

# Generative Technology for Human Emotion Recognition: A Scope Review

Fei Ma<sup>a,1</sup>, Yucheng Yuan<sup>b,1</sup>, Yifan Xie<sup>a</sup>, Hongwei Ren<sup>c</sup>, Ivan Liu<sup>d</sup>, Ying He<sup>e</sup>, Fuji Ren<sup>f</sup>, Fei Richard Yu<sup>a,\*</sup>, Shiguang Ni<sup>b,\*</sup>

<sup>a</sup>*Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China*

<sup>b</sup>*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China*

<sup>c</sup>*MICS Thrust, The Hong Kong University of Science and Technology (GZ), Guangzhou, China*

<sup>d</sup>*Department of Psychology, Faculty of Arts and Sciences, Beijing Normal University at Zhuhai, Zhuhai, China*

<sup>e</sup>*College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China*

<sup>f</sup>*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China*

## Abstract

Affective computing stands at the forefront of artificial intelligence (AI), seeking to imbue machines with the ability to comprehend and respond to human emotions. Central to this field is emotion recognition, which endeavors to identify and interpret human emotional states from different modalities, such as speech, facial images, text, and physiological signals. In recent years, important progress has been made in generative models, including Autoencoder, Generative Adversarial Network, Diffusion Model, and Large Language Model. These models, with their powerful data generation capabilities, emerge as pivotal tools in advancing emotion recognition. However, up to now, there remains a paucity of systematic efforts that review generative technology for emotion recognition. This survey aims to bridge the gaps in the existing literature by conducting a comprehensive analysis of over 320 research papers until June 2024. Specifically, this survey will firstly introduce the mathematical principles of different generative models and the commonly used datasets. Subsequently, through a taxonomy, it will provide an in-depth analysis of how generative techniques address emotion recognition based on different modalities in several aspects, including data augmentation, feature extraction, semi-supervised learning, cross-domain, etc. Finally, the review will outline future research directions, emphasizing the potential of generative models to advance the field of emotion recognition and enhance the emotional intelligence of AI systems.

**Keywords:** Emotion Recognition, Generative Technology, Autoencoder, Generative Adversarial Network, Diffusion Model, Large Language Model

\*Corresponding author.

Email addresses: mafei@gml.ac.cn (Fei Ma), yuan-yc23@mails.tsinghua.edu.cn (Yucheng Yuan), xielyifan@stu.xjtu.edu.cn (Yifan Xie), hren066@connect.hkust-gz.edu.cn (Hongwei Ren), ivanliu@bnu.edu.cn (Ivan Liu), heying@szu.edu.cn (Ying He), renfuji@uestc.edu.cn (Fuji Ren), yufei@gml.ac.cn (Fei Richard Yu), ni.shiguang@sz.tsinghua.edu.cn (Shiguang Ni)

<sup>1</sup>The two authors contribute equally to this work.

## 1. Introduction

In the field of affective computing [1, 2, 3], researchers are actively engaged in developing technologies that can simulate, recognize, and understand human emotions. The primary objective of these endeavors is to imbue computers with a profound level of emotional perception, thereby facilitating more intelligent and human-like interactive experiences [4, 5]. There are extensive applications of affective computing in numerous fields. For example, affective computing can be employed in healthcare to monitor patients' emotional changes, assist in diagnosing mental disorders, and evaluate treatment outcomes [6, 7]. In education, it is employed to develop emotionally intelligent educational systems that adjust teaching strategies based on students' emotional states, thereby enhancing learning effectiveness [8, 9]. In the field of autonomous driving, affective computing can monitor drivers' emotional conditions and provide timely warnings for risks such as fatigue driving [10, 11]. In marketing, it enables the analysis of consumers' emotional preferences, providing insights for businesses to optimize marketing strategies and customer service [12, 13].

As a core component of affective computing, emotion recognition [14, 15, 16, 17] focuses on identifying emotional states from the data that conveys human emotions. Due to the inherent heterogeneity of human emotional expressions, these data can come from various modalities, such as speech, facial images, text, and physiological signals. Notably, speech emotion recognition (SER) identifies the speaker's emotion by extracting features such as pitch, volume, and speech rate [18]. Facial emotion recognition (FER) predicts human emotional states by detecting and tracking facial feature points [19]. Textual emotion recognition (TER) focuses on identifying the emotional content expressed in written language, such as social media posts, customer reviews, or personal messages [20]. Physiological signals, such as electroencephalogram (EEG), electrocardiogram (ECG), heart rate variability (HRV), electrodermal activity (EDA), has become more prevalent in recent years due to their ability to objectively measure an individual's emotional state [21]. In addition, multimodal emotion recognition (MER) also receives increasing attention [22, 23]. Compared with emotion recognition based on a single modality, it comprehensively utilizes emotional information from different modalities to improve the performance of emotion recognition.

Over the past few decades, researchers have conducted extensive work on emotion recognition based on machine learning and deep learning [16, 24], among which generative technology-based methods have shown tremendous potential. In contrast to discriminative models that directly learn the decision boundaries between different emotion categories, generative methods focus on learning the intrinsic distribution and representation of the emotional data [25, 26]. Generative models, with their powerful generative capabilities, can generate samples highly similar to real emotional data, effectively enhancing the performance and generalization ability of emotion recognition. However, to the best of our knowledge, there is currently a lack of a systematical review that summarizes the work of generative technology for emotion recognition.

In view of this, this paper aims to provide a comprehensive overview of the research progress in generative technology for emotion recognition, based on more than 320 technical papers up to June 2024. The overall flow is shown in Figure 1. Specifically, this survey will focus on several representative models, such as Au-

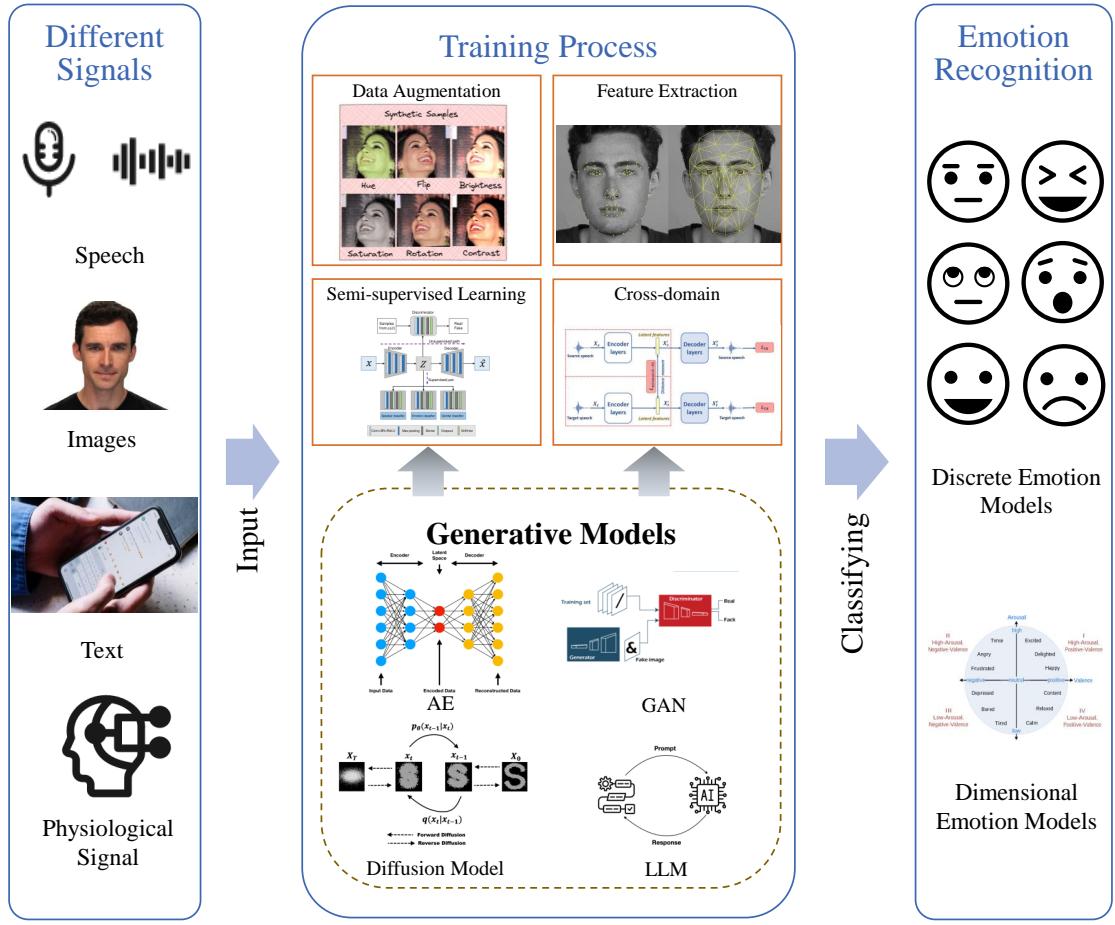


Figure 1: Schematic Diagram of Generation technology for Emotion Recognition. <sup>2</sup>

toencoder (AE) [27], Generative Adversarial Network (GAN) [28], Diffusion Model (DM) [29], and Large Language Model (LLM) [30, 31]. AEs enable the modeling of data distribution by learning a compressed representation of the input data and reconstructing the input from the compressed representation. Among them, the Variational Autoencoder (VAE) can generate new samples similar to the input data by introducing hidden variables and variational inference. GANs learn through the adversarial learning of the generator and the discriminator, where the generator tries to generate samples that are as similar as possible to the real data distribution, while the discriminator tries to differentiate between the generated samples and the real samples. The game between the two finally enables the generator to generate high-quality samples. DMs achieve sample generation by learning the process of gradual perturbation of the data distribution and by reversing the process, which has excellent performance in terms of generation quality and diversity. LLMs, such as the Generative Pre-trained Transformer (GPT) series [32], learn the statistical patterns and semantic representations of language by pre-training on vast amounts of text data. This pre-training process involves exposing the models to diverse and extensive corpora, allowing them to capture the intricate nuances, contextual dependencies, and latent structures inherent in natural language. In terms of the structure of this survey, we will elucidate the mathematical principles underlying these models to help readers gain a comprehensive understanding of the development and evolution of generative techniques.

Then, this survey will cover datasets commonly used in the field of generative technology for emotion recognition. For emotion recognition based on speech, facial images, and textual information, a variety of datasets will be discussed. Particularly, FER2013 [36] provides a wide range of facial expressions collected via the Google search engine and labeled for seven emotional states. AFEW [37], derived from movies, serves as a dynamic, real-world dataset capturing both auditory and visual expressions. IEMOCAP [38] consists of audio-visual recordings of acted emotional states. RAVDESS [39] includes both voice and video recordings of actors performing emotional expressions in a controlled environment. Additionally, CMU-MOSEI [40] integrates audio, video, and text annotations, making it one of the most versatile datasets available for studying emotion recognition based on multiple modalities. For emotion recognition based on physiological signals, several datasets will be introduced. For example, DEAP [41] includes EEG and peripheral physiological signals of participants as they respond to music videos, providing valuable insights into emotional responses. DREAMER [42] is another significant dataset that provides both EEG and ECG signals while subjects watch emotional video clips. By exploring these diverse datasets, readers will gain a better understanding of the characteristics and challenges associated with different emotion recognition tasks.

Based on the introduction of the principles of generative models and the datasets used, this review will show the research progress of generative models in the field of emotion recognition from various perspectives, including data augmentation, feature extraction, semi-supervised learning, cross-domain, and so on. The taxonomy is shown in Figure 2. Regarding data augmentation, it is an important way to improve the generalization performance of the model. However, traditional data augmentation methods, such as rotation and cropping, are difficult to portray the intrinsic attributes of emotional data [43]. Generative models open a new path for sentiment data augmentation. For example, in SER, researchers [44, 45] use GAN to generate samples similar to real speech emotion data, enhancing the model's ability to adapt to different speech conditions. In FER, researchers [46, 47] integrate GAN and VAE to generate facial expression images under different pose, light, and occlusion conditions, improving the generalization ability of the model. In TER, researchers [48, 49] utilize LLMs to generate text samples with different emotional characteristics, expanding the diversity of training data. Feature extraction here refers to the use of generative models to learn effective feature representations from emotional data [50]. For instance, AEs and GANs can be used to learn compact representations from speech and facial expressions. These learned features can be

---

Facial image is sourced from <https://zenodo.org/records/1188976>.

Feature extraction module is adapted from is from <https://medium.com/clique-org/how-to-create-a-face-recognition-model-using-facenet-keras-fd65c0b092f1>.

Dimensional emotion models can be found at [33].

Data augmentation schematic is from <https://www.baeldung.com/cs/ml-gan-data-augmentation>.

Semi-supervised learning schematic is provided by [34].

Cross-domain image is from [35].

Schematics of each of the four generative models are taken from <https://www.compthree.com/blog/autoencoder/>, <https://www.javatpoint.com/generative-adversarial-network>, <https://towardsdatascience.com/diffusion-models-made-easy-8414298ce4da>, <https://engineeringprompts.substack.com/p/frameworks-to-build-llm-applications>

used in downstream emotion recognition tasks to improve recognition performance [51, 52]. Semi-supervised learning [53, 54] is a paradigm for jointly training models using a large amount of unlabeled data and a small amount of labeled data, which is important for alleviating the scarcity of labeled data. Generative models provide new technical tools for semi-supervised emotion recognition. For example, some researchers try to apply GANs to semi-supervised learning by synthesizing unlabeled samples using a generator, then predicting pseudo-labels using a discriminator, and finally co-training classifiers with pseudo-labeled samples and real labeled samples [55, 56]. Cross-domain refers to the phenomenon where the performance of emotion recognition models significantly degrades in different domains [57, 58]. This is usually caused by the differences in data distribution between the source domain and the target domain. By capturing the commonalities between different domains, generative models can achieve cross-domain emotion recognition and improve the adaptability of models in new domains [35, 59].

Finally, this review will present the main findings and provide an outlook on future research directions based on the above systematic analysis. The main findings include: (i) Generative models are most widely used in FER compared to emotion recognition based on other modalities. (ii) Generative models are primarily applied to emotion recognition in the form of data augmentation and feature extraction, although the specific working mechanisms may differ. (iii) AEs and GANs are more widely used than other generative models. Looking ahead, there are still many directions worth exploring, which include: (i) Investigating how to combine DMs and Transformer architecture for emotion recognition. This is expected to be a hotspot for future research, as these have already achieved significant success in a number of domains and may provide new opportunities to improve emotion recognition performance [60, 61]. (ii) Combining techniques such as reinforcement learning [62] and federated learning [63] with generative models to better model the complex dependencies and temporal evolution of emotions. This highly promising direction has the potential to capture the intricate dynamics and context-dependent nature of emotional expressions, leading to more accurate and nuanced emotion recognition systems. (iii) Further expanding the combination of generative technology in emotion recognition, particularly in virtual reality (VR) and augmented reality (AR) applications [64, 65], to enhance user experiences. (iv) Based on the generative models discussed in this paper, it is possible to directly generate natural, realistic content with rich emotional expressions, such as emotionally charged speech, images, text, and videos [66, 67]. To sum up, this relatively unexplored area presents a wealth of opportunities for innovation and creative applications of generative models, with the potential to improve how emotions are detected, interpreted, and responded to in immersive environments. Through the above work, this survey aims to deepen the understanding of generative technology in the context of emotion recognition and provide inspiration and guidance for future research. Overall, the main contributions of this survey include:

(1) To the best of our knowledge, this paper is the first work to systematically review the research progress of generative technology in the field of emotion recognition, filling a gap in the existing literature.

(2) By analyzing more than 320 research papers, this paper gives a taxonomy of generative models for emotion recognition from different perspectives, including data augmentation, feature extraction, semi-supervised learning, cross-damain, etc. This

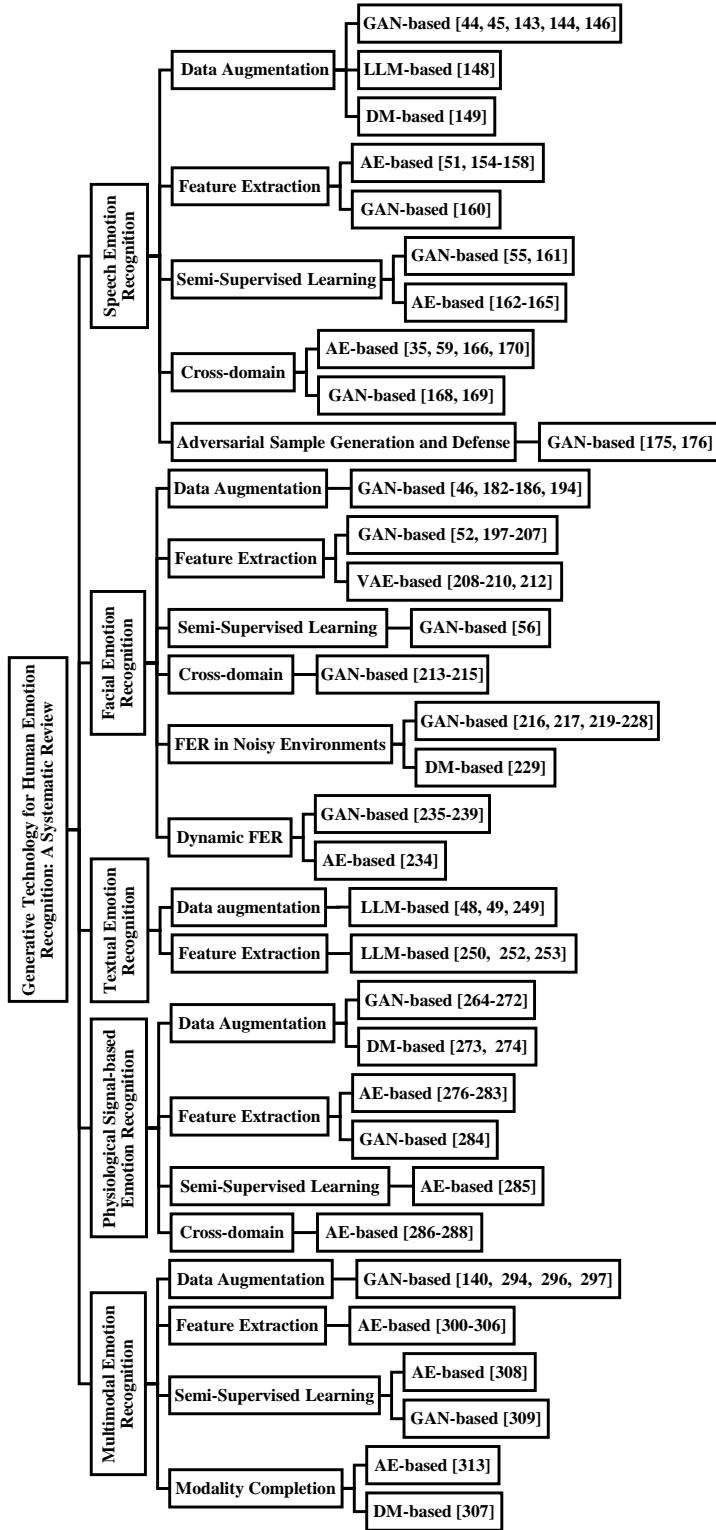


Figure 2: Taxonomy of This Survey.

in-depth analysis enables readers to gain a thorough understanding of the diverse applications and methodologies of generative technology in emotion recognition.

(3) We summarize the benchmark datasets used according to the different modalities, highlighting key characteristics such as sample size, number of subjects, and emotion categories. Furthermore, we present the performance of different generative models on emotion recognition based on these datasets.

(4) Finally, we discuss some findings of generative models for emotion recognition and further point out potential future research directions.

The rest of the paper is structured as follows: Section 2 introduces the difference between this review and other existing reviews, Section 3 gives the mathematical principles of different generative models, Section 4 describes the datasets used, Sections 5 - 9 present a series of work on SER, FER, TER, emotion recognition based on physiological signals, and MER based on generative models, respectively, Section 10 gives the main findings and future outlook, and Section 11 gives the conclusion. A list of abbreviations is given in Table 1.

Table 1: Main Acronyms.

Acronym	Full Form	Acronym	Full Form
SER	Speech Emotion Recognition	FER	Facial Emotion Recognition
TER	Textual Emotion Recognition	MER	Multimodal Emotion Recognition
AE	Autoencoder	DAE	Denoising Autoencoder
AAE	Adversarial Autoencoder	VAE	Variational Autoencoder
SAE	Stacked Autoencoder	CAE	Convolutional Autoencoder
GAN	Generative Adversarial Network	DCGAN	Deep Convolutional GAN
CGAN	Conditonal GAN	ACGAN	Auxiliary Classifier GAN
DANN	Domain Adversarial Neural Network	GCNN	Graph Convolutional Neural Network
DM	Diffusion Model	LLM	Large Language Model
DDPM	Denoising Diffusion Probabilistic Model	ViT	Vision Transformer
GPT	Generative Pre-trained Transformer	T5	Text-to-Text Transfer Transformer
MLP	Multilayer Perceptron	CNN	Convolutional Neural Networks
RNN	Recurrent Neural Network	LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficient	AU	Action Unit
EEG	Electroencephalogram	ECG	Electrocardiogram
EMG	Electromyogram	HRV	Heart Rate Variability
EDA	Electrodermal Activity	fMRI	functional Magnetic Resonance Imaging
AUC	Area Under Curve	CCC	Consistency Correlation Coefficient
ACC	Accuracy	UAR	Unweighted Average Recall

## 2. The Difference

Existing reviews of generative technology mainly focus on GAN-based generation of single modalities such as images, speech, and text. For example, Kammoun et al. [68] provide an overview of advances in GANs for face generation, focusing on their applications, architectures, training modes, and evaluation metrics. Wali et al. [69] categorize speech GANs according to application areas: speech synthesis, speech enhancement and conversion, and data augmentation in speech recognition and emotional speech recognition systems, and summarize commonly used datasets and evaluation metrics in speech GANs. In particular, the review by Hajarolasvadi et al. [70] covers various aspects of GAN-based human emotion synthesis, including facial expression synthesis and speech emotion synthesis.

