methods for Unsupervised ESA

Category	Methods		References
Emotion and sentiment clues-based methods	Sentiment words clues		[59, 74–76]
	Emoticons clues		[77]
	Target-sentiment word pairs clues		[78, 79]
	Clues between text and image		[80]
Self-supervised learning-based methods	Generative pre-training-based ESA	Casual language modeling	[81]
	Contrastive pre-training-based ESA	Cross-modal contrastive learning	[82, 83]
	Predictive pre-training-based ESA	Masked language modeling	[84]
		Replaced token detection	[85]
		Cross-modal matching	[86]

Leveraging massive amounts of unlabeled data becomes an opportunity for ESA. There are currently two technical routes. One route is to use the rich artificially defined emotion and sentiment clues without any machine learning or deep learning techniques, which are called emotion and sentiment clue-based methods. The other route is to use the self-supervised learning framework to implicitly capture semantic information from the well-designed pre-training tasks without labeled data, which are called self-supervised learning-based methods. We will introduce the algorithms of these two routes in the following subsections, and we summarize them in Table 1.

A. Emotion and Sentiment Clue-Based Methods

Each of the modalities we deal with contains a lot of emotional clues, such as the emotional words in the text, the facial expressions of humans in the image, and the accent in the speech. Compared with image and speech, text is more capable of expressing emotions independently [73]. Specifically, visual features (e.g., color histograms and visual attributes) lack direct emotional semantic clues. Similarly, emotional features that are purely extracted from speech are more ambiguous. Typically, there are two categories of emotion and sentiment clue-based ESA methods. One is the unsupervised algorithms that target only text modality, and the other is the unsupervised algorithms for multiple modalities that use the sentiment clues hidden in textual information to connect visual or audio features with sentiment labels.

The rich emotion and sentiment clues hidden in text are introduced in detail as follows.

1) Sentiment Words: The words that can directly express sentiment and emotion are called sentiment words. According to the sentiment category, the sentiment words can be divided into positive (such as “good” and “perfect”) and negative (such as “bad” and “terrible”). Also, according to the emotion category, they can be divided into happy (such as “delighted” and “joyful”), sad (such as “unhappy” and “disappointed”), angry (such as “overheated” and “rage”), and other emotions. There are a lot of research works to build the sentiment dictionaries, such as WordNet [74] and SentiWordNet [75]. Many dictionaries not only list the words of various emotions but also include the intensity of emotions [75]. In addition, the emotional inversion role of negation words in the text is considered [76].

2) Emoticons: Emoticon that is short for “emotion icon” refers to a pictorial representation or text format of facial expressions. Emoticons are now being widely used in our daily life to directly express our emotions. For example, “:(” conveys bad emotion. People prefer to use emojis instead of words to express their emotions, especially on social media. Many researchers use emojis to automatically construct large-scale naturally labeled corpora [77]. Similarly, emojis are explicit clues that often appear in text to express emotions.

3) Target–Sentiment Word Pairs: Sometimes it is reckless to directly determine the emotion of a text based on emotional words or emojis. For example, the word “well” does not express a positive meaning in all contexts. Some researchers [78] proposed that the collocation of the sentiment word (“well”) and its corresponding target (“play”) can better convey emotional clues.

Based on the above rich sentiment and emotion clues, most researchers weigh these emotional clues and sum them to determine the emotional polarity of the text [87]. Paltoglou and Thelwall [59] proposed an intuitive, less domain-specific, unsupervised, and lexicon-based approach to estimate the level of emotional intensity in order to make predictions. The approach is appropriate for subjectivity detection and sentiment polarity classification, which are two complementary tasks. The proposed algorithm outperforms supervised algorithms in the majority of experiments on the Twitter, MySpace, and Digg datasets. Hu et al. [60] proposed a method by counting the word frequency in the user description and predicted the sentiment by measuring the word’s sentiment. Other scholars use these textual emotion clues to simulate manually annotated sentiment labels, to automatically train emotion classifiers. Zeng et al. [79] used target–opinion word pairs as a supervision clue to learn a sentiment classifier instead of labeled training data. The target–sentiment word pairs are extracted by using dependency parsers and several simple handcrafted rules, and they are used as supervision clues, which are very flexible for sentiment classification tasks in different granularities. This method outperforms unsupervised baselines and obtains comparable results to supervised methods in the customer reviews domain and clinical narratives domain.

The advantage of this type of approach is that it is more practical, and the disadvantage is that heuristic rules need to be artificially specified and are not easy to extend to new datasets. What is more, because such models do not have rich and explicit sentiment supervisory clues to guide, the model's ability to learn emotional expression is weak.

For multimodal data, we can also use the emotional clues from the text to assist in identifying the emotions of other modalities, such as the emotions expressed in images. The success of the unsupervised techniques for sentiment analysis on social media images is based on the strong assumption that the (visual, textual) pair shares the same sentiment polarity. In detail, there is always text around the image, and the sentiments of the text and image are consistent. This also means that the emotional clues in the text can be used to identify the sentiment of the image. In [80], an unsupervised sentiment analysis framework is proposed to exploit relations among visual concepts and relevant contextual information for sentiment analysis of social images.

However, the common situations of metaphor, sarcasm, and implicit expression of sentiment are ignored under such a scenario, resulting in the mismatching problem of sentiment information between images and the corresponding texts. Thus, more social media resources, such as link information, user history, geolocation, and nationality, are supposed to be exploited for more accurate sentiment analysis.

B. Self-Supervised Learning-Based Methods

As a new type of machine learning method, self-supervised learning has received more and more attention. The so-called self-supervised learning is to obtain supervision clues through the input data itself without manual labels. Instead of obtaining manually labeled data, there are many ways to obtain pseudo "labeled" data, such as masking a portion of the data and then using the remaining visible portion to predict the masked portion; randomly modifying the data or adding noise to the data and then learning a restoration or denoising model to restore the corrupted data to the original data; augmenting the data; selecting and forming the semantically unchanged data as a positive example with the original data; and then randomly selecting from the other data to form a negative example. Compared with the model without self-supervised pretraining, the self-supervised learning-based ESA approaches can obtain better performance for downstream tasks since the pretrained model trained on a large scale of "labeled" data has better generalization ability.

Specifically, the self-supervised learning-based ESA can be divided into three categories.

1) Generative Pretraining-Based ESA: Generative pretraining-based models learn the representation in the pretraining phrase by autoregressively predicting the next target given the previous input. Specifically, generative pretrained transformer (GPT) utilizes casual language modeling, which is a well-known objective for training a language model, to pretrain the parameters. It is well worth trying to utilize the pretrained model, such as GPT3 [81], to conduct unsupervised ESA. Brown et al. [81] adopted the prompt, which is a text paragraph describing the task to enable the model to generate the answer.

2) Contrastive Pretraining-Based ESA: The core idea of contrastive pretraining is to learn how to distinguish between different data. During unsupervised contrastive pretraining, the unlabeled texts/images are clustered in the latent space, forming fairly good decision boundaries between different classes [82]. Contrastive pretraining-based models utilize the relationship between samples as the supervision information to train an ESA model. For example, contrastive language–image Pre-training (CLIP) [83] leverages the relationship between the image and its textual caption to learn the representation. This method is called cross-modal contrastive learning, which aims to pull close the representations for the positive text–image pairs and pull away the negative pairs. As this pretraining objective enables the model to output the relevance score between an image and a text, we can construct different prompts such as "it makes someone happy/sad/" and utilize the pretrained model to generate the relevance scores. The statement with the highest score is considered to be true.

3) Predictive Pretraining-Based ESA: There are various predictive pretraining objectives, including masked language modeling [84], replaced token detection [85], masked vision modeling [88], and cross-modal matching [86]. Masked language modeling is introduced by bidirectional encoder representations from transformers (BERT) [84], which makes the model predict the masked word based on the context. To apply BERT on unsupervised ESA, Shin et al. [89] constructed the prompt following GPT3. They used [MASK] as the placeholder and let BERT predict the corresponding word. The prompt could be like "it was [MASK]." To mitigate the mismatch between pretraining and fine-tuning of BERT, the replaced token detection is proposed in ELECTRA [85], which lets the model discriminate whether the token is replaced. The prompting method for ELECTRA [90] is similar to it for BERT. The difference lies in using different classification labels to replace the [MASK] token separately and making ELECTRA discriminate whether the filled word is replaced. If the output result is "original," this answer will be accepted. This method obtains 82.8% in terms of accuracy on SST-2. As for multimodal data, different from the mask language modeling, masked vision modeling samples and masks the visual features of vision regions or patches, and lets the vision-language model reconstruct the masked features. Also, cross-modal matching is presented to capture the inherent relationship between different modalities, which makes the model predict 1 for the matching pairs and 0 for others. Similar to CLIP, the matching scores can be calculated and the statement with the highest matching score is taken as the final answer. An example of masked language

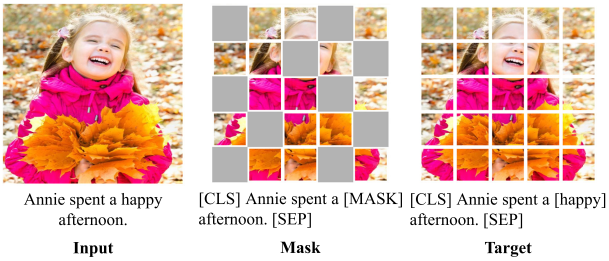

Annie spent a happy afternoon.
Input
[CLS] Annie spent a [MASK] afternoon. [SEP]
Mask
[CLS] Annie spent a [happy] afternoon. [SEP]
Target

Fig. 3. ***Example of masked language modeling and masked vision modeling in self-supervised learning.***

modeling and masked vision modeling for image–text pair data is shown in Fig. 3.

In recent years, some predictive pretraining models have also achieved good performance in fine-tuning downstream ESA tasks, such as SentiBERT [91], SentiLARE [92], and SKEP [93]. They build emotion-related pretraining tasks at the word level, by introducing word-level emotional knowledge, such as emotional words, part-of-speech, or part-of-speech parse trees, in the mask language model (MLM) task of BERT. However, they ignore the design of pretraining tasks at the sentence level. Also, SentiWSP [94] designs pretraining tasks at both word level and sentence level to enhance the model's ability to capture sentiment information in text and uses a word-level sentiment substitution detection task to enhance the discriminator's learning of sentiment information in the text through joint training of the generator and discriminator. The discriminator is then trained at the word level to improve its ability to capture the sentiment of the whole sentence through a contrastive learning framework.

Self-supervised learning was first used in NLP tasks and achieved impressive performance, and in recent years, it has also made amazing progress in CV tasks. Compared with other unsupervised learning methods for ESA, self-supervised learning-based algorithms are most worth trying.

IV. SEMISUPERVISED ESA

For ESA, due to subjectivity and ambiguity, it is challenging to construct a large-scale dataset with reliable annotation [4], [5]. Recent research utilizes major voting to annotate the emotional datasets. Many workers are employed to provide their emotional responses to given affective stimuli, and the class with the most votes is labeled as the ground truth. This strategy alleviates the subjectivity and ambiguity issue via the generally considered dominant emotion, but the reliability of the labels depends on a large number of workers. In this way, semisupervised ESA is proposed to be a promising direction due to the low demand for labeled data. On the one hand, training models in the semisupervised setting can significantly reduce the cost of annotation. On the other hand, semisupervised algorithms [115] progressively select high-confidence samples for training, which potentially alleviates the impact of unreliable data. Therefore, semisupervised ESA has attracted increasing attention.

SSL aims to design algorithms that learn from both labeled and unlabeled data. In terms of ESA, there are three commonly used strategies to train robust models in the SSL settings, i.e., graph-based methods, regularization-based methods, and pseudo-label-based methods, as shown in Fig. 4 and summarized in Table 2. Next, we will introduce these three groups of methods.

A. Graph-Based Methods

The graph-based SSL methods focus on the geometry relations induced by labeled and unlabeled data. Generally, the geometry relation is modeled by a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where nodes $\mathcal{V}$ are the samples and edges $\mathcal{E}$ are the similarities between the samples. After mining the intrinsic relations among samples, it is possible to propagate information from a few labeled data through the graph structure.

In the early days, Internet shopping was not popularized, and there was little rating data for the review of products. In order to automatically analyze the user's satisfaction with products, Goldberg and Zhu [95] adopted an SSL algorithm to make use of reviews with and without ratings. The algorithm is optimized based on the assumption

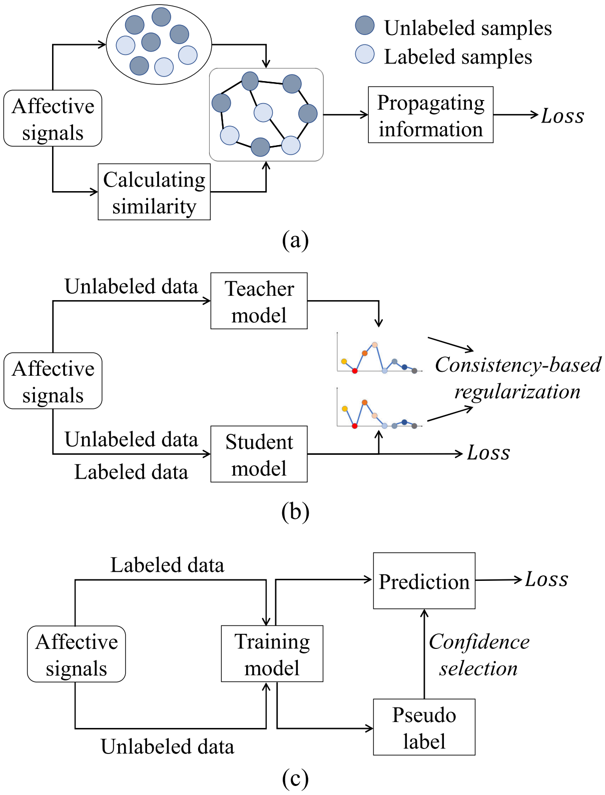


Fig. 4. ***Conceptual illustration of semisupervised ESA: (a) graph-based method, (b) regularization-based method, and (c) pseudo-label-based method.***

Table 2 Categorization and Representative Methods for Semisupervised ESA

Category	Methods		References
Graph-based strategy	Structured label propagation		[95–97]
Regularization-based strategy	Generative method	GAN/VAE-based constraint	[69, 70, 98, 99]
		Auxiliary identification task	[100]
		Multi-modal data complement	[101–103]
	Consistency regularization	Label-oriented teacher-student model	[104]
		Cross-modal consistency	[105]
Pseudo label-based strategy	Pseudo label calibration	Sentiment knowledge guidance	[62, 106–109]
		Ensemble models	[110]
	Robust representation learning	Improving sampling process	[111–113]
		Robust training process	[114]

that the function predicting ratings should be smooth with respect to the graph. Specifically, the smoothness is measured by the L_2 distance, which is constrained within the labeled data, within the unlabeled data, and between the labeled and unlabeled data. Different from document-level coarse-grained tasks, many researchers are interested in utilizing SSL algorithms to address the expensive cost of fine-grained word-level annotation. Rao and Ravichandran [96] compared the performance of three classical graph-based SSL algorithms, i.e., mincuts, randomized mincuts, and label propagation on the word-level sentiment polarity analysis. The experimental results demonstrate that mincuts and randomized mincuts have an advantage on the recall, but the precision is lower than label propagation. Sindhwani and Melville [97] constructed graph $\mathcal{G}$ with $(n + D)$ vertices, which consists of n documents and D words. The undirected edge (i, j) exists if the jth word appears in the ith document, and the weight is calculated by frequency. Then, the graph is optimized based on three rules: 1) the label of documents should be close to ± 1 value ($+1$ represents positive and -1 represents negative); 2) the label of words should be close to ± 1 value; and 3) the label of word i and document j should be similar if the edge weight of i and j is large. Overall, the graph can be completed and the sentiment value of the words can be predicted in the training process.

B. Regularization-Based Methods

Due to the challenge of obtaining a reliable large-scale affective dataset, the number of initial labeled samples is insufficient. Furthermore, in some modalities, such as EEG and speech, with weak, nonstationary, and multirhythm properties, the samples contribute differently to the emotional states [116]. Training on these datasets easily leads to an overfitting problem. Therefore, researchers design regularization strategies to improve the generalization ability of the models.

Latent-variable model is an effective generative SSL strategy with the ability to learn intrinsic representations with the help of the latent variables. Variational autoencoder (VAE) is a representative latent-variable model, which contains a probabilistic encoder and decoder. Kingma et al. [98] extended VAE to SSL combined with a classifier trained simultaneously. This scheme is leveraged to address the instability of the speech and EEG sentiment analysis [69], [70], [99]. Moreover, Latif et al. [100] integrated gender identification and speaker recognition as auxiliary tasks to further regularize the network. In addition, Kim et al. [101] designed a 3-D autoencoder for video emotion analysis to encourage the network to learn spatial–temporal representation. For multimodal emotion analysis, Du et al. [102] leveraged a generative model to address the missing modality problem. Lian et al. [103] introduced an intramodal interactive module to maintain the information from each modality by leveraging multimodal features to reconstruct the supervision from each modality.

Consistency-regularization methods apply a consistency loss term to specify the prior constraints [117], which is one of the mainstream SSL algorithms. Specifically, a popular assumption is that the predictions should be similar when the inputs are the perturbed versions of the same image [115]. The typical consistency method is the teacher–student model, which forces the predictions of the student model to be consistent with those of the teacher model. The constraint is formally defined as

$$\min_{x \in X} \mathcal{D}\left(h_{\text{stu}}\left(f_{\text{stu}}\left(x\right)\right), h_{\text{tea}}\left(f_{\text{tea}}\left(x\right)\right)\right) \tag{1}$$

where $f_{\text{stu}}(\cdot)$, $f_{\text{tea}}(\cdot)$, $h_{\text{stu}}(\cdot)$, and $h_{\text{tea}}(\cdot)$ represent the student feature extractor, teacher feature extractor, student classifier, and teacher classifier, respectively, and $\mathcal{D}(\cdot, \cdot)$ measures the distance between the distributions, which is often implemented by mean squared error. The student model is optimized according to the constraint. In turn, the teacher model is updated by the exponential moving average (EMA) strategy with the help of the parameters of the student model.

For ESA, Liang et al. [105] proposed a cross-modal distribution matching module, which assumes that the emotional state of the internal modality should be consistent with the text on the utterance level. Specifically, the module is designed for large-scale unlabeled data, and the maximum mean discrepancy (MMD) is leveraged to measure the consistency. The calculated MMD value is directly leveraged as a loss term to optimize the model.

To minimize the entropy of the prediction [118], Zhang and Etemad [104] sharpened the guessed distribution for the unlabeled data. In addition, consistency regularization is also integrated into the graph-based methods. After constructing the graph $\mathcal{G}$, a loss function is designed based on the consistency assumption that the data with close relation have a similar prediction [95], [97]. Combining the advantages of consistency constraints and geometry relations, these methods show competitive performance on ESA.

C. Pseudo Label-Based Methods

Due to its powerful performance and superior generalization ability, pseudo label becomes one of the most popular methods for SSL [115], [119], [120]. The classical algorithms generate pseudo labels by selecting the category that has the maximum probability of the predictions. However, directly adopting pseudo labels will suffer from confirmation bias [121]. Specifically, training with the wrongly predicted labels can impact the performance of the model and the error will be accumulated in this process. Therefore, researchers focus on designing reliable methods to measure the confidence of the pseudo labels [115], [122]. The commonly used cross-entropy-based loss for unlabeled data can be denoted as follows:

$$\mathcal{L}_{\text{unlab}} = \mathbb{1}_{\left[\text{prob}\left(q_u^w\right)\geq\tau\right]} \sum_{c=1}^{C} \mathbb{1}_{\left[c=q_u^w\right]} \log\left(h^c\left(f\left(x_u^s\right)\right)\right) \qquad (2)$$

where $q_u^w = \arg\max(f(x_u^w))$ denotes the pseudo label from the weakly augmentation of unlabeled sample x_u and τ is the fixed threshold to divide reliable or unreliable pseudo labels. Based on the methods, the high-confident unlabeled data can be selected and combined into the training process with labeled data.

Generating pseudo labels is a key step for these methods, and some strategies have been proposed for ESA. Considering the changing process of the sentiment, Rong et al. [106] integrated the sentiment state vector into the hidden layer to improve the recurrent neural networks (RNNs), which implicitly enables the model to generate more accurate pseudo labels. Benefiting from the pretrained model which understands semantic information, Kumar et al. [62] leveraged BERT to extend the vocabularies for each emotion based on the seed. Then, the pseudo labels are assigned according to these vocabularies. Moreover, some methods combine decisions from many models to obtain more reliable pseudo labels. Xiang and Zhou [110] trained a separate sentiment model for each topic cluster and replaced the single model with a mixture of sentiment models. Then, the mixed prediction is obtained based on the weight of each topic. For unlabeled data, the pseudo labels are generated from the ensemble decisions. Zhang and Singh [111] leveraged forward and backward segment learner to model the mapping between data and label, and the pseudo label is selected from the learner with a larger probability value. In the real world, the number of emotion and sentiment stimulation for classes is imbalanced, and thus, Li et al. [112] improved the sampling strategy to form new training subsets. In addition, Sintsova et al. [107] utilized the exponential operator to generate the weight of each class and then reweighted the distribution to cope with the imbalance problem. Besides the generation process, the measurement of pseudo-label confidence is also an important step. Hwang and Lee [113] automatically constructed lexicon when training with labeled data. Then, the confidence score of the unlabeled data is calculated according to the lexicon. For multimodal ESA, the current methods mainly pay attention to robust representation. Zhang et al. [108] leveraged early fusion to combine the audio and visual features. Li [109] designed a hierarchical fusion approach to leverage multimodal information, which is helpful to correctly model the uncertainty. In addition, Zhou et al. [114] addressed the wrongly predicted pseudo labels by curriculum learning. Specifically, the training process first leverages the strong and balanced emotion samples and subsequently utilizes the weak and imbalanced emotion ones.

