GANSER: A Self-Supervised Data Augmentation Framework for EEG-Based Emotion Recognition

Zhi Zhang, Yan Liu, and Sheng-hua Zhong

Abstract—Electroencephalography (EEG)-based affective computing has a scarcity problem. As a result, it is difficult to build effective, highly accurate and stable models using machine learning algorithms, especially deep learning models. Data augmentation has recently shown performance improvements in deep learning models with increased accuracy, stability and reduced overfitting. In this paper, we propose a novel data augmentation framework, named the generative adversarial network-based self-supervised data augmentation (GANSER). As the first to combine adversarial training with self-supervised learning for EEG-based emotion recognition, the proposed framework generates high-quality and high-diversity simulated EEG samples. In particular, we utilize adversarial training to learn an EEG generator and force the generated EEG signals to approximate the distribution of real samples, ensuring the quality of the augmented samples. A transformation operation is employed to mask parts of the EEG signals and force the generator to synthesize potential EEG signals based on the unmasked parts to produce a wide variety of samples. A masking possibility during transformation is introduced as prior knowledge to generalize the classifier for the augmented sample space. Finally, numerous experiments demonstrate that our proposed method can improve emotion recognition with an increase in performance and achieve state-of-the-art results.

Index Terms—EEG-based emotion recognition, data augmentation, generative adversarial networks

1 INTRODUCTION

As humans, we manifest emotions throughout our daily life. Understanding emotions is one of the most important aspects of measuring human development and growth, and therefore, it is an important factor for the emulation of human intelligence [1]. Thus, affective computing and automatic emotion recognition are important for the advancement of AI [2] and all the research fields that stem from it. Electroencephalography (EEG) measures oscillations in the brain, which reflect the synchronized activity of neurons. It is thought that the changes in these oscillations are correlated with the cognitive process and can be used to reveal important information about human emotional states. As a kind of physiological signal, EEG has the advantage of being difficult to hide or disguise. Compared with other physiological signals, it has excellent time resolution, which is similar to the nuanced changes in emotional states on a time scale. Owing to the rapid development of noninvasive, easy-to-use and inexpensive recording devices, EEG-based emotion recognition has received an increasing amount of attention in both research and application [3].

Nevertheless, EEG is also subject to several limitations. First, as an aggregate signal from the activity of millions of neurons, EEG suffers from a low signal-to-noise ratio (SNR) [4]. Second, EEG is generally recorded by using tens to hundreds of electrodes simultaneously, and the sampling time usually exceeds a few seconds in each trial. Thus, the original feature dimension of an EEG sample is not low. However, a typical dataset for cognitive neuroscience tasks usually contains only a few hundred to a few thousand samples (i.e., experimental trials). This leads to a very low initial ratio of samples to features. Third, EEG is a nonstationary signal, and its statistics vary over time. The inherent variabilities in brain anatomy, head size and dynamics across trials/subjects considerably limit the generalizability of EEG analyses across subjects, and even across trials within a single subject performing a single task. The second limitation makes the use of machine learning models difficult, while the other two exacerbate this difficulty.

In the past few years, deep learning methods have achieved breakthrough performances for EEG-based emotion recognition. Unfortunately, deep learning models are typically very complex, i.e., they have many free parameters (or degrees of freedom) to fit [5]. Thus, if we lack enough training data, cases of EEG-based emotion recognitions that have a low initial ratio of samples to features are particularly problematic, training such deep learning models risks overfitting these models to specific quirks of the training set. It also severely limits the generalizability of these models.

Manuscript received 14 July 2021; revised 16 April 2022; accepted 18 April 2022.

Address for correspondence: Sheng-hua Zhong.

Digital Object Identifier no. 10.1109/TAFFC.2022.3170369
Data augmentation is considered an effective technology for solving the data scarcity problem. It is usually a process of generating new realistic-like data by applying a transformation to the real data [6]. It also has the potential to increase the accuracy and stability of the classification or regression and is thus widely adopted in research fields such as resource allocation of Internet of Things (IoT) [7], [8]. To overcome the data scarcity problem, in this paper, we propose a generative adversarial network-based self-supervised data augmentation (GANSER) framework for EEG-based emotion recognition. The proposed framework is comprised of two networks, including the adversarial augmentation network (AAN) and multi-factor training network (MTN). In the AAN, we propose a masking transformation operation to mask parts of the EEG signals and then force the proposed generative adversarial network (GAN) to synthesize the potential EEG signals based on the remaining parts. UNet, the channel masking operation and STNet are employed to model the spatiotemporal features of the EEG signals, while adversarial training forces the generated EEG signals to approximate the distribution of real signals, ensuring the quality of the simulated EEG signals. Next, in the MTN, the simulated EEG signals are utilized for training the emotion recognition models as augmented samples. A multi-factor self-supervised learning loss (MSL) is proposed to introduce the masking possibility as prior knowledge to guide the model to generalize the classifier to the augmented sample space.

In summary, the research contributions of this paper are highlighted as follows. (i) This paper proposes a novel data augmentation framework named GANSER, opening up a new road to combine adversarial training with self-supervised learning to address the data scarcity bottleneck for EEG-based emotion recognition. (ii) This paper first proposes a novel adversarial augmentation network to synthesize multi-channel EEG data at the signal level, which can be adapted to arbitrary classifiers with more precise low-level simulation. Here, the masking transformation operation is designed and combined with the UNet network, the channel masking operation, and the adversarial STNet network to ensure the diversity and quality of the simulated EEG signals. (iii) This paper first introduces a multi-factor training network to measure and trade-off the quality and diversity of augmented samples and explore optimizing classifiers with diverse augmented samples while simultaneously inhibiting the side effect of low-quality samples. (iv) Numerous experiments are carried out, and the results show that our proposed deep framework significantly outperforms the existing state-of-the-art methods. Finally, we adopt a quantitative assessment approach for EEG analysis to evaluate the quality and diversity of the augmented samples, and the visualization results are provided for qualitative analysis.

2 RELATED WORK
2.1 EEG-Based Emotion Recognition
In recent years, boosted by the success of deep neural networks, deep learning-based emotion recognition [9], [10], [11], [12], [13] has received an increasing amount of attention in both research and applications, and these studies seek to explore end-to-end methods to tackle EEG-based emotion recognition.

In detail, deep neural networks, such as recurrent neural networks (RNNs), 2D/3D convolutional neural networks (CNNs), or both have been employed for feature extraction and classification. In 2016, Zhang et al. [14] proposed a spatial-temporal recurrent neural network (STRNN) to investigate both spatial and temporal dependencies of EEG signals and achieved the state-of-the-art performance. In 2018, Li et al. further [15] proposed a hybrid deep learning structure based on a CNN and an RNN for emotion recognition. The proposed method was effective in the trial-level emotion recognition task. In the same year, Salama et al. [16] employed 3D CNNs to classify human emotions. To input an EEG signal into a 3D CNN, they divided the 2D shape (channel \( \times \) time) of EEG data into 6-s segments and stacked them along the third axis. In a subsequent study, Moon et al. [17] pointed out the limitation that only signals or features from individual electrodes are considered and employed brain connectivity features to account for synchronous activations of different brain regions. In 2020, Moon et al. [18] further improved their research and introduced three different types of connectivity measures to model brain connectivity with a CNN. Furthermore, two data-driven methods have been proposed to construct the connectivity matrix and maximize classification performance.

In addition to RNN and CNN, recent work has also made significant improvements with the help of graph convolutional networks (GNNs) and generative adversarial networks. In 2018, Song et al. [9] proposed a dynamical graph convolutional neural network (DGCNN) to explore the intrinsic connections between EEG channels and implement the graph convolution operation on learned graphs. In the following work, Zhang et al. [19] combined graph convolution and regular convolution to learn features on graph-structured data and extract high-level information. Then, broad learning [20] was applied in the model to concatenate outputs of hierarchical layers so that the model could search for desirable results in broad feature spaces. In 2018, Luo et al. [21] found that EEG signals significantly varied depending on the individual, which makes it difficult to achieve satisfactory classification performance. They proposed a Wasserstein GAN-based framework to solve the domain shift problem by narrowing the gap between the probability distributions of different subjects. Recently, Moon et al. [18] proposed learning feature space mappings and performing individuality detachment to reduce subject-related information in EEG signals. The proposed method effectively discarded the subject-related information and performed well in emotion recognition tasks.