Emotion recognition, as a much-anticipated area in affective computing, has been reviewed by researchers in a number of aspects. For example, Deng and Ren

[20] provide a comprehensive overview of TER research, focusing on deep learning approaches, which categorizes TER methods based on different stages of implementation, such as word embeddings, architectures, and training levels. Zhao et al. [71] present the history of FER, emotion representation models, well-known datasets, and applications of FER in fields such as transportation, healthcare, and business. Khare et al. [17] present emotion models based on different modalities, stimuli used for emotion elicitation, and existing automatic emotion recognition systems. Cîrneanu et al. [72] discuss different neural network architectures used in FER, including Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and emphasize the key elements and performance of each architecture. Younis et al. [73] provide an overview of the advancements in emotion recognition from the broad perspective of machine learning, but lack a detailed analysis of the progress made using generative models.

Through the above analysis, it can be seen that existing works conduct separate reviews of generative technology and emotion recognition. However, due to the superior characteristics of generative models, they begin to be widely applied in the field of emotion recognition, such as data augmentation [74], feature extraction [75], etc. Nonetheless, so far, there is no work that systematically reviews generative models for emotion recognition. This paper aims to fill this gap by comprehensively summarizing the latest progress and applications of generative models in the field of emotion recognition, providing references and insights for future research. Table 2 shows how our work differs from other reviews.

Table 2: Differences Between Ours and Previous Reviews.

Reference	Focus on generative models	Related to emotion recognition	Focus on multiple modalities
Ours	✓	✓	✓
Kammoun et al. [68]	✓		
Wali et al. [69]	✓	✓	
Hajarolasvadi et al. [70]	✓		
Deng and Ren [20]		✓	
Zhao et al. [71]		✓	
Khare et al. [17]		✓	✓
Cîrneanu et al. [72]	✓	✓	
Younis et al. [73]		✓	✓

### 3. Generative Model

Generative models are a powerful class of machine learning models that aim to learn the underlying probability distribution of a given dataset. By capturing the intricate patterns and structures within the data, these models can generate new samples that closely resemble the original data distribution. The ability to create realistic and diverse examples has led to a wide range of applications across various fields [76, 77]. Generative models can be achieved in different approaches to model the probability distribution of data. Common generative modeling methods include AEs, GANs, DMs, and LLMs. These methods have their own unique characteristics in modeling complex data distributions and enhancing model generation capabilities [78]. The following describes each of these models in detail.

### 3.1. Autoencoder (AE)

AE [27] is an unsupervised generative model that learns the compressed representation of data and reconstructs the input data. It consists of an encoder that maps the original input data to a low-dimensional latent space and a decoder that maps the latent representation back to the original data space. Typical variants of AE include Adversarial Autoencoder (AAE) and Variational Autoencoder (VAE). Unlike traditional AE, AAE [79] introduces a discriminator network that evaluates the authenticity of reconstructed data by comparing the encoded latent representation with randomly sampled vectors from a prior distribution.

VAE [80] is a generative model that combines the ideas of AE and variational inference. It works as follows: First, the input data is passed through an encoder, which maps the input data to the mean and variance parameters in the latent space. Then, a hidden variable is sampled from the latent space, usually using a normal distribution. Next, the sampled hidden variable is passed through a decoder, which maps the hidden variable back to the space of the original input data to generate a reconstructed sample. To constrain the latent representation, VAE uses the Kullback-Leibler (KL) divergence as a regularization term to make the learned latent representation closer to a prior distribution. Therefore, the VAE’s loss function consists of a reconstruction error term and a KL divergence term, given by:

$$L = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + \text{KL}(q(z|x)||p(z)) \quad (1)$$

where  $\mathbb{E}_{q(z|x)}$  indicates the calculation of the expectation over the posterior distribution  $q(z|x)$  of the latent variable  $z$  generated by the encoder.  $\log p(x|z)$  refers to the reconstruction error term, which indicates the log-likelihood of generating data  $x$  given the latent variable  $z$ .  $\text{KL}(q(z|x)||p(z))$  represents the KL divergence term, which is used to evaluate the dissimilarity between the latent variable distribution generated by the encoder  $q(z|x)$  and the prior distribution  $p(z)$ .

### 3.2. Generative Adversarial Network (GAN)

GAN [28] is a generative model consisting of two components: a generator and a discriminator. They learn from each other in an adversarial process, where the generator aims to generate more realistic samples, and the discriminator aims to distinguish between the generated samples and real samples. In addition to the traditional GAN model, there are several variants of GANs, such as Conditional GANs (CGANs) [81], Deep Convolutional GANs (DCGANs) [82], Wasserstein GANs [83], and CycleGANs [84].

To be specific, the training process of GAN involves an iterative optimization procedure. The generator takes a random noise vector as input and generates a sample. The discriminator then takes both real and generated samples as input, and outputs a probability that represents the likelihood of the input being a real sample. The parameters of the generator and discriminator are updated separately to improve their performance. The generator’s loss function is designed to maximize the probability of the discriminator classifying the generated samples as real, which can be formulated as:

$$L_G = -\mathbb{E}_{z \sim p_z(z)}[\log(D(G(z)))] \quad (2)$$

where  $G$  represents the generator,  $D$  represents the discriminator,  $z$  is a random noise vector sampled from a prior distribution  $p_z(z)$  (e.g., a Gaussian distribution),

and  $\mathbb{E}$  denotes the expected value. On the other hand, the discriminator's loss function is designed to maximize the probability of correctly classifying both real and generated samples, which can be expressed as:

$$L_D = -\mathbb{E}_{x \sim p_{data}(x)}[\log(D(x))] - \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (3)$$

where  $x$  represents real samples from the training data distribution  $p_{data}(x)$ . This dynamic equilibrium results in the generator being able to generate samples that resemble real data closely, making GAN capable of generation tasks.

### 3.3. Diffusion Model (DM)

DMs [29] are a class of generative models which have gained significant attention in recent years due to their ability to generate high-quality and diverse samples. Unlike other generative models that directly learn the data distribution, DMs learn to transform a simple noise distribution into the desired data distribution through a gradual diffusion process. The core idea behind DMs is to define a forward diffusion process that gradually adds noise to the data, and then learn a reverse diffusion process that removes the noise step by step, eventually generating clean samples. The diffusion process is typically defined as a Markov chain, where each step corresponds to a small amount of Gaussian noise being added to the data.

The forward diffusion process can be described by a sequence of latent variables  $z_0, \dots, z_T$ , where  $z_0$  represents the clean data and  $z_T$  represents the fully noised data. The transition from  $z_t$  to  $z_{t+1}$  is modeled by a Gaussian transition kernel  $q(z_{t+1}|z_t)$ , which adds Gaussian noise with a predetermined variance schedule. The reverse diffusion process, also known as the denoising process, is learned by training a neural network to predict the noise that was added in each step of the forward diffusion process. Given a noisy sample  $z_t$  at time step  $t$ , the goal is to learn a function  $f_\theta(z_t, t)$  that predicts the noise  $\epsilon_t$ , such that the clean data can be recovered by subtracting the predicted noise from the noisy sample:

$$z_{t-1} = z_t - \sqrt{\beta_t} \cdot \epsilon_t \quad (4)$$

where  $\beta_t$  is a time-dependent noise scaling factor. The training objective of DMs is to minimize the weighted sum of the squared error between the predicted noise and the actual noise added in each step of the forward diffusion process. This is typically achieved using a variant of the variational lower bound, such as the evidence lower bound (ELBO) or the simplified loss proposed in [29].

During inference, the generation process starts with a sample from the noise distribution  $z_T$  and iteratively applies the learned denoising function  $f_\theta(z_t, t)$  to remove the noise and obtain increasingly cleaner samples. The generated sample at each step is obtained by:

$$z_{t-1} = z_t - \sqrt{\beta_t} \cdot f_\theta(z_t, t) \quad (5)$$

By repeating this process for a sufficient number of steps, DMs can generate high-quality samples that closely resemble the training data.

### 3.4. Large Language Model (LLM)

LLMs are a type of deep learning model based on the principles of generative modeling, aiming to learn the probability distribution of words in text data to

generate coherent and natural text sequences [30, 31]. The basic principle involves pre-training on large-scale text data to capture semantic relationships and syntactic structures between words, followed by fine-tuning on specific tasks to adapt to particular application scenarios [30]. They typically employ neural network architectures like the Transformer [85] model. During pre-training, these models utilize unsupervised learning on unlabeled text data to learn rich language representations [86]. In the fine-tuning stage, LLMs improve their performance on specific tasks through supervised learning [87]. LLMs, such as T5 (Text-to-Text Transfer Transformer) [87], XLnet [88], LLaMA [89], and GPT (Generative Pre-trained Transformer) [32], demonstrate significant potential in various natural language processing tasks, leading to advancements in text generation [90].

Through the above discussion, we explore the mathematical principles behind several representative generative models, including AE, GAN, DM, and LLM. These models exhibit unique characteristics in learning complex data distributions and generating realistic samples, providing powerful tools for applications in various domains, including emotion recognition [69, 77].

#### 4. Databases

This section focuses on the datasets commonly used in generative technology for emotion recognition. These datasets are crucial in this field because the performance of related algorithms largely depends on the quality and richness of the data. Table 3 lists these datasets, which capture and represent emotional information in various ways, enabling researchers to develop and evaluate generative models that can effectively understand and generate emotional content. Specifically, as shown in Table 3, these datasets can be divided into two main categories: spontaneous datasets and in-the-wild datasets [91]. Spontaneous datasets include expressions simulated by participants, where emotions are expressed naturally despite participants' awareness that they are being monitored. The acquisition environment for these datasets is typically a controlled laboratory setting. On the other hand, in-the-wild datasets capture emotions in real-world scenarios without the need for controlled or processed collection, and participants are filmed in their natural environments. In addition, "Modalities" in Table 3 refers to the modalities contained in the dataset, such as visual or physiological data. "Samples" indicates the type and number of samples in the dataset. "Subjects" refers to the number of individuals included in the dataset. "Categories" shows the categories for emotion recognition, indicating the range of emotions that the dataset covers. A slash indicates that the specific information is not explicitly mentioned in the dataset description. For example, the IAPS dataset [92] contains data in the form of visual modality, including over 9,000 images, and the emotion category information includes valence, arousal, and dominance. In the following sections, we will show the performance of different generative models for emotion recognition based on the datasets listed in Table 3.

Table 3: Common Databases for Emotion Recognition with Generative Models.

Database	Year	Modalities	Spontaneous/in-the-wild	Samples	Subjects	Categories
IAPS [92]	1997	Visual	In the Wild	More than 9000 images	/	Valence, Arousal, and Dominance
TFD [93]	2010	Visual	In the Wild	4178 images	/	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
SFEW [94]	2011	Visual	In the Wild	1739 images	330	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
FER2013 [36]	2013	Visual	In the Wild	35887 images	/	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
AffectNet [95]	2017	Visual	In the Wild	450000 images	/	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral, Contempt
RAF-DB [96]	2017	Visual	In the Wild	29672 images	/	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral, Contempt
JAFFE [97]	1998	Visual	Spontaneous	219 images	10	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
BU-3DFE [98]	2006	Visual	Spontaneous	2500 images	100	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
TFEID[99]	2007	Visual	Spontaneous	7200 images	40	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral, Contempt
Multi-PIE [100]	2008	Visual	Spontaneous	750000 images	337	Happiness, Surprise, Sadness, Anger, Fear, Disgust
BU-4DFE [101]	2008	Visual	Spontaneous	606 sequences	101	Happiness, Surprise, Sadness, Anger, Fear, Disgust
FAU Aibo [102]	2008	Audio	Spontaneous	9.2 hours of speech	31	Joyful, Surprised, Emphatic, Helpless, Touchy, Rest

**Table 3** continued from previous page

Database	Year	Modalities	Spontaneous/ In The Wild	Samples	Subjects	Category
CHB-MIT [103]	2009	EEG	Spontaneous	/	23	Arousal and Valence
USTC- NVIE[104]	2010	Visual	Spontaneous	236 images	215	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
CK+ [105]	2010	Visual	Spontaneous	593 images	123	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral, Contempt
FACES [106]	2010	Visual	Spontaneous	1026 videos	171	Happiness, Surprise, Sadness, Anger, Fear, Disgust
Oulu- CASIA[107]	2011	Visual	Spontaneous	480 sequences	80	Happiness, Surprise, Sadness, Anger, Fear, Disgust
SMIC [108]	2013	Visual	Spontaneous	164 sequences	16	Positive, Negative, Surprise
SEED [109]	2013	EEG	Spontaneous	/	15	Positive, Neutral, Negative
CFEE [110]	2014	Visual	Spontaneous	229 images	230	Happiness, Surprise, Sadness, Anger, Fear, Disgust
CASME II [111]	2014	Visual	Spontaneous	255 sequences	35	Happiness, Surprise, Disgust, Repression, and Others
ECG 200 [112]	2001	ECG	Spontaneous	200 samples	/	Happiness, Surprise, Sadness, Anger, Fear, Disgust
SAMM [113]	2016	Visual	Spontaneous	159 sequences	32	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Contempt
ADFES- BIV [114]	2016	Visual	Spontaneous	320 videos	12	Anger, Contempt, Disgust, Embarrassment, Fear, Pride, Joy, Neutral, Sad, Surprise

**Table 3 continued from previous page**

Database	Year	Modalities	Spontaneous/ In The Wild	Samples	Subjects	Category
BP4D+ [115]	2016	Visual	Spontaneous	about 1.4 million frames	140	12 Different Categories and 5-point Likert-type Scales
KDEF [116]	2018	Visual	Spontaneous	490 images	272	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
AudiofMRI Parallel Set [117]	2018	fMRI	Spontaneous	/	18	Sad, Happy, Excited, Surprise, Neutral, Angry, Distress, Frustrated
F2ED [118]	2019	Visual	Spontaneous	219719 images	119	54 Different Facial Expression Classes
Affwild [119]	2019	Visual	In the Wild	298 videos	/	Arousal and Valence
AFEW [37]	2012	Visual + Audio	In the Wild	1426 sequences	330	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
EMO-DB [120]	2005	Audio + Text	Spontaneous	500 utterances	10	Neutral, Anger, Fear, Joy, Sadness, Disgust, Boredom
eNTERF-ACE'05 [121]	2006	Visual + Audio	Spontaneous	1166 sequences	42	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
SAVIEE[122]	2008	Visual + Audio	Spontaneous	480 utterances	4	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
VAM [123]	2008	Audio + Text	Spontaneous	12 hours of data	47	Happiness, Surprise, Sadness, Anger, Fear, Disgust
MMI [124]	2010	Visual + Audio	Spontaneous	2900 videos	75	Happiness, Surprise, Sadness, Anger, Fear, Disgust
RECOLA [125]	2013	Visual + Audio	Spontaneous	1308 utterances	23	Valence and Arousal

**Table 3** continued from previous page

Database	Year	Modalities	Spontaneous/ In The Wild	Samples	Subjects	Category
CREMA-D [126]	2014	Visual + Audio	Spontaneous	7442 utterances	91	Happiness, Surprise, Sadness, Anger, Fear, Disgust
EMOVO [127]	2014	Audio + Text	Spontaneous	84 sentences	6	Happiness, Surprise, Sadness, Anger, Fear, Disgust
JTES [128]	2016	Audio + Text	Spontaneous	20,000 speech samples	100	Neutral, Angry, Joy, Sad
MSP-IMPROV [129]	2017	Visual + Audio	Spontaneous	8348 samples	12	Sadness, Happiness, Anger, Neutrality
DREAMER [42]	2017	EEG + ECG	Spontaneous	/	25	Arousal and Valence
RAVDESS [39]	2018	Visual + Audio	Spontaneous	7356 videos	24	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral, Contempt
URDU [130]	2018	Audio + Text	Spontaneous	400 utterances	38	Angry, Happy, Sad, Neutral
Emoti-W [131]	2017	Visual + Audio + Text	In the Wild	More than 100 movies	/	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
MELD [132]	2018	Visual + Audio + Text	In the Wild	more than 1400 dialogues	304	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
CMU-MOSEI[40]	2018	Visual + Audio + Text	In the Wild	videos and 23453 sentences	1000	Happiness, Surprise, Sadness, Anger, Fear, Disgust
OMG [133]	2018	Visual + Audio + Text	In the Wild	7371 utterances	/	Happiness, Sadness, Surprise, Fear, Anger, Disgust, Neutral
IEMOCAP [38]	2008	Visual + Audio + Text	Spontaneous	10039 samples	10	Arousal and Valence

**Table 3** continued from previous page

Database	Year	Modalities	Spontaneous/ In The Wild	Samples	Subjects	Category
DEAP [41]	2011	EEG + EOG + EMG + GSR	Spontaneous /	/	32	Arousal and Valence
UlmTSST [134]	2021	EDA + ECG + RESP + BPM	Spontaneous /	/	69	Arousal and Valence

## 5. Speech Emotion Recognition (SER)

SER is a crucial branch of emotion recognition that aims to identify and understand the emotional state of a speaker by analyzing the features in speech signals [18]. Speech signals contain rich emotional information, such as intonation, volume, and speaking rate, which can reflect the speaker’s emotional changes [135]. In recent years, generative models have contributed to the development of SER in multiple aspects. In the following sections, we will delve into the specific applications of generative models in SER. Section 5.1 will focus on discussing speech data augmentation based on generative models. Section 5.2 will explore the application of generative models for speech feature extraction. Section 5.3 will describe the application of generative models in conjunction with semi-supervised learning, Section 5.4 will present cross-domain SER based on generative models, and Section 5.5 will provide an introduction to the use of generative models for adversarial sample generation and defense.

### 5.1. Speech Data Augmentation

Data augmentation is a crucial topic in emotion recognition, as it addresses the problem of limited training data [43, 136]. In many real-world scenarios, acquiring and annotating large amounts of emotional data can be time-consuming, expensive, and labor-intensive. Moreover, emotional data may also suffer from the problem of class imbalance, where the number of labeled samples for certain emotional categories is relatively small, resulting in poor performance of the model on these categories [137]. The scarcity of labeled data described above can hinder the development of robust and generalizable emotion recognition models. Data augmentation techniques aim to mitigate this issue by artificially increasing the size and diversity of the training dataset, thereby improving the model’s ability to learn and generalize from the available data. Traditionally, researchers have proposed various strategies for data augmentation in SER [138]. These strategies include simple techniques such as noise addition, pitch shifting, time stretching, and speed perturbation. These methods aim to introduce variations in the speech signal while preserving the emotional content. However, these traditional approaches often rely on handcrafted rules and may not capture the complex and nuanced patterns of emotional speech.

In recent years, generative models have emerged as a promising approach for data augmentation in SER [44, 139]. By leveraging the power of generative models, researchers can create realistic and diverse emotional speech samples, effectively expanding the training dataset. Figure 3 illustrates a general architecture. Table 4 presents different generative methods used for data augmentation in SER and their performance, which is typically evaluated using metrics such as Accuracy (ACC) and Unweighted Average Recall (UAR), where ACC provides an overall measure of the model’s correctness, indicating its general performance, while UAR, calculates the average recall score across all classes, making it particularly useful when dealing with imbalanced datasets.

To be specific, generative models can effectively augment speech data by directly generating synthetic samples. For example, Chatziagapi et al. [44] adapt and improve the Balancing GAN architecture to generate realistic emotional data for minority classes. Heracleous et al. [139] propose a data augmentation framework that utilizes CycleGAN to transfer natural speech to emotional speech and employs

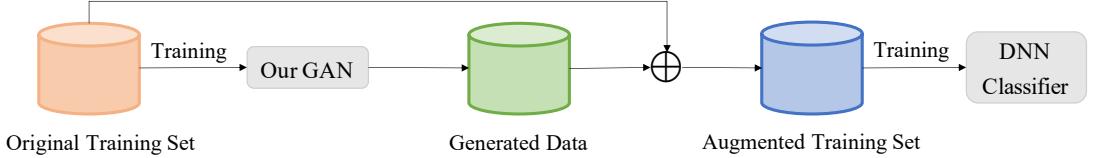


Figure 3: The pipeline of data augmentation from [140]: First, a generative model is employed to generate new data, then the original training set is combined with the generated data, creating an augmented training set for the emotion classification task.

a Vision Transformer (ViT) [141] as the classifier for SER. Wang et al. [142] proposes a GAN-based augmentation strategy with triplet loss to increase and stabilize the augmentation performance for the SER task.

In addition to directly performing augmentation in the data space for SER, some researchers attempt to conduct data augmentation in the feature space. This approach aims to manipulate and transform the features extracted from the raw speech data, generating new feature representations to increase the diversity and robustness of the training data. For example, in [45], a CycleGAN model is employed for data augmentation, and two feature selection networks, Fisher and Linear Discriminant Analysis, are used to enhance SER performance. Yi and Mak [143] combine GANs and AEs to design an adversarial data augmentation network to generate emotional feature vectors that share a common latent representation. Latif et al. [144] utilize a data augmentation technique called mixup [145] to enhance GAN’s ability in representation learning and synthesis of feature vectors. In another study, Sahu et al. [146] employ GANs to learn the distribution of low-dimensional representations of high-dimensional feature vectors and CGAN to learn the distribution of high-dimensional feature vectors conditioned on labels.

Particularly, it is worth noting that apart from directly performing data augmentation on the speech modality, another approach is to transfer information from other modalities to the speech modality for data augmentation, thereby assisting in improving the performance of SER. This cross-modality information transfer-based data augmentation method has received increasing attention from researchers in recent years. For instance, He et al. [147] employ GAN and VAE to convert facial images into spectrograms to increase the number of sample data, thereby enhancing SER. Ma et al. [148] use a LLM, GPT-4, to generate emotional text, then use Azure emotional TTS to synthesize emotional speech. Malik et al. [149] utilize an improved DM model to generate synthesized emotional data. They use text embeddings represented by bidirectional encoders from transformers (BERT) [150] to adjust the DM model to generate high-quality synthesized emotional samples in the voices of different speakers.

