Beyond Mimicking Under-Represented Emotions: Deep Data Augmentation with Emotional Subspace Constraints for EEG-Based Emotion Recognition

Zhi Zhang[1,2], **Shenghua Zhong**[1,*], **Yan Liu**[2]

[1] Shenzhen University, the College of Computer Science and Software Engineering, Shenzhen, 518052, Guangdong, China
[2] the Hong Kong Polytechnic University, the Department of Computing, Hong Kong, 999077, China
zhi271.zhang@connect.polyu.hk, csshzhong@szu.edu.cn, yan.liu@polyu.edu.hk

Abstract

In recent years, using Electroencephalography (EEG) to recognize emotions has garnered considerable attention. Despite advancements, limited EEG data restricts its potential. Thus, Generative Adversarial Networks (GANs) are proposed to mimic the observed distributions and generate EEG data. However, for imbalanced datasets, GANs struggle to produce reliable augmentations for under-represented minority emotions by merely mimicking them. Thus, we introduce Emotional Subspace Constrained Generative Adversarial Networks (ESC-GAN) as an alternative to existing frameworks. We first propose the EEG editing paradigm, editing reference EEG signals from well-represented to under-represented emotional subspaces. Then, we introduce diversity-aware and boundary-aware losses to constrain the augmented subspace. Here, the diversity-aware loss encourages a diverse emotional subspace by enlarging the sample difference, while boundary-aware loss constrains the augmented subspace near the decision boundary where recognition models can be vulnerable. Experiments show ESC-GAN boosts emotion recognition performance on benchmark datasets, DEAP, AMIGOS, and SEED, while protecting against potential adversarial attacks. Finally, the proposed method opens new avenues for editing EEG signals under emotional subspace constraints, facilitating unbiased and secure EEG data augmentation.

Introduction

With the rapid development of deep learning technology, artificial intelligence systems have developed fantastic perception and computing capabilities. Nonetheless, it is critical to enable AI to communicate emotionally with humans to meet their emotional and psychological needs. To this end, affective computing strives to develop systems that bridge the gap between human emotions and machine intelligence (Tao and Tan 2005). In recent years, researchers have increasingly focused on analyzing the EEG of different emotions, applying these research results to emotional artificial intelligence.

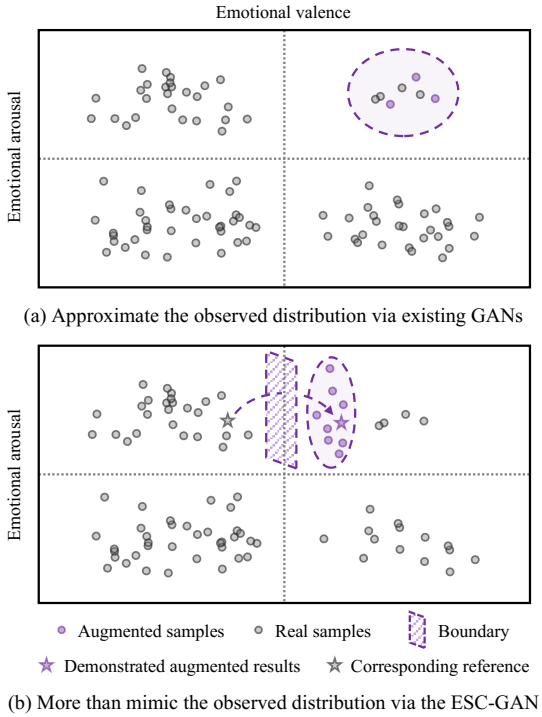
With the advancement of deep learning in image and natural language processing, attention is being paid to its application in EEG-based emotion recognition (Song et al. 2020; Zhang et al. 2020; Zhao, Yan, and Lu 2021; Tripathi et al. 2017; Thammasan, Fukui, and Numao 2017). However, a significant challenge is that deep learning algorithms require extensive training data to obtain high-performance models (Li et al. 2021). Compared with image and natural language data, EEG data acquisition is labor-intensive and time-consuming, resulting in limited available EEG data (Wei et al. 2015; Lin and Jung 2017), which restricts the potential of deep learning algorithms (Hu et al. 2019).

To tackle this problem, data augmentation is widely applied. Traditional data augmentation methods, such as geometric transformation and noise addition, are often argued to be unsuitable for EEG signals (Bao et al. 2021). If we rotate or shift the EEG data, the feature of time domain will be destroyed (Wang et al. 2018). If the EEG signal is noisy, the amplitude and data distribution of the original signal will be changed. (Bao et al. 2021).

As an alternative, Generative Adversarial Networks (GANs) have garnered widespread attention for their ability to mimic the observed distributions and generate real-like augmented data (Bao et al. 2021; Luo and Lu 2018). GANs depend heavily on extensive, diverse, and high-quality training examples (Zhao et al. 2020). However, several studies that used public datasets are observed to have a high imbalance of emotional data (Wirawan, Wardoyo, and Lelono 2022). For example, the AMIGOS dataset (Santamaria-Granados et al. 2018) is skewed toward higher arousal ratings, while the DEAP dataset (Koelstra et al. 2011) contains a higher proportion of positive emotions compared to negative ones (Wang et al. 2022). We concur that current GANs can generate real-like samples for well-represented majority emotions. However, for underrepresented minority categories, merely mimicking observed distributions can be problematic, potentially leading to issues, e.g., mode collapse (Zheng et al. 2020). As depicted in Fig. 1(a), if an emotion category lacks sufficient data, mimicking the observed distribution will increase sample density but not fulfill potential emotional subspaces like well-represented ones, hindering the development of unbiased emotion recognition systems.