2.2 Data Augmentation for EEG-Based Emotion Recognition
In the field of image processing and computer vision, there are two simple and direct ways to augment data: geometric transformations and noise addition operations. Krell and Su et al. suggested using rotational distortions, which are similar to affine/rotational distortions of images, to generate augmented EEG data [22]. Relying on relevant combinations and distortions of original trials, Lotte proposed three methods to obtain artificial EEG trials [23]. Different from
geometric transformations, Wang et al. generated new features by adding Gaussian noise with different standard deviations to the original EEG feature and applied several deep learning models to verify the effect [24]. Other data augmentation methods include sliding window, sampling, Fourier transform, and recombination of segmentations [5]. All of the abovementioned methods reported that the data scarcity problem had been alleviated, and the performance of the classifiers was improved [25].

Recently, generative adversarial networks (GANs) have revealed their potential in generating EEG signals that mimic real signals utilized in emotion recognition tasks [25, 26, 27] and a wide variety of applications [28, 29, 30, 31, 32]. A conditional version of the Wasserstein generative adversarial network (WGAN) was used to augment EEG data for emotion recognition in [26]. They tried different amounts of the augmented data, and they found that doubling the data led to the highest performance compared to other data amounts. An SVM classifier trained on an augmented dataset improved the performance by 2.97% (from 83.99% to 86.96%) on the SEED dataset. Luo et al. proposed using a conditional boundary equilibrium GAN (cBEGAN) to generate artificial differential entropy features from original EEG data, eye movement data and their concatenations for multimodal emotion recognition. The main advantage of this approach is that the proposed GAN has good stability and a very quick convergence speed [27]. Luo et al. proposed three methods for augmenting EEG training data to enhance the performance of emotion recognition models, including a conditional Wasserstein GAN, selective variational autoencoder, and selective WGAN [25]. They trained SVM and deep neural networks on original and augmented training datasets. The experimental results showed that the augmented training datasets enhanced the performance of EEG-based emotion recognition model.

Although many efforts have been made, the research on data augmentation for emotion recognition is far from complete. For example, while a human can easily determine whether images in an augmented dataset, e.g., a dataset containing cat images or other images, still resemble the original class, the same is not true of augmented signals. How to measure the quality and diversity of augmented samples and synthesize high-quality and diverse augmented samples deserves further exploration.

3 Proposed Method
3.1 Overall Framework
This paper designs a generative adversarial network-based self-supervised emotion recognition (GAN-SER) framework for EEG-based emotion recognition. Illustrated in Fig. 1a, the proposed framework is comprised of two networks, the adversarial augmentation network (AAN) and the multi-factor training network (MTN). Taking real EEG samples as input, the AAN is first designed to synthesize high-quality and diverse augmented EEG samples. Then, the EEG-based emotion recognition classifier can be learned on the augmented EEG samples and the self-supervised learning is finished under the guidance of the proposed MTN. In the remainder of this section, we will detail the network architectures of the AAN and MTN proposed in this paper.

3.2 Adversarial Augmentation Network
In the AAN, the masking transformation operation is first proposed to randomly cut out part of the data points of the given EEG. Then, a GAN is used to synthesize the EEG signals fitting the real data distribution based on the remaining EEG signals. By restoring the missing data points of the EEG signals, the proposed GAN is able to recognize the feature distribution of the source EEG signals and introduce new data points to generate new EEG signals.

Specifically, given a 32-channel EEG signal with a one-second length (sampled to 128 Hz) with a size of 32 × 128, we first follow the preprocessing of an existing work [33] to apply baseline removals, measuring the differences between the baseline signals and the given signal. Then, the results of 32 channels are transformed into mesh-like maps, guided by the existing work [34, 35].

For raw EEG signals, a single EEG channel is a one-dimensional (1D) time series, and multiple EEG channels correspond to a chain-like set of 1D time series. In this way, the EEG channel has two adjacent electrodes at most, structurally limiting the connections between brain regions. To solve this problem, we transform the chain-like 1D EEG channel set to a two-dimensional (2D) mesh-like EEG signal representation by mapping the EEG recordings with the position of the EEG signal acquisition electrode. In particular, the 2D mesh size is empirically selected based on the international standard for electrode placements. For the experiment in this paper, we follow the 10-20 system covering all EEG channels to map the electrode mapping on the 9 × 9 matrix. The mesh points not allocated for the EEG channel are assigned zero values. As a result, the given EEG signal can be denoted by \( e \in \mathbb{R}^{128 \times 9 \times 9} \).

As we described before, we design the masking transformation operation to cut out the partial signal values of the given \( e \) and force the following GAN to restore the missing parts, fitting the remaining information to include potential real-like samples different from the input EEG signal. In detail, we first randomly sample a matrix \( r \in \mathbb{R}^{128 \times 9 \times 9} \) of the same size as \( e \) with a uniform distribution \( U \sim [0, 1] \), and utilize \( r \) as the probability matrix representing the probabilities of signal values being masked. Then, the parameter \( \tau \) is sampled from the uniform distribution \( U \sim [\tau_{\text{min}}, \tau_{\text{max}}] \) as the threshold to determine which data point should be masked. In this way, the obtained EEG signals \( \delta(e, \tau) \) transformed from \( e \) based on the threshold \( \tau \) can be defined by:

\[
\delta(e_{ijk}, \tau) = \begin{cases} 
    e_{ijk} , & r_{ijk} > \tau \\
    0 , & r_{ijk} \leq \tau 
\end{cases}
\] (1)

Here, a large \( \tau \) means random masking ignores larger parts of the signal values of the source EEG samples. In this case, the feature distribution of the source EEG signals is difficult to preserve due to the limited remaining signal values. As a result, we can avoid learning an identity mapping and produce simulated EEG signals different from the original signals, ensuring the diversity of the augmented samples. Conversely, the generated EEG signal may be similar to the source EEG signal forced to fit the distribution of given signals. So, in this paper, we utilize \( \tau \) as an augmentation factor to represent the augmented sample’s diversity and difference from the original sample.
Then, based on \( \delta(e, \tau) \), we design a GAN to synthesize simulated EEG samples and ensure that the generated EEG signals fit the feature distribution of the real samples. Unlike the masking transformation operation, which focuses on integrating a diversity of synthesized EEG samples at the signal value level, the proposed GAN is responsible for learning the distribution of realistic EEG signals at the feature level. In this way, the generated augmented samples are further forced to preserve the natural features of real samples. Finally, realistic and diverse samples can lead to better classification performances for emotion recognition.

The designed GAN is composed of two networks, a generator \( G \) and a discriminator \( D \), optimized to minimize a two-player min-max game. The generator \( G \) is trained to generate the simulated EEG sample \( G(\delta(e, \tau)) \), taking the EEG signal \( e \) as input. The discriminator \( D \) is required to decide whether the given EEG signals are simulated or real, while \( G \) learns to fool the discriminator and tries to make the simulated samples as close facsimiles to the real ones. Due to the instability problems of the traditional training procedures of the GANs, in contrast to the GANs proposed for emotion recognition, we utilize a modified version of the Wasserstein GAN gradient penalty, i.e., WGAN-GP [36], for combining adversarial supervision and the random masking augmentation strategy. The loss function of \( G \) can be formulated as:

\[
L_G = -E_{e \sim P_e}[D(G(\delta(e, \tau)))],
\]

where \( P_e \) denotes the distribution of the given real EEG signals, and \( e \) represents an EEG signal sample from it. The goal of \( D \) is to minimize the loss function illustrated in:

\[
L_D = E_{e \sim P_e}[D(G(\delta(e, \tau)))] - E_{e \sim P_e}[D(e)] + \lambda_p E_{e \sim P_e}[(\|\nabla D(e)\|_2 - 1)^2],
\]

\( P_e \) is defined as sampling uniformly along straight lines between pairs of points sampled from the data distribution \( P_e \) and the generator distribution among \( G(\delta(e, \tau)) \). The gradient of the discriminator \( D \) is denoted by \( \nabla D(\hat{e}) \), and \( \lambda_p \) is a hyperparameter denoting the weight of the penalty term. In this way, the Wasserstein distance is used to compare the distributions of the generated samples and real samples, where the Lipschitz-continuous map ensures the property.

---

\( P_e \) is defined as sampling uniformly along straight lines between pairs of points sampled from the data distribution \( P_e \) and the generator distribution among \( G(\delta(e, \tau)) \). The gradient of the discriminator \( D \) is denoted by \( \nabla D(\hat{e}) \), and \( \lambda_p \) is a hyperparameter denoting the weight of the penalty term. In this way, the Wasserstein distance is used to compare the distributions of the generated samples and real samples, where the Lipschitz-continuous map ensures the property.