## 5.2. Speech Feature Extraction

Feature extraction plays a vital role in SER, as it involves deriving meaningful representations from speech signals to capture emotional cues effectively. Traditionally, handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs) [151] and prosodic features have been extensively utilized in SER systems. However, these handcrafted features have limitations in capturing subtle emotional variations and can be sensitive to noise and speaker variability [152]. To address these lim-

Table 4: Literature on Generative Models for Data Augmentation in SER.

Reference	Year	Augmentation Domain	Model	Dataset	Performance (%)
Chatziagapi et al. [44]	2019	Sample Space	GAN	IEMOCAP	UAR: 53.6
Heracleous et al. [139]	2022	Sample Space	CycleGAN	RAVDESS/JTES	ACC: 77/66.1
Wang et al. [142]	2022	Sample Space	GAN	IEMOCAP	UAR: 58.19
Shilandari et al. [45]	2022	Feature Space	CycleGAN	EMO-DB/SAVEE/IEMOCAP	ACC: 86.32/73.82/61.67
Yi and Mak [143]	2020	Feature Space	GAN	EMO-DB/IEMOCAP	ACC: 84.49/71.45
Latif et al. [144]	2020	Feature Space	GAN	IEMOCAP/MSP-IMPROV	UAR: 59.6/46.60
Sahu et al. [146]	2018	Feature Space	GAN + CGAN	IEMOCAP	UAR: 60.29
He et al. [147]	2020	Sample Space (Cross-modality)	GAN + VAE	Multi-PIE/CK+/MMI/Oulu-CASIA	ACC: 61.3
Ma et al. [148]	2023	Sample Space (Cross-modality)	GPT-4	IEMOCAP	ACC: 68.85
Malik et al. [149]	2023	Sample Space (Cross-modality)	DM	IEMOCAP/MSP-IMPROV/CREMA-D/RAVDESS	UAR: 61.22

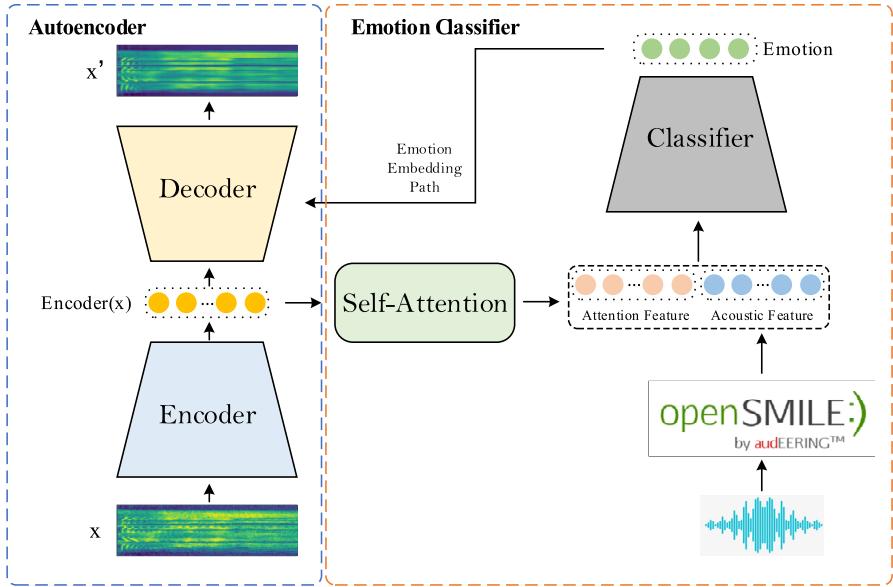


Figure 4: The framework designed for extracting speech features from [154]: It introduces instance normalization and an emotion embedding path to guide the AE in learning a priori knowledge from the label, enabling it to distinguish the most emotion-related features. The latent representation learned by the AE, enhanced with self-attention, is concatenated with acoustic features obtained using the openSMILE toolkit, and the resulting feature vector is then used for emotion classification.

itations, generative models, have emerged as a promising approach to learn and extract expressive representations directly from the raw speech signals [153]. Table 5 provides an overview of the literature on feature extraction and other applications in SER.

For example, Latif et al. [51] employ VAEs to learn latent representations of speech emotion and utilize Long Short-Term Memory (LSTM) networks for classifying phonological emotions. Zhang and Xue [154] propose a hybrid approach that combines the potential representations learned by an AE with acoustic features obtained using the openSMILE toolkit, as shown in Figure 4. The concatenated feature vectors are then used for emotion classification. Sahu et al. [155] leverage an Adversarial AE (AAE) framework to learn low-dimensional code vectors that retain class discriminability from the higher-dimensional feature space for SER. Ying et al. [156] employ Denoising AEs (DAEs) and AAEs in conjunction with unsupervised learning to extract features for model pre-training. Almotlak et al. [157] extend the

VAE framework to disentangle speaker-dependent information, such as identity and gender, from speaker-independent features like emotions, providing a more nuanced representation of speech data. Fonnegra and Díaz [158] use paralinguistic features and a deep convolutional Stacked AE (SAE) network to transform a set of 1582 INTERSPEECH 2010 features [159] extracted from speech signals into a higher-level representation for SER. Sahu et al. [160] use two methods to generate emotion-specific feature vectors with GAN variants: training a generator using samples from a mixture prior and providing a one-hot emotion vector to the generator to explicitly generate features for the specified emotion.

Table 5: Literature on Generative Models for Feature Extraction, Semi-supervised learning, Cross-domain, in SER.

Reference		Year	Model	Application	Dataset	Performance (%)
Latif et al. [51]		2017	VAE	Feature Extraction	IEMOCAP	ACC: 64.93
Zhang and Xue [154]		2021	AE	Feature Extraction	EmoDB/IEMOCAP	ACC: 95.6/71.2
Sahu et al. [155]		2018	AAE	Feature Extraction	IEMOCAP	UAR: 58.38
Ying et al. [156]		2021	DAE + AAE	Feature Extraction	IEMOCAP	ACC: 76.89
Almotlak et al. [157]		2020	VAE	Feature Extraction	OMG	ACC: 99.86
Fonnegra and Díaz [158]		2018	DCAE	Feature Extraction	eNTERFACE'05	ACC: 91.4
Sahu et al. [160]		2022	GAN + AE	Feature Extraction	IEMOCAP/MSP-IMPROV	ACC: 45.56
Zhao et al. [55]		2020	GAN	Semi-Supervised Learning	IEMOCAP	UAR: 58.7
Chang and Scherer [161]		2017	DCGAN	Semi-Supervised Learning	IEMOCAP	ACC: 49.80
Deng et al. [162]		2017	AE	Semi-Supervised Learning	FAU Aibo	UAR: 63.6
Xiao et al. [163]		2023	VAE	Semi-Supervised Learning	FAU Aibo	UAR: 42.9
Neumann and Vu [164]		2019	AE	Semi-Supervised Learning	IEMOCAP/MSP-IMPROV	UAR: 59.54/45.76
Latif et al. [34]		2020	AAE	Semi-Supervised Learning	IEMOCAP/MSP-IMPROV	ACC: 66.7/60.2
Zhou et al. [165]		2018	VAE	Semi-Supervised Learning	IEMOCAP	ACC: 76.7
Nasersharif et al. [35]		2023	AE	Cross-domain SER	EMODDB/EMOVO	ACC: 57.71
Xiao et al. [59]		2020	VAE	Cross-domain SER	IEMOCAP/MSP-IMPROV	UAR: 44.1/56.2
Das et al. [166]		2022	VAE	Cross-domain SER	IEMOCAP	ACC: 52.09
Latif et al. [167]		2022	ADDi	Cross-domain SER	EmoDB/IEMOCAP	UAR: 47.1/49.8
Su et al. [168]		2023	CycleGAN	Cross-domain SER	IEMOCAP/VAM/MSP-Podcast	ACC: 61.64
Su and Lee [169]		2021	CycleGAN	Cross-domain SER	IEMOCAP/MSP-IMPROV	ACC: 51.13/65.7

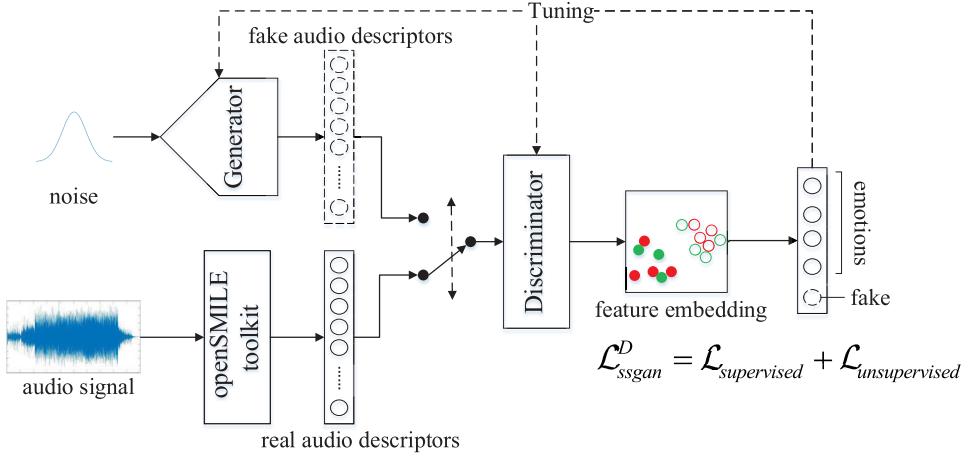


Figure 5: A semi-supervised learning framework in SER from [55]: A generator creates synthetic audio descriptors from noise. These descriptors, along with real ones from the openSMILE toolkit, are fed to a discriminator. The discriminator is trained with supervised and unsupervised loss functions to better distinguish real from fake audio cues.

### 5.3. Semi-Supervised Learning

Unlabeled data refers to samples that lack corresponding labels, which may be due to the failure to obtain appropriate labels during the data collection process or the inability to cover the entire dataset due to the high cost of manual labeling. In the field of emotion recognition, the presence of unlabeled data can hinder the performance of models in emotion classification tasks [170]. To address this issue, semi-supervised learning emerges as a promising approach that leverages both labeled and unlabeled data to improve the generalization and performance of emotion recognition models [55, 162]. Common semi-supervised learning methods include strategies based on label propagation, self-training, and generative models [53]. Among these approaches, generative models have proven to be particularly effective.

To be specific, Zhao et al. [55] propose a semi-supervised GAN for SER, which is designed to capture underlying knowledge from both labeled and unlabeled data, as shown in Figure 5. In their approach, a generator creates synthetic audio descriptors from noise, while a discriminator is trained to distinguish between real and fake audio cues using both supervised and unsupervised loss functions. The discriminator not only classifies input samples as real or fake but also learns to identify the emotional class of real samples. Chang and Scherer [161] present a multitask deep convolutional GAN (DCGAN) approach for emotional valence classification in speech, which leverages a multi-task annotated corpus and a large unlabeled conference corpus. In addition to GANs, AEs have also been widely used by researchers for semi-supervised learning in SER. For example, Deng et al. [162] propose a semi-supervised AE model for SER that combines both generative and discriminative perspectives, enabling the model to distill knowledge from unlabeled data and incorporate it into the supervised learning process, thus enhancing the overall performance of the SER system. Xiao et al. [163] propose a novel semi-supervised adversarial VAE that learns the distribution of shared hidden features of labeled and unlabeled data. Neumann and Vu [164] employ a recurrent sequence-to-sequence AE trained on unlabeled data to generate representations for the labeled data. These generated representations

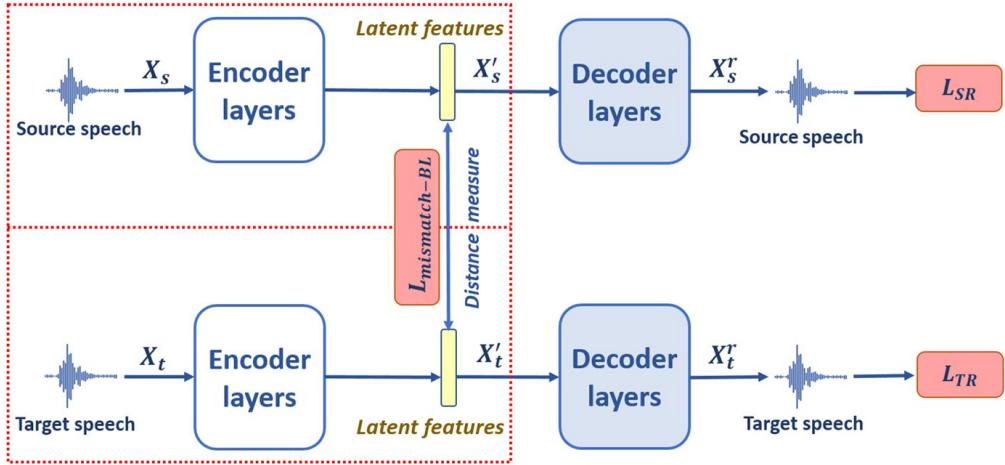


Figure 6: A framework [35] which leverages AEs to address the cross-domain problem in SER: The key idea is to utilize the latent representations learned by the AE to align and compare the features of the source and target speech domains.

serve as auxiliary information to the classifier during training, helping to improve the emotion recognition performance. Latif et al. [34] design an unsupervised AAE and combine it with a supervised classification network to achieve semi-supervised learning. Zhou et al. [165] extend the multi-path deep neural network to a generative model based on semi-supervised VAE, which enables simultaneous training on both labeled and unlabeled data.

#### 5.4. Cross-domain SER

Here, we discuss cross-domain SER, which refers to the ability of a model trained on one speech emotion domain (source domain) to be extended and applied to another different speech emotion domain (target domain). This cross-domain generalization is a major challenge faced by SER because different speech emotion domains may have significant differences, such as language, culture, speaking style, recording conditions, and so on [57, 58]. These differences form distribution inconsistencies between different domains, so that the performance of a model trained on one domain often decreases significantly on another domain. Generative models provide a new perspective for solving this problem. By introducing domain adaptation techniques, generative models can learn to map samples from the source dataset and target dataset to a shared feature space, thereby eliminating the distribution differences between them.

For example, Nasersharif et al. [35] employ separate AEs for each source and target domain dataset and introduce a domain adaptation loss in addition to the conventional loss to achieve domain-invariant feature extraction, as shown in Figure 6. Xiao et al. [59] design a framework by incorporating a VAE to extract generalized and domain-invariant representations from latent feature distributions. Das et al. [166] propose a VAE-based method with KL annealing and semi-supervised learning to achieve more consistent latent embedding distributions across datasets. Latif et al. [167] introduce an Adversarial Dual Discriminator (ADDi) network, which generates domain-invariant representations through a three-player adversarial game between a generator and two discriminators. Su et al. [168] introduce corpus-aware emotional CycleGAN, which incorporates a corpus-aware attention mechanism to

aggregate each source dataset and generate synthetic target samples. Su and Lee [169] propose a conditional cycle GAN to generate samples similar to the source domain and assign the original labels of the source samples.

### 5.5. Adversarial Sample Generation and Defense

Adversarial examples refer to samples that are created by adding carefully designed perturbations to the original samples, causing the model to misclassify them [171]. In the SER task, adversarial examples can be speech with specific noise or perturbations added, which can deceive the SER model and lead to incorrect emotion predictions [172]. The existence of adversarial examples highlights the vulnerability of SER models, as these models may overly rely on certain specific acoustic features while ignoring the overall features of emotional expression. Adversarial example defense refers to improving the ability of SER models to resist adversarial attacks, enabling them to maintain high emotion recognition accuracy even when faced with adversarial examples.

Generative models can play an important role in the generation and defense of adversarial examples in SER. On the one hand, generative models can be used to generate realistic adversarial examples. For example, Chang et al. [173] introduce STAANet, a novel method based on WaveNet [174] for generating sparse and transferable adversarial examples to spoof SER systems. Wang et al. [175] propose a GAN-based attack approach that can efficiently generate adversarial examples with pre-specified target labels. On the other hand, generative models can also be used to construct adversarial example defense systems. For example, Latif et al. [176] introduce a two-step defense strategy for SER systems. In their approach, perturbed utterances are first cleaned using GANs, and then a classifier is applied to the cleaned data. By removing the adversarial perturbations from the input samples, the classifier can make more accurate predictions and maintain robust performance. Chang et al. [177] propose a framework based on federated adversarial learning to protect data privacy and defend against adversarial attacks in SER systems. Their approach leverages the distributed nature of federated learning to train models on decentralized data while incorporating adversarial training techniques to enhance robustness against adversarial examples.

In summary, generative models have made significant contributions to SER by enabling data augmentation, feature extraction, semi-supervised learning, and adversarial sample generation and defense. These applications have greatly enhanced the performance of SER systems. As research in this area continues to advance, generative models are expected to play an increasingly crucial role in pushing the boundaries of SER.

## 6. Facial Emotion Recognition (FER)

FER is a research field that utilizes computer vision techniques to recognize and understand the emotional states of individuals by analyzing facial expressions in facial images or video frames [19, 178]. In recent years, generative models have made impressive progress in the field of FER, which is similar to SER, but presents some unique research focuses and development directions. This section will provide a detailed discussion on the applications of generative models in FER. Firstly, Section 6.1 will discuss facial data augmentation. Secondly, Section 6.2 will focus on facial

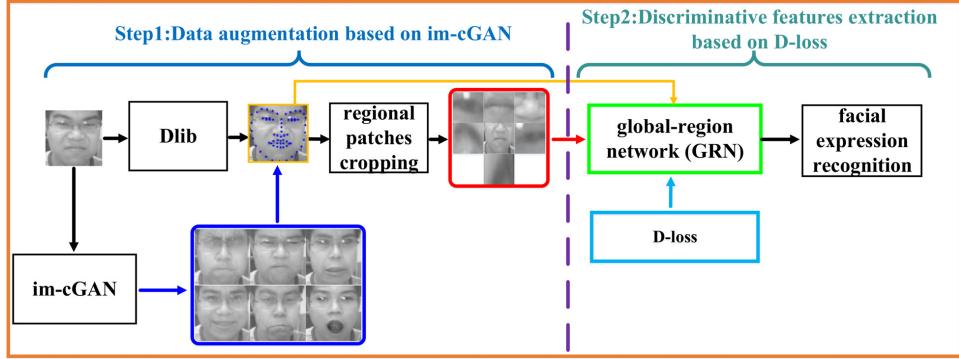


Figure 7: A facial data generation framework based on GANs from [180]: The original image undergoes facial landmarks detection using Dlib, after which the im-cGAN generates augmented images and regional patches. These patches are then cropped into specific regions. In the second step, discriminative features are extracted. The cropped regional patches are fed into a global-region network (GRN), which extracts discriminative features based on the D-loss function.

feature extraction. Next, Section 6.4 will introduce semi-supervised learning for FER. Furthermore, Section 6.5 will focus on FER in typical noisy environments. Finally, Section 6.3 will discuss dynamic FER.

### 6.1. Facial Data Augmentation

Similar to speech data, acquiring high-quality facial image data is an expensive and time-consuming task, which hinders the development of FER [95, 179]. With generative models, it becomes possible to generate synthetic facial images that align with the desired emotion labels, thus providing a data augmentation method to overcome the scarcity of labeled data. Table 6 shows the literature on generative models for data augmentation in FER.

Specifically, a series of GAN-based methods are proposed. Porcu et al. [46] reveal that GAN-based data augmentation substantially outperforms geometric transformations in improving emotion recognition accuracy of a VGG16 CNN-based FER system [181], highlighting the significance of large training datasets and the potential of combining GAN with less complex techniques to mitigate computational complexity. Zhu et al. [182] propose a data augmentation method using a framework combining a CNN classifier and a CycleGAN generator with least-squared adversarial loss to complement and complete the data manifold, find better margins between neighboring classes, and avoid the gradient vanishing problem. As shown in Figure 7, Sun et al. [180] propose an improved CGAN (im-cGAN) model trained on facial images with Action Units (AUs) to generate more labeled samples for data augmentation. Similarly, Wang et al. [183] introduce an end-to-end synthetic Compositional GAN (Comp-GAN) that employs six task-driven loss functions to generate more realistic and natural facial images. Kusunose et al. [184] propose a data augmentation method using StyleGAN2 [185] to generate artificial facial expression images for seven emotions, which are used as additional training data to train a VGG16-based emotion recognition model through transfer learning. Yang et al. [186] introduce an end-to-end deep learning framework for generating high-quality facial images. The framework is built upon DCGAN and incorporates an Ensemble Networks (Ens-Net) model for network integration. Han et al. [187] improve upon the image conversion model StarGAN V2 [188] by integrating the

Table 6: Literature on Generative Models for Data Augmentation in FER.

Reference	Year	Model	Dataset	Performance (%)
Porcu et al. [46]	2022	GAN	CK+	ACC: 83.3
Zhu et al. [182]	2018	CycleGAN	FER2013	ACC: 94.65
Sun et al. [180]	2023	CGAN	JAFFE/CK+/Oulu-CASIA/KDEF	ACC: 98.37/98.10/93.34/98.30
Wang et al. [183]	2019	GAN	F2ED	ACC: 44.19
Kusunose et al. [184]	2022	StyleGAN2	CFEE	ACC: 82.04
Yang et al. [186]	2022	GAN	FER2013/CK+/JAFFE	ACC: 82.1/84.8/91.5
Han et al. [187]	2023	starGAN	CK+/MMI	ACC: 99.20/98.14
Wang et al. [191]	2020	starGAN	KDEF/MMI	ACC: 95.97/98.30
Li et al. [47]	2023	GAN + VAE	RAF-DB	ACC: 74.43
Pons et al. [192]	2020	CycleGAN	USTC-NVIE	ACC: 51

Squeeze-and-Excitation Network (SENet) [189] into the generator. They also introduce hinge losses [190] to enhance the authenticity of the generated images. These modifications aim to produce more realistic and diverse facial expressions across multiple domains. Wang et al. [191] propose an improvement to the StarGAN model for data augmentation in FER by introducing context loss and an attention mechanism. These modifications enable the model to generate more realistic and expressive facial images with fine details in crucial regions, such as eyes and mouth.