D. Discussion

Here, we qualitatively discuss the advantages and disadvantages of the three types of representative SSL algorithms for ESA. First, the graph-based algorithms mine the association among samples, which clearly models the structure of the dataset [95], [97]. However, these methods usually have two limitations. First, as the vertices represent samples in the dataset, the space cost of n samples is $O(n^2)$. Second, when new samples are added, these methods have to reconstruct the graph by combining both the original and new data to better and comprehensively model the relations among the data. The regularization-based methods have made much progress for SSL in recent years [117], [121]. Essentially, these methods are designed to prevent models from overfitting on small-scale labeled datasets. Nevertheless, the regularization strategies often ignore the mapping between unlabeled data and the potential label. In light of this consideration, pseudo-label-based methods attract more and more attention due to their practicality and simplicity [115]. For ESA, because of the subjectivity and ambiguity, the low accuracy of pseudo labels and the small proportion of high-confidence data are two main challenges for the SSL setting. Current methods focus on extracting discriminative features [103], [108] or designing curriculum learning tasks [114], [116] to alleviate the problems. In the future, label smoothing guided by the polarity may be a promising direction to further improve the performance of SSL-based ESA.

V. WEAKLY SUPERVISED ESA

Weakly supervised ESA has received much attention since the dependence on labeled data is greatly reduced. In

Table 3 Categorization and Representative Methods for Weakly Supervised ESA

Category	Methods		References
Initialization-based method	Efficient Emotional Cue	Emotion related information	[63, 64]
		Emotional seed	[68, 123]
	Metric Diffusion	Prototype-based generation	[124–126]
	Probabilistic Model	Conditional generation model	[127–130]
	Feature Generation	Fine-grained reweighting	[131, 132]
Refinement-based method	Feature Calibration	Contrastive constraint	[133, 134]
		Robust architecture	[135, 136]
	Iterative Adjustment	Reliable sample selection	[49, 137, 139, 140]
		Iterative training process	[64, 140, 141]

detail, the weakly supervised methods are proposed to address the two practical challenges, as shown in Fig. 5. First, there exist potential relations among affective tasks, and thus, training multitasks simultaneously can prompt the robustness and performance of the model. For example, the aspect and opinion are often mined from the documents simultaneously [129], [130], [142]. Furthermore, detecting affective region is also optimized together with the emotion recognition [131]. However, there is no dataset simultaneously containing labels for these tasks. In order to provide the supervision information for each task, weakly supervised ESA has become an important research topic. Second, the noisy labels for ESA are easy to obtain, and thus, it is a promising topic to mine emotional cues from these data. There are many websites that provide reviews together with ratings, which can be divided into either positive or negative sentiment by a fixed threshold. Furthermore, the words also play an important role in weakly supervised ESA [126], [143]. There are some text-based tools such as SentiWordNet [144] and SentiStrength [145], which calculate the sentiment values of the words, and thus, many works leverage the value of the words in the sentence to obtain noisy labels.

In detail, current weakly supervised ESA works focus on the strategy of learning from inconsistent and noisy labels. The inconsistent labels refer to the inconsistency between fine-grained samples and coarse-grained supervision information, e.g., sentiment analysis at the word level with annotation at the document level, affective region at the pixel level with emotion label at the image level, and emotion analysis at the frame level with emotion label at the video level. The noisy labels are particularly common in ESA. For example, many methods obtain annotation from users' ratings. Specifically, the polarity label can be generated according to a fixed threshold. The higher ratings denote a positive opinion, and the lower rating denotes a negative experience. Nevertheless, review ratings are not reliable labels for the constituent sentences, which results in noisy supervision information.

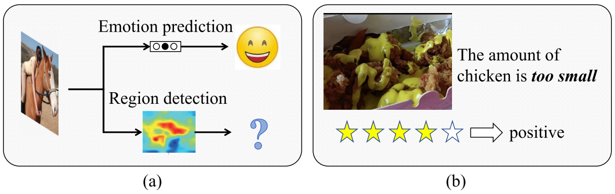


Fig. 5. *Illustration of the motivations for utilizing weakly supervised ESA methods. (a) Inadequate annotation for multiple tasks: both emotion prediction and emotional region detection are trained simultaneously, but only the emotion label is available. (b) Noisy labels: a food review is assigned with a positive label based on the rating, but actually, the comment is negative.*

Next, we will illustrate these methods in the following sections. These weakly supervised ESA methods can be divided into two groups: initialization-based methods and refinement-based methods, as summarized in Table 3. The initialization-based methods aim to generate supervision information with the help of the inexact supervision information, while the refinement-based algorithms leverage unreliable weak labels to train a robust model for ESA. Furthermore, some methods contain both initialization and refinement steps. To elaborate on the strategies of each group, we will introduce the strategies of these methods.

A. Initialization-Based Methods

On the one hand, ESA suffers from the lack of sufficient labeled data. On the other hand, for the multitask training scheme, part of the tasks usually has no supervision information. Therefore, there are many researchers considering directly initializing labels for training models. Note that compared with the semisupervised algorithms, these methods do not directly utilize labels such as semisupervised settings but have related supervision information, such as rating, emoticon, and hashtag. Here, we divide the initialization-based methods into the following four strategies.

1) Efficient Emotional Cue: The emotional cue is a common but efficient resource for initialization. First, the rating can be directly obtained from the websites. For instance, the ratings of drugs ranging from 1 to 5 are provided on the forum Askapatient (https://www.askapatient.com/). Then, using 3 as the boundary [64], the reviews with ratings lower than the boundary are assigned as negative, and the reviews with ratings higher than the boundary are considered positive.

Similarly, Wei et al. [63] leveraged the sentiment dictionary to extract keywords with distinct polarity. These keywords are initialized to be the supervision information for training visual emotion networks. Note that all these processes are automatically implemented. Second, the emotional cues, such as emoticon and hashtag, just require a small amount of labeling. Qadir and Riloff [123] annotated five seed hashtags for each emotion. The sentences are considered to have the same emotion as the hashtags, and classifiers are trained using these data. Then, the emotion of other hashtags can be calculated by the average prediction score of the sentences containing them. In this way, more and more hashtags can be assigned an emotion category and thus help to train the recognition model. Chen et al. [68] explored the role of emoticon for multimodal sentiment analysis and annotated 6k tweets with clean labels for initialization. Then, the corresponding emoticon and the sentiment representation are optimized in the expectation–maximization (EM) algorithm. In these methods, the labeled data are leveraged as the seed, which establishes the mapping between supervision cues and emotions. Then, more cues are integrated based on the co-occurrence to further improve the robustness of the model.

2) Metric Diffusion: Many methods leverage metric functions such as cosine similarity to diffuse emotion and sentiment labels to the unlabeled data. Specifically, the metric functions are used to measure the relations between prototypes such as a lexicon. Compared with the fully supervised methods, these methods only need a small number of prototypes for each class instead of the label for each instance. Therefore, the key factors of these methods are the predefined prototypes and the metric of the relations.

For text, the word is one of the most important pieces of information to construct the prototypes. Read and Carroll [124] leveraged the lexicon association, distributional similarity, and semantic spaces. Lexical association determines the relation of the textual data by considering the co-occurrence of the words with pointwise mutual information. Semantic spaces utilize the points from the high dimensions to represent concepts. In detail, the space contains four dimensions here, which are basic elements such as word stem, co-occurrence frequencies, embeddings of the document, and the mapping transformation like dimensionality reduction for the embeddings. Distributional similarity models the context as the group of grammatical relations of the words. Pereg et al. [125] proposed a sentiment analysis system, which allows users to change the prototypes. This process is called lexicon editing. When the system automatically extracts the aspects and opinions in the document, users can delete, add, or modify the lexicon items related to the aspects. This process ensures the discriminative and the correspondence of the lexicon items. Then, the similarity between items and the subset of the two sentiments is calculated to predict the label of the text. For images, the high-level features are more suitable as prototypes than the basic pixel-level ones. Borth et al. [126] proposed to describe the visual concepts in images by a vector consisting of ANPs. The values in the vector represent the probability that the image contains the ANP. Leveraging the ANP as prototypes and then calculating the similarity between the vectors is a promising way to initialize the label of the affective visual content.

3) Probabilistic Model: The classical statistic language modeling leverages the probabilistic model to represent the generalization of sequences of words [146], [147], [148]. Furthermore, researchers consider that the words and aspects in the document have close relation with the sentiment. Therefore, many methods have been proposed to incorporate sentiment into the probabilistic model.

He [127] added emotional prior information to the document-level analysis and proposed latent Dirichlet allocation (LDA) with Dirichlet prior modified method. Specifically, they considered that the overall sentiment (negative, positive, or neutral) of a document is first decided, and then, the sequence of words is generated on account of the prior information. Therefore, a Dirichlet prior distribution is leveraged to incorporate sentiment knowledge. On the other hand, sentiment information is usually used to mine opinions on the aspect level. Lin et al. [128] proposed a probabilistic modeling framework called a joint sentiment-topic model. In the training phase, the sentiment and topic are progressively detected from the text. First, the sentiment has a potential distribution based on the documents. Second, according to the sentiment label, a topic (aspect) can be further selected according to a certain distribution. Third, the sequence of words can be generated on account of the topic and sentiment. The distributions are also initialized in LDA and optimized by the Gibbs sampling process. Ramesh et al. [129] applied the technology of analyzing aspects and polarity to MOOC reviews and summarized effective suggestions for improving the quality of lessons. Zeng et al. [130] proposed a scheme that predicts an opinion word in the light of the target word. Specifically, given a review and a target word, the sentiment can be predicted by training a polarity classifier. Then, using the target word and sentiment as prior knowledge, the opinion word like “good” is generated by the opinion word classifier.

4) Feature Generation: The methods that belong to the feature generation group transform the features to obtain the label. This group of methods is usually utilized in multitask learning, which has no supervision information for one harder task. Therefore, many methods leverage other related easier tasks to train the model and transform the feature as the pseudo label of the task. To improve the performance in both tasks, the generated weakly supervised pseudo labels are used as auxiliary information to guide the training of the easier task.

Giving a single annotation for a sample requires relatively less cost. For instance, we can quickly assign an emotion label to an image or a document. However, giving annotation on the pixel level or word level requires a higher cost. Specifically, Peng et al. [149] demonstrated that the importance of the pixels in an image is different. For text, the words also have different contributions to the sentiment of the sentence. Therefore, the emotion and sentiment intensity in fine-grained affective signals is helpful for analysis. Yang et al. [131] proposed a coupled network to simultaneously detect the emotional salient region and recognize the emotion of an image. In detail, the features extracted after the convolutional layers are adopted for classification. The predicted results are leveraged as the weight for each channel to generate the sentiment salient map. The map contains the region that contributes significantly to classification. Then, the original feature is enhanced by multiplying with the sentiment map to better capture the spatial information. In this way, both tasks can be improved in the coupled networks. Similarly, Lee et al. [132] used attention weight to explore the contribution on word level, which achieves good performance on both detection and recognition tasks.

B. Refinement-Based Methods

Although it is difficult to obtain exact emotional annotations, many frameworks have been proposed to automatically obtain affective data with noisy labels [63], [126], [150]. Therefore, it is important to explore refinement strategies that train robust models with these weakly labeled data. Here, we will introduce the refinement-based methods for ESA into two groups: feature calibration and iterative adjustment.

1) Feature Calibration: To the target of learning distinctive representations to obtain reliable prediction, many methods design strategies based on the features to calibrate the network. For ESA, due to the existence of polarity and ambiguity, many feature-based methods have made significant progress.

Guan et al. [133] proposed to train an embedding space that maintains the general sentiment relation of the samples. Specifically, the samples are selected to form pairs and then design loss terms to reduce distances for same-label pairs and increase distances for opposite-label pairs. Furthermore, Yang et al. [134] considered that the emotional samples with the same polarity have a relatively closer relation. The polarity-based hierarchical emotion model is different from other recognition tasks. A triplet constraint is proposed to leverage the prior information. In detail, the commonly used emotion model can be separated according to the polarity. Therefore, a loss term with three types of pairs has been proposed to constrain the model. First, the samples with the same emotion have the closest relations. Second, the samples with different emotions but the same polarity are closer than the ones with different polarities. By adding emotional constraints to features with prior information, the network can be implicitly calibrated in the training process. From another perspective, due to ambiguity, an affective image may contain more than one emotion, and each emotion contributes differently to the image. Xue et al. [135] designed a nonextreme channel attention mechanism to alleviate the impact of noisy labels. This method enables the network to prevent overconfidence in one emotion and pay attention to the nondominant emotions. In addition, She et al. [136] proposed a multiple kernel network, which utilizes pooling layers to extract the informative features from a small-scale dataset with accurate annotation and a large-scale dataset without manual annotation. In this way, both the abundant visual patterns and emotionally discriminative concepts can be learned in the multiple kernel feature fusion module.

2) Iterative Adjustment: Due to the robustness of iterative adjustment, it is practical to learn with weakly annotated data step by step [4], [5], [49]. In this section, we will introduce the methods that adopt the iterative strategy in ESA from two aspects: data and task.

You et al. [137] proposed a strategy of progressively selecting reliable samples for refinement on a large-scale dataset without manual labeling. To achieve this goal, they first trained a model on the dataset with unreliable labels. Next, a subset of training instances is selected according to the prediction score by the model. Then, the model is further fine-tuned on the selected subset. The central of this method is the strategy of selecting a cleaner subset based on the entropy minimization [118]. For example, a sample with a predicted distribution of [0.9, 0.1] is more reliable than the sample with [0.6, 0.4]. In addition, multiple-instance learning is a classical method for learning with inexact supervision [138], [139]. Specifically, multiple-instance learning aims to predict the emotion of each segment, but we only have the label of the sample, which consists of many segments. Angelidis and Lapata [139] proposed to select the most negative and most positive segments into the training set according to the predicted distribution. Then, combine these segments via a gated recurrent unit (GRU)-based attention mechanism to output the prediction of a sample. Repeating the process iteratively, the model can be optimized end-to-end by the samples' loss. Differently, Zhang et al. [49] iteratively updated the weight of each segment describing the sample to replace the most confident segment. In this way, more segments can be integrated into the training process, which implicitly reduces the impact of unreliable segments. From the perspective of the task, Panda et al. [140] considered the hierarchy (i.e., Parrott's hierarchical emotion model) and proposed an effective curriculum-guided training strategy to gradually learn discriminative representations. Besides, Min [64] and Deriu et al. [141] first trained the model on a large-scale weakly labeled dataset and then fine-tuned it on a small-scale accurately labeled dataset. By training iteratively in this way, the model can learn robust representations and perform well on the emotion and sentiment datasets.

Table 4 Categorization and Representative Methods for Low-Shot ESA. The Second Column Indicates From Which the Methods Learn Prior Information

Category	Prior information	Dataset Usage	References
Data augmentation	Generated data	Internal	[151–154]
Pre-trained and fine-tuning	Resource-rich datasets	External	[63, 142, 155–160, 160–163]
Metric Learning	Unlabeled data	Internal	[164–170]
Multi-task learning	Relevant tasks	Internal	[158, 159, 171]
Meta-learning	Past learning experience	External	[65, 154, 172–175]
Embedding learning	Latent embeddings	External	[176–180]

C. Discussion

The initialization-based methods aim to generate labels by extra information (i.e., efficient emotion cue and seed of metric diffusion) or intra-assumption (i.e., distribution of probabilistic model or feature). The refinement-based methods design training strategies to reduce the impact of uncertain labels and learn useful information simultaneously. In general, the former methods are often utilized in situations where little supervision information can be provided, and the latter methods usually have noisy labels to train the models. Therefore, the initialization-based methods are more challenging to achieve better performance. In addition, for ESA, it is easy to collect a weakly annotated dataset. Therefore, we believe that combining initialization and refinement modules for an efficient learning framework is a promising direction.

VI. LOW-SHOT ESA

Typical deep learning-based ESA methods require a large amount of annotated data for parameter optimization and thus are data-hungry. However, it is usually time-consuming, laborious, error-prone, and high-qualified labelers that are required to provide sufficient data with well-annotated labels. As a result, this requirement can be violated for ESA in the wild. The main reasons are manifold. First, it is exhausting and expensive to provide accurate and fine-grained labels for massive data from a wide variety of modalities such as documents and videos. Second, research on emotion and sentiment is still in its early stages, which leads to a lack of large datasets dedicated to specific emotion classification tasks. Just as an example, the datasets collected for microexpression recognition usually contain a couple of hundreds of microexpression clips. Third, specific tasks, such as microexpression spotting, may require certain certifications to understand and mark the emotion states from the samples [181], [182]. It is thus challenging to invite enough well-qualified labelers for annotation. Finally, the clear individual differences in labeling and understanding emotion [183] as well as the inconsistent perception of fine-grained emotion intensity further hinder the acquisition of reliable emotion-related annotations.

To meet this challenge, recent studies pay special attention to the low-shot learning paradigm such as few shot and zero shot. Low-shot learning aims at finding solutions to a series of learning tasks from a small amount of data to avoid the overfitting problem. Take few-shot learning (FSL) [42] as an example. In each of the tasks, also known as episodes, few-shot learners are performed to solve the C-class classification problems using only K samples for each class, which is referred to as the C-way K-shot FSL problem. On the other hand, zero-shot learning (ZSL) [43] attempts to recognize unseen-emotional states without providing any labeled data of the unseen tasks. Usually, lateral information shall be provided, such as the attribute profile of the new tasks and the relationship between the seen and unseen tasks. Though both FSL and ZSL are designed to handle tasks with limited labeled data, they have different focuses. FSL focuses on recognizing new classes based on limited examples, while ZSL focuses on recognizing unseen classes using auxiliary information.

Modern deep models are usually with a large number of parameters. In the low-shot ESA scenarios, they are usually overparameterized and prone to overfit to limited samples [184]. In a comprehensive study about four distance metric learning-based FSL algorithms for (few-shot) FER, it is observed that the domain gap between the training and testing data highly influences the generalization ability of FSL algorithm [155]. A moderate domain shift is thus significant to safeguard the performance. Another evaluation on the performance of typical FSL methods for the general image recognition task can be found in [185].

To solve the dilemma caused by insufficient labeled data, different approaches bring in or share prior knowledge from different sources. Accordingly, existing low-shot ESA approaches can be briefly categorized into the following six groups, as shown in Table 4.

A. Data Augmentation

The approaches falling into this category respond to the problem of limited data from the viewpoint of data. It augments data for enriching the training set so that more supervised information with prior knowledge can be leveraged [151]. Wang et al. [152] focused on few-shot visual sentiment analysis and used noisy data of auxiliary datasets to guide FSL. A noise matrix is generated using the pretrained network on the auxiliary noisy data, which is then served as reweighing parameters. A model is first pretrained using the reweighted instance from the large noisy dataset, then fine-tuned using the clean instances, and finally retrained by relabeling the noisy data. In [153], an end-to-end compositional generative adversarial network (Comp-GAN) is proposed to synthesize face images with specified poses and facial expressions. The extended set of instances of diverse high-quality help to train a robust FER model with a good generalization ability. Data augmentation is also widely used in NLP for text sentiment analysis and stance detection. A masked-and-then-generation supervised learning scheme is used for data augmentation [154].

B. Pretrained and Fine-Tuning

The transfer learning paradigm is to build a model pretrained on a large auxiliary dataset such as ImageNet [186] designed for different tasks, such as image recognition, and then fine-tune it to the target dataset for ESA. Transfer learning greatly diminishes the need to collect a large amount of data and train a model from scratch. It allows leveraging the model pretrained in a resource-rich task for resourced-low tasks. The fine-tuning stage is designed to learn C classes using a few K examples for each class. In the fine-tuning stage, the network parameters are usually frozen and a classifier is updated using the data of new classes. Such a scheme is usually used as a baseline for few-shot ESA [155], which has shown good performance in addressing the data limitation issues for cross-lingual emotion detection by leveraging the knowledge from resource-rich languages [156]. More recently, models pretrained using large-scale auxiliary dataset show great potential. Xu et al. [142] discussed the posttraining of the pretrained BERT model for aspect-based sentiment analysis and review reading comprehension, which is formulated as a question-answering task. The training techniques, including masked language model and next sentence prediction (NSP), are introduced for domain knowledge posttraining. MLM and NSP are then fused with the task-aware two-pointer averaged cross-entropy losses to enhance both the domain and task knowledge. Hosseini-Asl et al. [157] regarded the feature extraction and prediction tasks in ASBA as the sequence generation task, which is implemented by a generative language model GPT2. The textual generation model learns to predict polarities and aspects without task-specific layers. The experiments show that the GPT2-based generative few-shot model outperforms the BERT with posttraining [142]. In [158] and [159], a general pretrained multitask network (FaceBehaviorNet) is trained for three facial behavior analysis tasks, including expressions recognition, continuous affect estimation, and facial action unit (AU) detection. It shows superior generalization performance on the task of compound expressions under the low-shot setting. Moreover, Wei et al. [63] established StockEmotion, a large-scale dataset from Web data with a size of over one million images and noisy fine-grained labels of emotion categories. A multimodal feature extraction network, Emotionnet, is trained on StockEmotion using the joint vision–text embedding losses. The zero-shot evaluations on the EMOTIC dataset suggest that a general model trained from scratch on StockEmotion has strong generalizability, even without using any EMOTIC instances for training [63]. The study [160] also indicates that multilingual BERT generalizes well when high-resource languages are transferred to low-resource languages for cross-lingual ZSL.

Zhong et al. [161] introduced a lightweight fine-tuning method to customize the transformer-based pretrained models, with a lightweight user-specific vector (token). During FSL, the majority of the parameters of the transformers are frozen and only the parameters of the user-specific token are updated. The method surpasses fine-tuning all parameters of the model in terms of both accuracy and efficiency [161].

Recent work reveals that the input prompt of advanced large-scale pretrained models such as GPT3 [81] provides a natural mechanism, termed few-shot "in-context" learning, for FSL without fine-tuning. An encouraging example of using GPT3 for few-shot sentiment analysis can be found in [162]. A comprehensive study of few-shot cross-lingual stance detection with prompt is provided in [163], where pattern-exploiting learning (PET) and corresponding prompt selection are evaluated.

C. Metric Learning

Instead of directly modeling the probability of one sample belonging to a specific emotion class, which may cause overfitting in an FSL setting, metric learning-based methods measure the similarities between the unlabeled samples with a few labeled samples, which reflects the probability of two inputs belonging to the same category.