---

**Fig. 1.** The overall framework of generative adversarial network-based self-supervised emotion recognition (GANSER).
of a uniformly continuous distribution. This design can limit the normal of the derivation from growing too large [36]. Utilizing Eqs. (2) and (3), and are optimized in turn. By optimizing the adversarial loss, and can distinguish the real distribution from the simulated distribution, while and improves the ability to construct samples closer to the real EEG signals.

Regarding the network architectures of and , unlike the existing studies that focus on generating feature representations or single-channel EEG signals, this paper aims to address the data augmentation problem for general emotion recognition methods and to synthesize high-resolution EEG samples.

For the generator , we aim to synthesize missing signal values of given EEG signals to augment new EEG samples. In this case, one of the natural ideas is to utilize an autoencoder to reconstruct EEG signals directly in a downsample and upscale fashion. However, EEG signals contain numerous details reflected by a time-series variance. It is challenging for existing autoencoder-based networks to generate low-distorted EEG signals because the downsampling of the autoencoder applied to high-resolution time sequences can lead to missing details, and the reconstructed signals can be smoothed. To address this bottleneck, in this paper, we integrate an adapted version of U-Net on the EEG signals as the proposed generator for EEG signal synchronization.

As shown in Fig. 1b, the proposed U-Net consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical architecture of an encoder convolutional neural network to extract and downsample the EEG signals and extract feature maps. It consists of the repeated application of one -convolutional layer, two -convolutional layers, and another -convolutional layer, each followed by a leaky rectified linear unit (LeakyReLU). At each convolutional layer, we set the stride to one and halve the number of feature channels. Meanwhile, the expansive path follows a decoder convolutional neural network architecture with skip connections to upsample feature maps to a high spatiotemporal resolution to generate new EEG samples based on the extracted features. Every step in the expansive path consists of a concatenation with the corresponding feature map from the contracting path and one convolutional layer with a stride of one, doubling the number of feature channels. Each convolutional layer is followed by the activation function. After the final convolutional layer, the preliminary synthesized results, with a size of , can be obtained. Furthermore, this paper finds that the EEG signals are sparse on the spatial dimension due to the limited number of electrodes placed on the cap, which creates challenges for U-Net to synthesize real-like EEG signals. Specifically, given a-channel EEG signal, we transform the channels into maps according to the location of the electrodes. In the locations where electrodes do not exist, signal values are unknown or unmeasured, and the signal values are represented as zero. Forcing the U-Net to fit unmeasured signal values in the locations where electrodes do not exist, signal values can go to 0.0. The generator is required to predict mutated low signal values into the original spatially dense and continuous signals. This produces a situation that varies with objective reality and inevitably affects the modeling of measured signal values made to fit the real distribution of the original EEG signals.

Thus, this paper first introduces the channel masking operation to improve the ability of U-Net to synthesize EEG signals. As the first step, we propose building the channel mask , a prior binary mask with a size of , and setting the values where electrodes exist; signals are measured to one while defining the other locations to zero. Then, we apply elementwise multiplication between the designed prior mask and the output of U-Net, i.e., the synthesized EEG signals, to artificially reset the signal values where electrodes do not exist to zero. In this way, the proposed generator only needs to focus on fitting the signal values where electrodes exist and can neglect the mutation of EEG signal values caused by the nonexistent electrodes.

As the last part, illustrated in Fig. 1b, we design a novel network architecture, STNet, to analyze the complex spatiotemporal features of the EEG signals and utilize STNet as the discriminator . In detail, the designed STNet is comprised of three two-dimensional convolutional layers, a separable convolutional layer and an inception block. As the first step, the input EEG signal is provided as the input to convolutional blocks to extract low-resolution features. The first convolutional block uses a convolutional layer, while the second and third convolutional blocks use a convolutional layer. Each convolutional layer halves the number of input feature channels, and the stride of all the convolutional layers is set to one. Following each convolutional layer, a scaled exponential linear unit (SELU) is utilized as an activation function. Then, because the recognition of specific emotions is only related to local patterns of spatial features or temporal features, we introduce a separable convolutional layer [37] to decouple the modeling of the spatiotemporal information. In detail, the separable convolutional layer factorizes a standard convolutional layer into a convolution and a pointwise convolution and splits the computation into two steps, a depthwise convolution and a pointwise convolution. The depthwise convolution applies a single convolutional filter per input channel, and the pointwise convolution is used to create a linear combination of the output of the depthwise convolution. The stride of both the depthwise convolution and pointwise convolution is set to one, and the pointwise convolution is followed by SELU. In this way, the separable convolutional layer can capture the spatial correlation and temporal correlations of the extracted feature maps. Recognizing a pattern of emotions requires analyzing EEG signals at different spatial scales, and thus, we introduce an inception block [38] containing three types of filters of different sizes to extract multiscale feature maps. In detail, this paper combines a convolutional layer, a convolutional layer and a convolutional layer to utilize their output filter banks concatenated into single feature maps forming the input of the next stage. At each convolutional layer, we set the stride to one and halve the number of feature channels. Finally, the feature maps extracted by convolutional layers of each EEG signal are reshaped to a vector feature, and two linear layers are utilized to map the feature vector to a scalar. The first linear layer has nodes and is followed by the SELU activation function, while the second linear layer has one node to predict the category of a given EEG signal.
3.3 Multi-Factor Training Network

By optimizing the adversarial loss, the trained generator of AAN can produce augmented samples varying in signal details but fitting the feature distribution of the real samples. For the next step, we train classifier $C$ and fine-tune the trained classifier, further utilizing the learned generator of AAN to generate augmented samples.

In this stage, finding ways to utilize augmented samples for supervision is a crucial problem. High valence (arousal) and low valence (arousal) are far from being clear-cut and distinguishable by an artificial threshold. Thus, if the shift of augmented EEG varies by wide limits, the threshold between high valence (arousal) and low valence (arousal) can be exceeded, and the augmented EEG would change to a different category of the original EEG. To resolve this bottleneck, we seek a self-supervised learning framework to supervise emotion recognition training based on augmented samples and uncertainty labels. It is important to acknowledge that in the field of computer vision, self-supervised learning frameworks already allow for a reasonable performance without the acquisition of large training sets and well-labeled training samples. For example, Dosovitskiy et al. [39] proposed learning a network to discriminate between a set of surrogate classes formed by applying various transformations to a randomly sampled “seed” image patch. Then, by learning to classify different transformed samples of seed images into the same surrogate categories, the proposed network can extract discriminative features favoring better classification performances.

Thus, inspired by the significant progress made by self-supervised learning, this paper proposes the MTN for EEG-based emotion recognition. As described before, the augmented samples are synchronized based on parts of the original EEG signal values, and thus, the augmented samples should preserve the feature distribution of the original samples to some extent, although they are not the same. Based on these observations, in a contrast to existing studies, which directly create a set of surrogate classes, this paper designs a set of surrogate confidences, measured by an augmentation factor $\tau$, learning to restrict the feature distribution difference between real samples and augmented samples under a given surrogate confidence. Specifically, in the case where the augmentation factor $\tau$ is large, limited signal values of augmented samples remain from the source EEG signal, and the generator $G$ cannot capture and preserve the feature of the original EEG signal during synchronization. Thus, the feature distribution of the augmented EEG signals should be constrained to fit the original EEG signals’ feature distribution with a low confidence. Conversely, if the augmentation factor $\tau$ is small, most original EEG signal values are preserved during augmentation. We should force the feature distribution of augmented samples close to the original samples with high confidence. Finally, we propose a multifactor self-supervised learning loss to assign different weights for restricting the feature distribution difference between augmented EEG signals and real samples based on the corresponding surrogate confidence. Combining the cross-entropy loss for supervising real samples’ training as usual, the total loss function can be formulated as:

$$L_C(\tau) = -\frac{1}{n} \sum_{i=0}^{n} y_i \log(C(\hat{e}_i)) + \frac{\lambda_u}{n} \sum_{i=0}^{n} (1 - \tau)[\|C_2(G(\delta(e_i, \tau_i))) - C_2(e_i)\|^2_2],$$

where $\lambda_u$ is the hyperparameter utilized to represent the importance of classifying augmented samples correctly.

In Eq. (4), the cross-entropy between the ground truth labels and the corresponding prediction results of the real EEG signals are formulated as $-\frac{1}{n} \sum_{i=0}^{n} y_i \log(C(\hat{e}_i))$. The feature distribution difference between the original EEG signals and the corresponding augmented signals is denoted by $\|C_2(G(\delta(e_i, \tau_i))) - C_2(e_i)\|^2_2$. Based on the augmentation factor $\tau_i$, we compute the surrogate confidence $(1 - \tau_i)$ to assign large weights to different augmented samples to narrow the feature distribution difference when most of the original EEG signal values are preserved in the augmented sample.