In addition to GAN-based methods, there are also some approaches that combine AEs for data augmentation in FER. For example, Li et al. [47] introduce a novel approach that integrates GAN and VAE in the latent space for data augmentation in FER. The proposed method aims to generate diverse facial images that carry rich semantic information, enhancing the quality and variety of the augmented dataset. Pons et al. [192] propose using CycleGAN to generate thermal facial images from visual spectrum images, thereby augmenting the training dataset for thermal emotion recognition. This can be seen as a cross-modality data augmentation approach.

### 6.2. Facial Feature Extraction

Traditional FER methods often rely on handcrafted features or predefined facial landmarks to describe facial expressions [19, 178]. These features typically include the locations of facial key points, geometric relationships, and local texture descriptors [193]. However, due to the complexity and diversity of facial expressions, these handcrafted features may not adequately capture the subtle differences between different expressions [194]. Moreover, the variations in facial features among different individuals also pose challenges to feature design [195, 196]. In recent years, with the development of deep learning, generative models provide new ideas for solving the feature extraction problem in FER by automatically learning compact and informative representations of facial expressions. Table 7 shows a series of research works, where CCC (Concordance Correlation Coefficient) assesses the agreement between predicted and actual values, taking into account both the correlation and the deviation from perfect agreement.

For example, Yang et al. [52] propose the Exchange-GAN model, which achieves the separation of expression-related and expression-irrelevant features through partial feature exchange and various constraints, such as adversarial loss, classification loss, content loss, and central loss. As illustrated in Figure 8, this approach effectively disentangles the facial expression information from identity-related features. Khemakhem and Ltifi [197] introduce a neural style transfer GAN (NST-GAN) to remove identity information from facial images. By converting each image into a

synthetic “average” identity, NST-GAN focuses on extracting expression-specific features while minimizing the influence of individual variations. Analogously, Yang et al. [198] propose a de-expression residue learning method by training a CGAN to generate neutral face images from input expressions, and then learn the expressive information residue from the intermediate layers of the generative model. Zhang and Tang [199] use a GAN to generate a neutral face from an emotional face, extract features from both faces using separate convolutional layers, and then subtract the neutral features from the emotional features to obtain pure “emotion features” for FER. Xie et al. [200] propose the two-branch divergent GAN, which consists of two generators and discriminators. One branch is dedicated to facial expression feature extraction and classification, while the other focuses on facial feature extraction and identity discrimination. This architecture allows for a more targeted approach to capturing expression-related information. Ali and Hughes [201] propose a novel disentangled expression learning GAN that decouples expression representation from identity components. By explicitly providing identity codes to the decoder, this approach enables the model to learn expression features independently of subject-specific characteristics. Similarly, Tiwary et al. [202] design an expression GAN to separate shape, skin color, and other identity-related information from specific muscle movements that are crucial for emotion recognition. Sima et al. [203] tackle the challenges of category overlap and facial expression variation due to individual differences using CGAN. By generating neutral expressions while retaining identity-related information, their method aims to normalize the facial images and facilitate more accurate expression recognition. Similar GAN-based approaches for addressing facial feature extraction include [204, 205, 206, 207].

In addition to the aforementioned methods based on GANs, several AE-based approaches are proposed to address the feature extraction problem in FER. For example, Kim et al. [208] design an AE network with CNNs to learn a contrastive representation of facial expressions by comparing a query image with a generative reference image, estimated from the given image itself. Wu et al. [209] propose Cross-VAE, an extension of conditional VAE that disentangles expression from identity by assuming orthogonal latent representations and employing a symmetric training procedure, ensuring disentangled and expressive encodings. Chatterjee et al. [210] propose a residual VAE model to significantly reduce class overlapping in the feature space by transforming facial images into latent vector. They [211] also apply various resampling techniques after feature extraction to solve the class imbalance problem. Zhou et al. [212] use Masked Autoencoders (MAEs) for model pre-training and improve the ability to extract facial features through masked-reconstruction method, as shown in Figure 8.

### 6.3. *Cross-domain FER*

Similar to cross-dataset SER, generative models can also provide effective solutions for cross-dataset FER through domain adaptation. For instance, Wang et al. [213] propose an unsupervised domain adaptation method based on GANs. Their approach generates new samples to fine-tune the pre-trained FER model, enabling it to better adapt to the target domain and achieve higher accuracy across different datasets. Fan et al. [214] employ CGANs to learn domain-invariant feature representations of facial expressions, allowing for effective knowledge transfer from

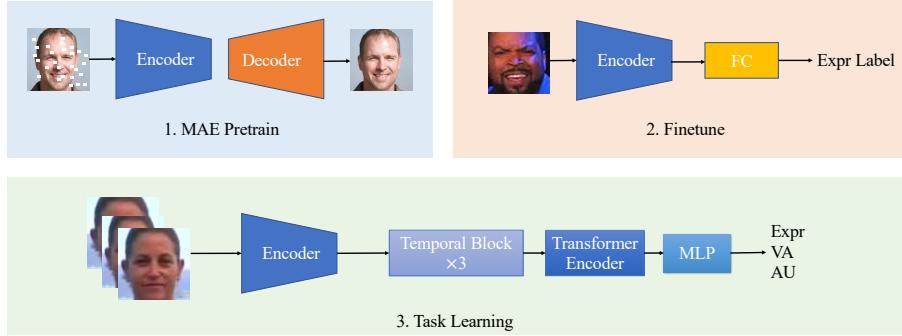


Figure 8: A Masked Autoencoder (MAE)-based facial feature extraction framework from [212]: It consists of a pre-trained ViT model using MAE on facial expression datasets, followed by fine-tuning the encoder with a fully connected layer and expression labels, and finally connecting the fine-tuned encoder with temporal blocks, a Transformer encoder, and a MLP for sequence task learning.

Table 7: Literature on Generative Models for Feature Extraction in FER.

Reference	Year	Model	Dataset	Performance (%)
Yang et al. [52]	2020	GAN	Multi-PIE/FACES/Oulu-CASIA	ACC: 91.08/95.24/86.33
Khemakhem and Ltifi [197]	2023	GAN	CK+/JAFFE/FER 2013	ACC: 98.14/95.12/85.93
Yang et al. [198]	2018	CGAN	CK+/MMI/BU-3DFE/BP4D+	ACC: 97.30/88.0/73.23/84.17
Zhang and Tang [199]	2021	GAN	CK+	ACC: 98.04
Xie et al. [200]	2021	GAN	CK+/TFEID	ACC: 97.53/97.20
Ali and Hughes [201]	2019	GAN	CK+/Oulu-CASIA /MMI	ACC: 97.28/89.17/72.97
Tiwary et al. [202]	2023	GAN	CK+/Oulu-CASIA /FER 2013	ACC: 94.67/92.57/92.85
Sima et al. [203]	2021	CGAN	CK/MMI	ACC: 93.52/93.60
Yang et al. [204]	2018	CGAN	CK+/Oulu-CASIA/BU-3DFE/BU-4DFE	ACC: 96.57/88.92/76.83/89.55
Wang [205]	2021	GAN	JAFFE/CK+/FER 2013	ACC: 96.6/95.6/72.8
Abiram et al. [206]	2021	StyleGAN	CK+	ACC: 96.97
Dharanya et al. [207]	2021	ACGAN	ADFES-BIV/CK+/KDEF	ACC: 93.6/97.39/96.33
Kim et al. [208]	2017	AE	CK+/MMI/Oulu-CASIA	ACC: 97.93/81.53/86.11
Wu et al. [209]	2020	VAE	CK+/Oulu-CASIA/RAF-DB	ACC: 94.96/86.87/84.81
Chatterjee et al. [210]	2022	VAE	Affectnet	ACC: 95
Zhou et al. [212]	2024	MAE	Affwild	CCC: 60.57

labeled source domains to unlabeled target domains. Zhang et al. [215] propose a domain adaptation model for FER, which utilizes unlabeled web facial images as an auxiliary domain to generate labeled facial images in the target domain using GANs with an attention transfer module, preserving structural consistency through pixel cycle-consistency and discriminative loss, and leveraging attention maps from the source domain classifier to enhance the target domain classifier’s performance.

#### 6.4. FER in Noisy Environments

In real-world applications, facial images are often affected by various noisy environments, such as occlusions, complex backgrounds, non-frontal views, and other factors that reduce the performance of FER [178]. In recent years, generative models demonstrate powerful capabilities in effectively addressing these noisy environments. Table 8 presents the literature on generative models for FER in noisy environments.

For example, to cope with occlusion and complex backgrounds, Du et al. [216] introduce an expression associative network that learns the associations between inter- and intra-class expressions, along with a GAN-based auxiliary module designed to suppress changes caused by occlusion, illumination, and pose variations. Peng et al. [217] propose a method based on ACGAN [218] to restore occluded facial images, thereby enhancing the data manifold and improving FER performance. Similarly,

Lu et al. [219] employ a Wasserstein GAN-based method to generate unoccluded facial images, enabling effective FER even in the presence of partial occlusion. To mitigate the influence of background clutter on FER, Tang et al. [220] introduce an expression CGAN that incorporates a novel mask loss. This loss function helps to reduce the impact of background information, allowing the model to focus more on the facial regions relevant to expression recognition. Building upon the idea of masking, Li et al. [221] propose a mask generation network (MGN) inspired by GANs. The MGN effectively filters out background noise and interference from facial images by generating masks that cover only the key areas of each face required for accurate expression classification. This targeted masking approach preserves the most informative regions while suppressing irrelevant information.

Apart from occlusion and complex backgrounds, the presence of non-frontal facial images in wild datasets poses another significant challenge for FER due to variations in facial posture. To address this issue, researchers have proposed different solutions based on generative models, aiming to generate high-quality frontal-face images from non-frontal views. For example, Han and He [222] introduce a GAN-based three-stage training algorithm for multi-view FER. They utilize a pre-trained classification model to construct a new GAN that effectively synthesizes frontal faces while preserving facial expression features. This approach ensures that the generated frontal images retain the emotional information necessary for accurate FER. Instead of explicitly generating frontal faces, Zhang et al. [223] propose a GAN-based method that leverages the geometric information inherent in facial shapes to achieve pose-invariant FER. This innovative perspective offers an alternative solution to the challenge of non-frontal images in wild datasets. Similar GAN-based methods for non-frontal FER also include [224, 225, 226].

Moreover, generative models can be employed to combat other noise factors that degrade the performance of FER. For example, Yang et al. [227] propose a GAN-based method for recognizing expressions in low-intensity facial images. The concatenated training critic combines both the general adversarial loss, which regulates the intensity of facial expressions, and the FER penalty, which directs the model to generate visually enhanced facial images. Nan et al. [228] propose a novel GAN-based feature-level super-resolution method for robust FER that reduces the risk of privacy leakage. This approach transforms low-resolution image features into more discriminative ones using a pre-trained FER model as a feature extractor. To address the issue of image degradation caused by under-display cameras (UDC), Wang et al. [229] employ a method based on DM to overcome inherent noise and distortion obstacles, thereby improving the accuracy of expression recognition, as shown in Figure 9.

### 6.5. Dynamic FER

The previously discussed FER mainly focuses on static images, i.e., single-frame facial expression recognition. However, in real-world scenarios, emotions often change dynamically, requiring consideration of the temporal evolution of facial expressions. Dynamic FER aims to analyze the changes in facial expressions across video sequences, capturing the temporal dynamic features of emotions [230]. Compared to static FER, dynamic FER can provide richer and more accurate emotional information, but it also presents greater challenges due to the complexity of temporal dynamics and the need for robust algorithms that can handle variations in facial

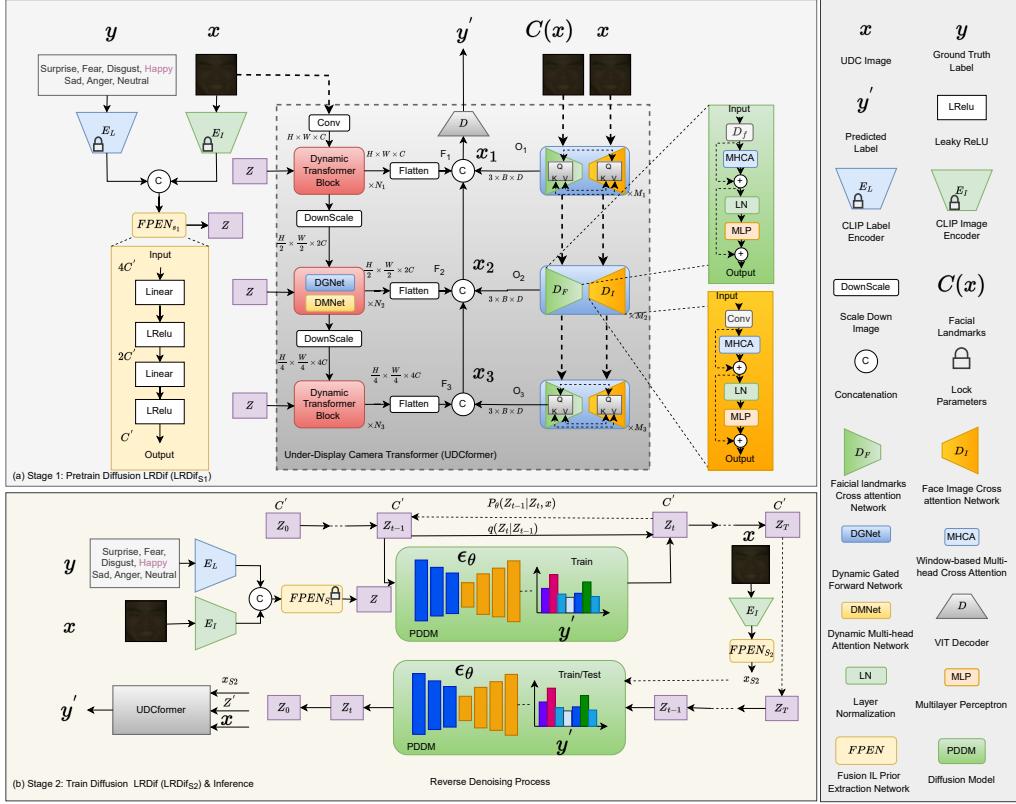


Figure 9: A DM-based framework for FER in noisy environments proposed by [229]: It overcomes image degradation by under-display cameras (UDC) through a two-stage training strategy involving a preliminary extraction network (FPEN) and a transformer network (UDCformer), enabling effective recovery of emotion labels from degraded UDC images.

movements, head poses, and lighting conditions across frames [231, 232, 233]. Table 9 shows the literature on generative models for dynamic FER, where UF1 stands for unweighted F1 score that combines precision and recall, providing a balanced measure of a model’s performance.

For example, Cai et al. [234] introduce a Masked Autoencoder (MAE) for learning robust and generic facial embeddings, as depicted in Figure 10. The MAE reconstructs spatio-temporal details of the face from densely masked facial regions, capturing both local and global aspects to encode transferable features. Priyanka A. Gavade et al. [235] propose the Taylor-Chicken Swarm Optimization-based Deep GAN (Taylor-CSO-based Deep GAN), which provides a new method for recognizing video expressions. The Deep GAN, trained by the Taylor CSO approach, is adopted to efficiently complete the recognition process on the basis of the feature matrix. Guo et al. [236] propose a novel approach that leverages GANs to generate inter-class optical flow images for FER. They calculate the difference between static fully expressive samples and neutral expression samples to obtain the optical flow information. By training the GAN on these optical flow images, the model learns to capture the subtle motion patterns associated with different facial expressions. Similarly, Liang et al. [237] propose an improved FER system that utilizes GANs to generate inter-class optical flow images by computing variations of optical flow. They utilize a self-attention mechanism within the GAN architecture, which enables the generator to integrate image information from all feature locations when generating image details. Other similar works based on GANs for dynamic FER include

Table 8: Literature on Generative Models for FER in Noisy Environments.

Type	Reference	Year	Model	Dataset	Performance (%)
Occlusion and Complex Backgrounds	Du et al. [216]	2021	GAN	RAF-DB/FER2013Plus	ACC: 87.03/90.07
	Peng et al. [217]	2020	ACGAN	FER2013	ACC: 73
	Lu et al. [219]	2019	Wasserstein GAN	RAF-DB/AffectNet	ACC: 83.49/59.73
	Tang et al. [220]	2019	CGAN	JAFFE	ACC: 80.32
	Li et al. [221]	2021	GAN	RAF-DB/FER2013Plus	ACC: 87.91/88.88
Non-frontal Views	Han and He [222]	2021	GAN	KDEF/Multi-PIE	ACC: 90.42/92.63
	Zhang et al. [223]	2020	GAN	Multi-PIE/BU-3DFE	ACC: 92.09/81.95
	Lai and Lai [224]	2018	GAN	Multi-PIE	ACC: 86.76
	Li et al. [225]	2019	GAN	Multi-PIE	ACC: 86.9
	Dong et al. [226]	2023	GAN	KDEF	ACC: 92.7
Other Factors	Yang et al. [227]	2021	GAN	BU-3DFE/BU-4DFE	ACC: 88.89/80.03
	Nan et al. [228]	2022	GAN	RAF-DB/SFEW2.0	ACC: 84.09/55.14
	Wang et al. [229]	2024	DM	UDC-RAF-DB/UDC-FERPlus	ACC: 88.55/84.89

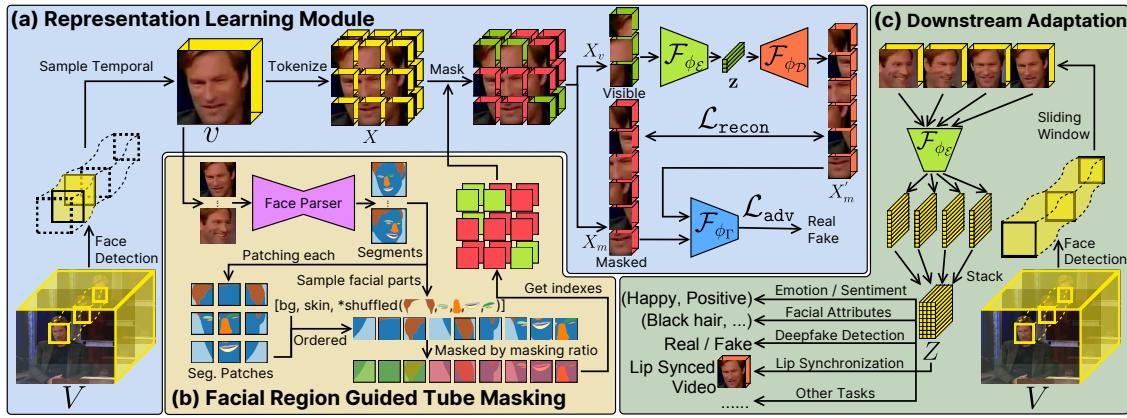


Figure 10: A MAE framework for dynamic FER from [234]: It begins by detecting faces in each video frame, followed by a temporal sampling and tokenization process that divides the detected faces into segments. These segments are then fed into a face parser, which identifies and segments various facial features, such as eyes, nose, and mouth. The segmented features are patched together to create a complete representation of the face. To enhance the learning process, the framework employs a selective masking strategy. It randomly masks some of the patches, creating a mix of visible and masked segments.

[238, 239].

In summary, generative models show tremendous advantages in the field of FER. First, generative models can effectively augment training data by generating realistic and diverse new facial images. Second, generative models demonstrate unique advantages in facial feature extraction. Moreover, generative models provide effective solutions for cross-domain FER. When dealing with noisy environments, generative models also exhibit promising performance. Finally, the effectiveness of generative models in dynamic FER is fully validated. In addition, it is worth noting that apart from these aspects, generative models can also play a role in semi-supervised scenarios for FER, although there are relatively fewer related works. For example, Chen et al. [56] propose an improved classification method based on semi-supervised GANs. They replace the output layer of the traditional unsupervised GAN with a Softmax layer, enabling the model to simultaneously learn from labeled and unlabeled data.

Table 9: Literature on Generative Models for Dynamic FER.

Reference	Year	Model	Dataset	Performance (%)
Cai et al. [234]	2023	MAE	CMU-MOSEI	ACC: 80.60
Gavade et al. [235]	2023	DGAN	CK+/RAVDESS/SAVEE	ACC: 82.14/80/79
Guo et al. [236]	2023	GAN	SAMM/AFEW	ACC: 56.98/60.20/44.72
Liong et al. [237]	2020	GAN	CASME II/SMIC/SAMM	ACC: 70.29
Li et al. [238]	2021	GAN	CASME II/SAMM/SMIC	UF1: 84.52, UAR: 84.65
Mazen et al. [239]	2021	CycleGAN	FER2013	ACC: 91.76

## 7. Textual Emotion Recognition (TER)

TER is a highly focused and challenging research direction in the field of natural language processing [240]. It is dedicated to automatically identifying and extracting the emotional information contained in textual data, providing insights into the emotional states and subjective attitudes conveyed by humans in their written expressions [241, 242]. The core task of TER is to identify the specific emotion categories expressed in the text, such as happiness, sadness, anger, surprise, etc. [241]. Compared to textual sentiment analysis [243, 244, 245, 246] (e.g., positive, negative, neutral), TER offers a more fine-grained characterization of emotional semantics, capturing richer and more subtle emotional expressions in the text [20, 247]. Table 10 illustrates the main differences between TER and textual sentiment analysis. In recent years, generative models have shown great potential in TER tasks. By learning the intrinsic patterns and distributions of textual data, generative models, especially LLMs or their predecessors, can generate emotionally expressive text, providing new perspectives and methods for emotion recognition [248]. Based on this, this section reviews a series of works of generative models in TER tasks.

Table 10: Difference between Emotion Recognition and Sentiment Analysis.

	Textual Emotion Recognition (TER)	Textual Sentiment Analysis
Type	Specific Emotion	Sentiment Polarity
Typical Categories	Happiness, Sadness, Anger, Fear, Surprise, Disgust	Positive, Neutral, Negative

For instance, in terms of data augmentation, Nedliko [48] utilizes ChatGPT to augment the dataset by translating the data into different languages for the task of TER. Similarly, Koptyra et al. [49] investigate the feasibility of employing ChatGPT for automatic emotional text generation and annotation in the context of TER. Pico et al. [249] study the application of text-generating LLMs for TER with the aim of generating more emotional knowledge, in the form of beliefs, that can be utilized by an emotional agent.

