Article

Data Augmentation for Audio–Visual Emotion Recognition with an Efficient Multimodal Conditional GAN

Fei Ma ¹, Yang Li ^{1,*}, Shiguang Ni ², Shao-Lun Huang ^{1,*} and Lin Zhang ¹¹ Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China; mf17@mails.tsinghua.edu.cn (F.M.); linzhang@tsinghua.edu.cn (L.Z.)² Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China; ni.shiguang@sz.tsinghua.edu.cn

* Correspondence: yangli@sz.tsinghua.edu.cn (Y.L.); shaolun.huang@sz.tsinghua.edu.cn (S.-L.H.)

Abstract: Audio–visual emotion recognition is the research of identifying human emotional states by combining the audio modality and the visual modality simultaneously, which plays an important role in intelligent human–machine interactions. With the help of deep learning, previous works have made great progress for audio–visual emotion recognition. However, these deep learning methods often require a large amount of data for training. In reality, data acquisition is difficult and expensive, especially for the multimodal data with different modalities. As a result, the training data may be in the low-data regime, which cannot be effectively used for deep learning. In addition, class imbalance may occur in the emotional data, which can further degrade the performance of audio–visual emotion recognition. To address these problems, we propose an efficient data augmentation framework by designing a multimodal conditional generative adversarial network (GAN) for audio–visual emotion recognition. Specifically, we design generators and discriminators for audio and visual modalities. The category information is used as their shared input to make sure our GAN can generate fake data of different categories. In addition, the high dependence between the audio modality and the visual modality in the generated multimodal data is modeled based on Hirschfeld–Gebelein–Rényi (HGR) maximal correlation. In this way, we relate different modalities in the generated data to approximate the real data. Then, the generated data are used to augment our data manifold. We further apply our approach to deal with the problem of class imbalance. To the best of our knowledge, this is the first work to propose a data augmentation strategy with a multimodal conditional GAN for audio–visual emotion recognition. We conduct a series of experiments on three public multimodal datasets, including eNTERFACE’05, RAVDESS, and CMEW. The results indicate that our multimodal conditional GAN has high effectiveness for data augmentation of audio–visual emotion recognition.

Citation: Ma, F.; Li, Y.; Ni, S.; Huang, S.-L.; Zhang, L. Data Augmentation for Audio–Visual Emotion Recognition with an Efficient Multimodal Conditional GAN. *Appl. Sci.* **2022**, *12*, 527. <https://doi.org/10.3390/app12010527>

Academic Editor: Andrea Prati

Received: 5 December 2021

Accepted: 1 January 2022

Published: 5 January 2022

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: audio–visual emotion recognition; data augmentation; multimodal conditional generative adversarial network (GAN); Hirschfeld–Gebelein–Rényi (HGR) maximal correlation

1. Introduction

The task of emotion recognition is to detect human affective states. It is crucial for affect-related human–machine interactions, which has attracted a lot of attention from researchers [1–8]. Audio–visual emotion recognition is a common type of emotion recognition, which combines the information of audio and visual modalities with multimodal learning [9,10]. Applications can be found in many domains, such as education [11], marketing [12], and gaming [13,14]. In recent years, we have witnessed tremendous progress in audio–visual emotion recognition by using deep neural networks (DNNs) [10,15–18]. However, these deep learning approaches may suffer from the following two challenges.

Firstly, data acquisition has become a major bottleneck for deep learning [19–21]. This is mainly due to two reasons: (i) Differently to traditional machine learning methods, deep learning requires larger amounts of the training data for learning tasks. (ii) The data collection process is often time-consuming and labor-intensive, especially for the multimodal data with different modalities. In consequence, the training data used for

deep learning may not be sufficient, which leads to overfitting of deep learning models. Secondly, the emotional data may be class-imbalanced, which means that the minority class contains significantly fewer samples than the majority class. The problem of class imbalance may result in deep learning models showing bias towards the majority class and completely ignoring the minority class in extreme cases [22,23], which can further deteriorate the effectiveness of deep learning for audio–visual emotion recognition.

To address the aforementioned issues, we propose an efficient data augmentation approach with a multimodal conditional generative adversarial network (GAN) for audio–visual emotion recognition. Data augmentation is shown to be an effective way to tackle problems of data insufficiency and class imbalance in many fields, such as speech [24–28], vision [29–34], and time series [35–38]. Different data augmentation techniques are proposed to increase the size of the training data, such as geometric transformations, random occlusion, and photometric transformations [35,39]. GAN was first proposed in [40] to generate new data and is used for data augmentation in many applications [28,30,31,41–44]. Compared to other data augmentation techniques, GAN has received increasing attention by bringing more variations to the training data through deep learning [45,46]. Our proposed multimodal conditional GAN is a generalization of existing GANs for data augmentation to improve the performance of audio–visual emotion recognition. Figure 1 shows the architecture of our data augmentation approach. We first use our multimodal conditional GAN to generate new data, then augment the training set with the generated data, and finally train a DNN classifier with the augmented training set for the emotion classification task.

Figure 1. The pipeline of data augmentation in our framework.

To be specific, our multimodal conditional GAN builds generators and discriminators for audio and visual modalities. Additional category information is used as their shared input to generate fake data of different categories. It is shown in [17,47–49] that in the real multimodal data, the audio modality and the visual modality are highly dependent, which is beneficial to emotion recognition. Inspired by this, we measure the correlation between the audio modality and the visual modality in the generated multimodal data based on Hirschfeld–Gebelein–Rényi (HGR) maximal correlation [50–52], which makes the generated multimodal data close enough to the real multimodal data to help improve the performance of audio–visual emotion recognition. In such a way, our multimodal conditional GAN can (1) effectively generate the multimodal data with audio and visual modalities, and (2) fully consider the correlation between the audio modality and the visual modality in the generated multimodal data. Then, the generated multimodal data are used to augment our training set. Finally, a DNN model is used as the classifier to perform the emotion classification task with the augmented training set. Three public multimodal datasets, including eNTERFACE’05 [53], RAVDESS [54], and CMEW [55–57], are used to conduct a series of experiments. The experiment results show that our multimodal conditional GAN can significantly enhance the performance of audio–visual emotion recognition.

To summarize, the main contributions of this paper lie in the following aspects:

- We design an efficient multimodal conditional GAN to augment the multimodal data with audio and visual modalities for emotion recognition.
- We model the correlation between the audio modality and the visual modality in the generated multimodal data to approximate the real multimodal data.
- We conduct experiments on three public multimodal datasets to show that our multimodal conditional GAN can be effectively used for data augmentation of audio–visual emotion recognition.