It is worth noting that existing data augmentation methods mostly provide augmented samples in an offline fashion, generating a preset number of samples and saving them as training samples for the first step. Then, during optimization, augmented samples are loaded and fed into models with the original samples. In this paper, inspired by self-supervised learning approaches, we attempt to explore an alternative strategy. On the one hand, the separated stages of augmentation and training are joined together as an end-to-end pipeline. Given a batch of samples, we first utilize the generator $G$ in AAN to augment the EEG signals and pair the real signals. Then, we utilize these samples to optimize the classifier $C$ with Eq. (4) in the current batch. By avoiding the cost of saving and reloading, the proposed method is more efficient. On the other hand, in this paper, augmented samples are regenerated between epochs in runtime. Benefiting from the randomness of the masking transformation operation, the augmented samples of the corresponding batch between epochs are different but both are sampled near the real distributions. In this way, instead of overfitting on a preset number of augmented samples, randomly synthesized EEG signals can approximate the distribution of real EEG signals without a number limitation. Finally, the proposed multifactor self-supervised learning loss can adaptively guide the feature extractor to extract distinguishable feature representations for simulated EEG signals and generalize the classifier to the augmented sample space.

Regarding the network architecture of $C$, because the designed STNet shows good effectiveness in EEG analysis, the classifier $C$ shares the design of STNet as the discriminator $D$, except for several modifications to fit the classification problem formulation of emotion recognition. In detail, as shown in Fig. 1b, the dashed box represents operations that exist in the classifier $C$ but do not exist in the discriminator $D$. In the classifier $C$, we employ the dropout [40] operation with $p = 0.5$ after the activation function of the convolutional layers to address the problem of overfitting. Here, $p$
denotes the probability of an element being zeroed. The number of nodes in the last linear layer is set to two or four, corresponding to the categories of emotions. Moreover, an additional softmax function is set to follow the last linear layer as the activation function to produce the final outputs representing a categorical distribution.

4 Experimental Results and Discussion

In this section, we validate our method on the DEAP [41], DREAMER [42] and SEED [43] benchmarks to compare our results to the current state-of-the-art methods and present and discuss our results. Details of the experimental settings are provided. Qualitative and quantitative analyses on augmented samples are performed. Ablation studies of GANSER are detailed.

4.1 Implementation Details

In this paper, we use PyTorch [44] to implement our networks based on eight NVIDIA Tesla V100 GPUs. For the networks $G$, $D$ and $C$, this paper adopts the Adam optimizer [45] to minimize the loss functions. Here, the coefficient $\beta_1$ used for computing running averages of the gradient is set to 0.9, and the coefficient $\beta_2$ used for computing running averages of the square is 0.99. In addition, the weight decay of the L2 penalty is set to 0.0005, and the batch size is 64. Table 1 reports other hyperparameters utilized in the proposed GANSER. Here, $lr_G$ denotes the learning rate of generator $G$, $lr_D$ corresponds to the learning rate of discriminator $D$ and $lr_C$ represents the learning rate of classifier $C$. In addition, $\lambda_p$ is a hyperparameter representing the weight of the penalty term in Eq. (3), and $\lambda_a$ denotes that the hyperparameter in Eq. (4) is utilized to represent the importance of classifying augmented samples correctly. The parameter $r$ is sampled from the uniform distribution $U \sim \left[ r_{\text{min}}, r_{\text{max}} \right]$. Moreover, in the proposed GANSER framework, the AAN is first trained to generate augmented EEG samples, and the MTN is then introduced to train the classifier for the emotion recognition task. Thus, as shown in Table 1, we report all hyperparameters utilized in these two stages.

| Stage | Epoch | $lr_G$  | $lr_D$  | $lr_C$  | $\lambda_p$ | $\lambda_a$ | $r_{\text{min}}$ | $r_{\text{max}}$ |
|-------|-------|---------|---------|---------|-------------|-------------|-----------------|-----------------|
| AAN   | 300   | 0.00001 | 0.00001 | -       | 1.0         | -           | 0.0             | 0.5             |
| MTN   | 300   | -       | -       | 0.00001 | -           | 0.5         | 0.5             | 0.9             |

For the EEG signals of each trial, the two preprocessing steps pre-given by the DEAP dataset are first employed. The recorded EEG signals are first downsampled to a 128 Hz sampling rate. Then, the obtained EEG signals are processed with a bandpass filter from 4 Hz to 45 Hz to remove physiological and power frequency noises [41]. In each trial of the preprocessed dataset, the contained EEG signals consist of a three-second-long baseline signal recorded in the relaxed state and a 60-second-long experimental signal recorded under stimulation. Furthermore, we use a nonoverlapping sliding window to separate the trial data into one-second-long chunks and construct the separated EEG signals as data samples. The sliding window size is set to 128 to separate one-second chunks under a sampling rate of 128 Hz. For the next step, to reduce the effect of the basic emotional state, following existing work, we remove a mean baseline value from each segment [57]. Finally, the total number of EEG samples from 40 trials is $40 \times 60 = 2400$.

In terms of the emotional rating value of each trial in the range of 1.0 to 9.0 in the arousal and valence domains, a median of 5.0 was used as the threshold to divide the rating value into two categories. In the cases where the emotional rating value is rated more than 5.0, the corresponding EEG signals are labeled as representing high arousal or valence. In contrast, for those less than or equal to 5.0, this paper labels the corresponding EEG signals as representing low arousal or valence. Finally, given EEG signals to predict corresponding labeled categories, the emotion recognition task is formulated as a binary classification problem.

In the experiments, data samples in the DEAP dataset are split into five folds at random, and a fivefold cross-validation is used to evaluate all the models. To evaluate our proposed method, we first utilize 80% of the randomly shuffled data samples as training data to train the AAN for 300 epochs. In the following step, we fix the AAN parameters and then use the pretrained AAN to generate augmented samples. Then, the proposed classifier is learned on each fold for 300 epochs, and we supplement the augmented samples generated by AAN for the fine-tuning of 300 epochs with the help of MTN. Finally, the fine-tuned models are utilized for evaluation. To assess the overall performance, the average classification accuracies over five folds are reported.

Illustrated in Table 2, we first compare our proposed GANSER with nine state-of-the-art studies on the DEAP dataset with respect to the emotion dimensions, including valence and arousal. These studies develop different network architectures and strategies for emotion recognition. For example, CNNLSTM [47] combined convolutional neural networks and long short-term memory networks to extract distinguishable features, and MMResLSTM [56] further utilized multimodal information to improve the classification performance.

| Method         | Valence | Arousal |
|----------------|---------|---------|
| AAN            | 0.85    | 0.78    |
| MTN            | 0.88    | 0.82    |
| GANSER         | 0.92    | 0.88    |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
In particular, TSCeption [49] is the latest-of-the-art of EEG-based emotion detection and attracts much attention in the community. However, they only report the algorithm’s performance on their self-collected dataset. Thus, following the official implementation from GitHub,1 we reproduce the introduced TSCeption and provide experimental results on the public dataset DEAP. The only change is that the sampling rate is changed from 256 Hz to 128 Hz, and the 4-second time window is changed to a 1-second time window. From Table 2, we find that the proposed method outperforms all state-of-the-art studies on both the valence and arousal dimensions. Although the designed classifier requires lightweight training parameters and only consists of convolutional layers and linear layers, the proposed method shows an exceptional classification performance of over 93% for two-dimensional classification tasks and considerably outperforms the second-best method by a margin of nearly 1.0%. The comparison results demonstrate the effectiveness of our proposed methods for EEG-based emotion recognition.

Furthermore, to verify the capability of our proposed method in the field of data augmentation for emotion recognition, we further compare GANSER with several GAN-based data augmentation frameworks. In particular, following the experimental setting of existing GAN-based methods [25], [52], we formulate the emotion recognition task as a four-category classification problem, which aims at distinguishing between the EEG signals of four categories: high valence and high arousal, high valence and low arousal, low valence and high arousal and low valence and high arousal. In Table 2, it can be found that the proposed GANSER significantly outperforms existing GAN-based data augmentation frameworks by a large margin of over 8.0%. We also find that even in the formulation of four-category classification, our proposed method can correctly classify nearly 90% of the EEG signals at valence and arousal dimensions simultaneously due to the well-designed AAN and MTN.