To overcome this challenge, we propose a novel paradigm for augmenting EEG signals in emotion recognition, as depicted in Fig. 1(b). (1) We propose sampling EEG signals from the majority category and using them as starting points for generating new data. These EEG signals are considered

*Corresponding author.



(a) Approximate the observed distribution via existing GANs

(b) More than mimic the observed distribution via the ESC-GAN

Figure 1: The paradigm differences between existing GAN-based methods and the proposed Emotional Subspace Constrained Generative Adversarial Networks (ESC-GAN).

reference EEG signals, providing information from various collected samples. We then learn to edit reference EEG signals to a well-constrained emotional subspace for augmenting under-represented emotions. (2) To encourage a diverse emotional subspace for under-represented distributions, we propose enlarging the difference between augmented samples, preventing the augmented EEG signals from becoming too homogeneous. (3) Inspired by related work (Karimi and Tang 2020), deep neural networks (DNNs) are susceptible to erroneous instances near decision boundaries (Pei et al. 2017; Karimi and Tang 2020). Therefore, we propose prioritizing the emotional subspace near the decision boundary to sample augmented signals and mitigate potential vulnerabilities.

Finally, the contributions of this work lie in three aspects: (1) We first propose the Emotional Subspace Constrained Generative Adversarial Network (ESC-GAN) as an alternative to existing GAN-based augmentation frameworks. (2) We first propose an editing paradigm that transforms reference EEG signals from well-represented emotions to under-represented emotional subspaces. Then, we introduce diversity-aware and boundary-aware losses to constrain the augmented subspace. Here, the diversity-aware loss encourages a diverse emotional subspace for under-represented distributions by enlarging the sample difference, while boundary-aware loss constrains the subspace near the decision boundary to defend against adversarial attacks. (3) Experiments demonstrate that employing ESC-GAN for data augmentation aids simple yet efficient classifiers in achieving satisfactory performance on the DEAP, AMIGOS, and SEED datasets. Moreover, ESC-GAN also yields benefits in defending against adversarial attacks.

Methodology

Model Architecture

This paper proposes a novel data augmentation framework for EEG-based emotion recognition, i.e., Emotional Subspace Constrained Generative Adversarial Network (ESC-GAN).

As shown in Fig 2, the ESC-GAN consists of a generator and two discriminators. The generator G is the core module of ESC-GAN. It takes the control signal t and the reference EEG signal e as input. Here, the control signal t refers to a user-specific emotional category, e.g., low arousal. For the reference EEG signal, existing studies (Russell and Mehrabian 1977; Mehrabian 1995, 1996; Russell 1980) show emotions can be described in nearly orthogonal dimensions, including valence and arousal. For under-represented emotions, we lack samples consistent with these emotions in both valence and arousal dimensions. Nevertheless, abundant EEG signals from well-represented emotions match under-represented emotions in one dimension. Thus, we select one such signal at random, designated as the reference EEG signal e, to serve as the starting point for augmentation.

Then, the generator G edits the reference EEG signal e to a new EEG signal with the user-specific emotional category t, thereby producing augmented samples. Specifically, G contains an encoder and a decoder. In the encoder, we utilize stacked convolutional layers to downsample the EEG signal and derive a compact feature representation. This representation is then combined with the control signal t and fed to the decoder. The decoder maps the feature representation according to t and upsamples the features to create a new EEG signal using stacked deconvolution layers.

The discriminators are designed to optimize the generator G in a zero-sum game with G. We present two discriminators: R, which distinguishes between generated EEG signals and real EEG signals, and C, which predicts the emotional category of a given EEG sample. The two discriminators share stacked convolutional layers that use strided convolutions to extract compact feature representations from EEG signals. The representations are then fed into unshared fully connected layers for predictions. Discriminator R utilizes its fully connected layers to distinguish real samples from generated samples, while discriminator C uses its own fully connected layers to classify the emotional category of the EEG signal.

Model Optimization

Then, we design loss functions to optimize the ESC-GAN as shown in Fig. 2, which includes controllable adversarial loss, diversity-aware loss, and boundary-aware loss.

Controllable Adversarial Loss For well-represented emotions, optimizing GAN to learn the distribution and produce real-like samples poses no issues because these emotions involve relatively adequate EEG signals to learn gener-

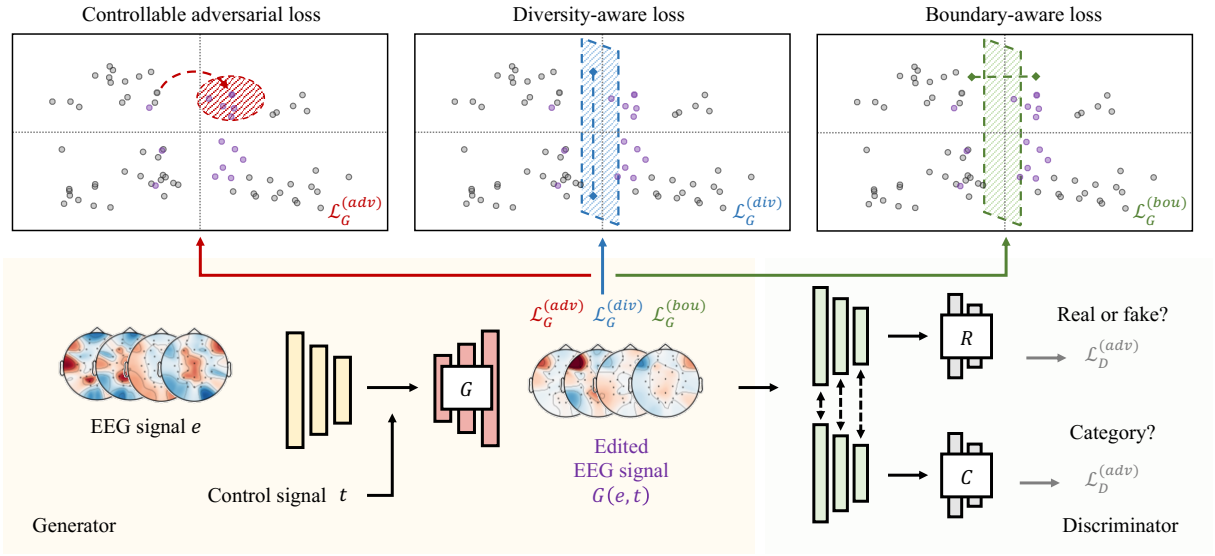


Figure 2: The overall framework of Emotional Subspace Constrained Generative Adversarial Networks (ESC-GAN).