In particular, prototopical network (Protonet) [187] is introduced to few-shot cross-subject cross-domain EEG emotion recognition in [164]. Moreover, Yang et al. [165] integrated the Siamese network [188], [189] with the self-attention mechanism for text sentiment analysis. In [166], Siamese networks are constrained via metric learning with carefully designed additional supervision information for efficient few-shot spontaneous speech emotion recognition. In addition, the study [167] treats the instance of different corpora of the same classes as the targeted classes in FSL and incorporates FSL to unsupervised DA (UDA). It uses a relation network-based [190] architecture to implement FSL for cross-domain speech emotion recognition. In a recent study, Zou et al. [168] aimed to classify compound expressions that are unseen based on the model trained only on seen basic expression datasets. Few-shot compound facial expression learning is performed by using a two-stage learning framework, with an emotion branch and a similarity branch. In the inference stage, the compound expression categories are output by the learned similarity branch. The study [169] investigates few-shot fine-grained emotion recognition using a small amount of physiological signals data. A Siamese network is proposed to learn the distance metric and a distance fusion module is used to make the final prediction. Zhang et al. [170] constructed a long interview video dataset of 50 patients for autism trait classification. Based on handcrafted features, distribution calibration that aims to calibrate the distribution of few-shot instances and the adaptive posterior learning model [191] that can be considered as a ProtoNet with a refinable class prototype are used for FSL.

D. Multitask Learning

Multitask learning [192] learns a group of relevant tasks jointly with a majority part of the model shared. More importantly, during the multitask learning process, by mining task-agnostic and task-specific information, common knowledge among tasks can be shared.

The pretrained FaceBehaviorNet [158], [159] has shown the power in few-shot ESA by multitask learning. Li and Shan [171] considered facial AU detection and FER in the MTL scenario. A meta net is used to weigh the AU and facial expression instances for adaptively transferring knowledge from the rich-resourced FER to the low-resourced AU detection.

E. Embedding Learning

Zhan et al. [65] proposed a structural embedding framework for zero-shot image emotion recognition. The structural embedding framework utilizes mid-level ANP features [126], which forms an intermediate embedding space to close the gap between extracted low-level visual features and expected high-level emotional states. In [172], a portable prediction head approach is designed to learn shared emotion embeddings for multilingual common representation. The portable prediction head approach is built by enforcing the multiway mapping model to produce a common emotion space. It can be well generalized to unseen language datasets (the alleged zero-shot setting) using either feedforward network or BERT.

Xu et al. [173] studied zero-shot speech emotion recognition and considered the emotional dimensions as attributes, which link the paralinguistic features to emotional states. After attribute learning, label learning further fulfills the maps from the attributes to the emotional states. A group of auditory affective descriptors (AADs) [193], including the per-emotion manually annotated, per-emotion semantic-embedding, and per-sample manually annotated AADs, are investigated. In [174], the samplewise learning and emotionwise learning strategies are developed to map the semantic-embedding prototypes, paralinguistic features, and the given labels, to predict emotional categories. The corresponding experimental results in these two studies validated the semantic-embedding prototypes from pretrained models for zero-shot speech emotion recognition.

A recent work [175] studies generalized zero-shot gesture emotion recognition. A semantically conditioned adversarial autoencoder is proposed to first produce latent representations, which model the gesture-based information learned from the fully supervised network, and then align the visual features with the corresponding word-level semantic features onto a latent space. During inference, the encoder outputs the corresponding semantic labels, which are matched with the class labels. To solve zero-shot stance detection, Liang et al. [154] proposed to distinguish the types (target-invariant/-specific) of stance features. The stance features are categorized into target-invariant and target-specific ones, and a hierarchical contrastive learning method is designed to characterize the correlation and differences between these kinds of features and further among stance labels.

F. Meta-Learning

Few-shot meta-learning, exemplified by model-agnostic meta-learning (MAML), is targeted at efficiently learning a model for a new task using only a few data and optimization iterations. In particular, MAML is a task-agnostic meta-learning algorithm. In MAML, the parameters of a model's parameters are optimized during the meta-learning phase, based on which the new task can be solved through fast adaptation with a handful of gradient updates on a small amount of data from that new task [194]. Zhao and Ma [176] used tensor decomposition to learn the low-rank embedding of sentences and designed an MAML-like approach for few-shot text emotion intensity value distribution learning. The study [177] introduces an MAML-based few-shot AU detection method for efficient model adaptation to new AUs and subjects using a few samples ($K = 1$ or $K = 5$) from the imbalanced AU distribution. It designs an MAML model with one general classifier for all AU detection tasks without task separation, where an instance is regarded as positive if at least one in the AU set of interest is detected. Another typical meta-learning method, namely, the neural process (NP), as a latent-variable neural network, is applied to personalize the stress evaluation based on the continuous electrocardiogram (ECG) and the galvanic skin responses (GSRs) for each subject [178]. Guibon et al. [179] studied metric learning-based meta-learning, which uses prototypical networks for episode training in the meta-learning process for text emotion recognition. In [180], an aspect-focused meta-learning (AFML) framework is designed, which can efficiently adapt the meta-trained model from the support set to learning the new concept for the aspect-specific instances in the query set without retraining the model.

G. Discussion

The great potential of the low-shot learning methods encourages the extension of their application to the scenario of learning from limited data, not just limited to the scenarios of few-shot or ZSL. In [195], the FSL approaches are introduced to microexpression recognition to bypass the issues caused by the lack of sufficient labeled data [196]. The feature learning stage is further divided into two stages, namely, the prior learning for generic feature extraction and target learning for high-level feature adjustment. A similar idea can be found in [197]. A deep residual prototypical network with the episodic training scheme is designed to map the input microexpression sequence into an embedding feature space to alleviate the data limitation and data bias issue existing in the microexpression recognition task. In addition, in the application of intelligent per-type robots, prototype mixture models [198] are used to detect new objects via voice interaction with users and then update the visual model.

One thing we need to mention here is the emerging issue of learning from (facial expression) data that are distributed across a varied set of local sites under different conditions. In [199], a few-shot federated learning paradigm is designed, where local models learn from a few labeled private facial expression data and are then aggregated in the central site to a global model. The combination of low-shot learning and federated learning is likely to spawn new research that will be useful in real-world scenarios as well.

VII. INCREMENTAL ESA

Though modern deep learning models achieve promising progress in ESA, they suffer from severe limitations when adapting to unseen emotion categories/instances. First, most ESA models only consider a limited number of emotion categories such as the seven basic expressions, namely, happiness, sadness, disgust, anger, fear, surprise, and neutral, defined by Ekman and Friesen [6]. These models cannot easily adapt to recognize new, fine-grained, or compound emotion states [12], which are not met when the models are learned. Second, existing approaches usually do not generalize well to recognize samples with distinct variations existing in appearance or with distributions significantly different from the training set, even when these samples belong to a category ever seen. Third, as society evolves, the sentiment states of a few words and phrases change. Thus, the ESA model shall adapt to these new meanings while preserving the current knowledge about other words and phrases. Fourth, in open-world environments, data keep constantly appearing and arriving, a traditional static supervised learning paradigm cannot handle this dynamic scenario.

To tackle these challenges, there has been an increasing interest in the study of incremental/continual ESA. The goal is to adapt the models to these new tasks efficiently, without catastrophic forgetting of old tasks and model retraining from scratch.

Incremental learning [200], continual learning [201], and continuous learning [202], which are also known as lifelong learning [203] before the deep learning era, are a long-standing research topic.[3] The work [204] proposes the efficient lifelong learning algorithm (ELLA), which preserves a shared repository for all task models formed of latent model components. A linear combination of shared latent model components from the knowledge repository is assumed to be able to represent the parameter vectors of models. The knowledge of the repository is then transferred to assist the new model learning. Afterward, the basis is updated with knowledge from the new task. Both linear and logistic regression are implemented in the provided experiments. Soon after, active curriculum selection is introduced to choose tasks for improving the performance of ELLA based on the information maximization heuristic [205], [206]. Chen et al. [207] defined and studied the lifelong sentiment classification problem. They chose the naive Bayesian text classification as the basis of the model and used the regularization term to effectively exploit the knowledge gained from past learning. There is a free and open tool [sentiment analysis and incremental learning (SAIL)] developed to update the parameters of the probabilistic models [208]. Random forest is updated in an incremental manner for customizing FER [66].

Table 5 Incremental ESA Approaches According to Three Incremental Settings. FER and TSA Stand for Facial Expression Recognition and Sentiment Analysis, Respectively

	TIL	DIL	CIL
FER	[210]	[204, 205, 206, 211]	[212, 213]
SA	[214–216]	[207, 217]	-

According to different experimental settings, existing approaches are usually categorized into three types: class-incremental learning (CIL), domain incremental learning (DIL), and task incremental learning (TIL) [209].

TIL usually learns a sequence of tasks and uses classification heads for individual tasks separately. During inference, the task identification shall be given in advance so that the corresponding head can be chosen. CIL approaches focus on learning new and unseen emotion categories without forgetting the knowledge about existing classes. The main focus is to obtain a unified classifier, without the task identification provided during inference. DIL handles the cases where new instances of the categories seen before emerge, with distinct distributions and domain gaps. There is no need to expand the classification head. A summary of existing incremental ESA approaches is listed in Table 5.

From the methodology point of view, existing incremental approaches can be grouped into active task selection, rehearsal-based method, and network-based method, which are shown in Fig. 6.

A. Active Task Selection

Active task selection assumes that the entire task sequences are known prior to the incremental learning process and the order of tasks can be changed during incremental learning. Algorithms can be designed to select the next tasks to learn in order to optimize the performance for the tasks to learn in the future. Ruvolo and

[3] The four concepts, namely, incremental learning, continual learning, lifelong learning, and continuous learning as mentioned, are highly consistent. They all refer to the machine learning paradigm, which allows models to learn and adapt to new tasks from new data, without forgetting how to perform previous tasks (the catastrophic forgetting phenomenon [201]). Thus, the four concepts are used interchangeably in this article. Nonetheless, from a more fine-grained perspective, they may still slightly differ in the specific evaluation protocols.

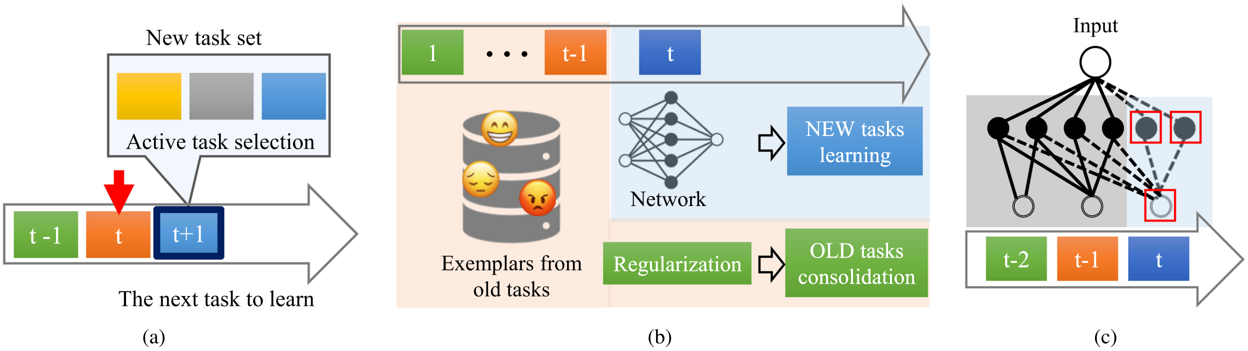


Fig. 6. *Illustration of three kinds of incremental ESA approaches. (a) Active learning-based approaches select the next task for maximizing performance. (b) Rehearsal-based approaches store a repository of exemplars from old tasks and design corresponding regularization terms to consolidate the old tasks during new task learning. (c) Network-based approaches freeze/penalize the shift of network weights and/or expand the network for learning new tasks.*

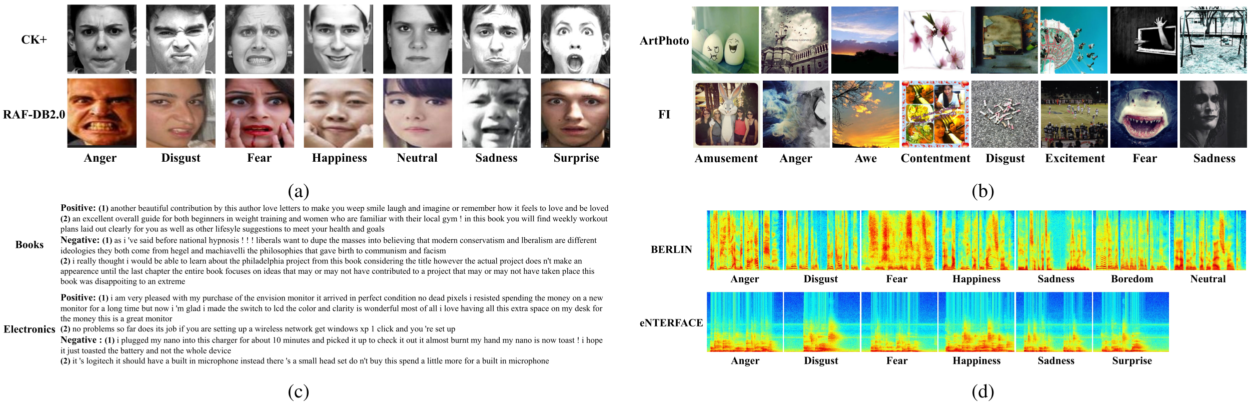
Eaton [205], [206] introduced active curriculum selection to determine which task to learn next for improving the performance of ELLA, based on a series of criteria, including the myopic information maximization criterion and the diversity heuristic. Moreover, according to the experimental findings in [214], the main challenge for continual sentiment classification is the bidirectional knowledge transfer between the old and new tasks. Not long after that, the study points out that continual sentiment classification approaches suffer from serious catastrophic forgetting when there is not much common knowledge shared among tasks [215]. In line with this understanding, a divide-and-conquer strategy is further proposed to measure the similarity between the new task and previous ones. For similar tasks, the main focus is on efficient knowledge transfer to improve the new task learning, while for dissimilar tasks, special attention is paid to forgetting avoidance [218]. A similar solution is also reported in [216] for continual aspect-based sentiment classification. In [219], capsules and dynamic routing are used to find out the similar enough old tasks for knowledge sharing, and a task mask scheme is developed to protect task-specific knowledge. Furthermore, a parameter-gate (p-gate) mechanism is designed to judge how useful they are to the new task for efficient knowledge transfer [217].

B. Rehearsal-Based Incremental Learning

Thuseethan et al. [227] used a random selection policy to build and update the exemplar samples and introduced a regularized term for the backbone for continual FER. To learn facial expression categories continuously from a stream of data, Zhu et al. [212] proposed an incremental facial expression recognition network (IExpressNet). IExpressNet keeps a set of exemplars that stands for the anchor of old classes by the herding exemplar selection algorithm [228]. The cross-entropy classification loss and the center loss are used to learn from new classes, while the distillation loss is then used to transfer knowledge from the seen classes to the unseen classes. An alternative is to use synthesized examples by, for example, generative adversarial network (GAN), VAE, and so on to consolidate old knowledge. This kind of method is usually referred to as pseudo-rehearsal-based methods. The study [213] aims at mimicking the neurocognitive phenomenon of imagined contact [229] for continual FER. It uses an autoencoder-based generative model, namely, the conditional adversarial autoencoder (CAAE)-based imagination model to perform imagination for particular subjects and augment learning. The growing dual memory (GDM) architecture with two growing when required (GWR) neural networks [230] is then adopted to perform incremental learning. A similar study can be found in [231], where a combined replay generative model [232] named Dreamnet is introduced to generate samples for exemplar-free continual learning.

C. Network-Based Incremental Learning

Network-based incremental learning methods address the catastrophic forgetting problem by designing a penalty term on the network weights to preserve old knowledge or expanding the subnetwork branches to learn new knowledge. Han et al. [233] focused on continual cross-cultural emotion recognition. Elastic weight consolidation (EWC) [201] is introduced to solve the catastrophic forgetting problem by regularizing the parameters of the network. Hung et al. [210] studied the continual learning along the functionalities, i.e., from face recognition to other tasks such as FER and gender prediction. The packing-and-expanding scheme is introduced to improve the PackNet [234] for continual learning. Ke et al. [214] proposed a dual-network structure consisting of a main continual learning network and an accessibility network to share knowledge among sentiment classification tasks. The main continual learning network maintains a knowledge base implemented using a GRU to store the knowledge from all tasks ever seen. The ac network decides which part of the past knowledge may contribute to learning the new task. A sister study utilizes two networks as well, which



Books

Positive: **(1)** another beautiful contribution by this author love letters to make you weep smile laugh and imagine or remember how it feels to love and be loved
(2) an excellent overall guide for both beginners in weight training and women who are familiar with their local gym ! in this book you will find weekly workout plans laid out clearly for you as well as other lifesyle suggestions to meet your health and goals
Negative: **(1)** as i 've said before national hypnosis ! ! ! liberals want to dupe the masses into believing that modern conservatism and lberalism are different ideologies they both come from hegel and machiavelli the philosophies that gave birth to communism and facism
(2) i really thought i would be able to learn about the philadelphia project from this book considering the title however the actual project does n't make an appearence until the last chapter the entire book focuses on ideas that may or may not have contributed to a project that may or may not have taken place this book was disappoiting to an extreme

Electronics

Positive: **(1)** i am very pleased with my purchase of the envision monitor it arrived in perfect condition no dead pixels i resisted spending the money on a new monitor for a long time but now i 'm glad i made the switch to lcd the color and clarity is wonderful most of all i love having all this extra space on my desk for the money this is a great monitor
(2) no problems so far does its job if you are setting up a wireless network get windows xp 1 click and you 're set up
Negative : **(1)** i plugged my nano into this charger for about 10 minutes and picked it up to check it out it almost burnt my hand my nano is now toast ! i hope it just toasted the battery and not the whole device
(2) it 's logitech it should have a built in microphone instead there 's a small head set do n't buy this spend a little more for a built in microphone

Fig. 7. ***Illustration of domain shift in different ESA tasks (please zoom in to see details). (a) FER: between lab-controlled environment (CK+ [220]) and in the wild (RAF-DB2.0 [221]). (b) Visual emotion analysis: between artistic paintings (ArtPhoto [9]) and the social images (FI [222]). (c) Textual sentiment classification: between the books domain and the electronics domain [223]. (d) Speech emotion recognition: between acted German (BERLIN [224]) and induced English (eNTERFACE [225]).***

attempt to balance the challenges of new task learning and knowledge retention for old tasks [235]. Furthermore, based on the BERT aspect sentiment classification model, a BERT adapter serves as a network carrier for DIL. A series of contrastive learning frameworks is introduced, including the alleged contrastive ensemble distillation, contrastive knowledge sharing, and contrastive supervised learning on the current task, to share knowledge among tasks and facilitate new task learning [236].

D. Discussion

Research on incremental ESA is still in the early stages. Somewhat surprisingly, incremental learning approaches have shown side benefits in bias mitigation. In most FER datasets, there is an imbalance of the distribution along with attributes such as age, gender, race, or skin color [221]. A recent study pioneers the use of a domain-incremental learning approach to mitigate the bias issue existing in the FER task to enhance the balance and fairness [211]. It is demonstrated that regularization-based continual learning methods contribute to fair and balanced expression recognition and AU detection. Likewise, popular incremental approaches, such as Packnet [234] and the compacting-picking-growing (CPG) method [237], are evaluated to tackle the long-tailed issues for large-scale facial expressions recognition.

One stream is to combine incremental learning with other machine learning paradigms. Dai et al. [238] considered both the sentiment categories and targets for (targeted) aspect-based sentiment analysis and incorporated incremental learning with multitask learning. A multitask network with category name embedding (CNE-net) is designed to reduce catastrophic forgetting. Incremental learning is implemented by sharing both encoder layers and decoder layers of all the tasks and fine-tuning using sample-target training data.

VIII. DOMAIN-ADAPTIVE ESA

For an unlabeled or sparsely labeled target domain, there might be another different but related source domain with sufficient labels. An intuitive idea is to transfer the learned knowledge from the source domain to the target domain. As shown in Fig. 7, such cross-domain transfer widely exists in ESA, such as the transfer between lab-controlled environments and in the wild application for FER, between artistic paintings and social images for visual emotion analysis, between the electronics domain and the kitchen domain for textual sentiment classification, and between acted German and induced English for speech emotion recognition. However, the existence of domain shift (also known as dataset bias) [55], [316] usually results in poor performance on the target domain when directly transferring the model that is learned on the source domain. Corresponding to the transfer examples in Fig. 7, the quantitative results of direct transfer (S) and training on the target (T) are shown in Table 6. It is clear that there is a significant accuracy drop between S and

Table 6 Quantitative Illustration of Domain Shift for Different ESA Tasks. The Domains Before and After ⟶, Respectively, Represent the Source Domain and the Target Domain. The Learned Model Is Tested on the Target Domain While Training on the Source Domain in "S" and Training on the Target Domain in "T." The Performance Is Evaluated by the Classification Accuracy (%) of Seven, Eight, Two, and Five Emotion Categories

Task	Model	DA setting	S	T	DA setting	S	T
Facial expression recognition	ResNet-50 [226]	CK+⟶RAF-DB2.0	36.9	85.6	RAF-DB2.0⟶CK+	82.3	86.1
Visual emotion classification	ResNet-101 [226]	ArtPhoto⟶FI	23.9	66.1	FI⟶ArtPhoto	29.1	43.7
Textual sentiment classification	BERT [84]	Electronics⟶Books	86.5	89.3	Books⟶Electronics	86.2	89.6
Speech emotion recognition	ResNet-18 [226]	BERLIN⟶eNTERFACE	29.2	80.0	eNTERFACE⟶BERLIN	44.0	84.1

Table 7 Categorization and Representative Methods for Domain-Adaptive ESA Under the Typical Settings: Single-Source, Single-Target, Unsupervised, Homogeneous, and Closed-Set DA

Category	Methods		References
Feature representation-level alignment	Shallow feature matching	Sample re-weighting	[71, 72]
		Feature transformation	[239–253]
	Domain-invariant deep feature learning	Discrepancy-based alignment	[246, 254–263]
		Adversarial alignment	[264–283]
		Self-supervision-based alignment	[67, 284–288]
Data pixel-level alignment	GAN-based data generation		[278, 289, 290]
Label space-level alignment	Pseudo label-based classifier training		[249, 291, 292]
	Classifier decision boundary refinement		[259, 277, 288]
	Category-level discrepancy-based alignment		[221, 256, 283, 291]
	Category-level adversarial discriminative alignment		[260, 265, 287, 292]

T, such as 23.9% versus 66.1% when transferring from ArtPhoto to FI for visual emotion classification.