Furthermore, in terms of feature extraction, Ghosal et al. [250] leverage COMET [251], a GPT-based model, to extract emotional textual features, such as mental states, events, and causal relations, for the task of TER. These features are then incorporated into their proposed framework, COSMIC, which aims to enhance the understanding of emotional dynamics in conversations by capturing the underlying commonsense knowledge that influences the emotional states of the interlocutors. Hama et al. [252] demonstrate that the LLMs can achieve excellent performance for emotion labels with a limited number of training samples. Additionally, they introduce a multi-step prompting technique to further enhance the discriminative capacity of the LLMs for different emotion labels. InstructERC [253] reformulates

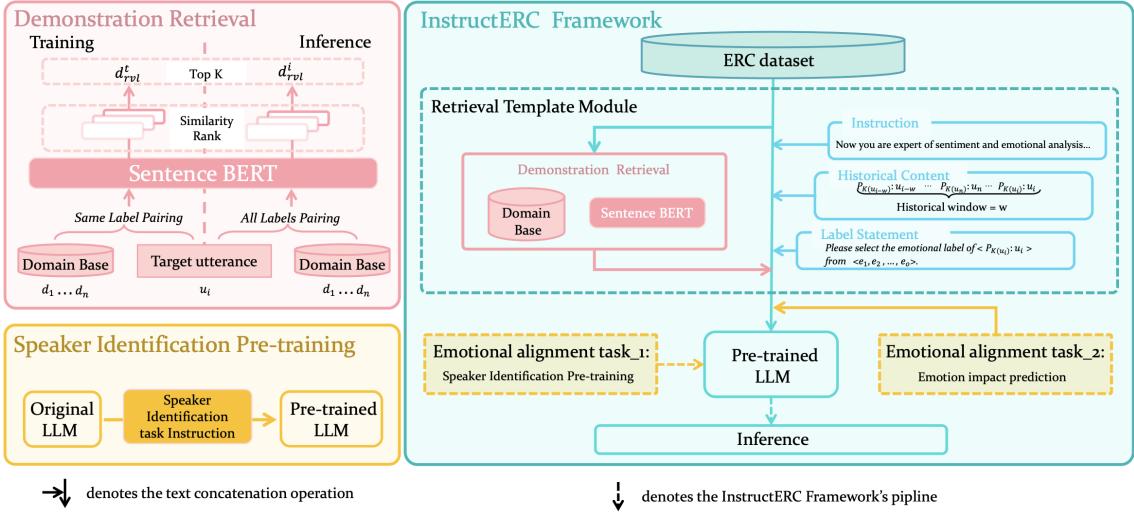


Figure 11: A LLM-based framework for TER proposed by [253]: It introduces an effective retrieval template module that explicitly integrates multi-granularity information. Moreover, it develops two additional emotion alignment tasks, which implicitly model the role relationships and future emotional tendencies.

the TER task by transitioning from a discriminative framework to a generative framework based on LLMs, as illustrated in Figure 11. CKERC [254] is a joint LLMs with commonsense knowledge framework for TER, which utilizes common-sense knowledge for LLM pre-training to finetune implicit clues information.

## 8. Physiological Signal-based Emotion Recognition

In addition to the emotion recognition based on speech, facial images, and text mentioned above, physiological signals have garnered significant attention in the field of emotion recognition because they directly reflect an individual’s bodily changes during emotional experiences. These changes are usually not influenced by an individual’s subjective consciousness and can therefore provide an objective reflection of emotional states [21, 255]. The physiological signals used in emotion recognition include EEG, ECG, HRV, EDA, and others, each with its own unique characteristics. For instance, EEG signals have the advantages of high temporal resolution and directly reflecting neural activity [256], enabling them to capture subtle changes in brain activity. As a result, EEG is widely used in emotion recognition research [257, 258]. On the other hand, ECG signals reflect changes in cardiac activity and are closely related to emotional states [259, 260]. HRV signals indicate the variability of heart rate, which can reflect the activity of the autonomic nervous system [261]. EDA signals reflect changes in skin conductance and are associated with emotional arousal [262]. These signals provide valuable insights into the physiological changes associated with different emotional experiences [255]. In recent years, generative models have shown great advantages in physiological signal-based emotion recognition [263], similar to their applications in the fields of SER, FER, and TER. These models have demonstrated their potential in various aspects, such as data augmentation, feature extraction, and cross-domain, providing new impetus for the development of this field. Table 11 presents the literature on generative models for physiological signal-based emotion recognition.

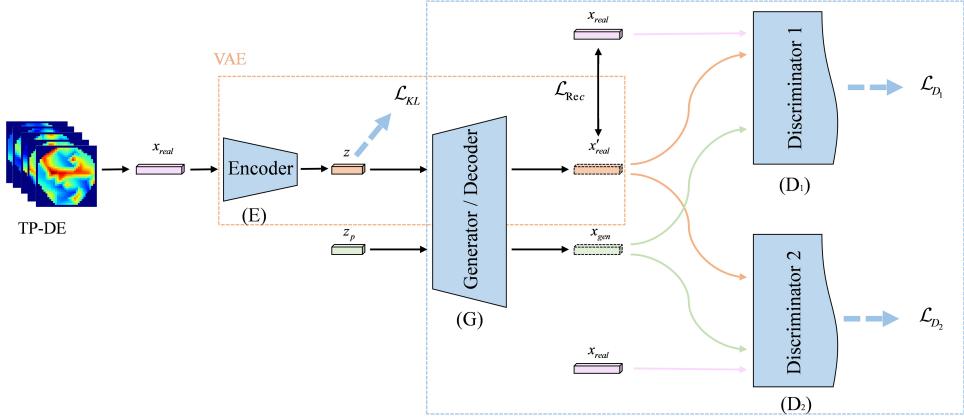


Figure 12: The framework of VAE-D2GAN model for data augmentation in physiological signal-based emotion recognition from [265]: It consists of four parts: an encoder, a generator, and two discriminators. The encoder and generator form the VAE. The encoder maps real samples to latent vectors, which the generator then uses to create artificial samples. The generator and the two discriminators form the GAN. The generator learns the real data distribution from the gradients provided by the discriminators, which distinguish real samples from generated ones.

Data augmentation is the primary application of generative models in physiological signal-based emotion recognition. Researchers conduct a series of studies in this direction. For example, Luo et al. [264] propose a Conditional Wasserstein GAN (CWGAN) framework for EEG data augmentation in emotion recognition, which generates high-quality realistic-like EEG data in differential entropy (DE) form, judged by three indicators, to supplement the data manifold and improve emotion classification performance. As shown in Figure 12, Bao et al. [265] propose VAE-D2GAN, a data augmentation model for EEG emotion recognition, which combines a VAE model with a dual discriminator GAN to learn the distributions of DE features under five classical frequency bands and generate diverse artificial EEG samples for training. Bhat and Hortal [266] present a model based on Wasserstein GAN with gradient penalty, capable of generating new data features to augment the original dataset. Zhang et al. [267] propose a GAN-based framework with self-supervised learning to generate EEG samples by learning an EEG generator, masking parts of EEG signals to synthesize potential signals, and utilizing masking possibility as prior knowledge to extract distinguishable features and generalize the classifier. Zhang et al. [268] propose an emotional subspace constrained GAN model that addresses the limitations of existing GANs in generating reliable augmentations for under-represented minority emotions in imbalanced EEG datasets by introducing an EEG editing paradigm and incorporating diversity-aware and boundary-aware losses to constrain the augmented emotional subspace. Similar works on EEG data augmentation based on GANs also include [269, 270, 271, 272]. Apart from GAN-based models, other generative approaches have also been explored. Siddhad et al. [273] employ a conditional DM to generate raw synthetic EEG data, while Tosato et al. [274] and Yi et al. [275] also utilize DMs for EEG data generation.

In addition to data augmentation, generative models can be employed for feature extraction in physiological signal-based emotion recognition, among which AE-based models are widely used. For example, Lan et al. [276] use an AE model to learn subject-specific salient frequency components from EEG signals' power spectral density, automatically identifying the most discriminative features without predefined

frequency band ranges. Rajpoot et al. [277] employ an unsupervised LSTM with channel-attention AEs to obtain subject-invariant latent vector subspace representations and a CNN with attention framework to perform EEG emotion recognition on the encoded latent space representations. Similar AE-based methods also include [278, 279, 280, 281, 282, 283]. In addition, some researchers attempt to combine GAN-based approaches. For example, Gu et al. [284] integrates generative adversarial learning into a hybrid model of Graph Convolutional Neural Network (GCNN) and LSTM, which utilizes a GAN to generate latent representations of EEG signals, aiming to extract more discriminative features and improve classification performance.

To effectively utilize both unlabeled and labeled data, semi-supervised learning approaches are explored. Notably, Zhang et al. [285] introduced a semi-supervised framework for EEG emotion recognition using an attention-based recurrent AE model. Their approach consists of two components: an unsupervised component that maximizes the consistency between the original and reconstructed input data, and a supervised component that minimizes the cross-entropy between the input and output labels. To address the cross-domain problem, researchers propose a series of methods. For example, Chai et al. [286] introduce a subspace-aligned AE framework to obtain a universal representation across training and testing domains for EEG emotion recognition. Wang et al. [287] propose a multimodal VAE to utilize the multi-modal data (EEG and eye movement data) from the source domain to assist in EEG emotion recognition in the target domain. Huang et al. [288] propose a GAN-based domain adaptation method with a knowledge-free mechanism, which converts the source domain in the data space into the target domain through a novel EEG content loss function.

Table 11: Literature on Generative Models for Physiological Signal-based Emotion Recognition.

Reference	Year	Model	Application	Modality	Dataset	Performance (%)
Luo et al. [264]	2018	CWGAN	Data Augmentation	EEG	SEED	ACC: 86.96
Bao et al. [265]	2021	VAE + GAN	Data Augmentation	EEG	SEED/SEED-IV	ACC: 92.5/82.3
Bhat and Hortal [266]	2021	Wasserstein GAN	Data Augmentation	EEG	DEAP	ACC: 71.50
Zhang et al. [267]	2022	GAN	Data Augmentation	EEG	DEAP/SEED/DREAMER	ACC: 94.38/97.71/85.28
Zhang et al. [268]	2024	GAN	Data Augmentation	EEG	DEAP/SEED	ACC: 96.68/97.14
Haradai et al. [269]	2018	GAN	Data Augmentation	EEG	ECG 200	ACC: 76.20
Pan and Zheng [270]	2021	GAN	Data Augmentation	EEG	DEAP	ACC: 90.26
Luo et al. [271]	2020	GAN	Data Augmentation	EEG	SEED/DEAP	ACC: 67.7/50.8
Kalashami et al. [272]	2022	CWGAN	Data Augmentation	EEG	DEAP	ACC: 71.9
Siddhah et al. [273]	2024	DM	Data Augmentation	EEG	DEAP	ACC: 71.50
Tosato et al. [274]	2023	DM	Data Augmentation	EEG	SEED	ACC: 92.98
Lan et al. [276]	2017	AE	Feature Extraction	EEG	SEED	ACC: 59.03
Rajpoot et al. [277]	2022	AE	Feature Extraction	EEG	DEAP/SEED/CHB-MIT	ACC: 69.5/76.7/72.3
Jiarayucharoensak et al. [278]	2014	SAE	Feature Extraction	EEG	DEAP	ACC: 49.52
Li et al. [279]	2019	VAE	Feature Extraction	EEG	DEAP/SEED	ACC: 72.43/84.29
Bethge et al. [280]	2022	CAE	Feature Extraction	EEG	SEED	ACC: 69
Liu et al. [281]	2020	SAE	Feature Extraction	EEG	DEAP/SEED	ACC: 92.86/96.77
Qing et al. [282]	2019	SAE	Feature Extraction	EEG	DEAP/SEED	ACC: 63.09/75
Li et al. [283]	2020	VAE	Feature Extraction	EEG	DEAP/SEED	ACC: 72.43/84.29
Gu et al. [284]	2023	GAN	Feature Extraction	EEG	DEAP/SEED	ACC: 94.78/97.28
Zhang et al. [285]	2021	AE	Semi-supervised Learning	EEG	SEED	ACC: 91.17
Chai et al. [286]	2016	AE	Cross-domain	EEG	SEED	ACC: 62.50
Wang et al. [287]	2022	VAE	Cross-domain	EEG	SEED/SEED-IV	ACC: 89.64/73.82
Huang et al. [288]	2022	AE	Cross-domain	EEG	DEAP	ACC: 63.85

## 9. Multimodal Emotion Recognition (MER)

Unlike unimodal emotion recognition mentioned in Sections 5 - 11 above, Multimodal Emotion Recognition (MER) is a research field that aims to identify emotional states by combining the information from different modalities, including speech, facial images, text, EEG, eye movements, and functional magnetic resonance imaging (fMRI), and so on [3, 16, 289, 290]. By integrating these heterogeneous data, MER can more comprehensively capture and analyze subtle differences in emotional expressions, thereby significantly improving the performance of emotion recognition. In recent years, generative models have been increasingly applied in the field of MER, showing potential in various aspects such as data augmentation, feature extraction, semi-supervised learning, and modality completion. Table 12 summarizes a series of representative works on generative models for MER.

Similar to unimodal emotion recognition, the primary application of generative models in MER is also data augmentation. For example, Ma et al. [140] design a multimodal CGAN based on Hirschfeld–Gebelein–Rényi (HGR) maximal correlation [291, 292, 293] to generate audio-visual data with different emotion categories. Luo et al. [294] implement a conditional boundary Equilibrium GAN [295] to generate artificial differential entropy features for EEG data, eye movement data, and their direct connections. Yan et al. [296] propose a CGAN-based framework that uses eye movement signals as the sole input to map information onto EEG-eye movement features. Chao et al. [297] employ a GAN framework with a cyclic loss to learn the bi-directional generative relationship between acoustic features and fMRI signals, enabling the derivation of fMRI-enriched acoustic features for emotion classification.

In addition to data augmentation, feature extraction is an important aspect of MER. For example, Yan et al. [298] propose a method that uses a bimodal deep AE and a fuzzy integral-based technique [299] to extract high-level fused features from EEG signals and eye movements, which are then input into a SVM classifier framework to obtain an emotion classification model. Similar works based on the bimodal deep AE also include [300, 301, 302, 303]. Nguyen et al. [304] propose a MER model that combines a two-stream AE with a LSTM network to perform end-to-end trainable joint learning of temporal and compact-representative features from visual and audio signals. Ma et al. [305] propose a deep framework with CNN and LSTM for MER that learns feature mappings from multimodal data by introducing a correlation loss based on HGR maximal correlation and a reconstruction loss of AEs to capture common and private information among different modalities, respectively. Zheng et al. [306] propose a novel Multi-channel Weight-sharing AE model with a convolutional encoder network for improved feature extraction and a scalable feature fusion method based on multi-head attention mechanism (CMHA) to fuse the output features of multiple encoders for MER.

Semi-supervised learning is another aspect where generative models are considered in MER. For example, Du et al. [308] propose a semi-supervised MER framework based on a multi-view VAE that imposes a mixture of Gaussians assumption on the posterior approximation of latent variables. This framework can leverage both labeled and unlabeled samples from multiple modalities and automatically learn the weight factor for each modality. Liang et al. [309] propose a semi-supervised learning approach for MER based on an improved GAN, CT-GAN [310], which utilizes unlabeled data from acoustic and visual modalities.

In addition to the aforementioned aspects, another direction that requires atten-

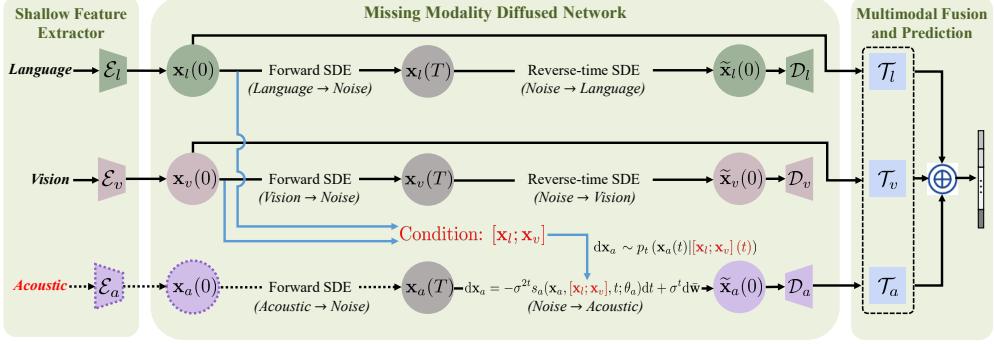


Figure 13: The IMDer framework for modality completion in MER from [307]: It handles incomplete data (e.g., missing acoustic modality) by extracting shallow features from the available modalities. Then the diffused network samples the missing modality data from a noise distribution and then recovers it through a reverse diffusion process. Finally, the recovered data is combined with the other modality data for multimodal fusion and prediction to regress emotional labels.

tion in MER is modality completion [311, 312]. It refers to situations where some modalities in multimodal data are missing or incomplete due to reasons such as device failure, signal loss, or data corruption. By completing the missing modalities, incomplete multimodal data can be fully utilized to ensure the performance of emotion recognition. Generative models provide an effective approach for modal completion. For example, Liu et al. [313] employ multiple AEs to generate high-quality missing modality data by reconstructing it from the available modality features and the invariant features learned through contrastive learning. As shown in Figure 13, Wang et al. [307] propose an Incomplete Multimodality-Diffused emotion recognition (IMDer) method that exploits a score-based DM to recover missing modalities while maintaining distribution consistency and semantic disambiguation by using available modalities as a condition to guide the diffusion-based recovering process.

Table 12: Literature on Generative Models for MER.

Reference	Year	Model	Application	Modalities	Dataset	Performance (%)
Ma et al. [140]	2022	CGAN	Data Augmentation	Speech + Image	eNTERFACE'05 / RAVDESS / CMFW	ACC: 49.48 / 65.90 / 46.19
Luo et al. [294]	2019	CGAN	Data Augmentation	EEG + Eye Movement	SEED / SEED-V	ACC: 90.33 / 68.32
Yan et al. [296]	2021	CGAN	Data Augmentation	EEG + Eye Movement	SEED / SEED-IV / SEED-V	ACC: 93.05 / 86.55 / 80.37
Chao et al. [297]	2018	GAN	Data Augmentation	Speech + fMRI	AudiofMRI Parallel Set / IEMOCAP	ACC: 49.58
Yan et al. [298]	2017	AE	Feature Extraction	EEG + Eye Movement	Self	ACC: 64.26
Guo et al. [300]	2019	AE	Feature Extraction	Eye movements + Eye images + EEG	SEED V	ACC: 79.63
Zhang [301]	2020	DAE	Feature Extraction	Image + EEG	DEAP	ACC: 85.71
Peng et al. [302]	2022	AE	Feature Extraction	Speech + Text	IEMOCAP / MELD / CMU-MOSI	ACC: 74.8 / 63.64 / 79.85
Hanmeh et al. [303]	2021	AE	Feature Extraction	Speech + Image + Text	UImTSST	ACC: 72.95
Nguyen et al. [304]	2021	AE	Feature Extraction	Speech + Image	RECOLA	ACC: 47.4
Ma et al. [305]	2019	AE	Feature Extraction	Speech + Image	eNTERFACE'05	ACC: 85.43
Zheng et al. [306]	2022	AE	Feature Extraction	Speech + Image + Text	IEMOCAP / MSP-IMPROV	ACC: 85.5 / 70.9
Du et al. [308]	2017	VAE	Semi-supervised Learning	EEG + Eye Movement	DEAP / SEED	ACC: 45.6 / 96.8
Liang et al. [309]	2019	GAN	Semi-supervised Learning	Speech + Image	IEMOCAP	UAR: 63.98
Liu et al. [313]	2024	AE	Modality Completion	Speech + Image + Text	IEMOCAP / MSP-IMPROV	ACC: 65.13 / 62.43
Wang et al. [307]	2024	DM	Modality Completion	Speech + Image	CMU-MOSI / CMU-MOSEI	ACC: 79.6 / 80.9

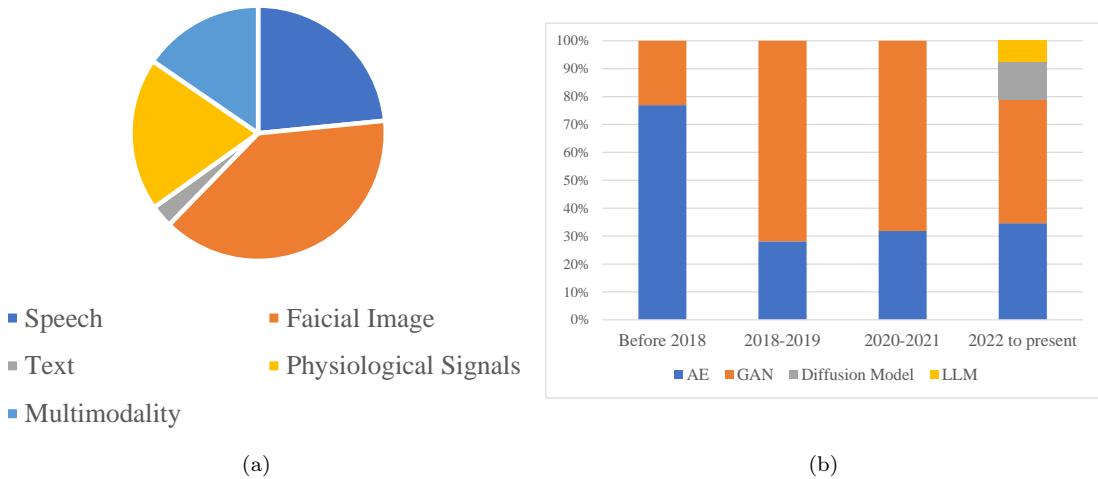


Figure 14: (a) The proportion of different modalities being used, (b) The proportion of different generative models used over time.