ative models. In this situation, we introduce controllable adversarial loss to approximate the real distribution via adversarial training while learning a mapping to the target emotional category.

In detail, the discriminator R is trained to distinguish between real and generated samples, while the discriminator C is optimized to classify the emotional categories corresponding to the EEG signals. Mathematically, the loss function of R and C can be formulated as:

$$\begin{aligned}\mathcal{L}_D^{adv} &= \mathbb{E}_{e\sim\mathbb{P}_e}\left[\mathcal{H}_{\text{cross}}\left(R(e),0\right)\right] \\ &+ \mathbb{E}_{e\sim\mathbb{P}_e}\left[\mathcal{H}_{\text{cross}}\left(R(G(e,t)),\tau\right)\right] \\ &+ \mathbb{E}_{e\sim\mathbb{P}_e}\left[\mathcal{H}_{\text{cross}}\left(C(e),y\right)\right]\end{aligned} \tag{1}$$

where the first two terms, formed by the cross-entropy loss $\mathcal{H}_{\text{cross}}(.,.)$, serve to differentiate between real samples e and fake samples $G(e,t)$ for R. In the last term, the cross-entropy loss is employed for C to classify the given EEG signals e. To prevent the model from making overconfidence predictions, we use the one-side label smoothing technique (Salimans et al. 2016) to smooth the labels of fake samples with the label τ $(0.5 < \tau < 1)$. In this way, R leans toward exhibiting moderate confidence in detecting generated samples and producing low cross-entropy loss during adversarial training. Thus, it avoids the instability of the optimization process and helps the convolutional layers extract stable features for classification.

The generator aims to fool the discriminator R into classifying the generated samples as real EEG signals and to force the discriminator R to classify the generated samples into the target emotional category. The loss function of G can be formulated as follows:

$$\begin{aligned}\mathcal{L}_G^{adv} &= \mathbb{E}_{e\sim\mathbb{P}_e}\|e - G(e,y)\|^2 \\ &+ \mathbb{E}_{e\sim\mathbb{P}_e}\left[\mathcal{H}_{\text{cross}}\left(R(G(e,t)),0\right)\right] \\ &+ \mathbb{E}_{e\sim\mathbb{P}_e}\left[\mathcal{H}_{\text{cross}}\left(C(G(e,t)),t\right)\right]\end{aligned} \tag{2}$$

where the first term denotes the L2 distance between the generated and real EEG signals. It means that if the control signal t is consistent with the category y of the given EEG signal, the generator will generate EEG signals similar to the given EEG signals. The second term aims at fooling R into classifying the generated samples as real EEG signals with cross-entropy loss. The last term is to produce EEG signals classified as the target emotional category t. Through adversarial training, the generator will eventually approximate the real distribution with the pattern of the target emotional category.

Diversity-aware Loss For under-represented emotions involving inadequate EEG signals, optimizing generated EEG signals to mimic their observed distributions may result in similar samples near the limited observed samples. The homogeneous samples hinder the augmented samples from extending under-represented emotions to build an unbiased decision boundary. Thus, we introduce an additional loss function to encourage a diverse emotional subspace to enhance the diversity of under-represented emotional distributions.

Here, we propose a new metric to measure the difference between augmented samples in the emotional subspace and maximize the difference between sample pairs to encourage the generator to generate diverse samples:

$$\mathcal{L}_G^{div} = \mathbb{E}_{e_i,e_j\sim\mathbb{P}_e,t_i,t_j\sim\mathbb{P}_t} e^{\frac{-|G(e_i,t_i)-G(e_j,t_j)|}{|F(G(e_i,t_i))-F(G(e_j,t_j))|}} \tag{3}$$

In detail, in mini-batch training, EEG signal sample pairs (e_i, e_j) are first sampled from a batch of samples. Subsequently, we employ the generator G to edit the pair of samples based on the provided control signal pair (t_i, t_j), yielding the generated EEG samples $(G(e_i,t_i), G(e_j,t_j))$. Here, drawing inspiration from the Radial Basis Function (RBF) kernel (Schölkopf 2000; Lin and Chen 2011) (as known as "squared exponential" kernel), we utilize an RBF-like measurement to compute the difference between the two generated EEG signals mapped in an infinite-dimensional space, to capture complex relationships between samples that might not be apparent in the original feature space.

Moreover, considering that different classes of EEG signals with distinct patterns should exhibit a substantially larger numerical difference than EEG signals sharing the same patterns, we maximize the ratio of sample difference ($|G(e_i,t_i) - G(e_j,t_j)|$) to feature difference ($|F(G(e_i,t_i)) - F(G(e_j,t_j))|$), between pairs of generated samples. As a result, the generator is optimized to highlight the difference in terms of the generated EEG signals according to the level of feature difference. Finally, the diversity-aware loss results in stronger diversity of the generated EEG signal samples to cover potential distribution.