Domain-adaptive ESA aims to bridge this domain shift by learning a model with high transferability based on the labeled source data and unlabeled or sparsely labeled target data. Three commonly considered domain shifts between the source and target domains include [55]:

1) covariate shift, i.e., $P_S(y \mid \mathbf{x}) = P_T(y \mid \mathbf{x})$ for all $\mathbf{x}$, but $P_S(\mathbf{x}) \neq P_T(\mathbf{x})$;
2) label shift, i.e., $P_S(\mathbf{x} \mid y) = P_T(\mathbf{x} \mid y)$ for all y, but $P_S(y) \neq P_T(y)$;
3) concept drift, i.e., $P_S(\mathbf{x}, y) \neq P_T(\mathbf{x}, y)$.

Based on different criteria, such as the availability of the target labeled data and source labeled data during training, the number of source domains and target domains, label set correlations, and data modalities, DA can be classified into multiple settings. Please refer to [55] for a more detailed DA taxonomy and [51] for general transfer learning. In this section, we first summarize the DA methods under the most typical setting: single-source, strongly supervised, single-target, unsupervised, homogeneous, closed-set, and target data available, that is, there is one source domain fully labeled, which is available during training; there is one target domain without any labels, but target data are available, the source and target data belong to the same modality, and all domains share the same emotion label set. According to the different levels of the alignment strategy, we divide existing methods into feature representation-level alignment, data pixel-level alignment, and label space-level alignment, as summarized in Table 7. After that, we introduce some other popular and important DA settings, as summarized in Table 8, including semisupervised DA where some labeled target data are available, multisource DA (MDA) where multiple source domains are available, multimodal DA where the source and target data are represented by multiple modalities, and cross-modal (or heterogeneous) DA where the source and target data belong to different modalities. Finally, we give a brief quantitative and qualitative comparison of these methods.

A. Feature Representation-Level Alignment

No matter in the shallow learning period or in the deep learning era, feature representation-level alignment is the most common strategy in domain-adaptive ESA.

1) Shallow Feature Matching: In the shallow learning period, the nondeep approaches mainly focus on the distribution matching of handcrafted features between the source and target domains, either by sample reweighting or by feature transformation.

Sample reweighting assigns different weights to the source samples based on their similarity to the target samples. The assumption is that the source samples with higher similarity play more important roles in the adaptation process. Xia et al. [71] proposed a principal component analysis (PCA)-based sample selection method, which can be

Table 8 Categorization and Representative Methods for ESA Based on Semisupervised DA, MDA, Multimodal DA, and Cross-Modal DA

Category			References
Semi-supervised domain adaptation	Task classifier training		[259, 293–295]
	Supervised shared subspace learning		[296, 297]
	Parameter adaptation		[298]
Multi-source domain adaptation	Domain alignment and pairing strategy	Each source and the target pair	[294, 299–303]
		The combined source and the target pair	[304–308]
		Selected sources and the target pair	[263, 309, 310]
		Multi-source discrimination	[306]
	Domain/sample weighting and sharing strategy		[263, 307, 309, 311]
	Task classifier training and fusion strategy	Single classifier training	[263, 300–304, 306–308, 311]
		Multiple classifier training	[241, 294, 299, 302, 305, 310, 312]
Multi-modal domain adaptation			[293, 295, 313]
Cross-modal domain adaptation			[314, 315]

viewed as a binary reweighting strategy. A subset of source data that are close to the target domain is selected and then used in classifier training. First, the latent concepts are extracted from the target domain by singular value decomposition. Second, the source samples are projected onto the latent space and the ones that are far away from the normal area are removed. Differently, Chu et al. [72] proposed a selected transfer machine by assigning different weights to the source samples in the loss function of the task classifier. Source sample reweighting and target classifier are simultaneously learned.

Feature transformation transforms the original features into a new embedding space where the knowledge learned from the source domain can be better transferred to the target domain. Existing transformation operations for ESA can be categorized into three groups: feature reduction and selection, feature alignment, and feature generation.

1) *Feature Reduction and Selection:* In [239], kernelized-PCA (KPCA) [317] is employed as a feature mapping method to minimize the marginal distribution difference between domains while making the variance of the instances as large as possible; another kernel-based feature mapping method is transfer component analysis (TCA) [318], which aims to reduce the marginal distribution discrepancy by minimizing the MMD (see the following discrepancy-based alignment for more details) and enforcing the scatter matrix as the constraint; feature clustering is also considered with information-theoretical learning [240] by optimizing two information-theoretical quantities. KPCA and TCA are also employed in [241]. Besides using MMD as a joint feature distribution regularization for measuring and alleviating the difference between different domains, another constraint is enforced to select the few but discriminative salient facial regions, either by sparse regularization [242] or by nonnegative weighting [243].
2) *Feature Alignment:* Besides the explicit features in a feature space, we can also align some implicit features, such as spectral features, statistic features, and subspace features. The domain-specific features and domain-independent features (pivots) are identified in spectral feature alignment [244]. A bipartite graph is constructed between the two groups and spectral clustering is performed to construct a low-dimensional representation. A similar framework is followed in [245], which also considers the labels when creating the representation. It is shown that this enables to learn customized representations with better sentiment classification performance.

   Statistic feature alignment is also investigated as a feature alignment method. A representative example is correlation alignment (CORAL) [246], in which the transformation matrix of the source features is constructed by the alignment of the second-order statistic covariance features.

   Subspace learning is another popular feature alignment strategy. In [239], a geodesic flow is constructed based on a geodesic flow kernel (GFK) [319] to link the subspaces of different domains on a Grassmann manifold. The source and target data are projected into each of the infinite subspaces and a resultant infinite-dimensional space is obtained. Subspace alignment [247] aims to align the PCA-generated bases of the subspace of different domains. Maximum independence DA [248] seeks to learn a domain-invariant subspace by maximizing the Hilbert–Schmidt independence between the projected samples and their respective domain features.

   Instead of directly applying existing subspace learning algorithms to the ESA task, some specific improvements are also designed. Based on labeled source samples and an unlabeled auxiliary set of target samples, transductive transfer regularized least-squares regression [249] is proposed to jointly learn a discriminative subspace and predict the emotion labels for the target samples. This method is further improved with an auxiliary set selection model [320]. Transfer linear subspace learning [250] is proposed to learn a common feature subspace between the source and target domains, where high transferable features are preserved, while low transferable ones are suppressed. Later, some improvements are incorporated, including feature selection and geometric structure regularization [251] as well as interclass and intraclass scatters [252].
3) *Feature Generation:* A sample regenerator is learned to generate new features for target samples that share the same or similar source feature distributions [253]. Being enforced to generate themselves for source features, the generator is decomposed of a kernel mapping and a linear projection. The minimization of MMD between the source and target features in the kernel space is regularized to align the feature distributions.

2) Domain-Invariant Deep Feature Learning: In the deep learning era, a two-stream conjoined architecture that enables end-to-end training of domain-invariant deep feature learning has dominated the UDA methods. One stream corresponds to the ESA task on the labeled source domain, while the other aims to bridge the domain shift by aligning the source and target feature representations. Based on the alignment component, we can categorize existing domain-invariant deep feature learning methods into different types, such as discrepancy-based alignment, adversarial discriminative alignment, self-supervision-based alignment, and so on. Moreover, the weight-sharing strategy, such as shared, partially shared, and unshared, is also different among these methods. The joint optimization of the task loss and alignment loss enables to learn domain-invariant features and a generalizable classifier that can also perform well on the target domain. The inference is

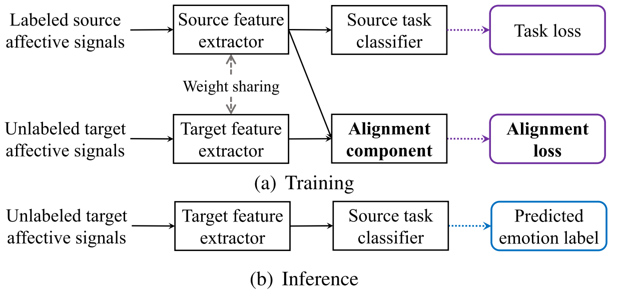


Fig. 8. ***Illustration of a general training and inference framework for learning domain-invariant deep features. (a) Difference between existing methods mainly lies in the alignment component with corresponding alignment loss and the weight-sharing strategy between the two domains. (b) During inference, the emotions of the target affective signals can be predicted by the target feature extractor and the source task classifier.***

simple and direct by combining the target feature extractor and source task classifier to obtain the predicted emotion labels. The illustration of a general training and inference framework of these methods is shown in Fig. 8. Please note that some methods might employ more than one alignment strategy. In such cases, we separately introduce them in each corresponding part.

The objective loss function of the two-stream conjoined architecture for domain-invariant feature learning can be summarized as

$$\mathcal{L}_{\text{DIFL}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{alignment}} \tag{3}$$

where $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{alignment}}$, respectively, correspond to the shared ESA task loss and the specific alignment loss between different methods. Let f_S and h_S denote the source feature extractor and source task classifier, respectively; the commonly employed cross-entropy loss, mean-squared error loss, and Kullback–Leibler (KL) divergence loss for emotion classification, emotion regression, and emotion distribution learning tasks are defined as

$$\mathcal{L}_{\text{CES}} = \mathbb{E}_{(\mathbf{x}_S, y_S)\sim P_S} \sum_{c=1}^{C} \mathbb{1}_{[c=y_S]} \log\left(\sigma\left(h_S^{(c)}\left(f_S\left(\mathbf{x}_S\right)\right)\right)\right) \tag{4}$$

where C is the number of emotion categories, σ is the softmax function, and $\mathbb{1}$ is an indicator function

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{(\mathbf{x}_S, y_S)\sim P_S} \sum_{k=1}^{N_E} \left(h_S\left(f_S\left(\mathbf{x}_S\right)\right)_k - \left(y_S\right)_k\right)^2 \tag{5}$$

where N_E is the number of dimensions of the employed DES model ($N_E = 3$ for the VAD model), and $(y_S)_k$ represents the emotion label of the kth dimension

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{(\mathbf{x}_S, y_S)\sim P_S} KL\left(y_S \parallel h_S\left(f_S\left(\mathbf{x}_S\right)\right)\right) \tag{6}$$

where $KL(\mathbf{p} \parallel \mathbf{q}) = \sum_{l=1}^{L}(\mathbf{p}_l \ln \mathbf{p}_l - \mathbf{p}_l \ln \mathbf{q}_l)$ is the function to compute the KL divergence.

Discrepancy-based alignment aims to align the feature representations between the source and target domains by minimizing the discrepancy that measures the distance between two feature distributions. The used explicit discrepancy metrics in domain-adaptive ESA include MMD and its variants [254], [255], [256], [257], CORAL [246], KL divergence [259], [260], and central moment discrepancy [261]. Some implicit discrepancy methods, such as adaptive batch normalization (AdaBN) [262], are also considered.

MMD that aims to compute the difference between the mean values of a smooth function on the two feature distributions [321] are widely investigated as a discrepancy metric. It has been proven that in the reproducing kernel Hilbert spaces, the MMD is zero if and only if the two distributions are equal [321]. The ordinary MMD is directly used in [254] and [256]. Three MMD values are combined to align the spatial-stream features, temporal-stream features, and concatenated features [255]. To overcome the sensitivity of kernel choice in MMD computation, Li et al. [258] employed multiple kernel variants of MMD (MK-MMD) to further reduce the domain discrepancy by selecting optimal multiple kernels. Both MMD and MK-MMD only consider marginal distribution alignment, He and Ding [257] combined joint MMD with ordinary MMD to simultaneously reduce the joint distribution discrepancy. CORAL [246] measures the discrepancy of second-order statistics of the source and target features, which is similar to the polynomial-kernel MMD. KL divergence [260] and its symmetric version [259] that measure the first-order statistics are adopted to explicitly minimize the distance between the two domains' embedding features. Central moment discrepancy is employed in [261] to align the central moment of each order instead of the weighted sum of all orders. A mixture of different discrepancy metrics, including L_2, cosine, MMD, Fisher linear discriminant, and CORAL, is designed [263]. Unsupervised criteria are employed to select an optimal subset of these metrics by estimating the "informativeness" of each metric.

Instead of measuring the discrepancy with detailed explicit metrics, Jiménez-Guarneros and Gómez-Gil [262] employed AdaBN to implicitly minimize the discrepancy between the source and target domains. Specifically, the batch normalization (BN) statistics, such as the moving average mean and variance of all BN layers, between different domains are aligned.

The source and target feature extractors in the above-mentioned methods usually share the same parameters to reduce complexity. As demonstrated in [322], the domain invariance may significantly weaken the discriminative power. How to relax the weight-sharing constraint to preserve the discriminative power with a balanced tradeoff is yet to be investigated. These methods also differ in the layers that the discrepancy works on, such as fully connected (FL) layers [246], [254], [255], [257], [258] and BN layers [262].

Adversarial discriminative alignment usually employs a domain discriminator to confuse different domains in an adversarial manner. It assumes that the domain labels of domain-invariant features cannot be easily distinguished by a discriminator. One milestone is domain-adversarial neural network (DANN) [323], where the domain classifier plays the role of a discriminator, trying to classify whether the features come from the source domain or the target domain. Cross-entropy loss is employed as the discriminator's objective. A gradient reversal layer (GRL) is designed to enable adversarial training. DANN-based adversarial discriminative alignment has been widely used in ESA, either directly used [264], [265] or with specific improvements [266], [267], [268], [269], [270], [271], [272], [273], [274], [275], [276], [277]. The first group of improvement is the input features to the domain classifier, such as the features with word attention [266], hierarchical attention [267], and sentence-aspect interactive attention [268]. In [269], BERT is encouraged to be domain aware and the domain-specific features are distilled by a specifically designed posttraining procedure in a self-supervised way. The posttrained BERT features are input to the DANN. Based on the emotion lateralization in neuroscience, Li et al. [270] employed four directed RNNs to traverse electrode signals for the left and right brain regions from both horizontal and vertical orientations, which aim to keep the intrinsic spatial dependence. The second group of improvement is the specific design of domain classifiers. Typically, multiple-domain classifiers are included, such as one for each aspect [271], one for each EEG channel [272], and one for the domain level plus the other for the subject level [273]. Li et al. [274] employed one global domain classifier to constrain the entire data distribution similarly and two local domain classifiers to narrow the left and right hemispheric data distributions separately. This method is further improved with a subject classifier [275]. Together with local and global attention, multiple local and one global domain classifiers are employed to highlight the transferable EEG brain regions and samples [276]. The third group of improvement is over GRL. For example, integrated adaptive strategy [277] is proposed to replace GRL to better explore the impact of syntactic graph structure transfer.

Apart from using a domain classifier with GRL and cross-entropy loss, another strategy is to employ a feature discriminator with GAN loss [278], [279] or its invariants [280], [281], [282]. Let f_T and d_F denote the target feature extractor and discriminator, respectively, and the feature-level GAN loss is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{GAN}_F} &= \mathbb{E}_{\mathbf{x}_S\sim P_S}\log d_F\left(f_S\left(\mathbf{x}_S\right)\right)\\ &\quad+\mathbb{E}_{\mathbf{x}_T\sim P_T}\log\left[1-d_F\left(f_T\left(\mathbf{x}_T\right)\right)\right]. \qquad (7)\end{aligned}$$

Wasserstein GAN is employed to overcome the gradients vanish and instability problems of traditional GAN [280]. Adversarial graph representation adaptation [281] is proposed for holistic-local feature co-adaptation. Specifically, the holistic-local features within each domain and across different domains are, respectively, propagated for exploring their interaction and feature co-adaptation. Instead of directly aligning the features, Latif et al. [282] proposed to adversarially align the reconstructed data from the encoded features in one domain and the raw data in the other domain.

Differently from the abovementioned methods, mutual information minimization (MIM) [283] is employed for domain-level adaptation by distilling the domain-invariant common knowledge and eliminating the domain-sensitive one in different domains.

Self-supervision-based alignment tries to learn domain-invariant features typically by combining some auxiliary self-supervised learning tasks with the ESA task. Glorot et al. [284] employed stacked denoising autoencoders to reconstruct the original affective signal from the corrupted version (e.g., adding a masking noise) by minimizing the denoising reconstruction error. The stacked features from the encoding output of intermediate layers are used as the domain-invariant features to train the task classifier. Instead of directly taking the combined source and target data into autoencoders, Deng et al. [67] proposed an adaptive denoising autoencoder, where the prior knowledge from the target domain is first learned and then regularized on the source domain. Feature learning and classifier learning are separated into two stages in the previous two methods. Furthermore, the feature learning is fully unsupervised, which does not consider the ESA task. Later, Deng et al. [285] designed an end-to-end learning paradigm based on Universum autoencoder, which simultaneously enables to discover the intrinsic structures in the input via reconstruction and exploit the prior knowledge from unlabeled data to regulate the task classifier through Universum learning. Yu and Jiang [286] designed two auxiliary binary prediction tasks in textual sentiment classification to classify whether a given sentence contains a positive or negative domain-independent sentiment word. Feature learning and sentiment classification are also jointly optimized. In [287], a self-supervised patch localization framework is designed to emphasize the local information by learning the differences between patches. Contrastive learning is investigated to extract domain-invariant and task-discriminative features [288]. Specifically, similar features are learned for the query and its positive pair, while discriminative features are learned by utilizing the negative samples in the pretext tasks.

B. Data Pixel-Level Alignment

Data pixel-level alignment usually combines the domain discriminator with a generator, which is used to generate fake source or target affective signals that cannot

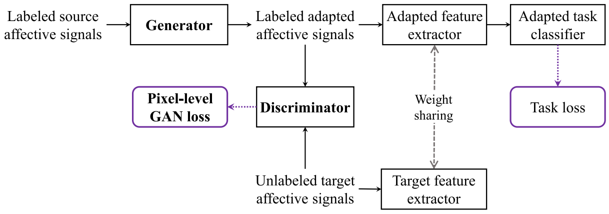

Labeled source affective signals
Generator
Labeled adapted affective signals
Adapted feature extractor
Adapted task classifier
Pixel-level GAN loss
Discriminator
Weight sharing
Task loss
Unlabeled target affective signals
Target feature extractor

Fig. 9. ***Illustration of a common adversarial generative framework with data pixel-level alignment for domain-adaptive ESA.***

be distinguished by the discriminator from the real data. Different from the feature representation-level alignment that aligns the source and target domains on the feature level, data pixel-level alignment tries to align the raw data of the two domains. Typically, such alignment is implemented through the GAN [324] and its variants, such as CycleGAN [325]. The adversarial learning is performed in a min–max game: the discriminator tries to correctly classify the real and generated affective signals, while the generator tries to confuse the discriminator by making the generated signals as real as possible. Let g and d denote the generator and discriminator, respectively, and the pixel-level GAN loss is defined as

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{\mathbf{x}_S \sim P_S} \log d\left(g\left(\mathbf{x}_S\right)\right) + \mathbb{E}_{\mathbf{x}_T \sim P_T} \log\left[1 - d\left(\mathbf{x}_T\right)\right]. \tag{8}$$

Zhao et al. [289] employed GAN with the source images as the generator's input to generate target-style adapted images. To overcome the underconstrained nature of GAN, a cycle-consistency constraint is employed in CycleEmotionGAN [278], [290]. After the adversarial learning, these GAN-based models can adapt source domain images such that they appear as if they were drawn from the target domain. To preserve the emotional information before and after image translation, some regularization constraints are often enforced, such as the emotional semantic consistency constraint [278], [289], [290]. The illustration of a common adversarial generative framework is shown in Fig. 9. The difference between different methods mainly lies in the input to the generator and discriminator, the weight-sharing strategy, and the regularization constraints.

C. Label Space-Level Alignment

Feature representation-level alignment and data pixel-level alignment mainly aim to address the covariant shift issue. Even if the source and target domains are aligned on the feature level or pixel level, the domain gap might still exist in the label space. For example, the class distribution is imbalanced [221], [291]; the classifier's decision boundary learned on the source might fall into the margin on the target domain [288]. Label space-level alignment aims to address such label shift challenges. Existing methods of this strategy can be divided into four categories: pseudo-label-based classifier training, classifier decision boundary refinement, category-level discrepancy-based alignment, and category-level adversarial discriminative alignment.

Pseudo label-based classifier training directly employs the target pseudo labels for classifier training. The pseudo target labels are obtained from the source classifier and then used to retrain the source classifier [291], [292]. In [249], the target pseudo labels are jointly optimized with the subspace learning. Then, a task classifier is trained on the target data and corresponding pseudo labels.

Classifier decision boundary refinement tries to refine the decision boundary of the classifier trained on the labeled source domain so that it can adapt better to the target domain. Since no target data are involved in the task classifier training, the predictions on the target data might be unreasonably biased to one class and some target instances might be distributed closely to the decision boundary. Methods to address this issue include entropy minimization [259], [277], mutual information maximization [288], self-ensemble bootstrapping [259], and variation of information minimization [262]. Through such optimizations, the unlabeled target data influence the training of the source classifier, which generally maximizes the margins between the target examples and the decision boundary and thus increases the confidence on the target predictions [259].

Category-level discrepancy-based alignment incorporates the class distribution prior information when computing the discrepancy between different domains. Class reweighted MMD discrepancy [221], [291] is proposed by resampling the class distribution in the source domain, which enables the source domain to share the same class distribution with the target domain. The weighting ratio is obtained by the class marginal distributions based on the source ground truth and target pseudo labels. Classwise MMD discrepancy [221], [256] first computes the MMD of different classes and then combines them together. The difference lies in how to obtain the target pseudo labels, either directly from the source classifier [221] or learned by semisupervised graph label propagation [256]. Differently, Li et al. [283] proposed semantic metric learning to minimize intraclass variations and maximize interclass variations to make different class centers better separated.

Category-level adversarial discriminative alignment performs adversarial alignment of the source and target domains for each category. One direct method is to employ a discriminator that is adversarially trained within each domain [292]. Another group of methods is based on maximum classifier discrepancy [260], [265], [287]. Two classifiers are adversarially trained with the feature extractor to consider the relationship between class boundaries and target samples. In this way, more discriminative features can be generated in the high-density region near the decision boundaries [260].