To the best of our knowledge, this is the first work to propose an efficient data augmentation approach with a multimodal conditional GAN for audio–visual emotion recognition. The rest of this paper is organized as follows. In Section 2, we show related works, including multimodal learning, and GAN and its application for data augmentation. In Section 3, we present our methodology in detail. In Section 4, a series of experiments are conducted to show the effectiveness of our multimodal conditional GAN. Finally, in Section 5, we draw conclusions and point out some future works.

2. Related Works

Our approach is closely related to multimodal learning, and GAN and its application for data augmentation. In this section, we review them in detail.

2.1. Multimodal Learning

Multimodal learning is a hot research topic in the academic community. It processes and relates the information from multiple modalities, and achieves success in many domains [58–62]. Representation and fusion are two key challenges of multimodal learning. The challenge of representation is to study how to learn good feature representations from the multimodal data [63–66]. It has two common types: joint and coordinated. Joint representations project different modalities into the same representation space. Coordinate representations learn projection functions from each modality and make them have similarity constraints. The challenge of fusion focuses on combining information from different modalities for prediction [67,68], which has three forms: early fusion, late fusion, and hybrid fusion. Early fusion combines the information of different modalities at the feature level, and late fusion works at the decision level. Hybrid fusion is between early fusion and late fusion.

Audio–visual emotion recognition is a specific scenario of multimodal learning by combining audio and visual modalities for emotion recognition. A number of previous works [10,15–18] has been proposed for audio–visual emotion recognition. For example, in [10], Zhang et al. use convolutional neural networks (CNNs) to extract audio and visual features, and fuse them in a deep belief network. In [15], Kim et al. design deep belief networks for unsupervised feature learning in audio–visual emotion recognition. In [16], multidirectional regression audio features and ridgelet transform-based face image features are captured, and a Bayesian sum rule is used as the fusion strategy. In [17], Ma et al. propose a deep learning approach in order to efficiently utilize common information between the audio modality and the visual modality for emotion recognition by correlation analysis. In [18], Peri et al. explore the disentanglement of multimodal signal representations for the primary task of emotion recognition and a secondary person identification task.

2.2. GAN

The concept of GAN was first proposed in [40] to generate new visual data via adversarial learning. It represents a two-player zero-sum game by adversarially training a generator and a discriminator. The compelling generation capability of GANs has attracted widespread attention from the research community [69–71]. Different variants of GANs are proposed. For example, in [72], CycleGAN is proposed for unpaired image-to-image translation by introducing a cycle consistency loss function. In [73], Ledig et al. propose SRGAN for image super-resolution by designing a perceptual loss function. In [74], Karras et al. propose StyleGAN for style transfer by introducing the separation of high-level attributes from stochastic variation in the generated data.

By supplementing the given dataset with the generated data, different forms of GANs are designed for data augmentation in many domains, such as medical image analysis [29,41,75,76], brain–computer interface [42,77], agriculture [43,78,79], and civil engineering [30,80]. In particular, researchers increasingly utilize GANs for data augmentation in the field of emotion recognition, including speech emotion recognition [27,28,81], facial expression recognition [31,82,83], EEG emotion recognition [44,84], and so on. These works have significantly improved the performance of unimodal emotion recognition.

However, as far as we know, there are few works using GANs for data augmentation to improve the performance of multimodal emotion recognition, especially audio–visual emotion recognition. This may be due to two facts: (1) Both audio modality and visual modality in multimodal data are high-dimensional. It is difficult to design GANs to effectively extract information from them and generate them simultaneously; (2) In addition, these two modalities are highly correlated in terms of their emotional information [17,85]. Elucidating how to measure their relationship in GANs is also not easy. In our framework, the proposed multimodal condition GAN can effectively tackle the above two issues.

In addition, GANs can be used to deal with class imbalance [86–89]. For example, in [86], Douzas et al. use a conditional GAN as an oversampling approach to generate synthetic data for the minority class. In [87], Ali-Gombe et al. use multiple fake classes to ensure a fine-grained generation and classification of the minority class instances. In [88], multiple generation modules are adopted for data augmentation of minority classes. In [89], a balanced semisupervised GAN and a balanced batch-sampling strategy are introduced to resolve class imbalance.

3. Methodology

3.1. Overview

In this paper, we design a multimodal GAN for data augmentation in the problem of audio–visual emotion recognition. The pipeline is shown in Figure 1. We first apply our proposed GAN to generate fake multimodal data that is as realistic as possible, then use the generated data to supplement the original training set. Finally, we train a DNN classifier to perform the classification task with the augmented training set. The architecture of our multimodal GAN is shown in Figure 2. It builds generators and discriminators for audio and visual modalities. Generators generate fake data to confuse discriminators, while discriminators are used to distinguish whether the input data is real or fake. They are implemented by parameterized neural networks, which are trained together.

Suppose in the training process, we are given the training set $D = \{(\mathbf{x}_A^{(i)}, \mathbf{x}_V^{(i)}, y^{(i)}) \mid y^{(i)} \in \mathcal{Y} = \{1, 2, \dots, |\mathcal{Y}|\}, i = 1, \dots, n\}$, where $\mathbf{x}_A^{(i)}$ and $\mathbf{x}_V^{(i)}$ represent the audio modality and the visual modality of the i -th sample, respectively. $y^{(i)}$ is their corresponding category label. For simplicity, the superscript will be omitted in the following. We denote the generator and discriminator of the visual modality as G_V and D_V , respectively. G_A and D_A are expressed in the same way. $\mathbf{z}_V$ and $\mathbf{z}_A$ are two independent Gaussian noise vectors that are used to generate the visual modality and audio modality, respectively.

3.2. Proposed Multimodal Conditional GAN

To train our multimodal condition GAN, we design loss functions from the following two aspects: (1) The high-dimensional multimodal data with audio and visual modalities should be generated appropriately; (2) The relationship between the audio modality and the visual modality in the generated multimodal data should be fully considered. In order to achieve these goals, we generate the audio modality and the visual modality simultaneously, and model their correlation in our framework. Specifically, to generate the visual modality in the multimodal data, the following loss function $L_V(D_V, G_V)$ is considered:

$$L_V(D_V, G_V) = \mathbb{E}[\log D_V(\mathbf{x}_V|y)] + \mathbb{E}[\log(1 - D_V(G_V(\mathbf{z}_V|y)))] \quad (1)$$

where $\mathbb{E}$ represents the expectation. The generator G_V takes the noise vector $\mathbf{z}_V$ and the category label y as input, and outputs the fake visual modality. The discriminator D_V can be regarded as a binary classifier. It takes the real or fake visual modality as the input together with the corresponding category label, and outputs a single scalar to distinguish whether the input visual modality is real or fake. G_V and D_V are optimized adversarially. Similarly, to generate the audio modality, we consider the following loss function $L_A(D_A, G_A)$:

$$L_A(D_A, G_A) = \mathbb{E}[\log D_A(\mathbf{x}_A|y)] + \mathbb{E}[\log(1 - D_A(G_A(\mathbf{z}_A|y)))] \quad (2)$$

It is worth pointing out that the category label is sent to G_V and G_A as the shared input. In this way, audio and visual modalities in the multimodal data are generated simultaneously, where they share the same emotional information.