Boundary-aware Loss Recent studies (Possas and Zhou 2017) have shown that adversarial attacks are more severe on imbalanced datasets. Thus, we propose to augment samples to improve the robustness against potential attacks for underrepresented emotions. Inspired by related work, DNNs are vulnerable to erroneous instances near their decision boundaries (Pei et al. 2017; Karimi and Tang 2020), prompting us to constrain the augmented subspace near the decision boundaries. Then, we can optimize models to classify augmented samples near the decision boundary precisely to mitigate vulnerabilities. However, how do we measure whether a sample is near the classification boundary?

In this paper, we propose a new metric to leverage discriminator C to determine to measure the distance of samples to the decision boundary in the emotional space. For a binary classification problem, when the generated EEG signal is far from the classification boundary, C will classify the generated signal into one category with a high probability. Conversely, when the generated EEG signal is on the classification boundary, C will classify it into two categories with a nearly equal probability of 50%. Based on these observations, we design the boundary-aware loss that minimizes the KL divergence between C's predictions and a uniform distribution $U(t)$:

$$\mathcal{L}_G^{bou} = \mathbb{E}_{e\sim\mathbb{P}_e, t\sim\mathbb{P}_t} KL(U(t)||C(G(e,t))) \tag{4}$$

where $KL(P||Q)$ is the KL divergence between two distributions P and Q, which is defined as $KL(P||Q) = \sum P(i)\log\frac{P(i)}{Q(i)}$. In this way, the boundary-aware loss can prevent the model from producing the generated EEG signal, which can be classified into a single category with a higher probability, thereby nudging the generated samples toward the classification boundary. It is worth noting that due to the constraint of Eq. (2), the augmented samples will not be exactly on the boundary, avoiding the generation of noise samples.

Classifier Optimization

After optimizing the ESC-GAN, we utilize the generator to produce augmented samples for optimizing the emotion recognition classifier. To ensure the quality of the augmented samples, we use the discriminators R and C to filter the generated samples. Specifically, we retain the generated samples that meet the following criteria for use as augmented samples in classifier training:

$$\begin{aligned} &\text{Condition } (1) : R(G(e,t)) < \delta \\ &\text{Condition } (2) : \operatorname{argmax} C(G(e,t)) == t \end{aligned} \tag{5}$$

Condition (1) mandates that the discriminator R not classify the generated sample as fake with a confidence greater than δ. We set δ to 0.5. This ensures that the distribution of the augmented samples is close to the real distribution. Condition (2) requires that the discriminator C classify the generated samples into the target category t, ensuring that the augmented samples contain patterns corresponding to the control signal t.

During the training process of the classifier $f(.)$, we use both augmented samples $\tilde{e}$ and real samples. We adopt the mix-up technique (Zhang et al. 2018) to sample EEG signal pairs from augmented samples and real samples, mixing them in a ratio of γ. The mixed EEG signal samples and their respective labels are subsequently used to optimize the classifier through cross-entropy:

$$\begin{aligned} \mathcal{L}_f = &- \mathbb{E}_{(e_i,e_j,y_i,y_j)} (\gamma\cdot y_i + (1-\gamma)\cdot y_j) \\ &\log(f(\gamma\cdot e_i + (1-\gamma)\cdot e_j)) \end{aligned} \tag{6}$$

where $(e_i, e_j) \sim \mathbb{P}_{\{e\}\cup\{\tilde{e}\}}$ represents pairs of EEG signals sampled from augmented samples and real samples, and $(y_i, y_j) \sim \mathbb{P}_{\{y\}\cup\{t\}}$ represents the pairs of corresponding labels for the EEG samples. The discriminator C, proposed in this study, has been demonstrated as an effective model for recognizing patterns relevant to emotional categories in EEG signals during adversarial training. Thus, the classifier shares similar model architectures as C.

Experiments

We conduct experiments on three widely used benchmark datasets, the DEAP dataset (Koelstra et al. 2011), the AMIGOS dataset (Santamaria-Granados et al. 2018), and the SEED dataset (Duan, Zhu, and Lu 2013). In the experiments, we use the five-fold cross-validation technique to partition training and test datasets. In each experiment, one-fifth of the continuous EEG segments are selected from each trial, forming the test dataset, while the remaining samples conduct the training dataset.

Emotion Recognition Performance

We first conduct comparative experiments to demonstrate the overall performance of our proposed approach, comparing it with other generative adversarial network-based data augmentation techniques and recent state-of-the-art algorithms. In this section, we report the average emotion classification accuracy in both valence and arousal dimensions on the DEAP and AMIGOS datasets. We also report the average accuracy of classifying emotions as positive, neutral, or negative, on the SEED dataset. As shown in Table 1, our method outperforms GAN-based data augmentation techniques and other state-of-the-art algorithms on the DEAP and AMIGOS datasets. Specifically, our performance surpasses Zhang *et al.*'s (Zhang, Zhong, and Liu 2022) work by 2% in classifying valence and arousal on the DEAP dataset, illustrating the effectiveness of our proposed method.