D. Semisupervised DA

In semisupervised DA, there are some target samples with labels, but the number is much smaller than the labeled source samples. One straightforward way is task classifier training, i.e., employing the target labels to train the task classifier together with the source labels after feature- or pixel-level alignment [259], [293]. A similar idea is adopted in [294] and [295], but the target labels are assigned higher weights during classifier training. Some other adaptation techniques are also incorporated in [294], such as soft parameter sharing and class refinement MMD, which also considers the target labels. Besides the source and target data, the source and target labels are jointly exploited in the supervised shared subspace learning process, either with graph topology [296] or by dictionary learning and metric learning [297], both of which use the pairwise constrained knowledge between the labeled source and target data. Another group of adaptation methods that require labeled target data is parameter adaptation. Duan et al. [298] applied meta-learning to accelerate the transfer process. Specifically, the MAML algorithm includes three steps: feature extractor pretraining on the labeled source data, meta training of meta-learner and base learner on sampled meta tasks, and adaptation onto target subject during meta test.

E. Multisource DA

In MDA, there are multiple source domains, which are essential for personalized ESA. We may consider simply combining the different sources into one domain and then directly employing the single-source DA methods. However, such source-combined DA does not guarantee better performance than just using the single best one [308]. This is probably caused by the interference between different sources during training, because of the domain shift among different sources. Existing MDA methods mainly focus on working on the following aspects.

1) Domain Alignment and Domain Pairing Strategy: The alignment between the source domains and the target domain as well as among different source domains plays a key role in successful MDA. Similar domain alignment strategies to single-source DA have been employed in MDA, such as shallow subspace learning [303], [304], discrepancy-based feature-level alignment [263], [294], [302], [305], [306], adversarial discriminative feature-level alignment [300], [301], [306], [309], adversarial generative feature-level alignment [307], reconstruction-based feature-level alignment [308], [310], adversarial generative pixel-level alignment [308], and classifier decision boundary refinement [299]. The main difference lies in the domain pairing strategy, i.e., how to select pairwise or groupwise domains for the alignment component. Some popular means include: each source and the target pair [294], [299], [300], [301], [302], [303], the combined source and the target pair [304], [305], [306], [307], [308], selected sources and the target pair [263], [309], [310], and multisource discrimination [306].

2) Domain/Sample Weighting and Selection Strategy: Different source domains have different discrepancies with the target domain and different samples in the same source domain have different similarities with the target samples. Assigning them different weights could generate better alignment and transferability. The weights of the source samples are first learned to reduce the marginal probability differences based on MMD and the weights of different sources are learned to reduce the conditional probability differences based on the smoothness assumption [311]. A multiarmed bandit controller is designed to learn an optimal trajectory and mixture of domains for transfer to the target based on upper confidence bound [263]. Without requiring the domain label, all labeled source samples are combined and a curriculum is learned dynamically to assign weights to different source samples based on their proximity to the target domain distribution [307]. Based on the few labeled target samples and pretrained source classifiers, the domains with top similarities are selected for alignment [309].

3) Task Classifier Training and Fusion Strategy: After different domains are aligned and the weights of different domains and samples are learned, we can train classifiers based on the labeled source domain and transfer them to the target domain.

1) *Single Classifier Training:* Some methods combine all the labeled source samples and train a single classifier [300], [301], [302], [303], [304], [306], [307], [308]. Instead of dealing with different source domains and source samples equally, the losses of corresponding sources and samples are weighted [311]. At each round during alignment, only the samples in the selected source domain based on a multiarmed bandit controller involve in the classifier training [263].

2) *Multiple Classifier Training:* Considering that each domain has specific class boundaries, some methods train a classifier for each source domain [241], [294], [299], [302], [305], [310], [312]. However, this might yield different or even conflicting target distributions. Based on the fact that some sources might be aligned better with the target, some weighting-based solutions are proposed to combine the predictions of different classifiers, such as point-to-set metric using Mahalanobis distance [305]. In [299], a domain-agnostic (shared) classifier and multiple-domain-specific (private) classifiers are first optimized with multitask learning; during adaptation, a target-specific classifier is learned by considering the similarity between the source and target domains; the final target prediction is the combination of domain-agnostic classifier and target-specific classifier. In [310], the target prediction is obtained by fusing the shared classifier with different source-private classifiers weighted by similarities between source- and target-private encoders. Transductive parameter transfer [241], [312] employs a different pipeline, which tries to transfer the classifier parameters from the source domain to the target: first, the classifier is pretrained for each source domain; second, a regression model is trained between the source features and classifier

parameters; and finally, we can predict the parameters of the target classifier based on the regression model and target features.

F. Multimodal and Cross-Modal DA

Multimodal DA focuses on the DA problem for multimodal emotion recognition, where more than one affective signal is available. The main challenges include: how to fuse different modalities and how to align different modalities together with different domains. One straightforward trial is to extract features for each modality separately and then fuse them on the feature level (early fusion) [313], which can be followed by different single-modal adaptation strategies, such as adversarial discriminative alignment. After early fusion, another FC layer is appended [293]. To diminish the heterogeneity gap, covariant multimodal attention is designed to learn a common feature representation for multiple modalities [326]. The attended features are then adaptively fused. Domain-invariant features are learned by jointly aligning single-modal features, fused features, and attention scores. In the disentangled sentiment representation adversarial network [327], multiple modalities are first aligned through a cross-modality attention layer to obtain the joint representation. The sentiment information from the joint representation is disentangled without domain-specific style information via adversarial training. To encourage the disentangled representation to keep useful information as much as possible, a reconstruction loss is constrained during the sentiment embedding learning. To deal with specific modality missing issues, a product of experts is employed [295]. Together with VAE reconstruction constraint and cycle-consistency constraint, discriminator-based adversarial alignment is performed on the reconstructed data between the source and target domains for each modality.

Different from multimodal adaptation, cross-modal DA tries to transfer the knowledge learned from the source modality to another different target modality. Based on the assumption that the emotional content of speech correlates with the facial expression of the speaker, Albanie et al. [314] employed cross-modal distillation to transfer the annotations from the facial domain to the speech domain. A strong teacher network is first learned for FER, and then, a student network is trained to reproduce the features of the teacher model. A similar cross-modal adaptation is adopted in [315]. A GAN conditioned on the source faces and noise is proposed to generate intermediate spectrograms that are adversarially aligned with the target spectrograms. Then, the classifier trained on the intermediate domain can be better transferred to the target domain.

G. Discussion

1) Quantitative Comparison: In order to give a general understanding on how existing domain-adaptive ESA methods perform, we conduct experiments to evaluate some representative methods for textual sentiment classification and FER.

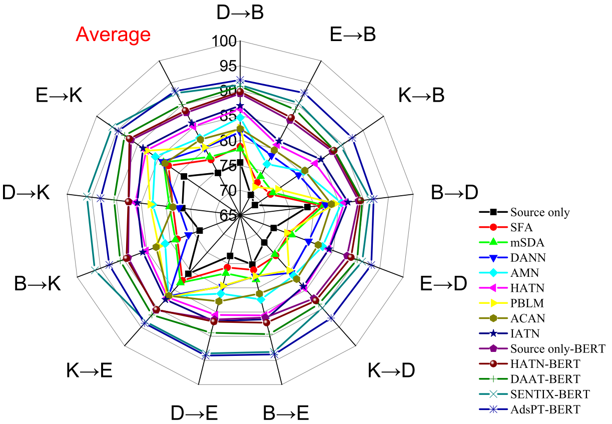


Fig. 10. ***Performance comparison of representative single-source UDA methods for textual sentiment classification on Amazon benchmark, measured by classification accuracy (%).***

a) Datasets: The Amazon benchmark dataset [223] that is widely used for textual sentiment binary classification contains four domains of product reviews on Amazon: books (B), DVD (D), electronics (E), and kitchen (K). We perform 12 adaptation tasks between every two of the four domains. Following [283], the employed datasets for FER include: RAF-DB2.0 [221], AffectNet [328], FER2013 [329], CK+ [220], MMI [330], Oulu-CASIA [331], and JAFFE [332]. The former three datasets contain unconstrained facial images, while the rest are mainly composed of laboratory-controlled ones. RAF-DB2.0 is selected as the source domain to transfer the learned knowledge to other datasets.

b) Compared methods: The compared UDA methods for textual sentiment classification include LSTM-based source only (direct transfer without adaptation), SFA [244], mSDA [223], DANN [323], AMN [266], HATN [267], PBLM [333], ACAN [260], IATN [268], BERT-based source-only, HATN-BERT, DAAT-BERT [269], SENTIX-BERT [334], and Adspt-BERT [301]. The compared methods for FER include ResNet-50-based source only [226], MMD [221], ECAN [221], AGRA [281], and JDMAN [283].

c) Results and analysis: The radar maps for performance comparison of different methods are shown in Figs. 10 and 11. From the results, we have the following brief observations.

1) The source-only method that does not take any adaptation actions achieves the worst performance in almost all adaptation settings. The existence of domain shift leads to the model's low transferability from the source to the target domain. This observation motivates the necessity of DA.

2) Better feature representation significantly matters in domain-adaptive ESA. The source-only-BERT method

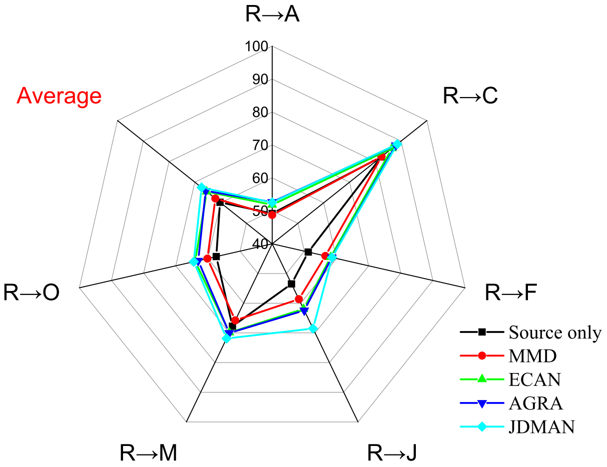


Fig. 11. ***Performance comparison of representative single-source UDA methods for FER from RAF-DB2.0 to different target benchmarks, measured by classification accuracy (%).***

outperforms the best non-BERT DA methods (i.e., 85.9 of LATN versus 88.3 of source-only-BERT).

3) The compared deep methods generally perform better than the shallow methods (e.g., IATN versus SFA).

4) If one source is more similar to the target than the other source (D⟶B versus K⟶B), the adaptation results will be better, which is consistent across different methods.

2) *Qualitative Comparison:* We compare the abovementioned domain-adaptive ESA methods from the following three perspectives.

a) On Different Levels of Alignment Strategies: Feature representation-level alignment is the most widely employed strategy and can be used in various types of ESA tasks. Data pixel-level alignment is often used for images since the generated intermediate domain for other modalities often makes no sense. For example, the generated text might be able to fool a discriminator, but its conveyed meanings are probably confused. Label space-level alignment highly depends on the accuracy of pseudo labels. If the pseudo labels can be well assigned, the adaptation performance can be boosted to a large extent by taking the label shift into consideration.

b) On Different Domain-Invariant Learning Strategies: Discrepancy-based alignment methods usually have good theoretical guarantees, add few or no parameters to the backbone, have higher computation efficiency, and can be easily optimized; they are not so applicable to large and complex domains with diversified affective signals. Adversarial discriminative alignment approaches can be supported by specific generalization bound and risk analysis, require a large amount of data to train, are relatively difficult to optimize, and do not always work well on small datasets. Self-supervision-based alignment methods do not have a strong theoretical guarantee and have competitive computation efficiency, data scalability, optimizability, and performance in-between discrepancy-based and adversarial strategies.

c) On Typical DA Setting and Other Settings: With some target labels, semisupervised domain adaption can be expected to outperform unsupervised adaptation. However, in practice, the target labels, even a small number, might be difficult to obtain. Multisource and multimodal DA can enrich the learned knowledge from more source domains and modalities, which might also bring conflict. How to effectively exploit the complementary information is worth investigating. As summarized in [4], multimodal affective signals have the advantages of data complementarity, model robustness, and performance superiority. Therefore, domain-adaptive multimodal ESA is an inspiring research topic but has been rarely investigated. On the one hand, compared to domain-adaptive single-modal ESA, multimodal DA encounters more challenges, such as cross-modality inconsistency and cross-modality imbalance [4]. There are also much fewer publicly available datasets to evaluate domain-adaptive multimodal ESA methods. On the other hand, unlike traditional cross-domain multimodal objective semantic understanding tasks, such as event recognition [343], cross-domain multimodal ESA needs to deal with the specific characteristics of emotions, e.g., subjectivity and complexity, as introduced in Section I. Because of the presence of affective gap, the intraclass variation for multimodal ESA is much larger. To better deal with such challenges, we might consider the following brave new ideas: 1) pretraining on large-scale unlabeled multimodal affective signals via self-supervised learning to enhance the features' representation ability; 2) incorporating emotion correlations, such as emotion hierarchy and emotion similarity, to design effective cross-domain alignment methods (e.g., emotion-aware domain-invariant disentanglement); and 3) exploring novel attention-based fusion strategies to combine different modalities with an optimal balance. Cross-modal DA has much potential in real applications, but it is more challenging to learn the correspondence and matching between different modalities.

IX. DOMAIN-GENERALIZABLE ESA

Domain-adaptive ESA requires that though without labels, the target data are accessible during adaptation training. In real-world ESA applications, such an assumption might not hold, i.e., the target data are unavailable before the deployment of the adaptation model. DG aims to address this issue by learning a model from one or more source domains that can generalize well to an unseen target domain. The difference between DA and DG is shown in Fig. 12.

Similar to DA, we can classify existing DG methods into three categories based on the different generalization levels: feature representation-level generalization, data pixel-level generalization, and label/classifier-level generalization. Table 9 categorizes the existing representative methods for domain-generalizable ESA.

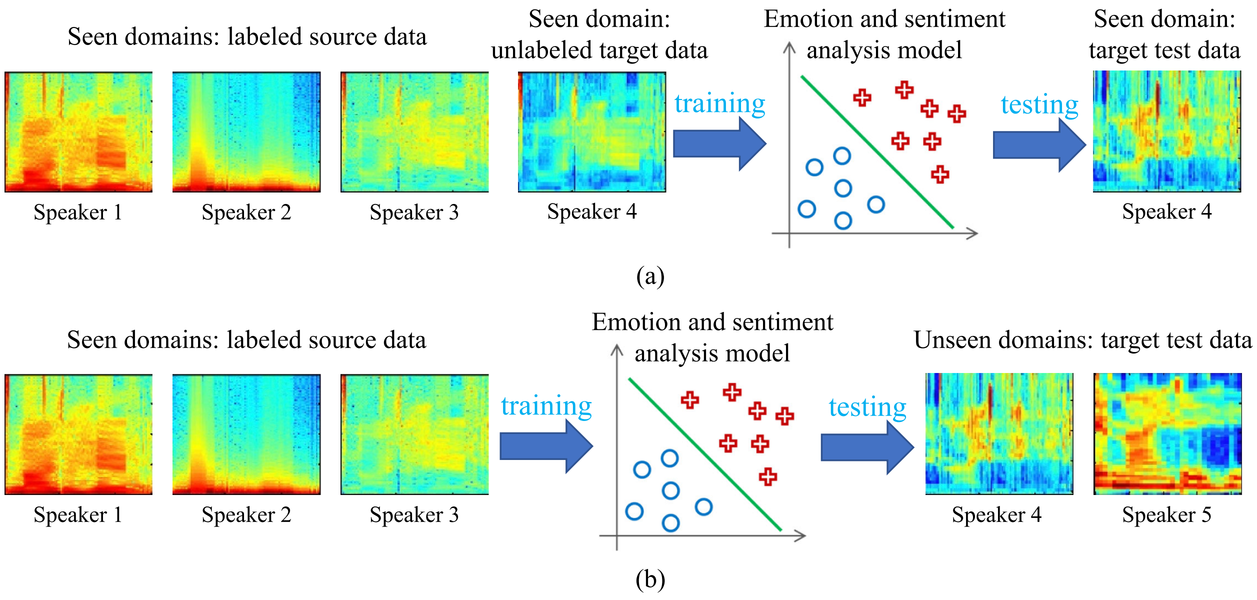

Seen domains: labeled source data
Seen domain: unlabeled target data
Emotion and sentiment analysis model
Seen domain: target test data
Speaker 1
Speaker 2
Speaker 3
Speaker 4
training
testing
Speaker 4
(a)
Seen domains: labeled source data
Emotion and sentiment analysis model
Unseen domains: target test data
Speaker 1
Speaker 2
Speaker 3
training
testing
Speaker 4
Speaker 5
(b)

Fig. 12. ***Comparison between domain-adaptive and domain-generalizable ESA taking speech emotion recognition as an example. The main difference lies in whether unlabeled target data are available during training. (a) Domain-adaptive ESA. (b) Domain-generalizable ESA.***

A. Feature Representation-Level Generalization

Learning generalizable feature representations is the most widely employed strategy for DG, especially for multisource DG. When multiple source domains are available, similar pipelines to DA can be employed to conduct feature-level generalization. The assumption is that if generalizable feature representation can be learned for the different sources, it is also generalizable to the unseen target domain. Typically, two categories of techniques have been used: domain-invariant feature learning and feature disentanglement.

The former category usually performs feature-level alignment among the source domains to learn domain-invariant features, such as discrepancy-based alignment [335], adversarial discriminative alignment [335], [336], [337], and self-supervised alignment [338]. Li et al. [335] tested two discrepancy-based alignments, i.e., MMD and CORAL. DANN-based alignment is employed in [335], [336], [337]. Besides the domain classifier, another subject classifier is designed [337]. Shen et al [338] assumed that the neural activities of different subjects are similar when receiving the same emotional stimuli. Based on this assumption, they employed contrastive learning to minimize the intersubject differences by maximizing the similarity in feature representations across subjects. The latter category tries to disentangle the features into domain-agnostic (shared) features and domain-specific (private) features, and then, the shared features can be used for generalization [339].

B. Data Pixel-Level Generalization

This kind of generalization method focuses on the manipulation of pixel-level data to assist generalizable feature learning. Existing methods can be classified into two categories: data augmentation and data generation.

Data augmentation is often used to regularize the training of deep learning models to avoid overfitting. The basic idea is to augment the original data with newly transformed data while preserving the labels. Popular augmentation operations include randomization and transformation. For example, standard Gaussian noise is added to perturb the training data [340]. Roy and Cambria [341] employed adversarial data augmentation [344] in an iterative procedure to augment the source data with examples from a fictitious target domain that is "hard" under the current model.

Data generation tries to generate new samples with diverse styles to boost the performance of generalization.

Table 9 Categorization and Representative Methods for Domain-Generalizable ESA

Category	Methods	References
Feature representation-level generalization	Domain-invariant feature learning	[335–338]
	Feature disentanglement	[339]
Data pixel-level generalization	Data augmentation	[340, 341]
	Data generation	[335]
Label/classifier-level generalization	Specific learning	[140, 335, 340–342]

In [335], Mixup [345] and group distributionally robust optimization (GroupDRO) [346] are used. Without the requirement to train generative models, Mixup generates new data by linear interpolation between two samples and between their corresponding labels. The weight of interpolation is sampled from a β distribution. GroupDRO leverages prior knowledge of spurious correlations to define groups over the training data.

C. Label/Classifier-Level Generalization

Specific learning strategies are also exploited to promote the generalization ability. A curriculum-guided coarse-to-fine learning [140] is employed to explore large-scale web images with diverse concepts. MetaReg is proposed to meta-learn the regularization parameters [342]. Representation self-challenging [347] is employed in [335] to discard the representation associated with the higher gradients at each epoch and focus the model to perform prediction with the remaining information during training. A triplet loss-based metric learning is designed [340]. Specifically, by adding a constraint to reduce the positive distance within the same emotion class, an improved version of the triplet loss is proposed to learn more generalizable features. A soft labeling formulation is proposed by considering the shift in label distributions across domains [341].

D. Discussion

According to the number of available source domains, existing methods can also be divided into multisource DG [335], [336], [337], [338], [339], [340], [342] and single-source DG [140], [341]. DG originates from the multisource setting with the motivation of leveraging MDA to learn domain-invariant feature representations that are believed to be able to generalize well to the unseen target domain. Single-source DG is more challenging and is related to the investigation of model robustness under image corruption. Generally, single-source DG methods are mainly based on data manipulation and specific learning strategies, while domain-invariant feature learning dominates the research on multisource DG.

X. APPLICATIONS

In this section, we will elaborate on several applications of ESA with label-efficient methods to alleviate the challenge of limited resources.

A. Emotional Comfort Assistant for the Elderly

According to the United Nations World Census, many countries in the world have entered an aging society. With the rapid development of economy and society in the 21st century, the problem of population aging in the world is becoming more and more prominent. With the decline of physical function and changes in lifestyle and family structure, the elderly are faced with both physical and mental challenges. Among them, the problem of mental health is particularly prominent and has become a social phenomenon that cannot be underestimated. This requires us to pay attention to the mental health of the elderly, offering them more companionship and care. Thus, it is necessary to develop service platforms and emotional comfort assistants for the elderly to improve their quality of life [348].

By using the ESA technology, the emotional comfort assistant collects information such as language, facial expressions, and voice during the dialog and interaction with the elderly, analyzes the current emotional state of the elderly [349], and generates appropriate reply utterances that can achieve the effect of emotional soothing, just like their children or close friends accompanying them. Hopefully, it becomes an important way for the elderly to vent their emotions and seek comfort. Also, there is a shard task [350] aiming to predict both valence and arousal scores in the speech of elderly individuals as three-class problems, motivating more research efforts in this field.

However, in order to achieve the above emotional comfort function, the label-efficient approach is needed to address some of the current challenges in resources to train the model. For example, although there are relevant public datasets for emotional support dialog systems [351], the content of the data is too general and not specific to the elderly. Therefore, it is challenging to successfully transfer the models trained on the existing data to the elderly conversation domain to generate responses that better match the expressions and language habits of the elderly. In addition, the current data are limited to the English language scenario, and how to meet the multilingual scenario of emotional comfort for the elderly in various countries around the world still needs to be addressed. Finally, it might be difficult to capture the facial expressions of the elderly, and how to transfer the FER model of young people to the elderly requires label-efficient approaches to solve the problem.