Figure 2. The structure of our multimodal conditional GAN. The audio modality and the visual modality are generated by G_A and G_V , respectively. Category information is used as their shared input to generate new data of different categories. Additionally, in the generated multimodal data, the correlation between the audio modality and the visual modality is modeled to approximate the real multimodal data.

In addition, it is shown in [17,47] that the audio modality and the visual modality in the real multimodal data are correlated, which contributes to emotional recognition. This factor inspires us to consider the relationship between different modalities in the generated multimodal data. Here, we adopt a novel correlation loss function based on the Hirschfeld–Gebelein–Rényi (HGR) maximal correlation [50–52] to capture the correlation between the audio modality and the visual modality:

$$L_{corr}(G_A, G_V) = -\mathbb{E}[G_A(\mathbf{z}_A|y)^T(G_V(\mathbf{z}_V|y))] + \frac{1}{2}\text{tr}(\text{cov}(G_A(\mathbf{z}_A|y))\text{cov}(G_V(\mathbf{z}_V|y))) \quad (3)$$

where the covariance matrix $\text{cov}(G_A(\mathbf{z}_A|y)) = \mathbb{E}[G_A(\mathbf{z}_A|y)G_A(\mathbf{z}_A|y)^T] - \mathbb{E}[G_A(\mathbf{z}_A|y)] \cdot \mathbb{E}[G_A(\mathbf{z}_A|y)]^T$, $\text{cov}(G_V(\mathbf{z}_V|y))$ is formulated in the same way, and $\text{tr}(\cdot)$ represents the trace operator. By optimizing the correlation function L_{corr} in our framework, we relate the audio modality and the visual modality in the generated multimodal data with high dependence, which makes the generated multimodal data as realistic as possible to help improve the performance of emotion recognition.

Finally, we use the following loss function to train our multimodal conditional GAN:

$$G_A^*, G_V^* = \arg \min_{G_A, G_V} \max_{D_A, D_V} L_A + L_V + \lambda \cdot L_{corr} \quad (4)$$

where λ is a coefficient to measure the importance of the correlation loss in our framework. Different modules are trained simultaneously. After training, optimized generators G_A^* and G_V^* are used to generate new multimodal data with audio and visual modalities for data augmentation. The procedure to update the parameters of our multimodal conditional GAN is summarized in Algorithm 1.

3.3. DNN Classifier

In fact, we can use the trained generators G_A^* and G_V^* to generate any number of new multimodal data. However, not all generated data can benefit audio–visual emotion recognition. This may be because some of the generated data does not have sufficient emotional information, thus failing to perform data augmentation. Therefore, inspired by [44,90], we design a strategy to evaluate the generated multimodal data. We first learn a DNN classifier with the original training set D , which is trained with the cross-entropy loss

function. Then, we use the trained classifier to predict the emotion labels of the multimodal data generated by our multimodal conditional GAN. If the predicted probability exceeds a certain threshold, we select this data as the final generated data.

After evaluating the generated data, we use the final generated data to supplement the original training set D as the augmented training set. On this set, we train a DNN classifier with the same architecture as the above DNN classifier. Then, we evaluate its classification performance on the test set.

Algorithm 1 Update the parameters of our multimodal conditional GAN.

Input: Coefficient λ , number of training iterations T , number of steps for discriminators K_d , number of steps for generators K_g .

for $t = 1, \dots, T$ **do**

for $k = 1, \dots, K_d$ **do**

 • Sample m noise samples $\{\mathbf{z}_A^{(1)}, \dots, \mathbf{z}_A^{(m)}\}$ and m noise samples $\{\mathbf{z}_V^{(1)}, \dots, \mathbf{z}_V^{(m)}\}$ from two independent prior distributions.

 • Sample m real examples $\{(\mathbf{x}_A^{(1)}, \mathbf{x}_V^{(1)}, y^{(1)}), \dots, (\mathbf{x}_A^{(m)}, \mathbf{x}_V^{(m)}, y^{(m)})\}$ from the training set D .

 • Update the discriminators D_A and D_V in Equations (1) and (2).

end for

for $k = 1, \dots, K_g$ **do**

 • Sample m noise samples $\{\mathbf{z}_A^{(1)}, \dots, \mathbf{z}_A^{(m)}\}$ and m noise samples $\{\mathbf{z}_V^{(1)}, \dots, \mathbf{z}_V^{(m)}\}$ from two independent prior distributions.

 • Update the generators G_A and G_V in Equations (1)–(3).

end for

end for

4. Experiments

4.1. Datasets

We conduct experiments on three public multimodal datasets: eNTERFACE'05, RAVDESS, and CMEW. eNTERFACE'05 [53] is an audio–visual emotion database that contains 42 subjects, coming from 14 different nationalities. Six basic emotions are considered in this dataset, including anger, disgust, fear, happiness, sadness, and surprise. The data of this dataset are all in video format. Following [17,91], we extract 30 segments from each video. Then, for each segment, we use 94 mel filter banks with a 40 ms Hanning window and a 10 ms overlapping to extract the log mel spectrogram with the size of $94 \times 94 \times 3$ as the audio modality. In addition, the human face with the size of $160 \times 160 \times 3$ is detected by the multitask cascaded convolutional neural network (MTCNN) [92] as the visual modality. In this way, segment samples are extracted from the original videos. Then, all samples are divided into three parts: training set, validation set, and test set. To verify that our model can cope with data insufficiency, we conduct experiments by empirically setting up training sets of three sizes: Setting1, where each emotion has 300 segment samples; Setting2, where each emotion has 1200 segment samples; Setting3, where each emotion has 2400 segment samples. It can be seen that Setting1 has the fewest samples, Setting2 has fewer, and Setting3 has the most. In each of the above three cases, our multimodal conditional GAN is trained for data augmentation. The number of segment samples per class in both the validation set and the test set is 990 to evaluate the classification performance of emotion recognition. Under Setting1, the ratio of training set, validation set, and test set is 10:33:33. The ratios of Setting2 and Setting3 are 40:33:33 and 80:33:33, respectively.

RAVDESS [54] is a validated multimodal emotional database. It consists of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. The data of this dataset are also in the form of video. The same six basic emotions as the eNTERFACE'05 dataset are studied here. Similar to the eNTERFACE'05 dataset, we also extract segment samples from original videos on the RAVDESS dataset and divide the samples into three parts: training set, validation set, and test set. On this dataset, we

conduct experiments on training sets with three different settings: Setting1, where each emotion has 150 segment samples; Setting2, where each emotion has 300 segment samples; Setting3, where each emotion has 1200 segment samples. For the validation and test sets, each class has 870 segment samples. The ratios of training set, validation set, and test set under Setting1, Setting2, and Setting3 are 5:29:29, 10:29:29, and 40:29:29, respectively.