### 4.3 Subject-Dependent Verification on DEAP Dataset

In the abovementioned experiment, we conducted a five-fold cross-validation over all the samples from the DEAP dataset in a variation from per-subject five-fold cross-validation settings, which split samples from a single subject into five folds for evaluating and reporting the average performance of the subjects. To further compare these methods fairly, we also report the average performance of per-subject five-fold cross-validation settings in the DEAP dataset. Here, the EEG signals of every given subject were randomly shuffled, following which they were divided into five folds. One fold served as the test set in turn, and the other four folds were used as the training set. Finally, the average classification accuracies over five folds were recorded as a subject’s performance, and we reported the average performance of all subjects in Table 3.

As shown in Table 3, we obtained the average classification accuracies of the proposed GANSER over 32 subjects, which was 93.86% at the valence dimension and 94.38% at

---

1. https://github.com/deepBrains/TSception

---

| Method | Acc. (%) |
|--------|----------|
|        | Valence  | Arousal  |
| SOTA   |          |          |
| cross-session split | CDCN [47] | 92.24 | 92.92 |
| 10-fold | PCRN [34] | 90.8 | 91.03 |
| random split(8:2) | CNN-LSTM [48] | 90.62 | 86.13 |
| 5-fold | 3D-CNN [16] | 87.44 | 88.49 |
| unknown | Merged1STM [49] | 84.89 | 83.85 |
| 5-fold | TSCeption [50] | 84.29 | 84.71 |
| 5-fold | C-RNN [15] | 72.06 | 74.12 |
| 5-fold | SFMS-Net [51] | 71.25 | 71.33 |
| 5-fold | ScalingNet [52] | 71.13 | 69.99 |
| GAN-based |  |          |
| 10-fold | MCLFS-GAN [53] |  | 81.32 |
| 5-fold | sWGAN [26] |  | 49.10 |
| Proposed | GANSER | 93.52 | 94.21 |

---

**TABLE 2**

| Method | Valence | Arousal |
|--------|---------|---------|
| SOTA   | BDAE [53] | 85.20 | 80.50 |
|        | LSTM-RNN [54] | 85.45 | 85.65 |
|        | ECLGCNN [55] | 90.45 | 90.60 |
|        | MM-ResLSTM [56] | 92.87 | 92.30 |
| Proposed | GANSER | 93.86 | 94.38 |

---

**TABLE 3**

| Method | Valence | Arousal |
|--------|---------|---------|
| SOTA   | BDAE [53] | 85.20 | 80.50 |
|        | LSTM-RNN [54] | 85.45 | 85.65 |
|        | ECLGCNN [55] | 90.45 | 90.60 |
|        | MM-ResLSTM [56] | 92.87 | 92.30 |
| Proposed | GANSER | 93.86 | 94.38 |
the arousal dimension. Compared with recent state-of-the-art methods, such as MM-ResLSTM [48], which achieve accuracies of 92.87% and 92.30%, the proposed method shows better performance with a margin of approximately 1-2%, demonstrating the effectiveness of the proposed methods under different experimental settings.

### 4.4 Experiments of Sample Size Effect on DEAP Dataset

The current framework also permits generating a fixed number of augmented examples in advance and providing pre-augmented EEG signals and the original EEG signals as the input to the models. Thus, in the following study, we conduct an experiment to further analyze the performance with varying numbers of augmented samples. We use the pre-trained generator in AAN to generate one, two and four augmented samples for every training sample to achieve one, two and four data augmentations. Then, in training epochs, we randomly sample the original signal-augmented signal pairs for self-supervised learning using Eq. (4). Finally, the classification accuracies over five folds are reported in Fig. 2 to demonstrate the performance gain and performance stability under different scales of data augmentation.

Here, the whiskers of the boxplots are the range of classification accuracies over five-folds, and the line inside the boxplots is the mean value of the classification accuracies. As shown in Fig. 2, we find that the performance increases with the increasing augmentation scales in terms of max accuracy, min accuracy and the mean value of the distribution, demonstrating that more augmentation signals can help further improve the overall performance.

### 4.5 Ablation Studies on DEAP Dataset

We further conduct an ablation study to investigate the performance gain brought by each key component in our model, including the GAN, the masking transformation operation, and the multifactor self-supervised learning loss (MSL), and provide the performance of the stripped-down versions by removing these components one by one.

When the MSL is ablated, the augmented samples are treated as an equivalent of the given training samples in order to optimize the cross-entropy based on the labels of the original EEG signals. In this way, we can evaluate the performance gain of the self-supervised learning design. Then, we further ablate our proposed model to evaluate MT and GAN. As we described before, MT is designed to mask parts of the EEG signals integrated for the diversity of the synthesized EEG samples at the signal value level. In contrast, GAN is utilized to force the generated EEG signals to approximate the distribution of real samples, ensuring the quality of the augmented samples. Thus, the ablation study here is used to evaluate the performance gain of the diversity-driven designed MT and the quality-driven designed GAN. For the ablated version of GANSER without MT, we use the original EEG signal \( e_{\text{ orig}} \) instead of the masked EEG signal \( \delta(e_{\text{ orig}}, r) \), as input to the generator \( G \). Then, the GAN is utilized to learn the generator to reconstruct new EEG signals based on the given EEG signal \( e \) and force the generated EEG signals \( G(e) \) to approximate the real distribution. Here, the discriminator \( D \) is required to decide whether the given EEG signals are simulated or real, while \( G \) learns to fool the discriminator and tries to make simulated samples close to the real ones. While the GAN is further removed, no data augmentation operation is employed, and the classifier is fine-tuned based on the given training samples. Finally, we follow the above experimental setting to train and evaluate stripped-down versions of our approach and report the average classification accuracies of different models in Table 4.

| Method | Valence | Arousal | Training Time (s) |
|--------|---------|---------|-------------------|
| GAN    | 91.39   | 92.22   | 0.028             |
| ✓      | 92.27   | 92.76   | 0.047             |
| ✓✓     | 93.28   | 93.12   | 0.047             |
| ✓✓✓    | 93.52   | 94.21   | 0.047             |

It has been shown that our proposed classifier is a strong and effective baseline even in the case where no data augmentation is utilized, and the classifier is fine-tuned based on the original training dataset. The proposed method can achieve performances of 91.39% and 92.22% at the valence and arousal dimensions, comparable to the latest-of-the-art designs based on complex network architectures or applying multimodal information. The design of both the GAN and masking transformation operation brings a performance gain of approximately 1.0% and obviously improves the classification accuracy of EEG signals. This phenomenon demonstrates that with the help of designed components, the framework synthesizes simulated EEG signals to favor learning the EEG signal patterns for emotion recognition. Finally, the additional Multi-factor Self-supervised Learning loss succeeds in enhancing the classification performance to over 93.5% by further considering the uncertainty of the augmented samples’
labels and providing further guidance for model training in a self-supervised learning mode.

The average training time on a batch of samples is also reported while utilizing the pretrained AAN to fine-tune the classifier. We find that the ablated method without GAN obtains a training time reduction of approximately 0.019 seconds compared with other methods, meaning that the data augmentation procedure only takes approximately 0.02 seconds to simulate a batch of samples. We also find that with the masking transformation (MT) operation and the multifactor self-supervised learning (MSL) loss, we can obtain a performance gain without significant extra time consumption.

### 4.6 Qualitative and Quantitative Analysis of Augmented Samples on DEAP Dataset

While the overall framework is introduced to improve emotion recognition accuracies, the well-designed adversarial augmentation network (AAN) is particularly crucial to synthesize real-like and diverse EEG signals. Thus, in this section, qualitative and quantitative experiments are carried out to assess the quality of the synthesized EEG samples and demonstrate the contribution of our designed AAN for EEG signal stimulation.

In terms of qualitative experiments, it is known that judging the quality and diversity of samples generated by GAN-based methods is a challenging task due to the difficulty of measuring the distribution difference between the real samples and generated samples. Following the existing work in the computer vision field, this paper adapts the Fréchet STNet distance (FSTD) to assess the quality of the synthesized EEG samples and demonstrate the contribution of our designed AAN.

In detail, we first train two proposed STNets on 80% of the training samples to distinguish different emotions at the assessed dimension, e.g., the valence dimension. Then, we utilize the learned STNet as the feature extractor and provide EEG samples as input to extract the output of the penultimate fully connected layer. Finally, we compute the sample means $\mu$, $\mu_g$, and the sample covariance matrices $\Sigma$, $\Sigma_g$ of the real samples and generate the samples' feature distributions, and the FSTD is then the Wasserstein distance between the two multivariate normal distributions $N(\mu, \Sigma)$ and $N(\mu_g, \Sigma_g)$:

$$FSTD = \|\mu - \mu_g\|^2 + \text{Tr} \left( \Sigma - \Sigma_g - 2\sqrt{\Sigma \Sigma_g} \right),$$

where a small FSTD value indicates a high similarity between the generated samples and the real data distribution.