B. Emotional Assistant for Car Driving

With the rapid development of society and economics and the continuous improvement of people's living standards, vehicles have generally entered thousands of households. With this, road congestion and traffic accidents are becoming more and more obvious. In first-tier cities, traffic congestion has become a "nightmare" in the mind of every car owner, leading to a common phenomenon that many car owners will become "road rage" and suffer from a terrible experience in the driving process. Thus, the design of an emotional assistant to comfort the emotional states of the drivers is vital to improving driving safety, comfort, and acceptance of intelligent vehicles.

To this end, the emotional assistant for car driving can collect the owner's voice and expression information and provide positive guidance for the owner's angry emotions, such as telling jokes, introducing topics of interest to the owner, and talking about the owner's next plan.

In addition, when the owner is in a state of fatigue or sadness, he/she is often reluctant to show emotions through words, which greatly relies on the system's recognition of the owner's facial expressions to provide positive emotional guidance. In recent decades, emotional assistance for car driving has attracted increasing attention [352]. Researchers exploit various signals to provide emotion analysis in automatic contexts, such as facial expression [353], [354], physiological reaction [355], [356], pose [357], [358], speech [359], [360], and behavior [361], [362]. Furthermore, some methods combine multimodal signals to obtain superior performance [363], [364]. However, it is difficult to obtain accurate physiological signals during driving, and speech may be missing for an extended period of time [365], [366]. Therefore, many researchers focus on FER to provide more reliable emotional assistance.

C. Caring for Mental Health

According to research studies, the number of people suffering from mental illness has increased dramatically worldwide after the 2019-nCoV epidemic. However, the current global mental health system is very weak. In low- and middle-income countries, even 76%–85% of people with mental disorders do not receive treatment. Against this backdrop, researchers hope to use artificial intelligence (AI) technology to alleviate such resource shortages.

The mental health care system provides patients with mental illnesses (such as depression and dryness) with a convenient and efficient outlet for their emotions, acting as a good listener and simply trying to detoxify the patients' negative emotions by means of empathy, or psychological strategies. Many researchers have begun to study how to use deep learning models to detect users' psychological disorders. Gui et al. [367] proposed to use reinforcement learning to screen depression-related posts and aggregated all representations with an RNN to diagnose depression in users. Also, Zogan et al. [368] identified users' depression by using hierarchical neural networks to model users' textual and social media behaviors. Delahunty et al. [369] chose to use user interview records and combined psychological questionnaires with user interview content to detect depression. As for treatment, it provides patients with comprehensive information about professional medical institutions when necessary, to facilitate their timely access to medical care.

However, due to patient privacy and possible patient resistance to discussing illness-related content, high-quality data collection on patients with mental illness is difficult, requiring our model to acquire as many patient interaction skills as possible under the setting of low data resources. To address such a data scarcity problem, some researchers have attempted to use semisupervised methods to obtain large-scale annotated data from a number of small, specialized datasets of high annotation quality. For depression research on social media, some previous methods [369], [370] attempt to use a small dataset of symptoms to train models to label the symptoms embedded in posts on social media. Also, Wang et al. [371] leveraged a small but professionally labeled cognitive distortion dataset to annotate cognitive errors on social media.

D. Intelligent Customer Service

For companies, customer service is very important. It brings great value to companies in terms of enhancing brand reputation, improving customer experience, identifying problems, and improving competitiveness. Chung et al. [372] proposed metrics of accuracy, credibility, and communication ability as dimensions to measure the quality of communication between customers and sales and investigate the impact of communication quality on customer satisfaction in the luxury retail environment. To achieve excellent customer service, the key is to fully understand customer needs and keep users in a good emotional state while solving their problems.

According to [373], customers expect AI to be considerate and have good communication skills similar to humans. Thus, in response to customers' various negative emotions, intelligent customer service systems should promptly switch to different tactics such as human politeness strategy [374] or combine different psychological strategies to stabilize customers' emotions. Through this, it makes users feel respected, understood, and valued, and enhances their experience through "warm" customer service.

However, there are some challenges related to data resources to realize intelligent customer service. For example, customer service data are naturally domain-related, and different domains have their specific monikers and after-sales issues. It is crucial to train a model with sufficient generalization capability on limited data resources. In addition, since the network and online signals of customer service platforms may fluctuate, the data collected from them may be missing, and our model should also be robust to the problem of missing data.

E. Online Education

Under the general environment of the 2019-nCoV epidemic prevention and control, online education has gradually come into the lives of students and teachers. In the learning environment of distance education, the separation of teachers and students in time and space results in a lack of effective communication between them, which becomes a problem that plagues the majority of educators.

Online education systems could provide support and guidance to students (with special attention to students' attitudes, feelings, beliefs, and emotions). Furthermore, they link emotions to cognitive development, i.e., how students feel as learners and how they feel about the subject they are studying. Also, they collect students' listening status through cameras during class or collect students' feedback comments after class to keep track of students' learning status and help teachers to more rationally arrange the teaching content. Using content analysis, lag sequential analysis, logistic regression, and grouped regression approaches, Liu et al. [375] attempted to uncover the relationship between discussion pacings, learners' cognitive presence, and learning achievements of participants on a Chinese online course platform.

In order to achieve this function, there is also the problem of DA. As a unique group, students have their own emotional expressions and communication styles, and students of different ages should be treated differently to achieve a comfortable and efficient emotional communication effect.

XI. FUTURE DIRECTIONS

As stated in the previous sections, significant progress has been made to improve the performance of LeESA. However, there are still several unsolved open issues and future directions that deserve more effort to investigate.

A. More Practical Settings

1) Combination of Different LeESA Paradigms: Existing methods of LeESA mainly focus on one specific label-efficient setting, which might fail to meet the complicated requirements in practice. For example, current domain-generalizable ESA methods usually assume that the source domains are fixed, the labeled data from different sources are provided once, and the emotion categories are predefined. The practical case might be that we have incremental source domains, labeled data, and emotion categories. In such cases, the combination of DG and incremental learning should be considered. On the other hand, more source domains and more labeled source data do not guarantee better generalization performances. Some theoretical analysis on the generalization upper bound would be helpful to determine whether the new source domain is beneficial or detrimental to the generalization performance. Combining incremental learning with other label-efficient learning paradigms, such as SSL and weakly supervised learning, is also unexplored.

Even in the same LeESA paradigm, the combinations of different settings might also be necessary. For example, in domain-adaptive ESA, there might be multiple source domains, multiple modalities, and different ESA tasks, and the source and target label sets are different [55]. Designing a general and universal framework that can deal with different settings would make it easily deployed in practice. When settings are changed, just a "ON–OFF" switch would work.

2) More Practical Learning and More Flexible Inference: Compared to the traditional machine learning paradigm, LeESA we focused on in this article has greatly relaxed the stringent requirements in data acquisition. Nonetheless, it is still far away from practical use. In the case of incremental learning, for example, although it is no longer required to acquire training data at once, there is still an assumption that data are provided in an organized manner according to different concepts. In concrete, the CIL setting organizes training data by classes, wherein each learning phase model learns only samples of the classes belonging to the respective phase. In the DIL setting, the classes of data are known and fixed. As for the TIL setting, sequential learning tasks are isolated from each other and the task IDs are provided during testing. Such a learning scenario is still an unattainable ideal in practice. Therefore, there is a strong need to investigate more practical learning settings. One potential direction is data incremental learning, where the data stream to learn is provided without any requirements related to the notion of task, class, or domain [376].

On the other hand, existing LeESA methods are typically based on the assumption that the testing data have the same format as the training data. As we know, multimodal ESA with effective fusion strategies performs better than single modality [4]. However, during inference, it might be difficult to collect the testing data from different modalities. Therefore, designing LeESA algorithms with multimodal data during training that enables flexible inference is a promising direction. In such cases, no matter which modalities are available during inference, the emotion and sentiment can be predicted. In this regard, one encouraging study about multimodal learning and single-modal prediction (MLSP) can be found in [377], which engages multimodal cues jointly during learning and enables making predictions using only one of these cues.

3) Model-Efficient, Hardware-Efficient, and Design-Efficient ESA: Besides the label efficiency, the model complexity and compute capacity are another two factors that we need to consider when deploying ESA applications on edge devices, such as mobile phones and autonomous vehicles. How to design deep neural networks for ESA to obtain compact models with balanced performance, how to codesign deep neural networks and hardware processors to accelerate the training and inference, and how to design optimal neural networks under constraints of given processors to improve the ESA performance are still open and have not been deeply touched. Based on recent progress in machine learning and multimedia computing, designing model-, hardware-, and design-efficient ESA models is an inspiring topic to explore. For example, Amiriparian et al. [378] designed an open-source, lightweight transfer learning framework, termed DeepSpectrumLite, for real-time speech and audio recognition on embedded devices.

4) Groupwise ESA: Recognizing the dominant emotion and sentiment for given stimuli is direct but ignores emotion's subjectivity, while personalized emotion analysis is ideal but impractical in real-world applications. Groupwise ESA, a tradeoff between dominant ESA and personalized ESA, would make more sense to balance the accuracy and practicability [379]. Groupwise ESA plays essential roles in advertisement and recommender systems. The main challenge is how to classify users into different groups. Various factors, such as interests, education background, and personality, might matter in the emotion perception process. Exploiting effective data mining strategies based on social network connections is probably a feasible solution.

To facilitate the development of the abovementioned practical settings, we need to build corresponding benchmark datasets with high quality and large scale. Some feasible channels include employing a hierarchical emotion model, exploiting crowdsourcing platforms, collecting multimodal data, and improving the reality of synthetic data.

B. New Methodologies for LeESA

1) Exploring Large-Scale Pretraining and Self-Learning for LeESA: The rapid development of large-scale pretraining and self-learning has been witnessed recently, such as BERT [84], GPT3 [81] in NLP and DINO [380], and MAE [58] in CV. First pretraining on an existing large-scale dataset to learn general representation and then fine-tuning the pretrained model to downstream tasks to explore task-specific representation is believed to perform well. This pipeline still requires a relatively large amount of labeled data in the downstream task (ESA in our case), which does not always hold in LeESA. Combining such pretraining and self-learning techniques with label-efficient learning methods is a promising direction. On the other hand, pseudo labels are widely used in LeESA, such as FSL and DA. However, how to generate high-quality pseudo labels is still challenging. It has been proven that the self-attention maps of pretrained large models contain rich semantic information, which has the potential in generating reliable pseudo labels without supervision. How to link such objective semantics with subjective emotions during pseudo-label generation is interesting.

2) Incorporating Knowledge From Multidisciplinary Studies: ESA is an interdisciplinary task that involves psychology, neuroscience, cognitive science, machine learning, and so on. Simply employing advanced machine learning techniques might be able to improve the performance to some extent, but it is difficult to reach humans' generalization capability. Exploring the studies in neuroscience and cognitive science on brain mechanisms that explain how emotion is evoked would significantly boost the performance in LeESA. Some simple mechanisms have been used, such as emotion lateralization [270] and bihemisphere structure [274], [275], which have been demonstrated to be effective. Breakthroughs would be made conditioned on how deep investigation of learning "common sense" mechanisms can go.

In psychology, different emotion theories, such as evolutionary theory, have been proposed to understand the how and why behind emotions. Some basic emotion correlations have been explored in ESA, such as emotion hierarchy (e.g., polarity-emotion hierarchy) [381], [382] and emotion similarity [383]. However, these emotion theories and correlations have not been deeply explored in LeESA.

3) Theoretical and Interpretable Analysis: Existing LeESA methods mainly focus on designing effective learning algorithms to improve performance but lack theoretical and interpretable analysis. Theoretical analysis can guide the design of learning algorithms [55]. For example, in semisupervised ESA, how much labeled data is required with the help of unlabeled data to reach the performance of traditional supervised methods? In class-incremental ESA, how to preserve the discriminative power for existing classes and meanwhile increase the ability to distinguish new classes? In domain-adaptive and generalizable ESA, what is the upper bound and does every source domain contribute?

Interpretable analysis can provide us with novel insights to understand how LeESA models work and why they fail [384]. For example, in weakly supervised ESA, we can display the filtered noise labels and analyze their characteristics; in domain-adaptive ESA, we can visualize the learned domain-invariant features to see whether different classes can be correctly classified, compare the generated intermediate domain for pixel data-level alignment methods to see whether they look more familiar with the target domain, and visualize the attention map before and after adaptation to test whether more discriminative local regions are paid more attention.

4) Unified Label-Efficient Framework for Different ESA Tasks: Existing LeESA methods mainly focus on one specific modality, such as facial expression, speech, or text. When developing new LeESA algorithms, it is difficult to compare with the latest and state-of-the-art techniques in different modalities [385]. This restriction also causes practical deployment challenging. What are the common and private properties among different ESA tasks? Can we explore those properties for a unified label-efficient framework that works for different modalities? Recent advance on large-scale pretraining makes this unified framework possible, which we believe will appear in the near future.

C. Ethical and Legal Restriction

1) Ethics: The data collection of explicit affective cues involves personal information, such as face and voice, which can reflect human identity. To obtain accurate emotion and sentiment labels, humans are also employed to annotate the collected data for both explicit affective cues and implicit affective stimuli. The collection and annotation protocols must be carefully designed to protect the involved humans. All the participants should be notified in advance about the target of the data collection and annotation to make sure that they join the process voluntarily. The dataset organizer is expected to filter all obscene languages or bad expressions, avoiding generating offensive speech, hate speech, and so on that are harmful to the users. During data distribution, some specifically designed consent forms are usually required for applicants and only necessary information can be shared based on the consent forms. When deploying the ESA models in applications, users should be notified about the purpose of such settings (e.g., security). Beneficial experiences on ethics can be learned from psychology [386].

2) Privacy: The ethics above mainly focuses on data collection, data distribution, and application deployment. Here, we discuss how to protect privacy during model training. Personal private and sensitive information in users' profiles is easy to leak out if storing them in a centralized way to train the ESA model. Currently, it is preferable to store them on individual devices and train a private model for each node without sharing the data. Federated learning provides such a mechanism for privacy protection [387]. Combining federated learning with label-efficient learning techniques is a reasonable solution. For example, in federated adversarial DA [388], models are trained separately on each source domain and a dynamic attention mechanism is employed to aggregate their gradients and thus update the target model. The limitation is that federated adversarial DA is vulnerable to privacy leakage attacks. Source data-free DA can better deal with the privacy issue without access to the source data and has been increasingly studied in several CV and NLP tasks [389]. However, source-free DA for ESA has still been rarely explored and calls for more attention.

3) Avoidance of Misuse: The rapid development of deep learning helps to improve the LeESA performance but meanwhile increases the risk of misuse. For example, fake faces and voices that are synthesized by deep generative models can be used to fool machines even humans, resulting in possible fraud. On the one hand, techniques to detect fake versus real identities should be better developed to minimize detection errors [390], [391]. From a technical point of view, we make it difficult and even impossible to use ESA techniques in an unauthorized way. On the other hand, the corresponding laws should be established internationally regarding the possible misuse [392]. From a legal perspective, we increase the severity of punishment to lower the misuse rate.

XII. CONCLUSION

In this article, we attempted to provide a comprehensive introduction to LeESA with a focus on representative and the latest methods. Based on different settings of training sample labels, emotion categories, and available data domains, we classified existing methods into seven label-efficient paradigms. We summarized and compared each paradigm with our own views included. Some promising applications and potential future directions are also discussed. Because of the multidisciplinary nature of the topic, we encourage interested readers to follow cited relevant surveys for a wider overview and detailed research papers for a deeper understanding. Although significant progress has been made, there is still a long way to go to enable machines to realize AEI with label efficiency.

Acknowledgment

The authors would like to sincerely thank the anonymous reviewers for their suggestive comments to help them improve this article. ■

REFERENCES

[1] M. Minsky, *The Society of Mind*. New York, NY, USA: Simon and Schuster, 1986.

[2] D. Schuller and B. W. Schuller, "The age of artificial emotional intelligence," *Computer*, vol. 51, no. 9, pp. 38–46, Sep. 2018.

[3] J. Z. Wang et al., "Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion," *Proc. IEEE*, early access, May 23, 2023, doi: 10.1109/JPROC.2023.3273517.

[4] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Process. Mag.*, vol. 38, no. 6, pp. 59–73, Nov. 2021.

[5] S. Zhao et al., "Affective image content analysis: Two decades review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6729–6751, Oct. 2022.

[6] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion," *J. Personality Social Psychol.*, vol. 17, no. 2, pp. 124–129, 1971.

[7] H. Schlosberg, "Three dimensions of emotion," *Psychol. Rev.*, vol. 61, no. 2, pp. 81–88, 1954.

[8] S. Zhao et al., "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 574–587, Oct. 2020.

[9] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 83–92.

[10] K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 860–868.

[11] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 147–160, Apr. 2018.

[12] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, pp. E1454–E1462, Apr. 2014.

[13] S. Li and W. Deng, "Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 884–906, Jun. 2019.

[14] X. Ben et al., "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5826–5846, Sep. 2022.

[15] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, Oct. 2012.

[16] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Sep. 2022.

[17] F. Z. Canal et al., "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Inf. Sci.*, vol. 582, pp. 593–617, Jan. 2022.

[18] M. Jampour and M. Javidi, "Multiview facial expression recognition, a survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2086–2105, Oct. 2022.

[19] X. Li et al., "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 563–577, Oct. 2018.

[20] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1253, 2018.

[21] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 49–67, Jan. 2023.

[22] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 845–863, Apr. 2022.

[23] G. Brauwers and F. Frasincar, "A survey on aspect-based sentiment classification," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 65:1–65:37, 2023.

[24] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 40:1–40:30, 2012.

[25] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020.

[26] R. Panda, R. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 68–88, Jan. 2023.

[27] S. Zhao et al., "Computational emotion analysis from images: Recent advances and future directions," in *Human Perception of Visual Information*. Cham, Switzerland: Springer, 2022,

pp. 85–113. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-81465-6

[28] A. Ortis, G. M. Farinella, and S. Battiato, "Survey on visual sentiment analysis," *IET Image Process.*, vol. 14, no. 8, pp. 1440–1456, Jun. 2020.

[29] F. Noroozi, C. A. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 505–523, Apr. 2021.

[30] M.-A. Mahfoudi, A. Meyer, T. Gaudin, A. Buendia, and S. Bouakaz, "Emotion expression in human body posture and movement: A survey on intelligible motion factors, quantification and validation," *IEEE Trans. Affect. Comput.*, early access, Dec. 2, 2022, doi: 10.1109/TAFFC.2022.3226252.

[31] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul. 2019.

[32] S. Saganowski, B. Perz, A. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Trans. Affect. Comput.*, early access, May 20, 2022, doi: 10.1109/TAFFC.2022.3176135.

[33] X. Li et al., "EEG based emotion recognition: A tutorial and review," *ACM Comput. Surv.*, vol. 55, no. 4, pp. 79:1–79:57, 2023.

[34] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, no. 1, pp. 3–14, Sep. 2017.

[35] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Inf. Fusion*, vol. 59, pp. 103–126, Jan. 2020.

[36] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Inf. Fusion*, vol. 91, pp. 424–444, Mar. 2023.

[37] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, Jun. 2014.

[38] Y. Bengio, A. C. Courville, and P. Vincent, "Unsupervised feature learning and deep learning: A review and new perspectives," 2012, *arXiv:1206.5538*.

[39] M. Akçakaya, B. Yaman, H. Chung, and J. C. Ye, "Unsupervised deep learning methods for biological image reconstruction and enhancement: An overview from a signal processing perspective," *IEEE Signal Process. Mag.*, vol. 39, no. 2, pp. 28–44, Mar. 2022.

[40] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[41] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 8, 2022, doi: 10.1109/TKDE.2022.3220219.

[42] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, May 2021.

[43] F. Pourpanah et al., "A review of generalized zero-shot learning methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4051–4070, Apr. 2023.

[44] Y. Hu, A. Chapman, G. Wen, and D. W. Hall, "What can knowledge bring to machine learning?—A survey of low-shot learning for structured data," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 3, pp. 1–45, Jun. 2022.

[45] V. Losing, B. Hammer, and H. Wersing, "Incremental on-line learning: A review and comparison of state of the art algorithms," *Neurocomputing*, vol. 275, pp. 1261–1274, Jan. 2018.

[46] E. Belouadah, A. Popescu, and I. Kanellos, "A comprehensive study of class incremental learning algorithms for visual tasks," *Neural Netw.*, vol. 135, pp. 38–54, Mar. 2021.

[47] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer, "Class-incremental learning: Survey and performance evaluation on image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5513–5533, May 2023.

[48] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[49] D. Zhang, J. Han, G. Cheng, and M.-H. Yang, "Weakly supervised object localization and detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5866–5885, Sep. 2022.

[50] S. Jialin Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[51] F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[52] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[53] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 51:1–51:46, 2020.

[54] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.

[55] S. Zhao et al., "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 473–493, Feb. 2022.

[56] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4396–4415, Apr. 2023.

[57] J. Wang et al., "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, early access, May 26, 2022, doi: 10.1109/TKDE.2022.3178128.

[58] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[59] G. Paltoglou and M. Thelwall, "Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–19, Sep. 2012.

[60] X. Hu, J. Tang, H. Gao, and H. Liu, "Unsupervised sentiment analysis with emotional signals," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 607–618.

[61] G. Jia and J. Yang, "S^2-VER: Semi-supervised visual emotion recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 493–509.

[62] A. Kumar, P. Gupta, R. Balan, L. B. M. Neti, and A. Malapati, "BERT based semi-supervised hybrid approach for aspect and sentiment classification," *Neural Process. Lett.*, vol. 53, no. 6, pp. 4207–4224, Dec. 2021.

[63] Z. Wei et al., "Learning visual emotion representations from web data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13103–13112.

[64] Z. Min, "Drugs reviews sentiment analysis using weakly supervised model," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Mar. 2019, pp. 332–336.

[65] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1151–1160.

[66] J. Gonzalez and L. Prevost, "Personalizing emotion recognition using incremental random forests," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 781–785.

[67] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1068–1072, Sep. 2014.

[68] F. Chen, R. Ji, J. Su, D. Cao, and Y. Gao, "Predicting microblog sentiments via weakly supervised multimodal deep learning," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 997–1007, Apr. 2018.

[69] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, and K. Lei, "Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 579–587.