Method		DEAP		AMIGOS		SEED
		Valence	Arousal	Valence	Arousal	
Recent state-of-the-art	Santamaria *et al.* (Santamaria-Granados et al. 2018)	-	-	75.00	76.00	-
	Siddharth *et al.* (Jung, Sejnowski et al. 2019)	71.09	72.58	83.02	79.13	-
	Topic *et al.* (Topic and Russo 2021)	76.61	77.72	87.39	90.54	88.45
	Hu *et al.* (Hu et al. 2021)	71.8	71.88	74.06	69.52	-
	Zhang *et al.* (Zhang, Zhang, and Wang 2022)	85.86	84.27	-	-	92.47
	Yang *et al.* (Yang et al. 2018)	90.26	90.98	-	-	-
	Tao *et al.* (Tao et al. 2020)	93.72	93.38	-	-	-
	Shen *et al.* (Shen et al. 2020)	94.22	94.58	-	-	94.74
	Feng *et al.* (Feng et al. 2022)	<u>95.04</u>	<u>95.52</u>	-	-	96.72
Imbalanced learning	Chawla *et al.* (Chawla et al. 2002)	94.41	94.65	<u>93.77</u>	94.38	95.85
	He *et al.* (He et al. 2008)	94.25	94.6	93.75	<u>94.42</u>	95.76
GAN-based augmentation	Luo *et al.* (Luo and Lu 2018)	73.89	78.17	-	-	86.96
	Liu *et al.* (Liu, Hao, and Guo 2023)	92.75	93.52	-	-	-
	Zhang *et al.* (Zhang, Zhong, and Liu 2022)	93.52	94.21	-	-	**97.7**
	Proposed	**96.33**	**96.68**	**94.40**	**95.23**	<u>97.14</u>

Table 1: Comparative experiments on the benchmark datasets DEAP and AMIGOS for classifying EEG signals in valence and arousal dimensions.

Method				DEAP	
				Valence	Arousal
w/o augmentation				94.17	94.75
reference paradigm	one-side smoothing	$\mathcal{L}_G^{\text{div}}$	$\mathcal{L}_G^{bou}$	Valence	Arousal
	✓	✓	✓	95.07	95.28
✓		✓	✓	96.11	<u>96.63</u>
✓	✓		✓	95.83	95.9
✓	✓	✓		<u>96.23</u>	96.11
✓	✓	✓	✓	**96.33**	**96.68**

Table 2: An ablation study on the DEAP dataset demonstrates performance gains brought about by the editing paradigm, one-side smoothing techniques in controllable adversarial loss $\mathcal{L}_G^{adv}$, diversity-aware loss $\mathcal{L}_G^{div}$, and boundary-aware loss $\mathcal{L}_G^{bou}$.

On the balanced SEED dataset, our model exhibits outstanding performance, ranking just behind the most recent state-of-the-art method, GANSER. Notably, without our proposed data augmentation framework, the model's performance on the SEED dataset is 95.22%. This indicates that our data augmentation framework contributes to a performance improvement of 1.92%, highlighting the generalization capability of our proposed model on balanced datasets.

Further, we apply the imbalanced learning techniques, SMOTE (Chawla et al. 2002) and ADASYN (He et al. 2008), under the same experimental setting. As shown in Table 1, these methods fall short compared to our proposed ESC-GAN, particularly on the DEAP dataset. It demonstrates compared to oversampling under-represented categories via linear combinations of observed data, ESC-GAN has advantages in modeling EEG signals and complementing imbalanced datasets.

We also conduct ablation studies to investigate the performance contribution of various designs in the proposed algorithm. First, we train a baseline classifier without using ESC-GAN augmented samples, while maintaining other settings. We then ablate various modules in the proposed algorithm, including the editing paradigm based on reference EEG signals, controllable adversarial loss, diversity-aware loss, and boundary-aware loss, retaining the others. For the ablation of the editing paradigm, we follow related work that generates EEG signal samples from random vectors, instead of reference EEG signals to produce new samples. To ablate the adversarial loss, we eliminate the one-side smoothing technique in controllable adversarial loss $\mathcal{L}_G^{adv}$, assessing the performance of the conditional adversarial generation network. When ablating the diversity-aware loss $\mathcal{L}_G^{div}$ and boundary-aware loss $\mathcal{L}_G^{bou}$, we remove the corresponding loss functions to optimize the generator. The performance of different ablated versions of the model is reported in Table 2.

It can be found that compared to the model performance without augmented samples, the ESC-GAN-based data augmentation results in approximately 2% performance gains, indicating the contribution of augmented samples. Then, in the ablated versions of the model that remove different modules, the model without the editing paradigm demonstrates the greatest performance decline. This phenomenon highlights the role of generating new EEG signals by editing reference EEG signals from well-presented emotions. Additionally, the removal of diversity-aware loss, boundary-aware loss, and one-side smoothing also leads to performance declines, demonstrating that these modules contribute to the final recognition performance, underscoring the importance of generating diverse augmented samples near classification boundaries.

Analysis on Augmented EEG Signals

In this subsection, we conduct experiments to investigate the quality, diversity, and distribution of samples generated by ESC-GAN. Following GAN-based related work for EEG data augmentation (Zhang, Zhong, and Liu 2022), we employ Kernel Maximum Mean Discrepancy (MMD) (Gretton et al. 2006), Inception Score (IS) (Salimans et al. 2016), and Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate the augmented samples generated by the optimized ESC-GAN.

Method			Kernel MMD ↓		IS ↑		FID ↓	
smooth loss	diversity loss	boundary loss	Valence	Arousal	Valence	Arousal	Valence	Arousal
	✓	✓	3.36	3.22	1.47	<u>1.49</u>	1000.28	974.43
✓	✓		**3.01**	**2.89**	**1.61**	1.48	**838.91**	**842.01**
✓		✓	4.67	35.77	1.19	1.21	1619.59	2745.02
✓	✓	✓	<u>3.15</u>	<u>3.11</u>	<u>1.52</u>	**1.50**	<u>947.87</u>	<u>921.52</u>

Table 3: Ablation experiments on the DEAP dataset demonstrate performance gains brought about by various modules in terms of the quality and diversity of augmented samples.