Illustrated in Table 5, FSTD scores of different stripped-down versions of our approach are compared in terms of the valence and arousal dimensions. The average FSTD scores are also provided to assess the overall performance. To explore the contribution of each design in the network architecture, we use the above experimental settings to train and evaluate different combinations of the UNet network architecture (UNet), the channel masking operation and the STNet architecture (STNet). Each ablated combination removed one of the components; particularly when UNet is removed, we utilize an autoencoder of the same parameter volume as an alternative. After eliminating the channel masking operation, the output of UNet is directly utilized as the generated EEG signals. While STNet is ablated, we replace the separable convolutional layer and the inception block with convolutional layers to produce output feature maps of the corresponding size.

As shown in Table 5, if any one of the proposed components is removed, the average FSTD scores will rise. This means that every proposed module contributes to the quality and diversity of the augmented samples, without which the feature distribution of generated samples would be different from real samples. In particular, we also find that STNet affects the FSTD score to the greatest extent in both the valence and arousal dimensions. Because it was favored by the designed separable convolutional layer and the inception block, STNet can model the complex spatiotemporal feature distribution of the EEG signals and force the generated samples to fit the real feature distribution. The lack of a channel masking operation also leads to a relatively sharp increase in FSTD scores in the valence dimension. This phenomenon is because the difference between channels in the EEG signal is crucial for valence recognition, and the channel masking operation can reduce the difficulty of synthesizing the difference between channels by giving prior information. Furthermore, arousal analysis is highly related to the signal value scale. In these cases, UNet preserves the modeled scale information to the greatest extent and thus improves the quality of the augmented samples.

The average training time on a batch of samples is also reported while utilizing the pretrained AAN to fine-tune the classifier. Compared with ablated methods, the proposed method not only improves the quality of the augmented samples but also shows a slight increase in training time while fine-tuning the classifier, demonstrating that the

---

**TABLE 5**

| Method   | FSTD score | Training Time (s) |
|----------|------------|-------------------|
|          | Valence    | Arousal | Average |          |
| UNet     | ✓          | ✓       | 96.78 | 80.81 | 88.80 | 0.045 |
| CM       | ✓          | ✓       | 103.62 | 72.26 | 87.94 | 0.047 |
| STNet    | ✓          | ✓       | 278.76 | 286.63 | 282.70 | 0.047 |
|          | ✓          | ✓       | 49.61 | 36.61 | 43.11 | 0.047 |

The Accuracies of Different Stripped-Down Versions of Our Approach to Evaluate the Performance Gain of the UNet Network Architecture (UNet), the Channel Masking (CM) Operation, and the STNet Architecture (STNet). The Average Training Time on a Batch of Samples is Provided.
proposed network architectures are efficient and straightforward for performance improvements.

To further highlight the data augmentation performance of our proposed framework, in this section, we carry out a visualization experiment on the DEAP dataset. Specifically, the experimental setting in this section is consistent with that in Section 4.2. After training is completed, the learned AAN is utilized to synthesize the EEG signals based on the given data samples. The original EEG signals and the generated simulated EEG signals are provided in Fig. 3. In detail, EEG signals of different categories are sampled for visualization, including low valence and high arousal (first column) and high valence and low arousal (second column). For each case, the original real EEG signals are presented in the first row, the second row depicts the EEG signals synthesized by our proposed method, and the third row shows the EEG signals generated by an ablated version of our approach that replaces the UNet network architecture with an autoencoder. Since the raw data of EEG signals are not conducive to visually representing the EEG signals’ characteristics, the topographic maps of EEG signals sampled at 0.0 s, 0.25 s, 0.5 s and 0.75 s are provided, where red denotes high values, and blue represents low values. As shown in Fig. 3, the activated areas shown by the real EEG signals and the corresponding simulated symmetric EEG signals proposed by our method are almost consistent at the spatial and temporal dimensions, demonstrating that the generated data have similar spatiotemporal data distributions as the real data. Meanwhile, we also observe a slight shift in the activation area and a change in the activation degree in the augmented EEG samples, indicating that our methods could synthesize diverse simulated samples and can avoid identity mapping. In contrast, we found that the EEG signal samples generated by the ablated model preserve the limited spatiotemporal features of the original EEG signals. It is also difficult to distinguish between different categories of samples, indicating that the GAN may suffer from mode collapse without a specially designed network.

In addition to sample-level visualization in the experiments mentioned above, we additionally conducted qualitative experiments to evaluate the generated EEG signals at the feature level to analyze the low-dimensional distribution of the augmented samples.

For the first step, we select a fold of test data in a five-fold cross-validation experiment of Section 4.2 and randomly sample 2,000 test EEG signals from the DEAP dataset. Then, we employ the optimized generator of AAN to simulate augmented EEG signals, providing 2,000 test EEG signals as input. For the next step, we use the STNet trained on the training data folds to describe the features of the EEG signals, where the output of the last convolution layer is utilized as the features of the EEG samples. Here, the valence features are extracted by the STNet to classify valence categories, while the arousal features are extracted by the STNet to classify the arousal categories. In this way, the features of the original EEG signals and the augmented EEG signals can be obtained. Then, due to the high-dimensional feature vectors of the extracted features, we utilize principal component analysis (PCA) to reduce the dimension of the extracted feature maps to 50. Finally, T-SNE [63] is applied to embed the outputs of PCA into two-dimensional space after employing a min-max normalization to normalize the two dimensions. We plot a low-dimensional representation of the EEG signals on a graph with different colors referring to the corresponding valence and arousal labels.

As shown in Figs. 4 and 5, the features of the original EEG signals are marked by circles, while diamonds denote the augmented EEG signals. Most augmented EEG signals close to the clusters share the same original valence/arousal label, demonstrating that the generated EEG signals fit the original real data distribution. We find that the generated EEG signals are distributed in the feature space in a large range, demonstrating the diversity of the generated EEG signals and interpolating the sparse distribution as too dense to determine the complex boundary of two categories.

4.7 Emotion Recognition Results on SEED Dataset

We then evaluate our proposed method on the widely used EEG-based emotion recognition dataset SEED [43]. The SEED dataset contains the EEG data of 15 subjects when they watched 15 4-minute-long Chinese movies, and 62 EEG electrodes are recorded via the 10-20 system. After watching the

![Fig. 3. A few qualitative results are showing the real EEG signals (the first line), augmented EEG signals (the second line) synthesized by the proposed method, and augmented EEG signals (the third line) produced by the proposed method w/o UNet.](image-url)
movie, participants are required to point out a keyword (positive, neutral, or negative) of the movie, and the average rating is calculated as the label of the film clip. Unlike the DEAP dataset, the SEED dataset also involves a session variance, where the data collected from each subject lasted for three different periods, corresponding to three sessions, with an interval of one week or more between the two sessions. Each session corresponds to 15 EEG data trials, and thus, a subject has 45 EEG data trials.

For the SEED dataset, the preprocessing is similar to the DEAP dataset. For the EEG signal of each trajectory, the two preprocessing steps given in the SEED dataset are first adopted. We first downsample the recorded EEG signal to a 200 Hz sampling rate. Then, the obtained EEG signal is processed by a 0.3-50 Hz bandpass filter to remove any physiological and power frequency noise. For the next step, we use a nonoverlapping sliding window to divide the experimental data into 1-second-long data blocks and use the separated EEG signals as data samples. Here, the sliding window size is set to 200 to separate them into one-second chunks at a sampling rate of 200 Hz. In particular, the SEED dataset does not provide a baseline signal in each trial, and thus, we regard the signal in the first three seconds as the baseline signal. Finally, to reduce the effect of the basic emotional state, we remove a mean baseline value from each segment following the procedures of the existing studies [57].

According to the emotional score value of each trial, all EEG signals are classified into one of three emotional states, positive, neutral, or negative, according to the label of the watched video film. The final classification can be considered a three-class classification task. Without changing the network architecture and other configurations, we directly implement our proposed method on the SEED dataset under the five-fold cross-validation setting and report the performance of our proposed method. Notably, the contained EEG data in the SEED dataset are collected via 62 EEG electrodes, and the EEG signals of 62 channels instead of 32 in the DEAP dataset are transformed into 9 maps according to the electrode locations based on the international 10-20 system. Thus, the representation of EEG signals in the SEED dataset is denser. Then, experiments are carried out to compare our proposed method with several state-of-the-art works.