## 10. Discussion

### 10.1. Summary Findings

Firstly, we conduct a statistical analysis of the above research papers, as shown in Figure 14, which includes two subfigures 14a and 14b. From subfigure 14a, it can be seen that among various modalities, facial images are the most commonly used by generative models for emotion recognition, reflecting that facial information can express emotions more intuitively and richly. Next are speech signals, while physiological signals, multimodal signals, and text modalities are less common. Subfigure 14b shows the usage trends of various generative models in emotion recognition research during different periods. It can be seen that AEs are most common initially, and over time, the use of GANs is increased. Recently, DMs and LLMs have also begun to gradually appear in research. This evolution reflects the progress of generative technology, with new and more complex models continuously being developed and applied to the field of emotion recognition.

Secondly, generative models have applications in various aspects of emotion recognition, including data augmentation, feature extraction, semi-supervised learning, cross-domain, etc., among which data augmentation and feature extraction are the most common. This may be because the acquisition and annotation of emotion data is quite expensive, especially for facial images and speech data. Utilizing generative models, more samples can be synthesized from existing data to expand the training set. At the same time, the feature representations learned by generative models can also be easily used for downstream emotion recognition tasks to improve performance.

Thirdly, although there are common integration points of generative models in emotion recognition based on different modalities, the specific usage still needs to consider the characteristics of each modality. For example, when generative models are used for data augmentation in SER, the temporal information of speech data can be preserved in the feature space or sample space, while in FER, it almost always operates directly in the sample space to generate image data. Meanwhile, in terms of enhancing robustness, SER pays attention to adversarial sample generation, while FER conducts more research directly from the perspective of dealing with noisy

environments such as occlusion and complex backgrounds during sample acquisition. These differences indicate that applying generative models to emotion recognition based on different modalities requires fully considering the unique attributes of their data and designing and optimizing models in a targeted manner.

Fourthly, it can be seen that among various emotion recognition models, those based on AE and GAN are used the most, possibly because they appeared earlier, and both theory and practice are relatively mature, with widespread applications. The use of DM and LLM is relatively less, possibly because as of now, the application of DM in emotion recognition is still in its infancy, while LLM is mainly used with text modalities. However, due to their excellent generative capabilities and ability to model prior knowledge, it is believed that they will play a greater role in the field of emotion recognition in the future.

Fifthly, it is worth mentioning that we also find relatively few literature on generative models in TER, which may be due to the fact that there has been less mining of fine-grained emotional attributes in text in the past. However, we also note that the application of LLM in TER is increasing. Its powerful semantic understanding and generative capabilities are expected to promote the further development of TER.

Finally, it can be seen that eye movement signals are often combined with other modalities for emotion recognition. However, since the ability of eye movement signals alone to comprehensively represent emotions is limited, they are usually not used independently for emotion recognition, but rather integrated with other modalities to improve overall performance, indicating that multimodal fusion is a major trend in emotion recognition.

## *10.2. Future Prospects*

In the above discussion, we explore various applications of generative models in the field of emotion recognition. Looking forward, research on using generative models for emotion recognition remains full of opportunities, including the following aspects:

Firstly, an exciting development direction is to combine advanced generative models such as DMs with Transformer architectures [60, 61]. DMs excel at generating high-quality, diverse data samples and have shown more remarkable results than AEs and GANs in an increasing number of domains [314, 315]. On the other hand, Transformer architectures, with their powerful sequence modeling capabilities and strong contextual modeling abilities, have brought about a paradigm shift [85]. Combining the two can produce a more powerful generative model capable of simultaneously processing multimodal data such as images, speech, and text, thereby capturing multiple aspects of human emotions more comprehensively and accurately. This combination can significantly enhance the performance of emotion recognition, enabling systems to understand more subtle and complex emotional expressions.

Secondly, combining generative models with other cutting-edge AI technologies such as reinforcement learning [62] and federated learning [63, 316] can further push the boundaries of emotion recognition. Reinforcement learning specializes in modeling and optimizing multi-stage decision-making processes, and emotion recognition and generative models are often dynamic processes involving multiple steps such as environment perception, strategy determination, and performance evaluation. Combining the two can achieve more interactive and adaptive emotional systems [317].

Furthermore, due to the privacy nature of emotional data, it is often difficult to collect on a large scale, which limits the improvement of model performance. Federated learning can allow distributed data to contribute to model training while protecting privacy [318, 319]. By training models on personal devices and then only uploading parameter updates instead of raw data, it is hoped that more powerful and personalized emotion models can be obtained.

Thirdly, with the continuous development of generative models, their applications in emotion recognition are expanding to new scenarios such as virtual reality (VR) and augmented reality (AR) [64, 65]. In these immersive environments, user's emotions and experiences are closely linked. For example, in VR games or social interactions, generating responsive scenes, characters, and dialogues in real-time based on the user's emotional state can create a more immersive experience [320]. In AR navigation or shopping, overlaying emotion-related information and recommendations can provide more thoughtful services [321]. Since VR/AR devices can collect multimodal user data such as gaze and actions, they provide richer clues for emotion recognition. Generative models can leverage this data to create more diverse and dynamic emotional responses.

Finally, this survey mainly discusses using generative models to assist emotion recognition, but another promising direction is to directly synthesize emotionally rich content, such as emotional speech, text, images, and videos [66, 67]. This has important application value in many fields. For example, in film and animation production, automatically generating dubbing and background music with specific emotional tones can reduce the cost of manual recording [322]. In intelligent customer service systems, empathetic and personalized responses can be synthesized in real-time based on the conversation content and user emotions, improving user satisfaction [323]. In the field of education, emotionally rich learning materials (such as stories and dialogues) can be automatically generated to create a more vivid and interesting learning experience [324]. This puts forward higher requirements for the design of generative models and requires more refined conditional control mechanisms. A beneficial prior work is the data augmentation discussed in this survey, i.e., using generative models to synthesize samples with different emotions to expand the dataset. In the future, data augmentation is expected to provide a data foundation and prior knowledge for further high-quality emotional content generation.

## 11. Conclusion

In this paper, we conduct the first systematic survey of the progress of generative technology in the field of human emotion recognition. By analyzing over 320 relevant papers, we delve into the extensive impact of generative models on various aspects and modalities of emotion recognition. This work fills the gap in existing reviews and provides valuable references for researchers in this field.

In Section 1, we introduce the background knowledge of emotion recognition and affective computing, clarifying the close relationship between the two. Emotion recognition is one of the key technologies for realizing affective computing and is crucial for building intelligent, natural, and friendly human-computer interaction systems. Generative models, with their powerful data modeling and generation capabilities, provide new ideas for emotion recognition in multiple aspects. Based on

this, we propose the taxonomy for the application of generative models in emotion recognition and clarify the main contributions of this review, which is to systematically summarize the current state of generative models in emotion recognition and to outlook future research directions.

In Section 2, we make a detailed comparison between this review and other relevant reviews. Although there have been some reviews on emotion recognition or generative models, no work has systematically discussed the combination of the two. This review fills this gap and helps researchers comprehensively understand the latest developments in this interdisciplinary field. In Section 3, in order to help readers better understand the subsequent content, we briefly introduce the mathematical principles of several common generative models, including AE, GAN, DM, and LLM. These models have their own characteristics and play important roles in different tasks of emotion recognition. Section 4 outlines the commonly used datasets for generative models in emotion recognition tasks, covering multiple modalities such as speech, facial images, text, and physiological signals.

Then, from multiple perspectives such as data augmentation, feature extraction, semi-supervised learning, and cross-domain, we elaborate on the applications of generative models in emotion recognition based on different modalities, including SER in Section 5, FER in Section 6, TER in Section 7, physiological signal-based emotion recognition in Section 11, and MER in Section 9. Through comprehensive analysis, we demonstrate that generative models can (1) generate high-quality emotional data to alleviate the problem of data scarcity, (2) learn high-level feature representations of emotional data to enhance the discriminative power of features, (3) utilize unlabeled data for semi-supervised learning to reduce the dependence on large amounts of labeled data, (4) generate cross-domain data to enhance the generalization ability of models, and other aspects. These applications greatly promote the development of emotion recognition. In Section 10, we analyze the characteristics of different generative models in emotion recognition based on different modalities, and emphasize the necessity of continuous innovation of generative models to further enhance their role in affective computing in the future.

In summary, generative models have brought new breakthroughs to emotion recognition, greatly expanding the depth and breadth of research. However, how to further tap the potential of generative models and build more intelligent, natural, and humanized affective computing systems still requires the joint efforts of academia and industry. We believe that with the continuous development of artificial intelligence technology, especially the increasing maturity of generative models, emotion recognition and the entire field of affective computing will usher in a broader development prospect and bring a better future for human-computer interaction. This review provides useful references and inspiration for achieving this goal.

## References

- [1] R. W. Picard, Affective computing, MIT press, 2000.
- [2] R. W. Picard, Affective computing: challenges, International Journal of Human-Computer Studies 59 (1-2) (2003) 55–64.
- [3] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information fusion 37 (2017) 98–125.

- [4] J. Tao, T. Tan, Affective computing: A review, in: International Conference on Affective computing and intelligent interaction, Springer, 2005, pp. 981–995.
- [5] R. A. Calvo, S. D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications, *IEEE Transactions on affective computing* 1 (1) (2010) 18–37.
- [6] L. Pepa, L. Spalazzi, M. Capecci, M. G. Ceravolo, Automatic emotion recognition in clinical scenario: a systematic review of methods, *IEEE Transactions on Affective Computing* 14 (2) (2021) 1675–1695.
- [7] R. Khanna, N. Robinson, M. O'Donnell, H. Eyre, E. Smith, Affective computing in psychotherapy, *Advances in Psychiatry and Behavioral Health* 2 (1) (2022) 95–105.
- [8] E. Yadegaridehkordi, N. F. B. M. Noor, M. N. B. Ayub, H. B. Affal, N. B. Hussin, Affective computing in education: A systematic review and future research, *Computers & Education* 142 (2019) 103649.
- [9] N. Mejbri, F. Essalmi, M. Jemni, B. A. Alyoubi, Trends in the use of affective computing in e-learning environments, *Education and Information Technologies* (2022) 1–23.
- [10] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, R. W. Picard, Driver emotion recognition for intelligent vehicles: A survey, *ACM Computing Surveys (CSUR)* 53 (3) (2020) 1–30.
- [11] L. Mou, Y. Zhao, C. Zhou, B. Nakisa, M. N. Rastgoo, L. Ma, T. Huang, B. Yin, R. Jain, W. Gao, Driver emotion recognition with a hybrid attentional multimodal fusion framework, *IEEE Transactions on Affective Computing* (2023).
- [12] Y. Liu-Thompkins, S. Okazaki, H. Li, Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience, *Journal of the Academy of Marketing Science* 50 (6) (2022) 1198–1218.
- [13] L. Gao, E. de Haan, I. Melero-Polo, F. J. Sese, Winning your customers' minds and hearts: disentangling the effects of lock-in and affective customer experience on retention, *Journal of the Academy of Marketing Science* 51 (2) (2023) 334–371.
- [14] N. Sebe, I. Cohen, T. Gevers, T. S. Huang, Multimodal approaches for emotion recognition: a survey, in: *Internet Imaging VI*, Vol. 5670, SPIE, 2005, pp. 56–67.
- [15] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, et al., A systematic review on affective computing: Emotion models, databases, and recent advances, *Information Fusion* 83 (2022) 19–52.
- [16] K. Ezzamel, H. Mahersia, Emotion recognition from unimodal to multimodal analysis: A review, *Information Fusion* (2023) 101847.

- [17] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, U. R. Acharya, Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations, *Information Fusion* (2023) 102019.
- [18] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, *Speech Communication* 116 (2020) 56–76.
- [19] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, A. C. Sobieranski, A survey on facial emotion recognition techniques: A state-of-the-art literature review, *Information Sciences* 582 (2022) 593–617.
- [20] J. Deng, F. Ren, A survey of textual emotion recognition and its challenges, *IEEE Transactions on Affective Computing* 14 (1) (2021) 49–67.
- [21] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: A review, *Electronic Notes in Theoretical Computer Science* 343 (2019) 35–55.
- [22] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, *IEEE transactions on pattern analysis and machine intelligence* 41 (2) (2018) 423–443.
- [23] N. Ahmed, Z. Al Aghbari, S. Girija, A systematic survey on multimodal emotion recognition using learning algorithms, *Intelligent Systems with Applications* 17 (2023) 200171.
- [24] A. Dzedzickis, A. Kaklauskas, V. Bucinskas, Human emotion recognition: Review of sensors and methods, *Sensors* 20 (3) (2020) 592.
- [25] T. Jebara, Machine learning: discriminative and generative, Vol. 755, Springer Science & Business Media, 2012.
- [26] A. Jabbar, X. Li, B. Omar, A survey on generative adversarial networks: Variants, applications, and training, *ACM Computing Surveys (CSUR)* 54 (8) (2021) 1–49.
- [27] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, *nature* 323 (6088) (1986) 533–536.
- [28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [29] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Advances in neural information processing systems* 33 (2020) 6840–6851.
- [30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.

- [31] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).
- [32] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
- [33] Z. Liu, A. Xu, Y. Guo, J. Mahmud, H. Liu, R. Akkiraju, Seemo: A computational approach to see emotions, 2018, pp. 1–12. doi:10.1145/3173574.3173938.
- [34] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, B. W. Schuller, Multi-task semi-supervised adversarial autoencoding for speech emotion recognition, *IEEE Transactions on Affective computing* 13 (2) (2020) 992–1004.
- [35] B. Nasersharif, M. Ebrahimpour, N. Naderi, Multi-layer maximum mean discrepancy in auto-encoders for cross-corpus speech emotion recognition, *Journal of Supercomputing* 79 (2023) 13031–13049. doi:10.1007/S11227-023-05161-Y/TABLES/3.  
URL <https://link.springer.com/article/10.1007/s11227-023-05161-y>
- [36] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: A report on three machine learning contests, in: *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3–7, 2013. Proceedings, Part III* 20, Springer, 2013, pp. 117–124.
- [37] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Collecting large, richly annotated facial-expression databases from movies, *IEEE MultiMedia* 19 (3) (2012) 34–41. doi:10.1109/MMUL.2012.26.
- [38] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, *Language resources and evaluation* 42 (2008) 335–359.
- [39] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, *PLoS one* 13 (5) (2018) e0196391.
- [40] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [41] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, *IEEE transactions on affective computing* 3 (1) (2011) 18–31.

- [42] S. Katsigiannis, N. Ramzan, Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices, *IEEE journal of biomedical and health informatics* 22 (1) (2017) 98–107.
- [43] K. Maharana, S. Mondal, B. Nemade, A review: Data pre-processing and data augmentation techniques, *Global Transitions Proceedings* 3 (1) (2022) 91–99.
- [44] A. Chatziagapi, G. Paraskevopoulos, D. Sgouropoulos, G. Pantazopoulos, M. Nikandrou, T. Giannakopoulos, A. Katsamanis, A. Potamianos, S. Narayanan, Data augmentation using gans for speech emotion recognition., in: *Interspeech*, 2019, pp. 171–175.
- [45] S. A, M. H, H. N, Effective feature selection in speech emotion recognition systems using generative adversarial networks (11 2022). doi:10.21203/RS.3.RS-2244414/V1.  
URL <https://europepmc.org/article/ppr/ppr570578>
- [46] S. Porcu, A. Floris, L. Atzori, Evaluation of data augmentation techniques for facial expression recognition systems, *Electronics* 9 (11) (2020) 1892.
- [47] Z. Li, Y. Wang, B. Guan, J. Yin, Semantic data augmentation for long-tailed facial expression recognition, in: *2023 8th International Conference on Computer and Communication Systems (ICCCS)*, IEEE, 2023, pp. 1052–1055.
- [48] A. Nedliko, Generative pretrained transformers for emotion detection in a code-switching setting, in: *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 2023, pp. 616–620.
- [49] B. Koptyra, A. Ngo, Ł. Radliński, J. Kocoń, Clarin-emo: Training emotion recognition models using human annotation and chatgpt, in: *International Conference on Computational Science*, Springer, 2023, pp. 365–379.
- [50] G. Harshvardhan, M. K. Gourisaria, M. Pandey, S. S. Rautaray, A comprehensive survey and analysis of generative models in machine learning, *Computer Science Review* 38 (2020) 100285.
- [51] S. Latif, R. Rana, J. Qadir, J. Epps, Variational autoencoders for learning latent representations of speech emotion: A preliminary study (12 2017).  
URL <https://arxiv.org/abs/1712.08708v3>
- [52] L. Yang, Y. Tian, Y. Song, N. Yang, K. Ma, L. Xie, A novel feature separation model exchange-gan for facial expression recognition, *Knowledge-Based Systems* 204 (2020) 106217. doi:10.1016/J.KNOSYS.2020.106217.
- [53] J. E. Van Engelen, H. H. Hoos, A survey on semi-supervised learning, *Machine learning* 109 (2) (2020) 373–440.
- [54] X. Yang, Z. Song, I. King, Z. Xu, A survey on deep semi-supervised learning, *IEEE Transactions on Knowledge and Data Engineering* 35 (9) (2022) 8934–8954.

- [55] H. Zhao, Y. Xiao, Z. Zhang, Robust semisupervised generative adversarial networks for speech emotion recognition via distribution smoothness, *IEEE Access* 8 (2020) 106889–106900. doi:[10.1109/ACCESS.2020.3000751](https://doi.org/10.1109/ACCESS.2020.3000751).
- [56] X. Chen, L. Xu, H. Wei, Z. Shang, T. Zhang, L. Zhang, Emotion interaction recognition based on deep adversarial network in interactive design for intelligent robot, *IEEE Access* 7 (2019) 166860–166868.
- [57] I. Redko, E. Morvant, A. Habrard, M. Sebban, Y. Bennani, A survey on domain adaptation theory: learning bounds and theoretical guarantees, arXiv preprint arXiv:2004.11829 (2020).
- [58] A. Farahani, S. Voghoei, K. Rasheed, H. R. Arabnia, A brief review of domain adaptation, *Advances in data science and information engineering: proceedings from IC DATA 2020 and IKE 2020* (2021) 877–894.
- [59] Y. Xiao, H. Zhao, T. Li, Learning class-aligned and generalized domain-invariant representations for speech emotion recognition, *IEEE Transactions on Emerging Topics in Computational Intelligence* 4 (2020) 480–489. doi:[10.1109/TETCI.2020.2972926](https://doi.org/10.1109/TETCI.2020.2972926).
- [60] W. Peebles, S. Xie, Scalable diffusion models with transformers, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
- [61] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, J. Zhu, All are worth words: A vit backbone for diffusion models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22669–22679.
- [62] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: A brief survey, *IEEE Signal Processing Magazine* 34 (6) (2017) 26–38.
- [63] L. Li, Y. Fan, M. Tse, K.-Y. Lin, A review of applications in federated learning, *Computers & Industrial Engineering* 149 (2020) 106854.
- [64] J. Marín-Morales, C. Llinares, J. Guixeres, M. Alcañiz, Emotion recognition in immersive virtual reality: From statistics to affective computing, *Sensors* 20 (18) (2020) 5163.
- [65] C. Papoutsi, A. Drigas, C. Skianis, Virtual and augmented reality for developing emotional intelligence skills, *Int. J. Recent Contrib. Eng. Sci. IT (IJES)* 9 (3) (2021) 35–53.
- [66] X. Wang, X. Li, Z. Yin, Y. Wu, J. Liu, Emotional intelligence of large language models, *Journal of Pacific Rim Psychology* 17 (2023) 18344909231213958.
- [67] Z. Lian, L. Sun, H. Sun, K. Chen, Z. Wen, H. Gu, S. Chen, B. Liu, J. Tao, Gpt-4v with emotion: a zero-shot benchmark for multimodal emotion understanding, arXiv preprint arXiv:2312.04293 (2023).

- [68] A. Kammoun, R. Slama, H. Tabia, T. Ouni, M. Abid, Generative adversarial networks for face generation: A survey, *ACM Computing Surveys* 55 (5) (2022) 1–37.
- [69] A. Wali, Z. Alamgir, S. Karim, A. Fawaz, M. B. Ali, M. Adan, M. Mujtaba, Generative adversarial networks for speech processing: A review, *Computer Speech & Language* 72 (2022) 101308.
- [70] N. Hajarolasvadi, M. A. Ramirez, W. Beccaro, H. Demirel, Generative adversarial networks in human emotion synthesis: A review, *IEEE Access* 8 (2020) 218499–218529.
- [71] X. Zhao, J. Zhu, B. Luo, Y. Gao, Survey on facial expression recognition: History, applications, and challenges, *IEEE MultiMedia* 28 (4) (2021) 38–44.
- [72] A.-L. Cirneanu, D. Popescu, D. Iordache, New trends in emotion recognition using image analysis by neural networks, a systematic review, *Sensors* 23 (16) (2023) 7092.
- [73] E. M. Younis, S. Mohsen, E. H. Houssein, O. A. S. Ibrahim, Machine learning for human emotion recognition: a comprehensive review, *Neural Computing and Applications* (2024) 1–47.
- [74] K. Nanthini, D. Sivabalaselvamani, K. Chitra, P. Gokul, S. KavinKumar, S. Kishore, A survey on data augmentation techniques, in: 2023 7th International Conference on Computing Methodologies and Communication (IC-CMC), IEEE, 2023, pp. 913–920.
- [75] P. Eigenschink, T. Reutterer, S. Vamosi, R. Vamosi, C. Sun, K. Kalcher, Deep generative models for synthetic sequential data: A survey, *IEEE Access* (2023).
- [76] A. Oussidi, A. Elhassouny, Deep generative models: Survey, in: 2018 International conference on intelligent systems and computer vision (ISCV), IEEE, 2018, pp. 1–8.
- [77] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, S. Z. Li, A survey on generative diffusion models, *IEEE Transactions on Knowledge and Data Engineering* (2024).
- [78] S. Bond-Taylor, A. Leach, Y. Long, C. G. Willcocks, Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models, *IEEE transactions on pattern analysis and machine intelligence* 44 (11) (2021) 7327–7347.
- [79] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, Adversarial autoencoders, *arXiv preprint arXiv:1511.05644* (2015).
- [80] D. P. Kingma, M. Welling, Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013).
- [81] M. Mirza, S. Osindero, Conditional generative adversarial nets, *arXiv preprint arXiv:1411.1784* (2014).