In detail, we report the MMD, IS, and FID of augmented samples generated by the proposed ESC-GAN and compare them with ablated models. These ablated models include versions without one-side smoothing, diversity-aware loss, and boundary-aware loss. As shown in Table 3, small Kernel MMD and FID values represent better quality and diversity of generated samples, while the opposite is true for the IS metric. We utilize bold text to indicate the best model and underline the second-best model. From Table 3, we can observe that the indicators improve when removing boundary-aware loss. This is because the indicators access quality and diversity of samples by comparing them with the observed dataset, while the boundary-aware loss does not contribute to approximating the observed distribution. Instead, the boundary-aware encourages the augmented sample near the decision boundaries to improve the classification ability of systems near the boundaries (to defend against adversarial attack). It leads to a drop in quality due to the difference from the observed data. However, jointly analyzing Table 2 and Table 3, the boundary-aware loss helps improve classification performance, demonstrating the inconsistency between the objectives of data augmentation and EEG signal simulation (to generate real-like EEG signals). Moreover, we can also observe that when removing diversity-aware loss and one-side smoothing loss, the quality and diversity of augmented samples decline. This phenomenon indicates that diversity loss contributes to the diversity of augmented samples, while one-side smoothing loss positively affects the stability of adversarial optimization, thus leading to higher-quality samples.

Emotion Recognition Performance against Hard Cases

To gain further insight, we conduct hard-case experiments to investigate the ability of ESC-GAN to handle imbalanced datasets with a cold-start problem, where under-represented emotions are significantly missing. Here, we first divide the DEAP dataset into the training set and test set at a ratio of 80% to 20%. As the emotional space can be split into four quadrants under the valence-arousal emotional model (Russell 1980; Kanimozhi and Raj 2015) (high valence-high arousal, high valence-low arousal, low valence-high arousal, and low valence-low arousal), we take turns selecting each quadrant as a control quadrant to simulate the under-representation of emotion distribution by removing parts of training samples (50%-90%) in that quadrant. We use the remaining training samples as the training set to train ESC-GAN and use them for data augmentation. The performance of the classifier is recorded, and the average performance with different quadrants as the control quadrant is reported in Fig. 3.

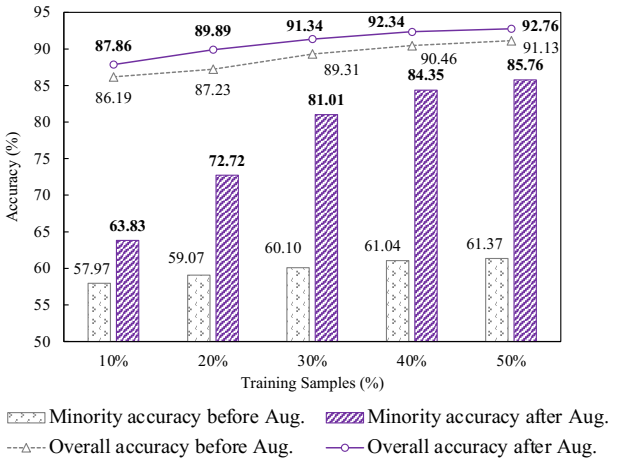


Figure 3: The hard-case experiments on the DEAP dataset by removing parts of training samples and keeping 10% to 50% training samples in the control quadrant.

Here, line plots denote the accuracy rate of the model classifying total category samples in the test set, while the bar plots report the accuracy rate of the model classifying test samples in the control quadrant, denoting the recognition performance on minority emotions. The proportion of training samples remaining in the control quadrant is taken as the horizontal axis. It can be found that regardless of the proportion of training samples retained, data augmentation can always help the model improve classification performance, although a lower proportion of training samples leads to worse classification performance. Compared with accuracy for the total category, the improvement brought by the proposed data augmentation is more evident in the minority emotions. The performance gain can even exceed a margin of 20%. It shows that the model effectively tackles the cold-start problem of minority emotions through data augmentation and finally contributes to the performance of the imbalanced dataset.

Emotion Recognition Performance against Adversarial Attacks

In this subsection, we further explore the ability of ESC-GAN-based data augmentation to defend against adversarial attacks. To our knowledge, the adversarial attack is consid-

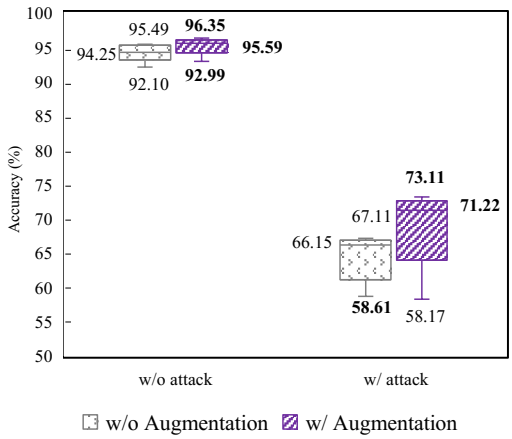


Figure 4: Performance of classifiers with and without ESC-GAN augmented samples in classifying emotional arousal on the AMIGOS dataset before and after adversarial sample attacks using the Fast Gradient Sign Method (Goodfellow, Shlens, and Szegedy 2014). The optimal performance is highlighted in bold.

ered one of the greatest threats deep learning models face. The attacker can deceive the attacked model into making incorrect predictions by operating specific perturbations on the input samples, while the permutation is imperceptible to humans. Nowadays, adversarial attack already poses a security risk for brain-computer interface applications. For example, in BCI-based driver drowsiness estimation, adversarial attacks may make a drowsy driver look alert, increasing the risk of accidents (Wu et al. 2021). Recent studies further show that adversarial attacks are more severe on imbalanced datasets (Possas and Zhou 2017), which is the case encountered in EEG-based emotion recognition, inspiring us to pay attention to prevent potential security problems.