CMEW [55–57] is an audio–visual emotion database collected from Chinese TV series and movies. It contains four emotional states, including neutral, sadness, happiness, and anger. Data preprocessing is similar to the eNTERFACE’05 dataset and the RAVDESS dataset. Training sets with three sizes are used to conduct experiments separately: Setting1, where each emotion has 300 segment samples; Setting2, where each emotion has 1200 segment samples; Setting3, where each emotion has 6000 segment samples. There are 3150 segment samples for each class of the validation set and the test set. The ratios of training set, validation set, and test set under Setting1, Setting2, and Setting3 are 2:21:21, 8:21:21, and 40:21:21, respectively.

4.2. Networks

In Section 4.1, we show that the audio modality with the size of $94 \times 94 \times 3$ and the visual modality with the size of $160 \times 160 \times 3$ are extracted from original videos. If we directly feed them into our multimodal condition GAN, it will be difficult for our model to learn useful information from audio and visual modalities due to the high dimensionality of the multimodal data. Inspired by [17,44], we utilize the fully connected layer before the softmax layer of ResNet-34 [93] to extract 512-dimensional features from the raw audio and visual modalities as the input of our framework. In other words, the dimensionality of both $\mathbf{x}_A$ and $\mathbf{x}_V$ is 512. In this way, audio and visual modalities are appropriately represented in our system. The dimensions of Gaussian noise vectors $\mathbf{z}_A$ and $\mathbf{z}_V$ are set to 50. In addition, the label y corresponding to $\mathbf{x}_A$ and $\mathbf{x}_V$ are converted into the label embedding, and then sent to the generators of our multimodal conditional GAN along with $\mathbf{z}_A$ and $\mathbf{z}_V$.

Generators G_A and G_V have the same architecture with four fully connected layers. The first layer is the input layer with the noise vector and the label embedding, the second layer has 128 nodes, the third layer has 256 nodes, and the last layer is the output layer with 512 nodes to generate fake audio modality or fake visual modality. The second and third layers are followed by the LeakyReLU function. Similarly, discriminators D_A and D_V have the same architecture. We first concatenate the label embedding with the input modality, which can be either a real modality or a fake one generated by generators. Then, we send them to the discriminator with four fully connected layers. The first layer is the input layer. Each of the following two layers has 256 LeakyReLU nodes, and the last layer has only one node to determine whether the input data is real or fake.

After generating data with our multimodal conditional GAN, we use classifiers to evaluate the generated data and perform the final emotion classification task. In [94–96], DNN is shown to be effective in learning feature representations from audio and visual modalities. Inspired by this, we adopt DNN to design our emotion classifier. To be specific, we first concatenate the audio modality and the visual modality into a feature vector, and then send it into the DNN classifier with four fully connected layers. The first layer is the input layer with 1024 nodes. Each of the following two layers has 128 LeakyReLU nodes. The number of nodes in the last layer is the number of emotion categories (six for eNTERFACE’05 and RAVDESS, and four for CMEW).

4.3. Implementation Details

As far as we know, we are the first to design a multimodal GAN for data augmentation in the problem of audio–visual emotion recognition. Under different experimental settings, we compare our approach (Ours) with the following methods to show the superiority of our framework: (1) Baseline (A), which only uses the audio modality of the original training set for emotion classification. (2) Baseline (V), which only uses the visual modality of the original training set for emotion classification. (3) Baseline (A+V), which only uses the multimodal data of the original training set for emotion classification without data

augmentation. In addition, some representative previous works of audio–visual emotion recognition are also compared, including (4) Attention Mechanism [97], (5) Canonical Correlation Analysis (CCA) [98], and (6) Tensor Fusion [99]. Moreover, we report the performance of (7) Ours (wo corr), which first uses our multimodal conditional GAN without the correlation loss to generate data, and then augment the original dataset with the generated data, finally performing emotion classification with the augmented training set. By comparing the performance of Ours and Ours (wo corr), we can show the importance of the proposed correlation loss in our framework.

In the training stage, we use the segment-level samples to perform classification. After obtaining the segment-level classification results, we use the majority rule to predict the video-level emotion labels. To ensure the reliability of the experimental result, we repeat each experiment five times and use the average classification accuracy (%) as the final result. For eINTERFACE'05, RAVDESS, and CMEW, we set λ to 0.001, 0.01, and 0.1, respectively. In addition, Adam optimizer [100] with the learning rate of 0.0002 is adopted, and the batch size is set to 100. It is worth noting that our architecture is a GAN conditional on the class label. It can generate samples of different categories, which can internally prevent mode collapse [40,101]. In addition, inspired by [102,103], the above hyperparameters and network weights are carefully set to deal with other training problems of our multimodal conditional GAN. Pytorch [104] is used to implement our architecture on an NVIDIA TITAN V GPU card.

4.4. Experiment Results

To show the effectiveness of our approach, we first compare the classification performance of our method with other methods on different datasets, including eINTERFACE'05, RAVDESS, and CMEW. On each dataset, three settings are used for experiments. Note that to take full advantage of the methods of Ours (wo corr) and Ours, false multimodal data with appropriate sizes should be generated for data augmentation. The classification accuracies of different methods on the test set are reported, as shown in Tables 1–3.

We find the following summarizations: (1) Our approach achieves the highest performance compared to other methods, which shows that the data generated using our multimodal conditional GAN can significantly benefit audio–visual emotion recognition. (2) Although the classification performance of Ours (wo corr) is significantly higher than that of baseline methods, it is still lower than that of Ours, which shows that the correlation between the audio modality and the visual modality is helpful for data augmentation to improve the performance of emotion recognition. (3) In most cases, the classification accuracies of Attention Mechanism, CCA, and Tensor Fusion are higher than those of Baseline (A), Baseline (V), and Baseline (A+V), but still obviously lower than those of Ours (wo corr) and Ours. In addition, when the scale of training data is relatively small, especially in the case of Setting1, the performance of Attention Mechanism, CCA, and Tensor Fusion may be lower than that of Baseline (A+V). These findings show that the above previous works are not as effective as our method to utilize the information from the size-limited training set for classification. (4) We can further find that as the size of the original training set becomes smaller, the classification performance gap between Ours and Baseline (A+V) becomes larger. For example, on the RAVDESS dataset, the accuracy gap between Ours and Baseline (A+V) is 8.10% and 6.37% in the case of Setting1 and Setting3, respectively. This indicates that the fewer the real data, the more obvious the advantages of our approach. (5) In most cases, classification accuracies on the eINTERFACE'05 dataset and the RAVDESS dataset are higher than those on the CMEW dataset, which is consistent with the experimental results of previous works [10,17].