[70] G. Zhang and A. Etemad, "Deep recurrent semi-supervised EEG representation learning for emotion recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–8.

[71] R. Xia, C. Zong, X. Hu, and E. Cambria, "Feature ensemble plus sample selection: Domain adaptation for sentiment classification," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 10–18, May 2013.

[72] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 3, pp. 529–545, Mar. 2017.

[73] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.

[74] A. Andreevskaia and S. Bergler, "Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses," in *Proc. Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2006, pp. 209–216.

[75] A. Esuli and F. Sebastiani, "SentiWordNet: A high-coverage lexical resource for opinion mining," *Evaluation*, vol. 17, pp. 1–26, Jan. 2007.

[76] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Computing Attitude and Affect in Text: Theory and Applications*. The Netherlands: Springer, 2006, pp. 1–10. [Online]. Available: https://link.springer.com/chapter/10.1007/1-4020-4102-0_1

[77] X. Zhou and W. Y. Wang, "MojiTalk: Generating emotional responses at scale," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1128–1137.

[78] Y. Zhao, B. Qin, and T. Liu, "Collocation polarity disambiguation using web-based pseudo contexts," in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 160–170.

[79] Z. Zeng et al., "A variational approach to unsupervised sentiment analysis," 2020, *arXiv:2008.09394*.

[80] Y. Wang, S. Wang, J. Tang, H. Liu, and B. Li, "Unsupervised sentiment analysis for social media images," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2378–2379.

[81] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[82] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[83] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[84] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[85] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–18.

[86] A. Singh et al., "FLAVA: A foundational language and vision alignment model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15617–15629.

[87] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proc. Int. Conf. Web Search Web Data Mining*, 2008, pp. 231–240.

[88] Y.-C. Chen et al., "Uniter: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.

[89] T. Shin, Y. Razeghi, R. L. Logan, E. Wallace, and S. Singh, "AutoPrompt: Eliciting knowledge from language models with automatically generated prompts," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 4222–4235.

[90] M. Xia, M. Artetxe, J. Du, D. Chen, and V. Stoyanov, "Prompting electra: Few-shot learning with discriminative pre-trained models," 2022, *arXiv:2205.15223*.

[91] D. Yin, T. Meng, and K.-W. Chang, "SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics," 2020, *arXiv:2005.04114*.

[92] P. Ke, H. Ji, S. Liu, X. Zhu, and M. Huang, "SentiLARE: Sentiment-aware language representation learning with linguistic knowledge," 2019, *arXiv:1911.02493*.

[93] H. Tian et al., "SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis," 2020, *arXiv:2005.05635*.

[94] S. Fan et al., "Sentiment-aware word and sentence level pre-training for sentiment analysis," 2022, *arXiv:2210.09803*.

[95] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in *Proc. 1st Workshop Graph Based Methods Natural Lang. Process.*, 2006, pp. 45–52.

[96] D. Rao and D. Ravichandran, "Semi-supervised polarity lexicon induction," in *Proc. 12th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2009, pp. 675–682.

[97] V. Sindhwani and P. Melville, "Document-word co-regularization for semi-supervised sentiment analysis," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 1025–1030.

[98] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[99] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2697–2709, 2020.

[100] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 992–1004, Apr. 2022.

[101] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 529–535.

[102] C. Du et al., "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 108–116.

[103] Z. Lian, B. Liu, and J. Tao, "SMIN: Semi-supervised multi-modal interaction network for conversational emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Jan. 7, 2022, doi: 10.1109/TAFFC.2022.3141237.

[104] G. Zhang, V. Davoodnia, and A. Etemad, "PARSE: Pairwise alignment of representations in semi-supervised EEG learning for emotion recognition," 2022, *arXiv:2202.05400*.

[105] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2852–2861.

[106] W. Rong, B. Peng, Y. Ouyang, C. Li, and Z. Xiong, "Semi-supervised dual recurrent neural network for sentiment analysis," in *Proc. IEEE 11th Int. Conf. Dependable, Autonomic Secure Comput.*, Dec. 2013, pp. 438–445.

[107] V. Sintsova, C. Musat, and P. Pu, "Semi-supervised method for multi-category emotion recognition in tweets," in *Proc. IEEE Int. Conf. Data Mining Workshop*, Dec. 2014, pp. 393–402.

[108] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schüller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5185–5189.

[109] Y. Li, "Semi-supervised learning for multimodal speech and emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2021, pp. 817–821.

[110] B. Xiang and L. Zhou, "Improving Twitter sentiment analysis with topic-based mixture modeling and semi-supervised training," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 434–439.

[111] Z. Zhang and M. P. Singh, "ReNew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 542–551.

[112] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, "Semi-supervised learning for imbalanced sentiment classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1826–1831.

[113] H. Hwang and Y. Lee, "Semi-supervised learning based on auto-generated lexicon using XAI in sentiment analysis," in *Proc. Conf. Recent Adv. Natural Lang. Process.-Deep Learn. Natural Lang. Process. Methods Appl.*, 2021, pp. 593–600.

[114] S. Zhou et al., "Inferring emotion from large-scale internet voice data: A semi-supervised curriculum augmentation based deep learning approach," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 6039–6047.

[115] K. Sohn et al., "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[116] X. Li, F. Shen, Y. Peng, W. Kong, and B.-L. Lu, "Efficient sample and feature importance mining in semi-supervised EEG emotion recognition," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 7, pp. 3349–3353, Jul. 2022.

[117] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[118] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 529–536.

[119] B. Zhang et al., "FlexMatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18408–18419.

[120] J. Li, C. Xiong, and S. C. H. Hoi, "CoMatch: Semi-supervised learning with contrastive graph regularization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9455–9464.

[121] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[122] Y. Xu et al., "Dash: Semi-supervised learning with dynamic thresholding," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11525–11536.

[123] A. Qadir and E. Riloff, "Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1203–1209.

[124] J. Read and J. Carroll, "Weakly supervised techniques for domain-independent sentiment classification," in *Proc. 1st Int. CIKM Workshop Topic-Sentiment Anal. Mass Opinion*, Nov. 2009, pp. 45–52.

[125] O. Pereg, D. Korat, M. Wasserblat, J. Mamou, and I. Dagan, "ABSApp: A portable weakly-supervised aspect-based sentiment extraction system," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP), Syst. Demonstrations*, 2019, pp. 1–6.

[126] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 223–232.

[127] Y. He, "Incorporating sentiment prior knowledge for weakly supervised sentiment analysis," *ACM Trans. Asian Lang. Inf. Process.*, vol. 11, no. 2, pp. 1–19, Jun. 2012.

[128] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[129] A. Ramesh, S. H. Kumar, J. Foulds, and L. Getoor, "Weakly supervised models of aspect-sentiment for online course discussion forums," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 74–83.

[130] Z. Zeng, W. Zhou, X. Liu, and Y. Song, "A variational approach to weakly supervised document-level multi-aspect sentiment classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2019, pp. 386–396.

[131] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7584–7592.

[132] G. Lee, J. Jeong, S. Seo, C. Kim, and P. Kang, "Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network," *Knowl.-Based Syst.*, vol. 152, pp. 70–82, Jul. 2018.

[133] Z. Guan, L. Chen, W. Zhao, Y. Zheng, S. Tan, and D. Cai, "Weakly-supervised deep learning for customer review sentiment classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 3719–3725.

[134] J. Yang, D. She, Y.-K. Lai, and M.-H. Yang, "Retrieving and classifying affective images via deep metric learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 491–498.

[135] L.-Y. Xue, Q.-R. Mao, X.-H. Huang, and J. Chen, "NLWSNet: A weakly supervised network for visual sentiment analysis in mislabeled web images," *Frontiers Inf. Technol. Electron. Eng.*, vol. 21, no. 9, pp. 1321–1333, Sep. 2020.

[136] D. She, M. Sun, and J. Yang, "Learning discriminative sentiment representation from strongly- and weakly supervised CNNs," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3s, pp. 1–19, Nov. 2019.

[137] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 381–388.

[138] T. Zhang, A. E. Ali, C. Wang, A. Hanjalic, and P. Cesar, "Weakly-supervised learning for fine-grained emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, early access, Mar. 10, 2022, doi: 10.1109/TAFFC.2022.3158234.

[139] S. Angelidis and M. Lapata, "Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3675–3686.

[140] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 579–595.

[141] J. Deriu et al., "Leveraging large amounts of weakly supervised data for multi-language sentiment classification," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1045–1052.

[142] H. Xu, B. Liu, L. Shu, and P. Yu, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics,*

Hum. Lang. Technol., 2019, pp. 2324–2335.

[143] L. Vadicamo et al., "Cross-media learning for image sentiment analysis in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 308–317.

[144] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2006, pp. 417–422.

[145] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010.

[146] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 932–938.

[147] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[148] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[149] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? Predicting the emotion stimuli map," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 614–618.

[150] W. Zhao et al., "Weakly-supervised deep embedding for product review sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 185–197, Jan. 2018.

[151] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019.

[152] L. Wang, X. Xu, F. Liu, X. Xing, B. Cai, and W. Lu, "Robust emotion navigation: Few-shot visual sentiment analysis by auxiliary noisy data," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2019, pp. 121–127.

[153] W. Wang et al., "Comp-GAN: Compositional generative adversarial network in synthesizing and recognizing facial expression," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 211–219.

[154] B. Liang, Z. Chen, L. Gui, Y. He, M. Yang, and R. Xu, "Zero-shot stance detection via contrastive learning," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2738–2747.

[155] A.-N. Ciubotaru, A. Devos, B. Bozorgtabar, J.-P. Thiran, and M. Gabrani, "Revisiting few-shot learning for facial expression recognition," 2019, *arXiv:1912.02751*.

[156] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: Transfer learning for emotion detection using cross lingual embedding," *Expert Syst. Appl.*, vol. 139, Jan. 2020, Art. no. 112851.

[157] E. Hosseini-Asl, W. Liu, and C. Xiong, "A generative language model for few-shot aspect-based sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics, NAACL*, 2022, pp. 770–787.

[158] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Face behavior a la carte: Expressions, affect and action units in a single network," 2019, *arXiv:1910.11111*.

[159] D. Kollias, V. Sharmanska, and S. Zafeiriou, "Distribution matching for heterogeneous multi-task learning: A large-scale face study," 2021, *arXiv:2105.03790*.

[160] S. Lamprinidis, F. Bianchi, D. Hardt, and D. Hovy, "Universal joy: A data set and results for classifying emotions across languages," in *Proc. Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2021, pp. 62–75.

[161] W. Zhong, D. Tang, J. Wang, J. Yin, and N. Duan, "UserAdapter: Few-shot user learning in sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics, ACL-IJCNLP*, 2021, pp. 1484–1488.

[162] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh, "Calibrate before use: Improving few-shot performance of language models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12697–12706.

[163] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, "Few-shot cross-lingual stance detection with sentiment-based pre-training," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 10729–10737.

[164] R. Ning, C. L. Philip Chen, and T. Zhang, "Cross-subject EEG emotion recognition using domain adaptive few-shot learning networks," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1468–1472.

[165] L. Yang, Y. Li, J. Wang, and N. N. Xiong, "FSLM: An intelligent few-shot learning model based on Siamese networks for IoT technology," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9717–9729, Jun. 2021.

[166] K. Feng and T. Chaspari, "Few-shot learning in emotion recognition of spontaneous speech using a Siamese neural network with adaptive sample pair formation," *IEEE Trans. Affect. Comput.*, early access, Sep. 3, 2021, doi: 10.1109/TAFFC.2021.3109485.

[167] Y. Ahn, S. J. Lee, and J. W. Shin, "Cross-corpus speech emotion recognition based on few-shot learning and domain adaptation," *IEEE Signal Process. Lett.*, vol. 28, pp. 1190–1194, 2021.

[168] X. Zou, Y. Yan, J.-H. Xue, S. Chen, and H. Wang, "When facial expression recognition meets few-shot learning: A joint and alternate learning framework," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 5367–5375.

[169] T. Zhang, A. E. Ali, A. Hanjalic, and P. Cesar, "Few-shot learning for fine-grained emotion recognition using physiological signals," *IEEE Trans. Multimedia*, early access, Apr. 7, 2022, doi: 10.1109/TMM.2022.3165715.

[170] N. Zhang, M. Ruan, S. Wang, L. Paul, and X. Li, "Discriminative few shot learning of facial dynamics in interview videos for autism trait classification," *IEEE Trans. Affect. Comput.*, early access, May 30, 2022, doi: 10.1109/TAFFC.2022.3178946.

[171] Y. Li and S. Shan, "Meta auxiliary learning for facial action unit detection," *IEEE Trans. Affect. Comput.*, early access, Dec. 14, 2021, doi: 10.1109/TAFFC.2021.3135516.

[172] S. Buechel, L. Modersohn, and U. Hahn, "Towards label-agnostic emotion embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9231–9249.

[173] X. Xu et al., "Rethinking auditory affective descriptors through zero-shot emotion recognition in speech," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 5, pp. 1530–1541, Oct. 2022.

[174] X. Xu, J. Deng, N. Cummins, Z. Zhang, L. Zhao, and B. W. Schuller, "Exploring zero-shot emotion recognition in speech using semantic-embedding prototypes," *IEEE Trans. Multimedia*, vol. 24, pp. 2752–2765, 2022.

[175] A. Banerjee, U. Bhattacharya, and A. Bera, "Learning unseen emotions from gestures via semantically-conditioned zero-shot perception with adversarial autoencoders," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3–10.

[176] Z. Zhao and X. Ma, "Text emotion distribution learning from small sample: A meta-learning approach," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3957–3967.

[177] J. Healey, "Sequential dependence and non-linearity in affective responses: A skin conductance example," in *Proc. Workshop Artif. Intell. Affective Comput.*, 2020, pp. 1–8.

[178] C. L. Stewart, A. Folarin, and R. Dobson, "Personalized acute stress classification from physiological signals with neural processes," 2020, *arXiv:2002.04176*.

[179] G. Guibon, M. Labeau, H. Flamein, L. Lefeuvre, and C. Clavel, "Meta-learning for classifying previously unseen data source into previously unseen emotional categories," in *Proc. 1st Workshop Meta Learn. Appl. Natural Lang. Process.*, 2021, p. 76.

[180] B. Liang et al., "Few-shot aspect category sentiment analysis via meta-learning," *ACM Trans. Inf. Syst.*, vol. 41, no. 1, pp. 22:1–22:31, 2023.

[181] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, Mar. 1999.

[182] P. Ekman. (2004). *Micro Expressions Training Tools*. [Online]. Available: https://www.paulekman.com/micro-expressions-training-tools/

[183] S. Hamann and T. Canli, "Individual differences in emotion processing," *Current Opinion Neurobiol.*, vol. 14, no. 2, pp. 233–238, 2004.

[184] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 32, pp. 15849–15854, Aug. 2019.

[185] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[186] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[187] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[188] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a Siamese time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1993, pp. 737–744.

[189] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, 2015.

[190] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[191] T. Ramalho and M. Garnelo, "Adaptive posterior learning: Few-shot learning with a surprise-based memory module," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[192] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75, Dec. 1997.

[193] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. Interspeech*, Sep. 2018, pp. 152–156.

[194] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[195] S. Zhao et al., "A two-stage 3D CNN based learning method for spontaneous micro-expression recognition," *Neurocomputing*, vol. 448, pp. 276–289, Aug. 2021.

[196] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 2258–2263.

[197] S. Zhao et al., "ME-PLAN: A deep prototypical learning with local attention network for dynamic micro-expression recognition," *Neural Netw.*, vol. 153, pp. 427–443, Sep. 2022.

[198] B. Yang, C. Liu, B. Li, J. Jiao, and Q. Ye, "Prototype mixture models for few-shot semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 763–778.

[199] D. Shome and T. Kar, "FedAffect: Few-shot federated learning for facial expression recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 4151–4158.

[200] A. Gepperth and B. Hammer, "Incremental learning algorithms and applications," in *Proc. Eur. Symp. Artif. Neural Netw.*, 2016.

[201] K. James et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[202] D. Maltoni and V. Lomonaco, "Continuous learning in single-incremental-task scenarios," *Neural Netw.*, vol. 116, pp. 56–73, Aug. 2019.

[203] S. Thrun and T. M. Mitchell, "Lifelong robot

learning," *Robot. Auton. Syst.*, vol. 15, nos. 1–2, pp. 25–46, Jul. 1995.

[204] P. Ruvolo and E. Eaton, "ELLA: An efficient lifelong learning algorithm," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 507–515.

[205] P. Ruvolo and E. Eaton, "Scalable lifelong learning with active task selection," in *Proc. AAAI Spring Symp., Lifelong Mach. Learn.*, 2013.

[206] P. Ruvolo and E. Eaton, "Active task selection for lifelong machine learning," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 862–868.

[207] Z. Chen, N. Ma, and B. Liu, "Lifelong learning for sentiment classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 750–756.

[208] S. Mishra, J. Diesner, J. Byrne, and E. Surbeck, "Sentiment analysis with incremental human-in-the-loop learning and lexical resource customization," in *Proc. 26th ACM Conf. Hypertext Social Media*, 2015, pp. 323–325.

[209] M. De Lange et al., "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.

[210] S. C. Y. Hung, J.-H. Lee, T. S. T. Wan, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, "Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2019, pp. 339–343.

[211] N. Churamani, O. Kara, and H. Gunes, "Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition," *IEEE Trans. Affect. Comput.*, early access, Jun. 9, 2022, doi: 10.1109/TAFFC.2022.3181033.

[212] J. Zhu, B. Luo, S. Zhao, S. Ying, X. Zhao, and Y. Gao, "IExpressNet: Facial expression recognition with incremental classes," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2899–2908.

[213] N. Churamani and H. Gunes, "CLIFER: Continual learning with imagination for facial expression recognition," in *Proc. 15th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Nov. 2020, pp. 322–328.

[214] Z. Ke, B. Liu, H. Wang, and L. Shu, "Continual learning with knowledge transfer for sentiment classification," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2020, pp. 683–698.

[215] Z. Ke, B. Liu, N. Ma, H. Xu, and L. Shu, "Achieving forgetting prevention and knowledge transfer in continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 22443–22456.

[216] Z. Ke, H. Xu, and B. Liu, "Adapting BERT for continual learning of a sequence of aspect sentiment classification tasks," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 4746–4755.

[217] Q. Qin, W. Hu, and B. Liu, "Using the past knowledge to improve sentiment classification," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 1124–1133.

[218] Z. Ke, B. Liu, and X. Huang, "Continual learning of a mixed sequence of similar and dissimilar tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18493–18504.

[219] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 3856–3866, 2017.

[220] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 94–101.

[221] S. Li and W. Deng, "A deeper look at facial expression dataset bias," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 881–893, Apr. 2022.

[222] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 308–314.

[223] M. Chen, Z. Xu, K. Q. Weinberger, and F. Sha, "Marginalized denoising autoencoders for domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1627–1634.

[224] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.

[225] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, 2006, pp. 1–8.

[226] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[227] S. Thuseethan, S. Rajasegarar, and J. Yearwood, "Deep continual learning for emerging emotion recognition," *IEEE Trans. Multimedia*, vol. 24, pp. 4367–4380, 2022.

[228] M. Welling, "Herding dynamical weights to learn," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Jun. 2009, pp. 1121–1128.

[229] R. J. Crisp and R. N. Turner, "Can imagined interactions produce positive perceptions? Reducing prejudice through simulated social contact," *Amer. Psychologist*, vol. 64, no. 4, pp. 231–240, 2009.

[230] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization," *Frontiers Neurorobotics*, vol. 12, p. 78, Nov. 2018.

[231] M. Mainsant, M. Solinas, M. Reyboz, C. Godin, and M. Mermillod, "Dream Net: A privacy preserving continual leaming model for face emotion recognition," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2021, pp. 01–08.

[232] M. Solinas et al., "Beneficial effect of combined replay for continual learning," in *Proc. 13th Int. Conf. Agents Artif. Intell.*, 2021, pp. 205–217.

[233] J. Han, Z. Zhang, M. Pantic, and B. Schuller, "Internet of Emotional People: Towards continual affective computing cross cultures via audiovisual signals," *Future Gener. Comput. Syst.*, vol. 114, pp. 294–306, Jan. 2021.

[234] A. Mallya and S. Lazebnik, "PackNet: Adding multiple tasks to a single network by iterative pruning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7765–7773.

[235] G. Lv, S. Wang, B. Liu, E. Chen, and K. Zhang, "Sentiment classification by leveraging the shared knowledge from a sequence of domains," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2019, pp. 795–811.

[236] Z. Ke, B. Liu, H. Xu, and L. Shu, "CLASSIC: Continual and contrastive learning of aspect sentiment classification tasks," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6871–6883.

[237] C.-Y. Hung, C.-H. Tu, C.-E. Wu, C.-H. Chen, Y.-M. Chan, and C.-S. Chen, "Compacting, picking and growing for unforgetting continual learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13647–13657.

[238] Z. Dai, C. Peng, H. Chen, and Y. Ding, "A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6955–6965.

[239] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for EEG-based emotion recognition: A comparative study on two public datasets," *IEEE Trans. Cognit. Develop. Syst.*, vol. 11, no. 1, pp. 85–94, Mar. 2019.

[240] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1275–1282.

[241] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 2732–2738.

[242] X. Jiang, Y. Zong, W. Zheng, J. Liu, and M. Wei, "Seeking salient facial regions for cross-database micro-expression recognition," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 1019–1025.

[243] T. Zhang et al., "Cross-database micro-expression recognition: A benchmark," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 2, pp. 544–559, Feb. 2022.

[244] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, "Cross-domain sentiment classification via spectral feature alignment," in *Proc. 19th Int. Conf. World Wide Web*, Apr. 2010, pp. 751–760.

[245] D. Bollegala, T. Mu, and J. Y. Goulermas, "Cross-domain sentiment classification using sentiment sensitive embeddings," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 2, pp. 398–410, Feb. 2016.

[246] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2058–2065.

[247] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.

[248] K. Yan, L. Kou, and D. Zhang, "Learning domain-invariant subspace using domain features and independence maximization," *IEEE Trans. Cybern.*, vol. 48, no. 1, pp. 288–299, Jan. 2018.