In this subsection, we conduct experiments to explore whether boundary-aware augmented samples contribute to a robust classification boundary and defend against adversarial sample attacks. In detail, experiments are conducted on the AMIGOS dataset with the typical Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014) as the attack algorithm. We first perform five-fold cross-validation on the AMIGOS dataset. Following the above-mentioned setting, we optimize ESC-GAN to augment the data and train the model to complete the emotion recognition task in classifying emotional arousal. Then, we use the FGSM with ϵ set to 0.01 to attack the trained models with adversarial samples separately.

We report the classification accuracy of the models with and without adversarial attack in Fig. 4. The whiskers of boxplots are the range of accuracies over five-fold cross-validation experiments, and the line inside the boxplots denotes the medium value of classification accuracies. We can find that regardless of attack, the models trained with augmented samples achieved better performance than those trained without augmented samples, indicating the stable

Adversarial Attacks	w/o Aug.	w/ Aug.
w/o attack	94.25	**95.59**
w/ FGSM attack	66.15	**71.22**
w/ PGD attack	63.3	**68.44**
w/ Deep Fool attack	1.24	**4.80**

Table 4: Experiments on the AMIGOS dataset show performance under various adversarial attacks with (w/) and without (w/o) data augmentation.

contribution of the ESC-GAN data augmentation method to defend against the adversarial attack.

In line with the experimental setting mentioned above, we further conduct emotion classification experiments on the AMIGOS dataset under other attack algorithms. For other hyperparameters during these attacks, we set ϵ to 0.01 for both FGSM and PGD attacks, set α to 0.001, and specify the number of steps as 10 for PGD. For the Deep Fool attack, the overshoot is configured to 0.001, with the number of steps set to 10. As shown in Table 4, we observe that the performance of the emotion classifier deteriorates when exposed to more sophisticated attack methods—specifically, iterative attacks like PGD—as compared to FGSM. The results also indicate a consistent performance gain across all types of adversarial attack algorithms when the proposed ESC-GAN data augmentation method is employed. This finding demonstrates that constraining the augmented data subspace close to the decision boundary effectively aids in defending against a diverse array of adversarial attacks.

Conclusion and Future Work

We observe that collecting vast amounts of EEG signals encompassing all emotions is challenging, and there are imbalanced EEG datasets involving under-presented emotions. Current GANs can generate real-like samples for well-represented majority emotions; however, augmenting the minority class by merely mimicking observed distributions is problematic. To tackle this problem, we propose an alternative to existing frameworks, Emotional Subspace Constrained Generative Adversarial Networks (ESC-GAN). Experiments on DEAP, AMIGOS, and SEED datasets demonstrate the superior performance of our proposed method compared to state-of-the-art approaches. Further experimental results demonstrate the effectiveness of the ESC-GAN on imbalanced problems in case studies. Moreover, ESC-GAN shows superiority in defending against adversarial attacks. Finally, the proposed method opens new avenues for editing EEG signals under emotional subspace constraints, facilitating unbiased and secure EEG data augmentation.

Acknowledgements

This research was funded by Natural Science Foundation of Guangdong Province (2023A1515012685, 2023A1515011296), Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant No. GML-KF-22-28, and Artificial Intelligence and Robotics (AIR) Group: Artifical Intelligence Art (P0033738).

References

Bao, G.; Yan, B.; Tong, L.; Shu, J.; Wang, L.; Yang, K.; and Zeng, Y. 2021. Data Augmentation for EEG-Based Emotion Recognition Using Generative Adversarial Networks. *Frontiers in Computational Neuroscience*, 15.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Duan, R.-N.; Zhu, J.-Y.; and Lu, B.-L. 2013. Differential entropy feature for EEG-based emotion classification. In *Conference on Neural Engineering*, 81–84.

Feng, L.; Cheng, C.; Zhao, M.; Deng, H.; and Zhang, Y. 2022. EEG-based emotion recognition using spatial-temporal graph convolutional LSTM with attention mechanism. *Journal of Biomedical and Health Informatics*, 26(11): 5406–5417.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *ArXiv Preprint*.

Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; and Smola, A. 2006. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19.

He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *International Joint Conference on Neural Networks*, 1322–1328.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hu, J.; Wang, C.; Jia, Q.; Bu, Q.; Sutcliffe, R.; and Feng, J. 2021. ScalingNet: extracting features from raw EEG data for emotion recognition. *Neurocomputing*, 463: 177–184.

Hu, X.; Chen, J.; Wang, F.; and Zhang, D. 2019. Ten challenges for EEG-based affective computing. *Brain Science Advances*, 5(1): 1–20.

Jung, T.-P.; Sejnowski, T. J.; et al. 2019. Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *Transactions on Affective Computing*, 13: 96–107.

Kanimozhi, A.; and Raj, V. C. 2015. A cognitive e-learning system using arousal valence emotional model. *Journal of Theoretical & Applied Information Technology*, 78(3).

Karimi, H.; and Tang, J. 2020. Decision boundary of deep neural networks: Challenges and opportunities. In *International Conference on Web Search and Data Mining*, 919–920.

Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2011. Deap: A database for emotion analysis; using physiological signals. *Transactions on Affective Computing*, 3(1): 18–31.

Li, W.; Huan, W.; Hou, B.; Tian, Y.; Zhang, Z.; and Song, A. 2021. Can emotion be transferred?—A review on transfer learning for EEG-Based Emotion Recognition. *Transactions on Cognitive and Developmental Systems*, 14(3): 833–846.

Lin, K.-P.; and Chen, M.-S. 2011. Efficient kernel approximation for large-scale support vector machine classification. In *International Conference on Data Mining*.