Next, we show the confusion matrices using Baseline (A+V), Ours (wo corr), and Ours in the case of Setting3 on the datasets of eINTERFACE'05, RAVDESS, and CMEW, as shown in Figures 3–5. It can be found that on the eINTERFACE'05 dataset, “disgust” is the easiest to be recognized. On the RAVDESS dataset, “happiness” is the easiest to be recognized. On the CMEW dataset, “anger” is the easiest to be recognized. This shows that different datasets have different emotional cues. In addition, we can further find that for a few

emotions, the classification accuracy using Baseline (A+V) may be higher than that using Ours (wo corr) or Ours, such as “disgust” on the RAVDESS dataset and “anger” on the CMEW dataset. However, for most emotions, the classification accuracy using Ours (wo corr) or Ours is higher. This shows that Ours (wo corr) or Ours can more effectively use the information of most emotions for the classification task.

Table 1. Classification accuracy (%) of different methods on the eNTERFACE’05 dataset. Best results are shown in bold.

Method	Setting1	Sample Size Setting2	Setting3
Baseline (A)	25.26	31.95	39.18
Baseline (V)	25.77	44.85	61.34
Baseline (A+V)	26.29	45.88	65.46
Attention Mechanism [97]	26.29	46.39	65.98
CCA [98]	26.80	45.88	66.49
Tensor Fusion [99]	27.32	46.91	67.53
Ours (wo corr)	29.38	48.97	69.07
Ours	31.95	50.52	71.13

Table 2. Classification accuracy (%) of different methods on the RAVDESS dataset. Best results are shown in bold.

Method	Setting1	Sample Size Setting2	Setting3
Baseline (A)	24.86	34.10	39.88
Baseline (V)	28.90	37.57	54.34
Baseline (A+V)	34.10	45.09	59.53
Attention Mechanism [97]	32.37	45.66	61.27
CCA [98]	33.06	44.51	60.69
Tensor Fusion [99]	33.53	43.35	60.12
Ours (wo corr)	39.88	50.28	64.16
Ours	42.20	52.02	65.90

Table 3. Classification accuracy (%) of different methods on the CMEW dataset. Best results are shown in bold.

Method	Setting1	Sample Size Setting2	Setting3
Baseline (A)	30.71	36.19	40.24
Baseline (V)	31.43	38.10	41.67
Baseline (A+V)	33.10	40.71	47.38
Attention Mechanism [97]	33.57	41.19	47.86
CCA [98]	34.05	41.43	48.10
Tensor Fusion [99]	32.14	40.00	47.14
Ours (wo corr)	36.90	44.76	50.47
Ours	39.29	46.43	52.14

(a)

(b)

Figure 3. Cont.

Figure 3. The confusion matrices of different methods on the eINTERFACE'05 dataset. (a) Baseline (A+V), (b) Ours (wo corr), (c) Ours.

Figure 4. Cont.

(b)

(c)

Figure 4. The confusion matrices of different methods on the RAVDESS dataset. (a) Baseline (A+V), (b) Ours (wo corr), (c) Ours.

(a)

(b)

Figure 5. Cont.

Figure 5. The confusion matrices of different methods on the CMEW dataset. (a) Baseline (A+V), (b) Ours (wo corr), (c) Ours.

In the above discussion, we show that our approach can efficiently help improve the classification accuracy on the test set. Inspired by [105], we next study the impact of the data generated by our multimodal conditional GAN on the training speed. On the datasets of eINTERFACE'05, RAVDESS, and CMEW, we plot the training loss with epochs of the DNN classifier in different methods, as shown in Figure 6. It is worth pointing out that the DNN classifier in Baseline (A+V) only uses the original training set for training, while the DNN classifier in the method of Ours (wo corr) or Ours uses the real data and generated data together for training. It can be found that the training loss of our approach decreases the fastest with epochs, which means that its convergence is the fastest, the convergence of Ours (wo corr) is the second fastest, and the convergence of Baseline (A+V) is the slowest. This shows that in addition to helping improve the accuracy of emotion classification, our framework can effectively speed up the training of the emotion classifier.

When analyzing the experimental results in Tables 1–3, we mention that when using Ours or Ours (wo corr), it is necessary to generate new data of appropriate size. Here, we discuss what scale of generated data is appropriate under different settings. We name the ratio of the size of the generated data to the size of the original training set as the augmentation ratio. It is used to measure the scale of the generated data compared to the original training set. Then, we analyze the classification performance of the methods of Ours and Ours (wo corr) with six augmentation ratios, including 0.5, 1, 2, 3, 4, and 5. The augmentation ratio of 2 means that the size of the generated data is twice the size of the original training set. Other ratios are represented in the same way. The experimental results on the datasets of eINTERFACE'05, RAVDESS, and CMEW are shown in Tables 4–6, respectively.

Table 4. Classification accuracy (%) with different augmentation ratios on the eNTERFACE'05 dataset. Best results are shown in bold.

Sample Size	Method	Augmentation Ratio					
		0.5	1	2	3	4	5
Setting1	Ours (wo corr)	27.32	28.35	27.84	29.38	28.87	27.84
	Ours	28.35	29.38	31.96	31.44	30.93	29.38
Setting2	Ours (wo corr)	46.39	48.45	48.45	48.97	47.94	46.91
	Ours	48.45	49.48	50.00	50.52	50.00	48.97
Setting3	Ours (wo corr)	67.53	69.07	68.56	68.04	68.56	67.01
	Ours	70.10	71.13	70.62	69.59	70.10	69.07

(a)

(b)

Figure 6. Cont.

Figure 6. Training loss of different methods on the datasets of eINTERFACE'05, RAVDESS, and CMEW. (a) Baseline (A+V), (b) Ours (wo corr), (c) Ours.

Table 5. Classification accuracy (%) with different augmentation ratios on the RAVDESS dataset. Best results are shown in bold.

Sample Size	Method	Augmentation Ratio					
		0.5	1	2	3	4	5
Setting1	Ours (wo corr)	37.57	38.73	39.88	39.31	39.31	38.15
	Ours	39.88	40.46	42.20	41.62	41.04	41.04
Setting2	Ours (wo corr)	47.40	49.71	50.29	49.13	47.98	48.55
	Ours	49.71	52.02	52.02	51.45	50.87	50.29
Setting3	Ours (wo corr)	62.43	64.16	63.58	63.01	63.01	62.43
	Ours	63.58	65.90	65.32	64.74	64.16	64.16

Table 6. Classification accuracy (%) with different augmentation ratios on the CMEW dataset. Best results are shown in bold.