- [82] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015).
- [83] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: International conference on machine learning, PMLR, 2017, pp. 214–223.
- [84] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2223–2232.
- [85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
- [86] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).
- [87] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of machine learning research 21 (140) (2020) 1–67.
- [88] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).
- [89] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
- [90] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).
- [91] H. Guerdelli, C. Ferrari, W. Barhoumi, H. Ghazouani, S. Berretti, Macro-and micro-expressions facial datasets: A survey, Sensors 22 (4) (2022) 1524.
- [92] P. J. Lang, M. M. Bradley, B. N. Cuthbert, et al., International affective picture system (iaps): Technical manual and affective ratings, NIMH Center for the Study of Emotion and Attention 1 (39-58) (1997) 3.
- [93] J. Susskind, A. Anderson, G. E. Hinton, The toronto face dataset, Tech. rep., Technical Report UTML TR 2010-001, U. Toronto (2010).
- [94] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark, in: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, 2011, pp. 2106–2112.
- [95] A. Mollahosseini, B. Hasani, M. H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild, IEEE Transactions on Affective Computing 10 (1) (2017) 18–31.

- [96] L. Shan, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, *IEEE Transactions on Image Processing* 28 (1) (2018) 356–370.
- [97] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets, in: Proceedings Third IEEE international conference on automatic face and gesture recognition, IEEE, 1998, pp. 200–205.
- [98] L. Yin, X. Wei, Y. Sun, J. Wang, M. J. Rosato, A 3d facial expression database for facial behavior research, in: 7th international conference on automatic face and gesture recognition (FGR06), IEEE, 2006, pp. 211–216.
- [99] L.-F. Chen, Y.-S. Yen, Taiwanese facial expression image database, Brain Mapping Laboratory, Institute of Brain Science, National Yang-Ming University, Taipei, Taiwan (2007).
- [100] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, *Image and vision computing* 28 (5) (2010) 807–813.
- [101] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, A high-resolution spontaneous 3d dynamic facial expression database, in: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, 2013, pp. 1–6.
- [102] A. Batliner, S. Steidl, E. Nöth, Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus (2008).
- [103] A. H. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment, Ph.D. thesis, Massachusetts Institute of Technology (2009).
- [104] S. Wang, Z. Liu, S. Lv, Y. Lv, G. Wu, P. Peng, F. Chen, X. Wang, A natural visible and infrared facial expression database for expression recognition and emotion inference, *IEEE Transactions on Multimedia* 12 (7) (2010) 682–691.
- [105] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression, in: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE, 2010, pp. 94–101.
- [106] N. C. Ebner, M. Riediger, U. Lindenberger, Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation, *Behavior research methods* 42 (2010) 351–362.
- [107] G. Zhao, X. Huang, M. Taini, S. Z. Li, M. Pietikäinen, Facial expression recognition from near-infrared videos, *Image and vision computing* 29 (9) (2011) 607–619.
- [108] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg), IEEE, 2013, pp. 1–6.

- [109] R.-N. Duan, J.-Y. Zhu, B.-L. Lu, Differential entropy feature for eeg-based emotion classification, in: 2013 6th international IEEE/EMBS conference on neural engineering (NER), IEEE, 2013, pp. 81–84.
- [110] S. Du, Y. Tao, A. M. Martinez, Compound facial expressions of emotion, *Proceedings of the national academy of sciences* 111 (15) (2014) E1454–E1462.
- [111] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, Casme ii: An improved spontaneous micro-expression database and the baseline evaluation, *PloS one* 9 (1) (2014) e86041.
- [112] R. T. Olszewski, Generalized feature extraction for structural pattern recognition in time-series data, Carnegie Mellon University, 2001.
- [113] A. K. Davison, C. Lansley, N. Costen, K. Tan, M. H. Yap, Samm: A spontaneous micro-facial movement dataset, *IEEE transactions on affective computing* 9 (1) (2016) 116–129.
- [114] T. S. Wingenbach, C. Ashwin, M. Brosnan, Validation of the amsterdam dynamic facial expression set–bath intensity variations (adfec-biv): A set of videos expressing low, intermediate, and high intensity emotions, *PloS one* 11 (1) (2016) e0147112.
- [115] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al., Multimodal spontaneous emotion corpus for human behavior analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3438–3446.
- [116] M. G. Calvo, A. Fernández-Martín, G. Recio, D. Lundqvist, Human observers and automated assessment of dynamic emotional facial expressions: Kdef-dyn database validation, *Frontiers in psychology* 9 (2018) 397727.
- [117] F. Pereira, B. Lou, B. Pritchett, S. Ritter, S. J. Gershman, N. Kanwisher, M. Botvinick, E. Fedorenko, Toward a universal decoder of linguistic meaning from brain activation, *Nature communications* 9 (1) (2018) 963.
- [118] W. Wang, Q. Sun, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, Y. Fu, A fine-grained facial expression database for end-to-end multi-pose facial expression recognition, *arXiv preprint arXiv:1907.10838* (2019).
- [119] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, S. Zafeiriou, Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond, *International Journal of Computer Vision* 127 (6) (2019) 907–929.
- [120] F. Burkhardt, A. Paeschke, M. Rolfs, W. F. Sendlmeier, B. Weiss, et al., A database of german emotional speech., in: *Interspeech*, Vol. 5, 2005, pp. 1517–1520.
- [121] O. Martin, I. Kotsia, B. Macq, I. Pitas, The enterface'05 audio-visual emotion database, in: 22nd international conference on data engineering workshops (ICDEW'06), IEEE, 2006, pp. 8–8.

- [122] S. Haq, P. J. Jackson, J. Edge, Audio-visual feature selection and reduction for emotion classification, in: Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia, 2008.
- [123] M. Grimm, K. Kroschel, S. Narayanan, The vera am mittag german audio-visual emotional speech database, in: 2008 IEEE international conference on multimedia and expo, IEEE, 2008, pp. 865–868.
- [124] M. Valstar, M. Pantic, et al., Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Vol. 10, Paris, France., 2010, pp. 65–70.
- [125] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the recola multimodal corpus of remote collaborative and affective interactions, in: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, 2013, pp. 1–8.
- [126] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, Crema-d: Crowd-sourced emotional multimodal actors dataset, IEEE transactions on affective computing 5 (4) (2014) 377–390.
- [127] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, et al., Emovo corpus: an italian emotional speech database, in: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), European Language Resources Association (ELRA), 2014, pp. 3501–3504.
- [128] E. Takeishi, T. Nose, Y. Chiba, A. Ito, Construction and analysis of phonetically and prosodically balanced emotional speech database, in: 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, 2016, pp. 16–21.
- [129] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, E. M. Provost, Msp-improv: An acted corpus of dyadic interactions to study emotion perception, IEEE Transactions on Affective Computing 8 (1) (2017) 67–80. doi:10.1109/TAFFC.2016.2515617.
- [130] S. Latif, A. Qayyum, M. Usman, J. Qadir, Cross lingual speech emotion recognition: Urdu vs. western languages, in: 2018 International conference on frontiers of information technology (FIT), IEEE, 2018, pp. 88–93.
- [131] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, T. Gedeon, From individual to group-level emotion recognition: Emotiw 5.0, in: Proceedings of the 19th ACM international conference on multimodal interaction, 2017, pp. 524–528.
- [132] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations, arXiv preprint arXiv:1810.02508 (2018).
- [133] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, S. Wermter, The omg-emotion behavior dataset, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–7.

- [134] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, B. W. Schuller, The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress, in: Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, 2021, pp. 5–14.
- [135] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, T. Alhussain, Speech emotion recognition using deep learning techniques: A review, *IEEE access* 7 (2019) 117327–117345.
- [136] A. Mumuni, F. Mumuni, Data augmentation: A comprehensive survey of modern approaches, *Array* 16 (2022) 100258.
- [137] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, *Progress in artificial intelligence* 5 (4) (2016) 221–232.
- [138] T. Ko, V. Peddinti, D. Povey, S. Khudanpur, Audio augmentation for speech recognition., in: *Interspeech*, Vol. 2015, 2015, p. 3586.
- [139] P. Heracleous, S. Fukayama, J. Ogata, Y. Mohammad, Applying generative adversarial networks and vision transformers in speech emotion recognition, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 13519 LNCS (2022) 67–75. doi:10.1007/978-3-031-17618-0\_6.
- [140] F. Ma, Y. Li, S. Ni, S.-L. Huang, L. Zhang, Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional gan, *Applied Sciences* 12 (1) (2022) 527.
- [141] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020).
- [142] S. Wang, H. Hemati, J. Guðnason, D. Borth, Generative data augmentation guided by triplet loss for speech emotion recognition, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2022-Septe* (2022) 391–395. doi:10.21437/Interspeech.2022-10667.  
URL <https://arxiv.org/abs/2208.04994v1>
- [143] L. Yi, M.-W. Mak, Improving speech emotion recognition with adversarial data augmentation network, *IEEE transactions on neural networks and learning systems* 33 (1) (2020) 172–184.
- [144] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, B. W. Schuller, Augmenting generative adversarial networks for speech emotion recognition, *arXiv preprint arXiv:2005.08447* (2020).
- [145] H. Zhang, M. Cisse, Y. N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (10 2017).  
URL <https://arxiv.org/abs/1710.09412v2>

- [146] S. Sahu, R. Gupta, C. Espy-Wilson, On enhancing speech emotion recognition using generative adversarial networks, arXiv preprint arXiv:1806.06626 (2018).
- [147] G. He, X. Liu, F. Fan, J. You, Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 912–913.
- [148] Z. Ma, W. Wu, Z. Zheng, Y. Guo, Q. Chen, S. Zhang, X. Chen, Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition (9 2023).  
URL <https://arxiv.org/abs/2309.10294v1>
- [149] M. I. Malik, S. Latif, R. Jurdak, B. W. Schuller, A preliminary study on augmenting speech emotion recognition using a diffusion model, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2023-August (2023) 646–650. doi:10.21437/Interspeech.2023-1080.  
URL <https://arxiv.org/abs/2305.11413v1>
- [150] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [151] S. Davis, P. Mermelstein, Comparison of parametric representations for mono-syllabic word recognition in continuously spoken sentences, IEEE transactions on acoustics, speech, and signal processing 28 (4) (1980) 357–366.
- [152] Mustaqeem, S. Kwon, A cnn-assisted enhanced audio signal processing for speech emotion recognition, Sensors 20 (1) (2019) 183.
- [153] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, B. Schuller, Survey of deep representation learning for speech emotion recognition, IEEE Transactions on Affective Computing 14 (2) (2021) 1634–1654.
- [154] C. Zhang, L. Xue, Autoencoder with emotion embedding for speech emotion recognition, IEEE Access 9 (2021) 51231–51241. doi:10.1109/ACCESS.2021.3069818.
- [155] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, C. Espy-Wilson, Adversarial auto-encoders for speech based emotion recognition, arXiv preprint arXiv:1806.02146 (2018).
- [156] Y. Ying, Y. Tu, H. Zhou, Unsupervised feature learning for speech emotion recognition based on autoencoder, Electronics 10 (17) (2021) 2086.
- [157] H. Almotlak, C. Weber, L. Qu, S. Wermter, Variational autoencoder with global-and medium timescale auxiliaries for emotion recognition from speech, in: Artificial Neural Networks and Machine Learning–ICANN 2020: 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15–18, 2020, Proceedings, Part I 29, Springer, 2020, pp. 529–540.

- [158] R. D. Fonnegra, G. M. Díaz, Speech emotion recognition integrating paralinguistic features and auto-encoders in a deep learning model, in: International Conference on Human-Computer Interaction, Springer, 2018, pp. 385–396.
- [159] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, S. Narayanan, The interspeech 2010 paralinguistic challenge, in: Proc. INTERSPEECH 2010, Makuhari, Japan, 2010, pp. 2794–2797.
- [160] S. Sahu, R. Gupta, C. Espy-Wilson, Modeling feature representations for affective speech using generative adversarial networks, *IEEE Transactions on Affective Computing* 13 (2022) 1098–1110. doi:10.1109/TAFFC.2020.2998118.
- [161] J. Chang, S. Scherer, Learning representations of emotional speech with deep convolutional generative adversarial networks, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 2746–2750.
- [162] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Semisupervised autoencoders for speech emotion recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26 (1) (2017) 31–43.
- [163] Y. Xiao, Y. Bo, Z. Zheng, Speech emotion recognition based on semi-supervised adversarial variational autoencoder, Proceedings - 2023 IEEE 10th International Conference on Cyber Security and Cloud Computing and 2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud, CSCloud-EdgeCom 2023 (2023) 275–280doi:10.1109/CSCLOUD-EDGECOM58631.2023.00054.
- [164] M. Neumann, N. T. Vu, Improving speech emotion recognition with unsupervised representation learning on unlabeled speech, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 7390–7394.
- [165] S. Zhou, J. Jia, Q. Wang, Y. Dong, Y. Yin, K. Lei, Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.
- [166] S. Das, N. N. Lønfeldt, A. K. Pagsberg, L. H. Clemmensen, Towards transferable speech emotion representation: On loss functions for cross-lingual latent representations, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2022-May (2022) 6452–6456. doi:10.1109/ICASSP43922.2022.9746450.
- [167] S. Latif, R. Rana, S. Khalifa, R. Jurdak, B. W. Schuller, Self supervised adversarial domain adaptation for cross-corpus and cross-language speech emotion recognition, *IEEE Transactions on Affective Computing* (2022).
- [168] B. H. Su, C. C. Lee, Unsupervised cross-corpus speech emotion recognition using a multi-source cycle-gan, *IEEE Transactions on Affective Computing* 14 (2023) 1991–2004. doi:10.1109/TAFFC.2022.3146325.

- [169] B.-H. Su, C.-C. Lee, A conditional cycle emotion gan for cross corpus speech emotion recognition, in: 2021 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2021, pp. 351–357.
- [170] S. Parthasarathy, C. Busso, Semi-supervised speech emotion recognition with ladder networks, *IEEE/ACM transactions on audio, speech, and language processing* 28 (2020) 2697–2709.
- [171] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, *arXiv preprint arXiv:1412.6572* (2014).
- [172] Y. Gong, C. Poellabauer, Crafting adversarial examples for speech paralinguistics applications (11 2017). doi:10.1145/3306195.3306196.  
URL <http://arxiv.org/abs/1711.03280><http://dx.doi.org/10.1145/3306195.3306196>
- [173] Y. Chang, Z. Ren, Z. Zhang, X. Jing, K. Qian, X. Shao, B. Hu, T. Schultz, B. W. Schuller, Staa-net: A sparse and transferable adversarial attack for speech emotion recognition, *arXiv preprint arXiv:2402.01227* (2024).
- [174] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, et al., Wavenet: A generative model for raw audio, *arXiv preprint arXiv:1609.03499* 12 (2016).
- [175] D. Wang, L. Dong, R. Wang, D. Yan, J. Wang, Targeted speech adversarial example generation with generative adversarial network, *IEEE Access* 8 (2020) 124503–124513. doi:10.1109/ACCESS.2020.3006130.
- [176] S. Latif, R. Rana, J. Qadir, Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness, *arXiv* (2018).  
URL <https://arxiv.org/pdf/1811.11402v2.pdf>
- [177] Y. Chang, S. Laridi, Z. Ren, G. Palmer, B. W. Schuller, M. Fisichella, Robust federated learning against adversarial attacks for speech emotion recognition (3 2022).  
URL <https://arxiv.org/abs/2203.04696v1>
- [178] S. Li, W. Deng, Deep facial expression recognition: A survey, *IEEE transactions on affective computing* 13 (3) (2020) 1195–1215.
- [179] C. Dalvi, M. Rathod, S. Patil, S. Gite, K. Kotecha, A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions, *Ieee Access* 9 (2021) 165806–165840.
- [180] Z. Sun, H. Zhang, J. Bai, M. Liu, Z. Hu, A discriminatively deep fusion approach with improved conditional gan (im-cgan) for facial expression recognition, *Pattern Recognition* 135 (2023) 109157. doi:10.1016/J.PATCOG.2022.109157.
- [181] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

- [182] X. Zhu, Y. Liu, J. Li, T. Wan, Z. Qin, Emotion classification with data augmentation using generative adversarial networks, in: Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22, Springer, 2018, pp. 349–360.
- [183] W. Wang, Q. Sun, Y. Fu, T. Chen, C. Cao, Z. Zheng, G. Xu, H. Qiu, Y. G. Jiang, X. Xue, Comp-gan: Compositional generative adversarial network in synthesizing and recognizing facial expression, MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia (2019) 211–219doi:10.1145/3343031.3351032.  
URL <https://dl.acm.org/doi/10.1145/3343031.3351032>
- [184] T. Kusunose, X. Kang, K. Kiuchi, R. Nishimura, M. Sasayama, K. Matsumoto, Facial expression emotion recognition based on transfer learning and generative model, ICSAI 2022 - 8th International Conference on Systems and Informatics (2022). doi:10.1109/ICSAI57119.2022.10005478.
- [185] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
- [186] D. YANG, S. HUANG, S. WANG, P. ZHAI, Y. LI, L. ZHANG, Ee-gan:facial expression recognition method based on generative adversarial network and network integration, Journal of Computer Applications 42 (2022) 750. doi:10.11772/J.ISSN.1001-9081.2021040807.  
URL <http://www.joca.cn/EN/10.11772/j.issn.1001-9081.2021040807>
- [187] B. Han, M. Hu, The facial expression data enhancement method induced by improved stargan v2, Symmetry 2023, Vol. 15, Page 956 15 (2023) 956. doi:10.3390/SYM15040956.  
URL <https://www.mdpi.com/2073-8994/15/4/956/htm><https://www.mdpi.com/2073-8994/15/4/956>
- [188] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions (2015).
- [189] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
- [190] D. Liu, X. Ouyang, S. Xu, P. Zhou, K. He, S. Wen, Saanet: Siamese action-units attention network for improving dynamic facial expression recognition, Neurocomputing 413 (2020) 145–157.
- [191] X. Wang, J. Gong, M. Hu, Y. Gu, F. Ren, Laun improved stargan for facial emotion recognition, IEEE Access 8 (2020) 161509–161518.
- [192] G. Pons, A. El Ali, P. Cesar, Et-cyclegan: Generating thermal images from images in the visible spectrum for facial emotion recognition, in: Companion

- publication of the 2020 international conference on multimodal interaction, 2020, pp. 87–91.
- [193] P. Ekman, W. V. Friesen, Facial action coding system, *Environmental Psychology & Nonverbal Behavior* (1978).
  - [194] S. Minaee, M. Minaei, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, *Sensors* 21 (9) (2021) 3046.
  - [195] B. C. Ko, A brief review of facial emotion recognition based on visual information, *sensors* 18 (2) (2018) 401.
  - [196] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using cnn with attention mechanism, *IEEE Transactions on Image Processing* 28 (5) (2018) 2439–2450.
  - [197] F. Khemakhem, H. Ltifi, Neural style transfer generative adversarial network (nst-gan) for facial expression recognition, *International Journal of Multimedia Information Retrieval* 12 (2023) 1–12.  
doi:10.1007/S13735-023-00285-6/FIGURES/10.  
URL <https://link.springer.com/article/10.1007/s13735-023-00285-6>
  - [198] H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2168–2177.
  - [199] T. Zhang, K. Tang, An efficacious method for facial expression recognition: Gan erased facial feature network (ge2fn), in: *Proceedings of the 2021 13th International Conference on Machine Learning and Computing*, 2021, pp. 417–422.
  - [200] S. Xie, H. Hu, Y. Chen, Facial expression recognition with two-branch disentangled generative adversarial network, *IEEE Transactions on Circuits and Systems for Video Technology* 31 (2021) 2359–2371. doi:10.1109/TCSVT.2020.3024201.
  - [201] K. Ali, C. E. Hughes, Facial expression recognition using disentangled adversarial learning, *arXiv preprint arXiv:1909.13135* (2019).
  - [202] G. Tiwary, S. Chauhan, K. K. Goyal, Facial expression recognition using expression generative adversarial network and attention cnn, *International Journal of Intelligent Systems and Applications in Engineering* 11 (7s) (2023) 447–454.
  - [203] Y. Sima, J. Yi, A. Chen, Z. Jin, Automatic expression recognition of face image sequence based on key-frame generation and differential emotion feature, *Applied Soft Computing* 113 (2021) 108029.

- [204] H. Yang, Z. Zhang, L. Yin, Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 294–301.
- [205] J. Wang, Improved facial expression recognition method based on gan, *Scientific Programming* 2021 (2021) 1–8.
- [206] R. N. Abiram, P. Vincent, P. Vincent, Identity preserving multi-pose facial expression recognition using fine tuned vgg on the latent space vector of generative adversarial network, *Math. Biosci. Eng* 18 (4) (2021) 3699–3717.
- [207] V. Dharanya, A. N. J. Raj, V. P. Gopi, Facial expression recognition through person-wise regeneration of expressions using auxiliary classifier generative adversarial network (ac-gan) based model, *Journal of Visual Communication and Image Representation* 77 (2021) 103110.
- [208] Y. Kim, B. Yoo, Y. Kwak, C. Choi, J. Kim, Deep generative-contrastive networks for facial expression recognition, *arXiv preprint arXiv:1703.07140* (2017).
- [209] H. Wu, J. Jia, L. Xie, G. Qi, Y. Shi, Q. Tian, Cross-vae: Towards disentangling expression from identity for human faces, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2020-May (2020) 4087–4091. doi:10.1109/ICASSP40776.2020.9053608.
- [210] S. Chatterjee, A. K. Das, J. Nayak, D. Pelusi, Improving facial emotion recognition using residual autoencoder coupled affinity based overlapping reduction, *Mathematics* 10 (3) (2022) 406.
- [211] S. Chatterjee, S. Maity, K. Ghosh, A. K. Das, S. Banerjee, Majority biased facial emotion recognition using residual variational autoencoders, *Multimedia Tools and Applications* 83 (5) (2024) 13659–13688.
- [212] W. Zhou, J. Lu, C. Ling, W. Wang, S. Liu, Enhancing emotion recognition with pre-trained masked autoencoders and sequential learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4666–4672.
- [213] X. Wang, X. Wang, Y. Ni, et al., Unsupervised domain adaptation for facial expression recognition using generative adversarial networks, *Computational intelligence and neuroscience* 2018 (2018).
- [214] Y. Fan, J. C. Lam, V. O. Li, Unsupervised domain adaptation with generative adversarial networks for facial emotion recognition, in: *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, 2018, pp. 4460–4464.
- [215] F. Zhang, T. Zhang, Q. Mao, L. Duan, C. Xu, Facial expression recognition in the wild: A cycle-consistent adversarial attention transfer approach, in: *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 126–135.