[249] W. Zheng, Y. Zong, X. Zhou, and M. Xin, "Cross-domain color facial expression recognition using transductive transfer subspace learning," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 21–37, Jan. 2018.

[250] P. Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, Apr. 2019.

[251] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 373–382, Jul. 2020.

[252] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 307–318, 2020.

[253] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning a target sample re-generator for cross-database micro-expression recognition," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 872–880.

[254] W. Xia, W. Zheng, Y. Zong, and X. Jiang, "Motion attention deep transfer network for cross-database micro-expression recognition," in *Proc. Int. Conf. Pattern Recognit. Workshops Challenges*, 2021, pp. 679–693.

[255] B. Song, Y. Zong, K. Li, J. Zhu, J. Shi, and L. Zhao, "Cross-database micro-expression recognition based on a dual-stream convolutional neural network," *IEEE Access*, vol. 10, pp. 66227–66237, 2022.

[256] Y. Peng, W. Wang, W. Kong, F. Nie, B.-L. Lu, and A. Cichocki, "Joint feature adaptation and graph adaptive label propagation for cross-subject emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 1941–1958, Oct. 2022.

[257] Y. He and G. Ding, "Deep transfer learning for image emotion analysis: Reducing marginal and joint distribution discrepancies together," *Neural Process. Lett.*, vol. 51, no. 3, pp. 2077–2086, Jun. 2020.

[258] H. Li, Y.-M. Jin, W.-L. Zheng, and B.-L. Lu, "Cross-subject emotion recognition using deep adaptation networks," in *Proc. Int. Conf. Neural Inf. Process.*, 2018, pp. 403–413.

[259] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "Adaptive semi-supervised learning for cross-domain sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3467–3476.

[260] X. Qu, Z. Zou, Y. Cheng, Y. Yang, and P. Zhou, "Adversarial category alignment network for cross-domain sentiment classification," in *Proc.*

Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., 2019, pp. 2496–2508.

[261] M. Peng, Q. Zhang, Y.-G. Jiang, and X. Huang, "Cross-domain sentiment classification with target domain specific information," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2505–2513.

[262] M. Jiménez-Guarneros and P. Gómez-Gil, "Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition," *Pattern Recognit. Lett.*, vol. 141, pp. 54–60, Jan. 2021.

[263] H. Guo, R. Pasunuru, and M. Bansal, "Multi-source domain adaptation for text classification via DistanceNet-bandits," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 7830–7838.

[264] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.

[265] K.-M. Ding, T. Kimura, K.-I. Fukui, and M. Numao, "EEG emotion enhancement using task-specific domain adversarial neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[266] Z. Li, Y. Zhang, Y. Wei, Y. Wu, and Q. Yang, "End-to-end adversarial memory network for cross-domain sentiment classification," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2237–2243.

[267] L. Zheng, W. Ying, Z. Yu, and Y. Qiang, "Hierarchical attention transfer network for cross-domain sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5852–5859.

[268] K. Zhang, H. Zhang, Q. Liu, H. Zhao, H. Zhu, and E. Chen, "Interactive attention transfer network for cross-domain sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 5773–5780.

[269] C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware BERT for cross-domain sentiment analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4019–4028.

[270] Y. Li et al., "A novel bi-hemispheric discrepancy model for EEG emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 2, pp. 354–367, Jun. 2021.

[271] Z. Li, X. Li, Y. Wei, L. Bing, Y. Zhang, and Q. Yang, "Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4589–4599.

[272] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.

[273] S. Rayatdoost, Y. Yin, D. Rudrauf, and M. Soleymani, "Subject-invariant EEG representation learning for emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 3955–3959.

[274] Y. Li, W. Zheng, Z. Cui, T. Zhang, and Y. Zong, "A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1561–1567.

[275] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 494–504, Apr. 2021.

[276] Y. Li, B. Fu, F. Li, G. Shi, and W. Zheng, "A novel transferability attention neural network model for EEG emotion recognition," *Neurocomputing*, vol. 447, pp. 92–101, Aug. 2021.

[277] K. Zhang et al., "Graph adaptive semantic transfer for cross-domain sentiment classification," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 1566–1576.

[278] S. Zhao et al., "Emotional semantics-preserved and feature-aligned CycleGAN for visual emotion adaptation," *IEEE Trans. Cybern.*, vol. 52, no. 10, pp. 10000–10013, Oct. 2022.

[279] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 732–737.

[280] Y. Luo, S.-Y. Zhang, W.-L. Zheng, and B.-L. Lu, "Wgan domain adaptation for EEG-based emotion recognition," in *Proc. Int. Conf. Neural Inf. Process.*, 2018, pp. 275–286.

[281] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9887–9903, Dec. 2022.

[282] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Apr. 22, 2022, doi: 10.1109/TAFFC.2022.3167013.

[283] Y. Li, Y. Gao, B. Chen, Z. Zhang, L. Zhu, and G. Lu, "JDMAN: Joint discriminative and mutual adaptation networks for cross-domain facial expression recognition," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3312–3320.

[284] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 513–520.

[285] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.

[286] J. Yu and J. Jiang, "Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 236–246.

[287] Y. Yin, L. Lu, Y. Wu, and M. Soleymani, "Self-supervised patch localization for cross-domain facial action unit detection," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.

[288] T. Li, X. Chen, S. Zhang, Z. Dong, and K. Keutzer, "Cross-domain sentiment classification with contrastive learning and mutual information maximization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 8203–8207.

[289] S. Zhao, X. Zhao, G. Ding, and K. Keutzer, "EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1319–1327.

[290] S. Zhao et al., "CycleEmotionGAN: Emotional semantic consistency preserved CycleGAN for adapting image emotions," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 2620–2627.

[291] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3092–3099.

[292] Y. Ji, Y. Hu, Y. Yang, and H. T. Shen, "Region attention enhanced unsupervised cross-domain facial emotion recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 4190–4201, Apr. 2023.

[293] H. Li, Y. Kim, C.-H. Kuo, and S. S. Narayanan, "Acted vs. improvised: Domain adaptation for elicitation approaches in audio-visual emotion recognition," in *Proc. Interspeech*, Aug. 2021, pp. 3395–3399.

[294] C. Zhao, S. Wang, and D. Li, "Multi-source domain adaptation with joint learning for cross-domain sentiment classification," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105254.

[295] Y. Wang, S. Qiu, D. Li, C. Du, B.-L. Lu, and H. He, "Multi-modal domain adaptation variational autoencoder for EEG-based emotion recognition," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 9, pp. 1612–1626, Sep. 2022.

[296] K. Xia, X. Gu, and B. Chen, "Cross-dataset transfer driver expression recognition via global discriminative and local structure knowledge exploitation in shared projection subspace," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1765–1776, Mar. 2021.

[297] T. Ni, C. Zhang, and X. Gu, "Transfer model collaborating metric learning and dictionary learning for cross-domain facial expression recognition," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 5, pp. 1213–1222, Oct. 2021.

[298] T. Duan et al., "Meta learn on constrained transfer learning for low resource cross subject EEG classification," *IEEE Access*, vol. 8, pp. 224791–224802, 2020.

[299] F. Wu and Y. Huang, "Sentiment domain adaptation with multiple sources," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 301–310.

[300] H. Zhao, S. Zhang, G. Wu, J. M. Moura, J. P. Costeira, and G. J. Gordon, "Adversarial multiple source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8568–8579.

[301] H. Wu and X. Shi, "Adversarial soft prompt tuning for cross-domain sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2438–2447.

[302] J. Cao, X. He, C. Yang, S. Chen, Z. Li, and Z. Wang, "Multi-source and multi-representation adaptation for cross-domain electroencephalography emotion recognition," *Frontiers Psychol.*, vol. 12, Jan. 2022, Art. no. 809459.

[303] X. Gu, W. Cai, M. Gao, Y. Jiang, X. Ning, and P. Qian, "Multi-source domain transfer discriminative dictionary learning modeling for electroencephalogram-based emotion recognition," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 6, pp. 1604–1612, Dec. 2022.

[304] D. Bollegala, D. Weir, and J. Carroll, "Cross-domain sentiment classification using a sentiment sensitive thesaurus," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1719–1731, Aug. 2013.

[305] J. Guo, D. Shah, and R. Barzilay, "Multi-source domain adaptation with mixture of experts," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4694–4703.

[306] C. Lu, Y. Zong, W. Zheng, Y. Li, C. Tang, and B. W. Schuller, "Domain invariant feature learning for speaker-independent speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 2217–2230, 2022.

[307] S. Zhao et al., "Curriculum CycleGAN for textual sentiment domain adaptation with multiple sources," in *Proc. Web Conf.*, Apr. 2021, pp. 541–552.

[308] C. Lin, S. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 2661–2668.

[309] Y. Wang, J. Liu, Q. Ruan, S. Wang, and C. Wang, "Cross-subject EEG emotion classification based on few-label adversarial domain adaption," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115581.

[310] L.-M. Zhao, X. Yan, and B.-L. Lu, "Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 863–870.

[311] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye, "A two-stage weighting framework for multi-source domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 505–513.

[312] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 357–366.

[313] Y. Yin, B. Huang, Y. Wu, and M. Soleymani, "Speaker-invariant adversarial domain adaptation for emotion recognition," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 481–490.

[314] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 292–301.

[315] C. Athanasiadis, E. Hortal, and S. Asteriadis, "Audio–visual domain adaptation using conditional semi-supervised generative

adversarial networks," *Neurocomputing*, vol. 397, pp. 331–344, Jul. 2020.

[316] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, Jun. 2011, pp. 1521–1528.

[317] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[318] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[319] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2066–2073.

[320] Y. Zong, W. Zheng, Z. Cui, G. Zhao, and B. Hu, "Toward bridging microexpressions from different domains," *IEEE Trans. Cybern.*, vol. 50, no. 12, pp. 5047–5060, Dec. 2020.

[321] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, Jan. 2012.

[322] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[323] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Apr. 2016.

[324] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[325] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[326] F. Qi, X. Yang, and C. Xu, "A unified framework for multimodal domain adaptation," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 429–437.

[327] Y. Zhang, Y. Zhang, W. Guo, X. Cai, and X. Yuan, "Learning disentangled representation for multimodal cross-domain sentiment analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 21, 2022, doi: 10.1109/TNNLS.2022.3147546.

[328] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019.

[329] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," in *Proc. Int. Conf. Neural Inf. Process.*, 2013, pp. 117–124.

[330] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 317–321.

[331] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image Vis. Comput.*, vol. 29, no. 9, pp. 607–619, Aug. 2011.

[332] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," in *Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit.*, Mar. 1998, pp. 200–205.

[333] Y. Ziser and R. Reichart, "Pivot based language modeling for improved neural domain adaptation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 1241–1251.

[334] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, and L. He, "SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 568–579.

[335] Y. Li, H. Chen, J. Zhao, H. Zhang, and J. Li, "Benchmarking domain generalization on EEG-based emotion recognition," 2022, *arXiv:2204.09016*.

[336] Z. Wang, Q. Wang, C. Lv, X. Cao, and G. Fu, "Unseen target stance detection with adversarial domain generalization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8.

[337] J.-M. Zhang et al., "A cross-subject and cross-modal model for multimodal emotion recognition," in *Proc. Int. Conf. Neural Inf. Process.*, 2021, pp. 203–211.

[338] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition," *IEEE Trans. Affect. Comput.*, early access, Apr. 4, 2022, doi: 10.1109/TAFFC.2022.3164516.

[339] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, "Reducing the subject variability of EEG signals with adversarial domain generalization," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 30–42.

[340] S.-W. Lee, "Domain generalization with triplet network for cross-corpus speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 389–396.

[341] A. Roy and E. Cambria, "Soft labeling constraint for generalizing from sentiments in single domain," *Knowl.-Based Syst.*, vol. 245, Jun. 2022, Art. no. 108346.

[342] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, "MetaReg: Towards domain generalization using meta-regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1006–1016.

[343] X. Ma, T. Zhang, and C. Xu, "Deep multi-modality adversarial networks for unsupervised domain adaptation," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2419–2431, Sep. 2019.

[344] R. Volpi, H. Namkoong, O. Sener, J. C. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 5339–5349.

[345] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2018.

[346] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2019.

[347] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 124–140.

[348] J. Abdi, A. Al-Hindawi, T. Ng, and M. P. Vizcaychipi, "Scoping review on the use of socially assistive robot technology in elderly care," *BMJ Open*, vol. 8, no. 2, Feb. 2018, Art. no. e018815.

[349] H. Chen, Y. Zhao, T. Zhao, J. Chen, S. Li, and S. Zhao, "Emotion recognition of the elderly living alone based on deep learning," in *Proc. IEEE Int. Conf. Consum. Electron.*, Sep. 2021, pp. 1–2.

[350] B. W. Schuller et al., "The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proc. Interspeech*, Oct. 2020, pp. 2042–2046.

[351] S. Liu et al., "Towards emotional support dialog systems," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 3469–3483.

[352] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard, "Driver emotion recognition for intelligent vehicles: A survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–30, May 2021.

[353] K. Ihme, C. Dömeland, M. Freese, and M. Jipp, "Frustration in the face of the driver: A simulator study on facial muscle activity during frustrated driving," *Interact. Stud.*, vol. 19, no. 3, pp. 487–498, Dec. 2018.

[354] Z. Ma, M. Mahmoud, P. Robinson, E. Dias, and L. Skrypchuk, "Automatic detection of a driver's complex mental states," in *Proc. Int. Conf. Comput. Sci. Appl.*, 2017, pp. 678–691.

[355] H. Rahman, S. Barua, and B. Shahina, "Intelligent driver monitoring based on physiological sensor signals: Application using camera," in *Proc. IEEE 18th Int. Conf. Intell. Transp. Syst.*, Sep. 2015, pp. 2637–2642.

[356] N. Munla, M. Khalil, A. Shahin, and A. Mourad, "Driver stress level detection using HRV analysis," in *Proc. Int. Conf. Adv. Biomed. Eng. (ICABME)*, Sep. 2015, pp. 61–64.

[357] H. Gao, A. Yüce, and J.-P. Thiran, "Detecting emotional stress from facial expressions for driving safety," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 5961–5965.

[358] M. Oehl, F. W. Siebert, T.-K. Tews, R. Höger, and H.-R. Pfister, "Improving human–machine interaction—A non invasive approach to detect emotions in car drivers," in *Proc. HCI Int. Conf.*, 2011, pp. 577–585.

[359] B. W. Schuller, "Speaker, noise, and acoustic space adaptation for emotion recognition in the automotive environment," in *Proc. ITG Conf. Voice Commun.*, Oct. 2008, pp. 1–4.

[360] A. Tawari and M. M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4605–4608.

[361] P. E. Paredes, F. Ordonez, W. Ju, and J. A. Landay, "Fast & furious: Detecting stress with a car steering wheel," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–12.

[362] O. Karaduman, H. Eren, H. Kurum, and M. Celenk, "An effective variable selection algorithm for aggressive/calm driving detection via CAN bus," in *Proc. Int. Conf. Connected Vehicles Expo (ICCVE)*, Dec. 2013, pp. 586–591.

[363] L. Malta, C. Miyajima, N. Kitaoka, and K. Takeda, "Analysis of real-world driver's frustration," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 109–118, Mar. 2011.

[364] G. Rigas, Y. Goletsis, and D. I. Fotiadis, "Real-time driver's stress event detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 1, pp. 221–234, Mar. 2012.

[365] K. Zaman, Z. Sun, S. M. Shah, M. Shoaib, L. Pei, and A. Hussain, "Driver emotions recognition based on improved faster R-CNN and neural architectural search network," *Symmetry*, vol. 14, no. 4, p. 687, Mar. 2022.

[366] W. Li et al., "CogEmoNet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 3, pp. 667–678, Jun. 2022.

[367] T. Gui et al., "Cooperative multimodal approach to depression detection in Twitter," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 110–117.

[368] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, Jan. 2022.

[369] F. Delahunty, R. Johansson, and M. Arcan, "Passive diagnosis incorporating the PHQ-4 for depression and anxiety," in *Proc. 4th Social Media Mining Health Appl.*, 2019, pp. 40–46.

[370] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan, "Improving the generalizability of depression detection by leveraging clinical questionnaires," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 8446–8459.

[371] B. Wang, Y. Zhao, X. Lu, and B. Qin, "Cognitive distortion based explainable depression detection and analysis technologies for the adolescent internet users on social media," *Frontiers Public Health*, vol. 10, pp. 1–12, Jan. 2023.

[372] M. Chung, E. Ko, H. Joung, and S. J. Kim, "Chatbot e-service and customer satisfaction regarding luxury brands," *J. Bus. Res.*, vol. 117, pp. 587–595, Sep. 2020.

[373] C. Pelau, D.-C. Dabija, and I. Ene, "What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry," *Comput. Hum. Behav.*, vol. 122, Sep. 2021, Art. no. 106855.

[374] V. Srinivasan and L. Takayama, "Help me please: Robot politeness strategies for soliciting help from humans," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2016, pp. 4945–4955.

[375] Z. Liu, X. Kong, S. Liu, Z. Yang, and C. Zhang,

"Looking at MOOC discussion data to uncover the relationship between discussion pacings, learners' cognitive presence and learning achievements," *Educ. Inf. Technol.*, vol. 27, no. 6, pp. 8265–8288, Jul. 2022.

[376] M. De Lange and T. Tuytelaars, "Continual prototype evolution: Learning online from non-stationary data streams," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8230–8239.

[377] L. Tian, X. Hong, C. Fan, Y. Ming, M. Pietikäinen, and G. Zhao, "Sparse Tikhonov-regularized hashing for multi-modal learning," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3793–3797.

[378] S. Amiriparian, T. Hübner, V. Karas, M. Gerczuk, S. Ottl, and B. W. Schuller, "DeepSpectrumLite: A power-efficient transfer learning framework for embedded speech and audio processing from decentralized data," *Frontiers Artif. Intell.*, vol. 5, Mar. 2022, Art. no. 856232.

[379] E. A. Veltmeijer, C. Gerritsen, and K. V. Hindriks, "Automatic emotion recognition for groups: A review," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 89–107, Jan. 2023.

[380] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9630–9640.

[381] S. Zhao et al., "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 303–311.

[382] L. Xu, Z. Wang, B. Wu, and S. Lui, "MDAN: Multi-level dependent attention network for visual emotion analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9469–9478.

[383] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 526–540, Oct. 2018.

[384] S. Wang, Y. Zhang, B. Lin, and B. Li, "Interpretable emotion analysis based on knowledge graph and OCC model," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2022, pp. 2038–2045.

[385] S. Zhao, S. Wang, M. Soleymani, D. Joshi, and Q. Ji, "Affective computing for large-scale heterogeneous multimedia data: A survey," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 15, no. 3s, pp. 93:1–93:32, 2019.

[386] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual*. Gainesville, FL, USA: NIMH, Center for the Study of Emotion & Attention, 2005.

[387] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[388] X. Peng, Z. Huang, Y. Zhu, and K. Saenko, "Federated adversarial domain adaptation," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[389] Y. Fang, P.-T. Yap, W. Lin, H. Zhu, and M. Liu, "Source-free unsupervised domain adaptation: A survey," 2023, *arXiv:2301.00265*.

[390] R. Wang et al., "DeepSonar: Towards effective and robust detection of AI-synthesized fake voices," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1207–1216.

[391] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194.

[392] S. A. Javadi, C. Norval, R. Cloete, and J. Singh, "Monitoring AI services for misuse," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2021, pp. 597–607.

ABOUT THE AUTHORS

Sicheng Zhao (Senior Member, IEEE) received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2016.

He was a Visiting Scholar with the National University of Singapore, Singapore, from 2013 to 2014; a Research Fellow with Tsinghua University, Beijing, China, from 2016 to 2017; a Postdoctoral Research Fellow with the University of California at Berkeley, Berkeley, CA, USA, from 2017 to 2020; and a Postdoctoral Research Scientist with Columbia University, New York, NY, USA, from 2020 to 2022. He is currently a Research Associate Professor with Tsinghua University. His research interests include affective computing, multimedia, and computer vision.

Xiaopeng Hong (Senior Member, IEEE) received the Ph.D. degree from the Harbin Institute of Technology (HIT), Harbin, China, in 2010.

He had been a Distinguished Research Fellow with Xi'an Jiaotong University, Xi'an, China, and an Adjunct Professor with the University of Oulu, Oulu, Finland. He is currently a Professor with HIT. He has authored over 50 papers in journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), and AAAI Conference on Artificial Intelligence (AAAI). His current research interests include visual surveillance, incremental learning, and affective computing.

Dr. Hong's studies about subtle facial movement analysis have been reported by international media such as MIT Technology Review and have been awarded the 2020 IEEE Finland Section Best Student Conference Paper.

Jufeng Yang (Member, IEEE) received the Ph.D. degree from Nankai University, Tianjin, China, in 2009.

He was a Visiting Scholar with the Vision and Learning Laboratory, University of California at Merced, Merced, CA, USA, from 2015 to 2016. He is currently a Full Professor with the College of Computer Science, Nankai University. His research interests include computer vision, machine learning, multimedia, affective computing, image retrieval, fine-grained classification, and medical image recognition.

Yanyan Zhao received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2011.

She was a Visiting Scholar with the University of California at Berkeley, Berkeley, CA, USA, from 2012 to 2013. She is currently an Associate Professor with the Faculty of Computing, Harbin Institute of Technology. She has authored over 20 papers in top-tier publications and conferences. Her research interests include natural language processing and sentiment analysis.

Guiguang Ding (Senior Member, IEEE) received the Ph.D. degree from Xidian University, Xi'an, China, in 2004.

In 2006, he was a Postdoctoral Research Fellow with the Department of Automation, Tsinghua University, Beijing, China. He is currently a Professor with the School of Software, Tsinghua University. He has published over 100 scientific papers in major journals and conferences. His current research interests include multimedia information retrieval, computer vision, and machine learning.

Dr. Ding served as a Leading Guest Editor for *Neural Processing Letters* (NPL) and *Multimedia Tools and Applications* (MTAP), the Special Session Chair for IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2021, IEEE International Conference on Multimedia and Expo (ICME) 2019 and 2020, and Pacific Rim Conference on Multimedia (PCM) 2017, and a reviewer for over 20 prestigious international journals and conferences.