Sample Size	Method	Augmentation Ratio					
		0.5	1	2	3	4	5
Setting1	Ours (wo corr)	35.00	36.43	36.90	36.67	35.95	35.71
	Ours	37.14	38.81	39.29	38.57	38.33	38.33
Setting2	Ours (wo corr)	44.76	44.05	43.57	42.86	42.62	42.62
	Ours	46.43	46.19	45.95	45.24	44.52	44.05
Setting3	Ours (wo corr)	50.48	50.24	50.24	49.76	49.29	48.81
	Ours	51.90	52.14	51.67	51.19	50.24	50.00

We can find that on each setting, there is an augmentation ratio with the highest classification accuracy for Ours or Ours (wo corr). We can call this augmentation ratio the most appropriate augmentation ratio. In most cases, when the augmentation ratio is lower than the most appropriate augmentation ratio and increases, the corresponding classification accuracy also increases, while when the augmentation ratio is higher than the most appropriate augmentation ratio and increases, the corresponding classification accuracy decreases. For example, on Setting2 of the eINTERFACE'05 dataset, the most appropriate

augmentation ratio is 3 for Ours. When the augmentation ratio increases from 0.5 to 1, 2, and then to 3, the corresponding classification accuracy also increases. However, when the augmentation ratio increases from 3 to 4 and then to 5, the corresponding classification accuracy becomes smaller. The reason behind this phenomenon may be that when the augmentation ratio is relatively small, the generated data is not enough for data augmentation effectively. Therefore, with the increase of augmentation ratio, the corresponding classification performance is improved. However, when the augmentation ratio is too large, the generated data may obscure the information of the original training set, which weakens the effect of data augmentation. Driven by these factors, we emphasize that the augmentation ratio should be selected appropriately for data augmentation.

From Tables 4–6, it can be further found that when the scale of the original training set is relatively small, the augmentation ratio should be set larger, i.e., more generated data can be used for data augmentation, while when the scale of the original data is relatively large, the augmentation ratio should be set smaller. For example, on the eNTERFACE'05 dataset, the sizes of Setting1 and Setting2 are relatively small, and it is more appropriate to set the augmentation ratio to 2 or 3, while Setting3 has more samples, and it is better to set the augmentation ratio to 1. The RAVDESS dataset and the CMEW dataset are similar to the eNTERFACE'05 dataset. In addition, we can see that the classification performances of Ours and Ours (wo corr) have a similar trend with the augmentation ratio, but the classification performance of Ours is higher than that of Ours (wo corr) under each setting. For example, on the eNTERFACE'05 dataset, the most appropriate augmentation ratios of Ours or Ours (wo corr) are 3 and 2, respectively, for Setting2 and Setting3. However, under these two settings, the classification accuracy of Ours is 1.55% and 2.06% higher than that of Ours (wo corr), respectively, which again shows the importance of capturing the correlation between the audio modality and the visual modality in our framework.

Furthermore, we show that our multimodal conditional GAN can deal with the problem of class imbalance of the training data. We conduct experiments on the datasets of eNTERFACE'05, RAVDESS, and CMEW. As shown in previous works [88,106], the majority-to-minority class ratio is high in real-world class imbalance problems. Inspired by this, we consider two different training sets: Case1 and Case2. For the training set of Case1, there are 300 samples for a class, and 1200 samples for each of other classes. For the training set of Case2, there are 150 samples for a class, and 1200 samples for each of other classes. It can be seen that the majority-to-minority class ratio of Case 2 is higher than that of Case 1. We use our multimodal conditional GAN or the multimodal GAN without the correlation loss to augment the unbalanced training set to make an augmented training set containing 2400 samples for each class, which is class-balanced. Then, the augmented training set is further used to train a DNN classifier. In both Case1 and Case2, we also perform experiments with other methods. Additionally, to compare with the class-balanced case, we report the classification performance of different methods on the training set with 1200 samples per class (for Ours (wo corr) or Ours, each class is augmented to 2400 samples). It is worth pointing out that on each dataset, the validation and test sets are still set as in Section 4.1, which means that they are still class-balanced. Under this setting, we report the classification accuracy of the test set on each dataset to evaluate the performance of our framework to cope with the class imbalance of the training data, as shown in Tables 7–9.

We can observe that with the increase of the majority-to-minority class ratio, the classification accuracy of different methods decreases. However, compared with other methods, our approach achieves the highest classification performance in various settings of class imbalance, indicating that our framework can effectively deal with the problem of class imbalance. In addition, it can be found that compared with different baselines, the method of Ours (wo corr) can also improve the performance of emotion classification, but it is weaker than Ours, which indicates that our proposed correlation loss is useful in the scenario of class imbalance.

Table 7. Classification accuracy (%) of different methods when the training set is class-imbalanced on the eNTERFACE'05 dataset. Best results are shown in bold.

Method	Class Balance	Class Imbalance	
		Case1	Case2
Baseline (A)	31.95	31.44	30.41
Baseline (V)	44.85	38.14	34.54
Baseline (A+V)	45.88	39.69	36.60
Attention Mechanism [97]	46.39	41.24	35.57
CCA [98]	45.88	39.18	36.08
Tensor Fusion [99]	46.91	40.72	37.11
Ours (wo corr)	48.45	43.30	41.75
Ours	49.48	44.33	42.78

Table 8. Classification accuracy (%) of different methods when the training set is class-imbalanced on the RAVDESS dataset. Best results are shown in bold.

Method	Class Balance	Class Imbalance	
		Case1	Case2
Baseline (A)	39.88	37.57	36.99
Baseline (V)	54.35	51.45	50.29
Baseline (A+V)	59.53	54.34	53.75
Attention Mechanism [97]	61.27	55.49	52.60
CCA [98]	60.69	56.07	53.18
Tensor Fusion [99]	60.12	52.60	51.45
Ours (wo corr)	64.16	59.54	58.95
Ours	65.90	61.85	60.11

Table 9. Classification accuracy (%) of different methods when the training set is class-imbalanced on the CMEW dataset. Best results are shown in bold.

Method	Class Balance	Class Imbalance	
		Case1	Case2
Baseline (A)	36.19	33.57	29.76
Baseline (V)	38.10	34.76	31.42
Baseline (A+V)	40.71	35.48	34.76
Attention Mechanism [97]	41.19	36.19	33.81
CCA [98]	41.43	36.10	33.10
Tensor Fusion [99]	40.00	35.24	32.86
Ours (wo corr)	44.05	39.05	37.62
Ours	46.19	40.95	39.04

5. Conclusions

In this paper, we propose an efficient data augmentation framework with a multimodal conditional GAN for audio–visual emotion recognition. To be specific, our framework builds generators and discriminators for audio and visual modalities. Then, we use the category information as their shared input to make our multimodal conditional GAN generate new multimodal data of different categories. In addition, the correlation between the audio modality and the visual modality in the generated multimodal data is fully considered in our framework. Finally, we use the generated data to supplement the original training set for the emotion classification task. To the best of our knowledge, we are the first to propose an efficient multimodal condition GAN to augment the multimodal data for audio–visual emotion recognition.