- [216] Y. Du, D. Yang, P. Zhai, M. Li, L. Zhang, Learning associative representation for facial expression recognition, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 889–893.
- [217] Z. Peng, J. Li, Z. Sun, Emotion recognition using generative adversarial networks, in: 2020 International Conference on Computer Engineering and Intelligent Control (ICCEIC), IEEE, 2020, pp. 77–80.
- [218] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: International conference on machine learning, PMLR, 2017, pp. 2642–2651.
- [219] Y. Lu, S. Wang, W. Zhao, Y. Zhao, Wgan-based robust occluded facial expression recognition, *IEEE Access* 7 (2019) 93594–93610.
- [220] H. Tang, W. Wang, S. Wu, X. Chen, D. Xu, N. Sebe, Y. Yan, Expression conditional gan for facial expression-to-expression translation, in: 2019 IEEE international conference on image processing (ICIP), IEEE, 2019, pp. 4449–4453.
- [221] H. Li, M. Sui, F. Zhao, Z. Zha, F. Wu, Mvt: Mask vision transformer for facial expression recognition in the wild (6 2021).  
URL <https://arxiv.org/abs/2106.04520v2>
- [222] Z. Han, H. Huang, Gan based three-stage-training algorithm for multi-view facial expression recognition, *Neural Processing Letters* 53 (2021) 4189–4205.  
*doi:10.1007/S11063-021-10591-X*.  
URL <https://dl.acm.org/doi/10.1007/s11063-021-10591-x>
- [223] F. Zhang, T. Zhang, Q. Mao, C. Xu, Geometry guided pose-invariant facial expression recognition, *IEEE Transactions on Image Processing* 29 (2020) 4445–4460.
- [224] Y.-H. Lai, S.-H. Lai, Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 263–270.
- [225] D. Li, Z. Li, R. Luo, J. Deng, S. Sun, Multi-pose facial expression recognition based on generative adversarial network, *IEEE Access* 7 (2019) 143980–143989.
- [226] J. Dong, Y. Zhang, L. Fan, A multi-view face expression recognition method based on densenet and gan, *Electronics* 2023, Vol. 12, Page 2527 12 (2023) 2527. *doi:10.3390/ELECTRONICS12112527*.  
URL <https://www.mdpi.com/2079-9292/12/11/2527/htmhttps://www.mdpi.com/2079-9292/12/11/2527>
- [227] H. Yang, K. Zhu, D. Huang, H. Li, Y. Wang, L. Chen, Intensity enhancement via gan for multimodal face expression recognition, *Neurocomputing* 454 (2021) 124–134. *doi:10.1016/J.NEUROCOMPUTING.2021.05.022*.

- [228] F. Nan, W. Jing, F. Tian, J. Zhang, K. M. Chao, Z. Hong, Q. Zheng, Feature super-resolution based facial expression recognition for multi-scale low-resolution images, *Knowledge-Based Systems* 236 (2022) 107678. doi: 10.1016/J.KNOSYS.2021.107678.
- [229] Z. Wang, K. Zhang, R. Sankaranarayana, Lrdif: Diffusion models for under-display camera emotion recognition, arXiv preprint arXiv:2402.00250 (2024).
- [230] S. M. Saleem, S. R. Zeebaree, M. B. Abdulrazzaq, Real-life dynamic facial expression recognition: a review, in: *Journal of Physics: Conference Series*, Vol. 1963, IOP Publishing, 2021, p. 012010.
- [231] X. Pu, K. Fan, X. Chen, L. Ji, Z. Zhou, Facial expression recognition from image sequences using twofold random forest classifier, *Neurocomputing* 168 (2015) 1173–1180.
- [232] E. G. Krumhuber, L. Skora, D. Küster, L. Fou, A review of dynamic datasets for facial expression research, *Emotion Review* 9 (3) (2017) 280–292.
- [233] D. Deng, Z. Chen, Y. Zhou, B. Shi, Mimamo net: Integrating micro-and macro-motion for video emotion recognition, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, 2020, pp. 2621–2628.
- [234] Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofghi, R. Haffari, M. Hayat, Marlin: Masked autoencoder for facial video representation learning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1493–1504.
- [235] P. A. Gavade, V. S. Bhat, J. Pujari, Improved deep generative adversarial network with illuminant invariant local binary pattern features for facial expression recognition, *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 11 (2023) 678–695. doi:10.1080/21681163.2022.2103450.  
URL <https://www.tandfonline.com/doi/abs/10.1080/21681163.2022.2103450>
- [236] W. Guo, X. Zhao, S. Zhang, X. Pan, Learning inter-class optical flow difference using generative adversarial networks for facial expression recognition, *Multimedia Tools and Applications* 82 (2023) 10099–10116. doi:10.1007/S11042-022-13360-7/TABLES/7.  
URL <https://link.springer.com/article/10.1007/s11042-022-13360-7>
- [237] S.-T. Liang, Y. S. Gan, D. Zheng, S.-M. Li, H.-X. Xu, H.-Z. Zhang, R.-K. Lyu, K.-H. Liu, Evaluation of the spatio-temporal features and gan for micro-expression recognition system, *Journal of Signal Processing Systems* 92 (2020) 705–725.
- [238] J. Chen, Y. Fu, Y. Jin, T. Liu, Dffcn: Dual flow fusion convolutional network for micro expression recognition, in: *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III* 28, Springer, 2021, pp. 76–87.

- [239] F. M. A. Mazen, A. A. Nashat, R. A. A. A. Seoud, Real time face expression recognition along with balanced fer2013 dataset using cyclegan, *International Journal of Advanced Computer Science and Applications* 12 (6) (2021).
- [240] N. Alswaidan, M. E. B. Menai, A survey of state-of-the-art approaches for emotion recognition in text, *Knowledge and Information Systems* 62 (8) (2020) 2937–2987.
- [241] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, P. Agrawal, Understanding emotions in text using deep learning and big data, *Computers in Human Behavior* 93 (2019) 309–317.
- [242] A. Yadollahi, A. G. Shahraki, O. R. Zaiane, Current state of text sentiment analysis from opinion to emotion mining, *ACM Computing Surveys (CSUR)* 50 (2) (2017) 1–33.
- [243] S. M. Mohammad, Sentiment analysis: Detecting valence, emotions, and other affectual states from text, in: *Emotion measurement*, Elsevier, 2016, pp. 201–237.
- [244] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8 (4) (2018) e1253.
- [245] P. Nandwani, R. Verma, A review on sentiment analysis and emotion detection from text, *Social Network Analysis and Mining* 11 (1) (2021) 81.
- [246] B. Liu, *Sentiment analysis and opinion mining*, Springer Nature, 2022.
- [247] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, S. Yu, A survey on deep learning for textual emotion analysis in social networks, *Digital Communications and Networks* 8 (5) (2022) 745–762.
- [248] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, 2018.
- [249] A. Pico, E. Vivancos, A. García-Fornes, V. J. Botti, Exploring text-generating large language models (llms) for emotion recognition in affective intelligent agents., in: *ICAART* (1), 2024, pp. 491–498.
- [250] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, S. Poria, Cosmic: Commonsense knowledge for emotion identification in conversations, *arXiv preprint arXiv:2010.02795* (2020).
- [251] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, Y. Choi, Comet: Commonsense transformers for automatic knowledge graph construction, *arXiv preprint arXiv:1906.05317* (2019).
- [252] K. Hama, A. Otsuka, R. Ishii, Emotion recognition in conversation with multi-step prompting using large language model, in: *International Conference on Human-Computer Interaction*, Springer, 2024, pp. 338–346.

- [253] S. Lei, G. Dong, X. Wang, K. Wang, S. Wang, Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework, arXiv preprint arXiv:2309.11911 (2023).
- [254] Y. Fu, Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation, arXiv preprint arXiv:2403.07260 (2024).
- [255] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, X. Yang, A review of emotion recognition using physiological signals, *Sensors* 18 (7) (2018) 2074.
- [256] S. M. Alarcao, M. J. Fonseca, Emotions recognition using eeg signals: A survey, *IEEE transactions on affective computing* 10 (3) (2017) 374–393.
- [257] P. Zhong, D. Wang, C. Miao, Eeg-based emotion recognition using regularized graph neural networks, *IEEE Transactions on Affective Computing* 13 (3) (2020) 1290–1301.
- [258] Y. Li, J. Huang, H. Zhou, N. Zhong, Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks, *Applied Sciences* 7 (10) (2017) 1060.
- [259] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, A. A. Aziz, Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review, *Sensors* 21 (15) (2021) 5015.
- [260] P. Sarkar, A. Etemad, Self-supervised ecg representation learning for emotion recognition, *IEEE Transactions on Affective Computing* 13 (3) (2020) 1541–1554.
- [261] H. Ferdinando, L. Ye, T. Seppänen, E. Alasaarela, Emotion recognition by heart rate variability, *Australian Journal of Basic and Applied Science* 8 (14) (2014) 50–55.
- [262] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, D. Puig, Feature extraction and selection for emotion recognition from electrodermal activity, *IEEE Transactions on Affective Computing* 12 (4) (2019) 857–869.
- [263] S. Qiu, Y. Chen, Y. Yang, P. Wang, Z. Wang, H. Zhao, Y. Kang, R. Nie, A review on semi-supervised learning for eeg-based emotion recognition, *Information Fusion* (2023) 102190.
- [264] Y. Luo, B.-L. Lu, Eeg data augmentation for emotion recognition using a conditional wasserstein gan, in: 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2018, pp. 2535–2538.
- [265] G. Bao, B. Yan, L. Tong, J. Shu, L. Wang, K. Yang, Y. Zeng, Data augmentation for eeg-based emotion recognition using generative adversarial networks, *Frontiers in Computational Neuroscience* 15 (2021) 723843. doi:10.3389/FNCOM.2021.723843/BIBTEX.

- [266] S. Bhat, E. Hortal, Gan-based data augmentation for improving the classification of eeg signals, in: Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference, 2021, pp. 453–458.
- [267] Z. Zhang, S.-h. Zhong, Y. Liu, Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition, *IEEE Transactions on Affective Computing* (2022).
- [268] Z. Zhang, S. Zhong, Y. Liu, Beyond mimicking under-represented emotions: Deep data augmentation with emotional subspace constraints for eeg-based emotion recognition, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 2024, pp. 10252–10260.
- [269] S. Haradal, H. Hayashi, S. Uchida, Biosignal data augmentation based on generative adversarial networks, in: *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, IEEE, 2018, pp. 368–371.
- [270] B. Pan, W. Zheng, Emotion recognition based on eeg using generative adversarial nets and convolutional neural network, *Computational and mathematical methods in medicine* 2021 (2021). doi:10.1155/2021/2520394.  
URL <https://pubmed.ncbi.nlm.nih.gov/34671415/>
- [271] Y. Luo, L. Z. Zhu, Z. Y. Wan, B. L. Lu, Data augmentation for enhancing eeg-based emotion recognition with deep generative models, *Journal of Neural Engineering* 17 (2020) 056021. doi:10.1088/1741-2552/ABB580.  
URL <https://iopscience.iop.org/article/10.1088/1741-2552/abb580>  
<https://iopscience.iop.org/article/10.1088/1741-2552/abb580/meta>
- [272] M. P. Kalashami, M. M. Pedram, H. Sadr, et al., Eeg feature extraction and data augmentation in emotion recognition, *Computational intelligence and neuroscience* 2022 (2022).
- [273] G. Siddhad, M. Iwamura, P. P. Roy, Enhancing eeg signal-based emotion recognition with synthetic data: Diffusion modeel approach, arXiv preprint arXiv:2401.16878 (2024).
- [274] G. Tosato, C. M. Dalbagno, F. Fumagalli, Eeg synthetic data generation using probabilistic diffusion models, arXiv preprint arXiv:2303.06068 (2023).
- [275] Y. Yi, Y. Xu, B. Yang, Y. Tian, A weighted co-training framework for emotion recognition based on eeg data generation using frequency-spatial diffusion transformer, *IEEE Transactions on Affective Computing* (2024).
- [276] Z. Lan, O. Sourina, L. Wang, R. Scherer, G. Müller-Putz, Unsupervised feature learning for eeg-based emotion recognition, in: *2017 International Conference on Cyberworlds (CW)*, IEEE, 2017, pp. 182–185.
- [277] A. S. Rajpoot, M. R. Panicker, et al., Subject independent emotion recognition using eeg signals employing attention driven neural networks, *Biomedical Signal Processing and Control* 75 (2022) 103547.

- [278] S. Jirayucharoensak, S. Pan-Ngum, P. Israsena, et al., Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation, *The Scientific World Journal* 2014 (2014).
- [279] X. Li, Z. Zhao, D. Song, Y. Zhang, C. Niu, J. Zhang, J. Huo, J. Li, Variational autoencoder based latent factor decoding of multichannel eeg for emotion recognition, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2019, pp. 684–687.
- [280] D. Bethge, P. Hallgarten, T. Grosse-Puppendahl, M. Kari, L. L. Chuang, O. Özdenizci, A. Schmidt, Eeg2vec: Learning affective eeg representations via variational autoencoders, in: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2022, pp. 3150–3157.
- [281] J. Liu, G. Wu, Y. Luo, S. Qiu, S. Yang, W. Li, Y. Bi, Eeg-based emotion classification using a deep neural network and sparse autoencoder, *Frontiers in Systems Neuroscience* 14 (2020) 43.
- [282] C. Qing, R. Qiao, X. Xu, Y. Cheng, Interpretable emotion recognition using eeg signals, *Ieee Access* 7 (2019) 94160–94170.
- [283] X. Li, Z. Zhao, Latent factor decoding of multi-channel eeg for emotion recognition through autoencoder-like neural networks, *Frontiers in neuroscience* 14 (2020) 496999.
- [284] Y. Gu, X. Zhong, C. Qu, C. Liu, B. Chen, A domain generative graph network for eeg-based emotion recognition, *IEEE Journal of Biomedical and Health Informatics* 27 (2023) 2377–2386. doi:10.1109/JBHI.2023.3242090.
- [285] G. Zhang, A. Etemad, Deep recurrent semi-supervised eeg representation learning for emotion recognition, in: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2021, pp. 1–8.
- [286] X. Chai, Q. Wang, Y. Zhao, X. Liu, O. Bai, Y. Li, Unsupervised domain adaptation techniques based on auto-encoder for non-stationary eeg-based emotion recognition, *Computers in biology and medicine* 79 (2016) 205–214.
- [287] Y. Wang, S. Qiu, D. Li, C. Du, B.-L. Lu, H. He, Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition, *IEEE/CAA Journal of Automatica Sinica* 9 (9) (2022) 1612–1626.
- [288] D. Huang, S. Zhou, D. Jiang, Generator-based domain adaptation method with knowledge free for cross-subject eeg emotion recognition, *Cognitive Computation* 14 (4) (2022) 1316–1327.
- [289] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, *IEEE Journal of selected topics in signal processing* 11 (8) (2017) 1301–1309.
- [290] J. Zhang, Z. Yin, P. Chen, S. Nicelle, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, *Information Fusion* 59 (2020) 103–126.

- [291] H. O. Hirschfeld, A connection between correlation and contingency, in: Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 31, Cambridge University Press, 1935, pp. 520–524.
- [292] H. Gebelein, Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung, ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik 21 (6) (1941) 364–379.
- [293] A. Rényi, On measures of dependence, Acta mathematica hungarica 10 (3-4) (1959) 441–451.
- [294] Y. Luo, L.-Z. Zhu, B.-L. Lu, A gan-based data augmentation method for multimodal emotion recognition, in: Advances in Neural Networks–ISNN 2019: 16th International Symposium on Neural Networks, ISNN 2019, Moscow, Russia, July 10–12, 2019, Proceedings, Part I 16, Springer, 2019, pp. 141–150.
- [295] D. Berthelot, T. Schumm, L. Metz, Began: Boundary equilibrium generative adversarial networks, arXiv preprint arXiv:1703.10717 (2017).
- [296] X. Yan, L.-M. Zhao, B.-L. Lu, Simplifying multimodal emotion recognition with single eye movement modality, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1057–1063.
- [297] G.-Y. Chao, C.-M. Chang, J.-L. Li, Y.-T. Wu, C.-C. Lee, Generating fmri-enriched acoustic vectors using a cross-modality adversarial network for emotion recognition, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 55–62.
- [298] X. Yan, W.-L. Zheng, W. Liu, B.-L. Lu, Identifying gender differences in multimodal emotion recognition using bimodal deep autoencoder, in: Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24, Springer, 2017, pp. 533–542.
- [299] M. SUGENO, Theory of fuzzy integrals and its applications, Doctoral Thesis, Tokyo Institute of Technology (1974).  
URL <https://cir.nii.ac.jp/crid/1572261549724574208>
- [300] J.-J. Guo, R. Zhou, L.-M. Zhao, B.-L. Lu, Multimodal emotion recognition from eye image, eye movement and eeg using deep neural networks, in: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2019, pp. 3071–3074.
- [301] H. Zhang, Expression-eeg based collaborative multimodal emotion recognition using deep autoencoder, IEEE Access 8 (2020) 164130–164143.
- [302] P. Shixin, C. Kai, T. Tian, C. Jingying, An autoencoder-based feature level fusion for speech emotion recognition, Digital Communications and Networks (2022).

- [303] S. Hamieh, V. Heiries, H. Al Osman, C. Godin, Multi-modal fusion for continuous emotion recognition by using auto-encoders, in: Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, 2021, pp. 21–27.
- [304] D. Nguyen, D. T. Nguyen, R. Zeng, T. T. Nguyen, S. N. Tran, T. Nguyen, S. Sridharan, C. Fookes, Deep auto-encoders with sequential learning for multimodal dimensional emotion recognition, *IEEE Transactions on Multimedia* 24 (2021) 1313–1324.
- [305] F. Ma, W. Zhang, Y. Li, S.-L. Huang, L. Zhang, An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, pp. 1144–1149.
- [306] J. Zheng, S. Zhang, Z. Wang, X. Wang, Z. Zeng, Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition, *IEEE Transactions on Multimedia* (2022).
- [307] Y. Wang, Y. Li, Z. Cui, Incomplete multimodality-diffused emotion recognition, *Advances in Neural Information Processing Systems* 36 (2024).
- [308] C. Du, C. Du, J. Li, W.-l. Zheng, B.-l. Lu, H. He, Semi-supervised bayesian deep multi-modal emotion recognition, *arXiv preprint arXiv:1704.07548* (2017).
- [309] J. Liang, S. Chen, Q. Jin, Semi-supervised multimodal emotion recognition with improved wasserstein gans, in: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2019, pp. 695–703.
- [310] X. Wei, B. Gong, Z. Liu, W. Lu, L. Wang, Improving the improved training of wasserstein gans: A consistency term and its dual effect, *arXiv preprint arXiv:1803.01541* (2018).
- [311] N. Jaques, S. Taylor, A. Sano, R. Picard, Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction, in: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2017, pp. 202–208.
- [312] A. Geetha, T. Mala, D. Priyanka, E. Uma, Multimodal emotion recognition with deep learning: advancements, challenges, and future directions, *Information Fusion* 105 (2024) 102218.
- [313] R. Liu, H. Zuo, Z. Lian, B. W. Schuller, H. Li, Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities, *IEEE Transactions on Affective Computing* (2024).
- [314] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, *Advances in neural information processing systems* 34 (2021) 8780–8794.

- [315] M. Stypułkowski, K. Vougioukas, S. He, M. Zięba, S. Petridis, M. Pantic, Dif-fused heads: Diffusion models beat gans on talking-face generation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5091–5100.
- [316] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, Y. Gao, A survey on federated learning, *Knowledge-Based Systems* 216 (2021) 106775.
- [317] G. Bilquise, S. Ibrahim, K. Shaalan, Emotionally intelligent chatbots: A systematic literature review, *Human Behavior and Emerging Technologies* 2022 (1) (2022) 9601630.
- [318] S. Latif, S. Khalifa, R. Rana, R. Jurdak, Federated learning for speech emotion recognition applications, in: 2020 19th ACM/IEEE international conference on information processing in sensor networks (IPSN), IEEE, 2020, pp. 341–342.
- [319] H. Zhao, H. Chen, Y. Xiao, Z. Zhang, Privacy-enhanced federated learning against attribute inference attack for speech emotion recognition, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
- [320] A. Felnhofer, O. D. Kothgassner, M. Schmidt, A.-K. Heinze, L. Beutl, H. Hlavacs, I. Kryspin-Exner, Is virtual reality emotionally arousing? investigating five emotion inducing virtual park scenarios, *International journal of human-computer studies* 82 (2015) 48–56.
- [321] A. Valente, D. S. Lopes, N. Nunes, A. Esteves, Empathic aurea: Exploring the effects of an augmented reality cue for emotional sharing across three face-to-face tasks, in: 2022 IEEE conference on virtual reality and 3D user interfaces (VR), IEEE, 2022, pp. 158–166.
- [322] S. Ji, X. Yang, Muser: Musical element-based regularization for generating symbolic music with emotion, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 12821–12829.
- [323] D. Leocádio, L. Guedes, J. Oliveira, J. Reis, N. Melão, Customer service with ai-powered human-robot collaboration (hrc): a literature review, *Procedia Computer Science* 232 (2024) 1222–1232.
- [324] X. Hong, A. Sayeed, K. Mehra, V. Demberg, B. Schiele, Visual writing prompts: Character-grounded story generation with curated image sequences, *Transactions of the Association for Computational Linguistics* 11 (2023) 565–581.