Illustrated in Table 6, we compare our proposed GANSER with different state-of-the-art studies on the SEED dataset. These studies developed different network architectures and strategies for emotion recognition. In particular, Res-Net18-1D-(T-then-S) [58] and CNN+DE [60] are CNN-based methods reported under five-fold cross-validation,
which is naturally consistent with our results. DGCNN [9] does not report their performance on the 5-fold cross-validation but releases official implementation on Github.\(^2\) Thus, following the official implementation, we reproduce the DGCNN under the setting of 5-fold cross-validation and report the performance for comparison. From Table 6, we find that the proposed method outperforms all the state-of-the-art studies. Although the designed classifier requires lightweight training parameters and only consists of convolutional layers and linear layers, the proposed method shows a stellar classification performance of over 97% for three-class classification tasks. The comparison results demonstrate the effectiveness of our proposed methods for EEG-based emotion recognition.

Then, we further supplement the comparison experiments on the SEED dataset with other dataset splitting methods. We evaluate our proposed GANSER under the random training-test split setting following existing works [61], [62] while fixing other configurations. Here, 80% of data are selected as training, and 20% are used for test. The experiment result is reported in Table 6. It can be found our method shows a satisfactory performance of 97.71% under the random training-test split setting over the performance of 97.70% under 5-fold cross-validation. The experimental results further demonstrate that our proposed method is robust to different splitting methods and is able to archive the state-of-the-art no matter what splitting method is utilized.

### 4.8 Emotion Recognition Results on DREAMER Dataset

The DREAMER [42] dataset contains the EEG data of 23 subjects, which were collected via 14 EEG electrodes from the subjects when they were watching 18 film clips. Each film clip lasted 65 seconds long to 393 seconds long. The length of each trial in the DREAMER dataset is different because each film clip lasts from 65 seconds long to 393 seconds long. As a result, we obtain a different number of EEG samples for each trial of the DREAMER dataset.

In terms of the emotional rating value of each trial, a median of 3.0 is used as the threshold to divide the trials according to the levels of valence and arousal. That is, the label is low when the rating is less than or equal to 3.0, and the label is high when the rating is greater than 3.0. In this way, the recognition task is actually a binary classification problem for each emotion dimension. Without changing the network architecture and other configurations, we directly implemented our proposed method on the DREAMER dataset and reported the performance of our proposed method as a baseline. Notably, the contained EEG data in the DREAMER dataset were collected via 14 EEG electrodes, and the EEG signals of 14 channels instead of 32 in the DEAP dataset are transformed into 9 × 9 maps according to the electrode locations based on the international 10-20 system. Thus, the representation of the EEG signals in the DREAMER dataset is sparser. Then, experiments were carried out to compare our proposed method with several state-of-the-art works.

In this experiment, following the experimental settings of existing works, data samples in the DREAMER dataset are split into ten folds at random, and a tenfold cross-validation is used to evaluate all the models. Notably, such an experimental setting is different from the configuration of existing studies reported on the DEAP dataset, where a five-fold cross-validation is generally used. To evaluate our proposed method, we first utilize 80% of the randomly shuffled data samples as training data to train the AAN for 300 epochs. Then, the proposed classifier is learned on training folds for 300 epochs, and we supplement the augmented samples generated by AAN with fine-tuning of 300 epochs with the help of MTN. Finally, the fine-tuned models are utilized for evaluation of the corresponding test folds. To assess the

| Method                  | Acc. (%) |
|-------------------------|----------|
| 5-fold                  | 93.42    |
| ResNet18-1D-(T-then-S)  | 90.41    |
| SOTA                    |          |
| 5-fold                  | 97.70    |
| CNN+DE [60]             | 96.68    |
| DGCNN [9]               | 96.67    |
| random split (8:2)      | 95.97    |
| SAE-DNN [61]            | 97.77    |
| random split (8:2)      | 97.71    |
| Proposed random split (8:2) | 97.70 |

For the EEG signals of each trial, we used a nonoverlapping sliding window to separate the trial data into one-second-long chunks and constructed the separated EEG signals as data samples. The sliding window size is set to 128 to separate the data into one-second chunks under a sampling rate of 128 Hz. For the next step, to reduce the effect of the basic emotional state, following the processes in existing studies, we removed a mean baseline value from each epoch [57].

\(2.\) https://github.com/xueyunlong12589/DGCNN
overall performance, the average classification accuracies over ten test folds are reported.

Illustrated in Table 7, we first compare our proposed GANSER with nine state-of-the-art studies on the DREAMER dataset with respect to the emotion dimensions, including valence and arousal. These studies developed different network architectures and strategies for emotion recognition. For example, Zhang et al. [14] proposed a spatiotemporal recurrent neural network (STRNN) to investigate both the spatial and temporal dependencies of EEG signals and achieved state-of-the-art performance. In the experiment, the STRNN [14] is implemented on the DREAMER dataset to see how much the designed RNN architecture can improve the discriminant ability. Furthermore, the parallel convolutional recurrent neural network (PCRNN) reported by [33] is compared, which introduces baseline signals into the preprocessing and proposes a hybrid neural network combining CNN and RNN to learn the compositional spatial-temporal features of the EEG signals. In [66], the authors proposed a 3D representation of the EEG segment to combine features of signals from different frequency bands while preserving the spatial information among the channels. Then, the performance of the decision tree (DT) and the multilayer perceptron (MLP) is reported to demonstrate the discriminant ability of the proposed EEG representation. Furthermore, the authors of [66] introduced the continuous CNN (ContCNN) to utilize the combination of features of multiple bands to complement each other and achieve comparable results. The reported performance of DT, MLP and ContCNN are all considered for the comparison experiment. From Table 7, we find that the proposed method outperforms most state-of-the-art studies on both the valence and arousal dimensions. In particular, although the designed classifier requires lightweight training parameters and only consists of convolutional layers and linear layers, the proposed method shows a stellar classification performance of approximately 85% for two-dimensional classification tasks and considerably outperforms all the state-of-the-art methods in the valence dimension. In the experiment, we conduct 10-fold cross-validation on the total samples in the DREAMER dataset that is different from the experiments applying 10-fold cross-validation on every subject and report the average performance of all the subjects.

### TABLE 7

| Method | Acc. (%) | Valence | Arousal |
|--------|----------|---------|---------|
| 10-fold BioCNN [64] | 55.94 | 81.41 |
| 10-fold STRNN [14] | 70.80 | 80.30 |
| 10-fold CRAM [65] | 73.37 | 74.71 |
| 10-fold DT [66] | 75.53 | 75.74 |
| 10-fold PCRNN [33] | 79.93 | 81.48 |
| 10-fold DGCNN [9] | 80.64 | 80.15 |
| 10-fold gForest [67] | 81.42 | 82.83 |
| 10-fold MLP [66] | 83.64 | 83.71 |
| 10-fold ContCNN [66] | 84.54 | 84.84 |
| Proposed | 10-fold GANSER | 85.28 | 84.16 |

5 LIMITATIONS

Some recent works [10], [21], [68] of EEG-based emotion recognition have considered individual independent generalization abilities. Following the experimental setting in [21], we conducted an individual independent evaluation to evaluate our proposed method. As shown in Table 8, it is important to acknowledge that the proposed method achieves trivial improvements under the individual independent setting. This is because the current classifier, STNet, lacks a special design to model the distribution shift between the subjects. In our follow-on work we plan to extend our current work for simulating different subjects’ EEG signals, permitting the modification of subject identification information while preserving the emotion features to generate augmented samples. As a result, we could train robust classifiers with simulated samples and recognize cross-subject patterns without modifying the design of the classifier.

### TABLE 8

| Method | Acc. (%) | Valence | Arousal |
|--------|----------|---------|---------|
| SOTA | SE-LeNet [68] | 66.23 | 68.50 |
| | WGANDA [21] | 67.99 | 66.85 |
| | ATDD-LSTM [10] | 69.06 | 72.97 |
| | ScalingNet [51] | 71.65 | 71.32 |
| Proposed | Before Aug. | 48.72 | 54.54 |
| | After Aug. | 49.36 | 55.12 |

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel generative adversarial network-based self-supervised data augmentation (GANSER) framework consisting of an adversarial augmentation network (AAN) and a multifactor training network (MTN). By learning to synthesize real-like EEG signals based on collected EEG datasets, the proposed method resolves the data scarcity bottleneck for EEG-based emotion recognition. In a variation from existing studies, this paper first proposes combining adversarial training with self-supervised learning to ensure the quality and diversity of the augmented samples. The design of the AAN employs the masking transformation operation to mask parts of the EEG signals and introduces uncertainties to control the diversity of the augmented samples. Additionally, we force a well-designed GAN to generate high-quality and high-diversity simulated EEG samples where the UNet with a channel masking operation and STNet are employed to capture the complex spatiotemporal features of the EEG signals. Furthermore, to effectively utilize simulated EEG signals, we introduce MTN, where a multifactor self-supervised learning loss is proposed to utilize the masking possibilities of the masking transformation operation as prior knowledge and guide the feature extraction process of the simulated EEG signals for generalizing the classifier for the augmented sample space.