A series of experiments on the datasets of eNTERFACE'05, RAVDESS, and CMEW show that compared with different baselines, our framework achieves the highest classifica-

tion accuracy for audio–visual emotion recognition, even when the training data is scarce. In addition, experiments show that the data generated by our multimodal conditional GAN can improve the training speed of DNN-based emotion classifiers. Moreover, the results further point out that the size of the data generated using our multimodal conditional GAN should be set appropriately for data augmentation. Finally, we exploit our approach to cope with the problem of class imbalance. Experiments show that in this scenario, the classification performance of our approach is still significantly higher than that of different baselines.

Despite the progress of our framework, it may have the following limitations: (1) Our approach is used for audio–visual emotion recognition. It does not consider combining information from other modalities, such as text [107] and physiological signals [108,109], which also contribute to emotion recognition. (2) Our architecture focuses on designing a multimodal condition GAN for data augmentation. Previous works such as Attention Mechanism, CCA, and Tensor Fusion are not considered to be integrated in our framework. As shown in Section 4.4, these previous works can also perform emotion recognition to a certain extent. (3) The labels of all samples are assumed to be correctly annotated to train our multimodal architecture. However, in reality, the data collection may suffer from label noise [110–112]. To further validate the effectiveness of our approach, we will tackle the above limitations in the future. Firstly, we will generalize our architecture to utilize more modalities. Secondly, we will combine our approach with the above previous works to further improve the performance of our framework. Finally, we will consider coping with label noise in our architecture.

Author Contributions: Conceptualization, F.M. and Y.L.; methodology, F.M. and Y.L.; software, F.M.; validation, F.M. and Y.L.; formal analysis, F.M., Y.L. and S.N.; investigation, F.M.; resources, Y.L., S.N., S.-L.H. and L.Z.; data curation, F.M.; writing—original draft preparation, F.M.; writing—review and editing, F.M., Y.L., S.N. and S.-L.H.; visualization, F.M. and Y.L.; supervision, L.Z.; project administration, Y.L., S.N. and S.-L.H.; funding acquisition, S.N., S.-L.H. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The work of Yang Li was supported by the National Natural Science Foundation of China under Grant 62001266. The work of Shiguang Ni was supported by the National Key R&D Program of China under Grant 2020YFC0833402. The work of Shao-Lun Huang was supported in part by the National Natural Science Foundation of China under Grant 61807021, in part by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146. The work of Lin Zhang was supported by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146.

Data Availability Statement: We conducted experiments on three datasets, including eINTERFACE'05, RAVDESS, and CMEW. eINTERFACE'05 can be found in <http://www.enterface.net/results/>, accessed on 4 December 2021; RAVDESS can be found in <https://smartlaboratory.org/ravdess/>, accessed on 4 December 2021; CMEW can be found in <https://mace-cream.github.io/en/>, accessed on 4 December 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GAN	Generative Adversarial Network
HGR	Hirschfeld-Gebelein-Rényi
DNN	Deep Neural Network
CNN	Convolutional Neural Network
MTCNN	Multi-Task Cascaded Convolutional Neural Network
CCA	Canonical Correlation Analysis

References

- Picard, R.W. *Affective Computing*; MIT Press: Cambridge, MA, USA, 2020.
- Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [\[CrossRef\]](#)
- Canedo, D.; Neves, A.J.R. Facial Expression Recognition Using Computer Vision: A Systematic Review. *Appl. Sci.* **2019**, *9*, 4678. [\[CrossRef\]](#)
- Singh, M.I.; Singh, M. Emotion Recognition: An Evaluation of ERP Features Acquired from Frontal EEG Electrodes. *Appl. Sci.* **2021**, *11*, 4131. [\[CrossRef\]](#)
- Zhang, H.; Huang, H.; Han, H. A Novel Heterogeneous Parallel Convolution Bi-LSTM for Speech Emotion Recognition. *Appl. Sci.* **2021**, *11*, 9897. [\[CrossRef\]](#)
- Luna-Jiménez, C.; Cristóbal-Martín, J.; Kleinlein, R.; Gil-Martín, M.; Moya, J.M.; Fernández-Martínez, F. Guided Spatial Transformers for Facial Expression Recognition. *Appl. Sci.* **2021**, *11*, 7217. [\[CrossRef\]](#)
- Koromilas, P.; Giannakopoulos, T. Deep Multimodal Emotion Recognition on Human Speech: A Review. *Appl. Sci.* **2021**, *11*, 7962. [\[CrossRef\]](#)
- Ngai, W.K.; Xie, H.; Zou, D.; Chou, K.L. Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources. *Inf. Fusion* **2022**, *77*, 107–117. [\[CrossRef\]](#)
- Chen, S.; Jin, Q. Multi-modal conditional attention fusion for dimensional emotion prediction. In Proceedings of the 2016 ACM on Multimedia Conference, New York, NY, USA, 6–9 June 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 571–575.
- Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Tian, Q. Learning Affective Features with a Hybrid Deep Model for Audio–Visual Emotion Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 3030–3043. [\[CrossRef\]](#)
- Wang, C.H.; Lin, H.C.K. Emotional Design Tutoring System Based on Multimodal Affective Computing Techniques. *Int. J. Distance Educ. Technol. (IJDET)* **2018**, *16*, 103–117. [\[CrossRef\]](#)
- Seng, K.P.; Ang, L.M. Video analytics for customer emotion and satisfaction at contact centers. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *48*, 266–278. [\[CrossRef\]](#)
- Karpouzis, K.; Yannakakis, G.N. *Emotion in Games*; Springer: Berlin/Heidelberg, Germany, 2016.
- Isbister, K. *How Games Move Us: Emotion by Design*; MIT Press: Cambridge, MA, USA, 2016.
- Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
- Hossain, M.S.; Muhammad, G. Audio-visual emotion recognition using multi-directional regression and Ridgelet transform. *J. Multimodal User Interfaces* **2016**, *10*, 325–333. [\[CrossRef\]](#)
- Ma, F.; Zhang, W.; Li, Y.; Huang, S.L.; Zhang, L. Learning better representations for audio-visual emotion recognition with common information. *Appl. Sci.* **2020**, *10*, 7239. [\[CrossRef\]](#)
- Peri, R.; Parthasarathy, S.; Bradshaw, C.; Sundaram, S. Disentanglement for audio-visual emotion recognition using multitask setup. In Proceedings of the ICASSP 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 6344–6348.
- Stonebraker, M.; Ilyas, I.F. Data Integration: The Current Status and the Way Forward. *IEEE Data Eng. Bull.* **2018**, *41*, 3–9.
- Polyzotis, N.; Roy, S.; Whang, S.E.; Zinkevich, M. Data lifecycle challenges in production machine learning: A survey. *ACM SIGMOD Rec.* **2018**, *47*, 17–28. [\[CrossRef\]](#)
- Roh, Y.; Heo, G.; Whang, S.E. A survey on data collection for machine learning: A big data-ai integration perspective. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 1328–1347. [\[CrossRef\]](#)
- Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 1–54. [\[CrossRef\]](#)
- Oksuz, K.; Cam, B.C.; Kalkas, S.; Akbas, E. Imbalance problems in object detection: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3388–3415. [\[CrossRef\]](#)
- Ko, T.; Peddinti, V.; Povey, D.; Seltzer, M.L.; Khudanpur, S. A study on data augmentation of reverberant speech for robust speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5220–5224.
- Sahu, S.; Gupta, R.; Espy-Wilson, C. On enhancing speech emotion recognition using generative adversarial networks. *arXiv* **2018**, arXiv:1806.06626.
- Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
- Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. *Data Augmentation Using GANs for Speech Emotion Recognition*; Interspeech: Graz, Austria, 2019; pp. 171–175.
- Shilandari, A.; Marvi, H.; Khosravi, H. Speech Emotion Recognition using Data Augmentation Method by Cycle-Generative Adversarial Networks. Available online: <https://www.preprints.org/manuscript/202104.0651/v1> (accessed on 4 December 2021).