Lin, Y.-P.; and Jung, T.-P. 2017. Improving EEG-based emotion classification using conditional transfer learning. *Frontiers in Human Neuroscience*, 11: 334.

Liu, Q.; Hao, J.; and Guo, Y. 2023. EEG Data Augmentation for Emotion Recognition with a Task-Driven GAN. *Algorithms*, 16(2): 118.

Luo, Y.; and Lu, B.-L. 2018. EEG data augmentation for emotion recognition using a conditional Wasserstein GAN. In *Engineering in Medicine and Biology Society*, 2535–2538.

Mehrabian, A. 1995. Framework for a comprehensive description and measurement of emotional states. *Genetic, Social, and General Psychology Monographs*.

Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14: 261–292.

Pei, K.; Cao, Y.; Yang, J.; and Jana, S. 2017. Deepxplore: Automated whitebox testing of deep learning systems. In *Symposium on Operating Systems Principles*, 1–18.

Possas, R.; and Zhou, Y. 2017. Effectiveness of Adversarial Attacks on Class-Imbalanced Convolutional Neural Networks. In *Neural Information Processing*, 333–342.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1161.

Russell, J. A.; and Mehrabian, A. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3): 273–294.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29.

Santamaria-Granados, L.; Munoz-Organero, M.; Ramirez-Gonzalez, G.; Abdulhay, E.; and Arunkumar, N. 2018. Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). *Access*, 7: 57–67.

Schölkopf, B. 2000. The kernel trick for distances. *Advances in Neural Information Processing Systems*.

Shen, F.; Dai, G.; Lin, G.; Zhang, J.; Kong, W.; and Zeng, H. 2020. EEG-based emotion recognition using 4D convolutional recurrent neural network. *Cognitive Neurodynamics*, 14: 815–828.

Song, T.; Liu, S.; Zheng, W.; Zong, Y.; and Cui, Z. 2020. Instance-adaptive graph for EEG emotion recognition. In *Association for the Advancement of Artificial Intelligence*, volume 34, 2701–2708.

Tao, J.; and Tan, T. 2005. Affective computing: A review. In *Affective Computing and Intelligent Interaction*, 981–995.

Tao, W.; Li, C.; Song, R.; Cheng, J.; Liu, Y.; Wan, F.; and Chen, X. 2020. EEG-based emotion recognition via channel-wise attention and self attention. *Transactions on Affective Computing*.

Thammasan, N.; Fukui, K.-i.; and Numao, M. 2017. Multimodal fusion of EEG and musical features in music-emotion recognition. In *Association for the Advancement of Artificial Intelligence*, volume 31.

Topic, A.; and Russo, M. 2021. Emotion recognition based on EEG feature maps through deep learning network. *Engineering Science and Technology, an International Journal*, 24(6): 1442–1454.

Tripathi, S.; Acharya, S.; Sharma, R.; Mittal, S.; and Bhattacharya, S. 2017. Using deep and convolutional neural networks for accurate emotion classification on DEAP data. In *Association for the Advancement of Artificial Intelligence*, volume 31, 4746–4752.

Wang, F.; Zhong, S.-h.; Peng, J.; Jiang, J.; and Liu, Y. 2018. Data augmentation for EEG-based emotion recognition with deep convolutional neural networks. In *MultiMedia Modeling*, 82–93.

Wang, Z.; Wang, Y.; Hu, C.; Yin, Z.; and Song, Y. 2022. Transformers for EEG-based emotion recognition: A hierarchical spatial information learning model. *Sensors Journal*, 22(5): 4359–4368.

Wei, C.-S.; Lin, Y.-P.; Wang, Y.-T.; Jung, T.-P.; Bigdely-Shamlo, N.; and Lin, C.-T. 2015. Selective transfer learning for EEG-based drowsiness detection. In *International Conference on Systems, Man, and Cybernetics*, 3229–3232.

Wirawan, I. M. A.; Wardoyo, R.; and Lelono, D. 2022. The challenges of emotion recognition methods based on electroencephalogram signals: A literature review. *IJECE*, 12(2): 1508.

Wu, D.; Xu, J.; Fang, W.; Zhang, Y.; Yang, L.; Xu, X.; Luo, H.; and Yu, X. 2021. Adversarial attacks and defenses in physiological computing: A systematic review. *ArXiv Preprint*.

Yang, Y.; Wu, Q.; Qiu, M.; Wang, Y.; and Chen, X. 2018. Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In *International Joint Conference on Neural Networks*, 1–7.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, T.; Cui, Z.; Xu, C.; Zheng, W.; and Yang, J. 2020. Variational pathway reasoning for EEG emotion recognition. In *Association for the Advancement of Artificial Intelligence*, volume 34, 2709–2716.

Zhang, Y.; Zhang, Y.; and Wang, S. 2022. An attention-based hybrid deep learning model for EEG emotion recognition. *Signal, Image and Video Processing*, 1–9.

Zhang, Z.; Zhong, S.-h.; and Liu, Y. 2022. GANSER: A self-supervised data augmentation framework for EEG-based emotion recognition. *Transactions on Affective Computing*.

Zhao, L.-M.; Yan, X.; and Lu, B.-L. 2021. Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In *Association for the Advancement of Artificial Intelligence*, volume 35, 863–870.

Zhao, S.; Liu, Z.; Lin, J.; Zhu, J.-Y.; and Han, S. 2020. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33: 7559–7570.

Zheng, M.; Li, T.; Zhu, R.; Tang, Y.; Tang, M.; Lin, L.; and Ma, Z. 2020. Conditional Wasserstein generative adversarial network-gradient penalty-based approach to alleviating imbalanced data classification. *Information Sciences*, 512: 1009–1023.