By applying the designed framework on the benchmark datasets for emotion recognition, DEAP, SEED, and
DREAMER, the experimental results show that the designed model can exploit the natural features of real EEG signals to synthesize high-quality and diverse simulated EEG signals and ultimately improve the classification performance. In the future, we will seek to explore a data augmentation framework in which the influence of subject-independent variance is further controlled and modified.

REFERENCES

[1] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, “Affective computing and sentiment analysis,” in A Practical Guide to Sentiment Analysis. Berlin, Germany: Springer, 2017, pp. 1–10.

[2] M. Minsky, The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind, New York, NY, USA: Simon & Schuster, 2007.

[3] P. Zhong, D. Wang, and C. Miao, “EEG-based emotion recognition using regularized graph neural networks,” IEEE Trans. Affect. Comput., to be published, doi: 10.1109/TAFFC.2020.2994199.

[4] Y. Roy, H. Banville, A. Albuquerque, A. Gramfort, T. H. Falk, and J. Fahrbert, “Deep learning-based electroencephalography analysis: A systematic review,” J. Neural Eng., vol. 16, no. 30, 2019, Art. no. 501001.

[5] E. Lashgari, D. Liang, and U. Maoz, “Data augmentation for deep-learning-based electroencephalography,” J. Neurosci. Methods, vol. 346, 2020, Art. no. 108885.

[6] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, “Adaptive data augmentation for image classification,” in Proc. IEEE Int. Conf. Image Process., 2016, pp. 3688–3692.

[7] X. Liu, X. B. Zhai, W. Lu, and C. Wu, “QoS-guarantee resource allocation for multibeam satellite industrial Internet of Things with noma,” IEEE Trans. Ind. Informat., vol. 17, no. 3, pp. 2052–2061, Mar. 2019.

[8] X. Liu and X. Zhang, “Rate and energy efficiency improvements for connected IoT with simultaneous transfer,” IEEE Internet Things J., vol. 6, no. 4, pp. 5971–5980, Aug. 2019.

[9] T. Song, W. Zheng, P. Song, and Z. Cui, “EEG emotion recognition using dynamical graph convolutional neural networks,” IEEE Trans. Affect. Comput., vol. 11, no. 3, pp. 532–541, July–Sept 2020.

[10] X. Du et al., “An efficient LSTM network for emotion recognition from multichannel EEG signals,” IEEE Trans. Affect. Comput., to be published, doi: 10.1109/TAFFC.2020.3051332.

[11] W. Tao et al., “EEG-based emotion recognition via channel-wise attention and self attention,” IEEE Trans. Affect. Comput., to be published, doi: 10.1109/TAFFC.2020.3025777.

[12] H. Becker, J. Fleureau, P. Guillot, F. Wendling, I. Merlet, and L. Albera, “Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources,” IEEE Trans. Affect. Comput., vol. 10, no. 2, pp. 244–257, Apr. –Jun. 2019.

[13] G. Zhang, M. Yu, Y.-F. Liu, G. Zhao, D. Zhang, and W. Zheng, “SparseDGCNN: Recognizing emotion from multichannel EEG signals,” IEEE Trans. Affect. Comput., to be published, doi: 10.1109/TAFFC.2021.3051332.

[14] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, “Spatial-temporal recurrent neural network for emotion recognition,” IEEE Trans. Cybern., vol. 49, no. 3, pp. 839–847, Mar. 2019.

[15] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, “Emotion recognition from multi-channel EEG data through convolutional recurrent neural network,” in Proc. IEEE Int. Conf. Bioinf. Biomed. Eng., 2016, pp. 352–359.

[16] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. W. Shalaby, “EEG-based emotion recognition using 3D convolutional neural networks,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 8, pp. 329–337, 2018.

[17] S.-E. Moon, S. Jang, and J.-S. Lee, “Convolutional neural network approach for EEG-based emotion recognition using brain connectivity and its spatial information,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2018, pp. 2556–2560.

[18] S.-E. Moon and J.-S. Lee, “Madenet: Disentangling individuality of EEG signals through feature space mapping and detachment,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 2020, pp. 2661–2665.

[19] T. Zhang, X. Wang, X. Xu, and C. P. Chen, “GCN-net: Graph convolutional broad network and its application in emotion recognition,” IEEE Trans. Affect. Comput., vol. 13, no. 1, pp. 379–388, Jan.–Mar. 2019.

[20] X. Gong, T. Zhang, C. L. P. Chen, and Z. Liu, “Research review for broad learning system: Algorithms, theory, and applications,” IEEE Trans. Cybern., pp. 1–29, Mar. 2021.

[21] Y. Luo, S.-Y. Zhang, W.-L. Zheng, and B.-L. Lu, “WGAN domain adaptation for EEG-based emotion recognition,” in Proc. Int. Conf. Neural Inf. Process. 2020, pp. 82–93.

[22] M. M. Krell and S. K. Kim, “Data augmentation for EEG-based emotion recognition with deep convolutional neural networks,” in Proc. Int. Conf. Multimedia Modell., 2018, pp. 1–10.

[23] Y. Luo, L.-Z. Zhu, Z.-Y. Wan, and B.-L. Lu, “Data augmentation for enhancing EEG-based emotion recognition with deep generative models,” J. Neural Eng., vol. 17, no. 5, 2020, Art. no. 56021.

[24] C. Szegedy, W. Liu, Y. Jia, Y.春天, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

[25] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, “Discriminative unsupervised feature learning with convolutional adversarial networks,” in Proc. Int. Conf. Neural Inf. Process. 2017, pp. 1251–1258.

[26] S. Wang, S.-H. Zhong, J. Peng, J. Jiang, and Y. Liu, “Data augmentation for EEG-based emotion recognition with deep convolutional neural networks,” in Proc. Int. Conf. Neural Inf. Process. 2019, pp. 1–8.

[27] S. Koelstra et al., “Deap: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices,” IEEE J. Biomed. Health Inform., vol. 22, no. 1, pp. 98–107, Jan. 2018.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
et al. (2020) described a convolutional recurrent attention model for subject-independent EEG signal classification.

H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel (2021) presented BioCNN: A hardware inference engine for EEG-based emotion detection.

D. Zhang, L. Yao, K. Chen, and J. Monaghan (2019) proposed a convolutional recurrent attention model for subject-independent EEG signal analysis.

Y. Luo et al. (2020) investigated the role of empirical mode decomposed EEG signals responsive to music.

Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma, and G. Chen (2020) explored the application of LSTM-based deep convolutional neural networks for emotion recognition.

Z. H. Cheah, H. Nisar, V. V. Yap, C.-Y. Lee, and G. Sinha (2020) analyzed the impact of deep residual networks and VGG for classification of EEG signals using LSTM networks.

J. Cheng et al. (2018) developed an emotion recognition method based on generative adversarial networks.

C. Shahnaz et al. (2018) utilized multi-channel EEG via deep forest.

Y. Luo et al. (2017) investigated the use of generative adversarial networks for EEG-based emotion recognition.

Y. Yang et al. (2016) presented a convolutional recurrent attention model for subject-independent EEG signal classification.

D. A. K. Bedi, R. K. Sunkaria, and A. Garg (2015) explored the use of deep recurrent attention models for emotion recognition.

S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi (2015) described an emotion recognition model using LSTM recurrent neural network.

J. Liu et al. (2015) proposed a multi-reservoirs EEG signal feature sensing and recognition method based on generative adversarial networks.

W. Liu, W.-L. Zheng, and B.-L. Lu (2014) examined the use of emotion recognition using multimodal deep learning.

S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi (2014) introduced an emotion recognition model using LSTM recurrent neural network.

H. A. Gonzalez, S. Muzaffar, J. Yoo, and I. M. Elfadel (2013) presented BioCNN: A hardware inference engine for EEG-based emotion detection.

L. Van der Maaten and G. Hinton (2008) explored the use of t-SNE for visualization of high-dimensional data.