29. Madani, A.; Moradi, M.; Karargyris, A.; Syeda-Mahmood, T. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In Proceedings of the Medical Imaging 2018: Image Processing. International Society for Optics and Photonics, Houston, TX, USA, 11–13 February 2018; Volume 10574, p. 105741M.
30. Gao, Y.; Kong, B.; Mosalam, K.M. Deep leaf-bootstrapping generative adversarial network for structural image data augmentation. *Comput.-Aided Civ. Infrastruct. Eng.* **2019**, *34*, 755–773. [\[CrossRef\]](#)
31. Zhu, X.; Liu, Y.; Li, J.; Wan, T.; Qin, Z. Emotion classification with data augmentation using generative adversarial networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 349–360.
32. Abayomi-Alli, O.O.; Damaševičius, R.; Misra, S.; Maskeliūnas, R. Cassava disease recognition from low-quality images using enhanced data augmentation model and deep learning. *Expert Syst.* **2021**, *38*, e12746. [\[CrossRef\]](#)
33. Abayomi-Alli, O.O.; Damaševičius, R.; Maskeliūnas, R.; Misra, S. Few-shot learning with a novel Voronoi tessellation-based image augmentation method for facial palsy detection. *Electronics* **2021**, *10*, 978. [\[CrossRef\]](#)
34. Abayomi-Alli, O.O.; Damaševičius, R.; Misra, S.; Maskeliūnas, R.; Abayomi-Alli, A. Malignant skin melanoma detection using image augmentation by oversampling in nonlinear lower-dimensional embedding manifold. *Turk. J. Electr. Eng. Comput. Sci.* **2021**, *29*, 2600–2614. [\[CrossRef\]](#)
35. Wen, Q.; Sun, L.; Yang, F.; Song, X.; Gao, J.; Wang, X.; Xu, H. Time series data augmentation for deep learning: A survey. *arXiv* **2020**, arXiv:2002.12478.
36. Abayomi-Alli, O.O.; Sidekerskienė, T.; Damaševičius, R.; Silka, J.; Połap, D. Empirical Mode Decomposition Based Data Augmentation for Time Series Prediction Using NARX Network. In *International Conference on Artificial Intelligence and Soft Computing*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 702–711.
37. Iwana, B.K.; Uchida, S. An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* **2021**, *16*, e0254841. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Bandara, K.; Hewamalage, H.; Liu, Y.H.; Kang, Y.; Bergmeir, C. Improving the accuracy of global forecasting models using time series data augmentation. *Pattern Recognit.* **2021**, *120*, 108148. [\[CrossRef\]](#)
39. Kaur, P.; Khehra, B.S.; Mavi, E.B.S. Data Augmentation for Object Detection: A Review. In Proceedings of the 2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS), Lansing, MI, USA, 9–11 August 2021; pp. 537–543.
40. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680. Available online: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afcf3-Paper.pdf> (accessed on 4 December 2021).
41. Lee, M.B.; Kim, Y.H.; Park, K.R. Conditional generative adversarial network-based data augmentation for enhancement of iris recognition accuracy. *IEEE Access* **2019**, *7*, 122134–122152. [\[CrossRef\]](#)
42. Fahimi, F.; Dosen, S.; Ang, K.K.; Mrachacz-Kersting, N.; Guan, C. Generative adversarial networks-based data augmentation for brain-computer interface. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4039–4051. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Qin, X.H.; Wang, Z.Y.; Yao, J.P.; Zhou, Q.; Zhao, P.F.; Wang, Z.Y.; Huang, L. Using a one-dimensional convolutional neural network with a conditional generative adversarial network to classify plant electrical signals. *Comput. Electron. Agric.* **2020**, *174*, 105464. [\[CrossRef\]](#)
44. Luo, Y.; Zhu, L.Z.; Wan, Z.Y.; Lu, B.L. Data augmentation for enhancing EEG-based emotion recognition with deep generative models. *J. Neural Eng.* **2020**, *17*, 056021. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Antoniou, A.; Storkey, A.; Edwards, H. Data augmentation generative adversarial networks. *arXiv* **2017**, arXiv:1711.04340.
46. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [\[CrossRef\]](#)
47. Ma, F.; Huang, S.L.; Zhang, L. An Efficient Approach for Audio-Visual Emotion Recognition With Missing Labels And Missing Modalities. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
48. Ma, F.; Zhang, W.; Li, Y.; Huang, S.L.; Zhang, L. An end-to-end learning approach for multimodal emotion recognition: Extracting common and private information. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1144–1149.
49. Liang, Y.; Ma, F.; Li, Y.; Huang, S.L. Person Recognition with HGR Maximal Correlation on Multimodal Data. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 1–8.
50. Hirschfeld, H.O. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*; Cambridge University Press: Cambridge, UK, 1935; Volume 31, pp. 520–524.
51. Gebelein, H. Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *ZAMM-J. Appl. Math. Mech. Für Angew. Math. Und Mech.* **1941**, *21*, 364–379. [\[CrossRef\]](#)
52. Rényi, A. On measures of dependence. *Acta Math. Hung.* **1959**, *10*, 441–451. [\[CrossRef\]](#)
53. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The enterface'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops 2006, Atlanta, GA, USA, 3–7 April 2006; p. 8.
54. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Ma, F.; Gu, W.; Zhang, W.; Ni, S.; Huang, S.L.; Zhang, L. Speech Emotion Recognition via Attention-based DNN from Multi-Task Learning. In Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, Shenzhen, China, 4–7 November 2018; pp. 363–364.

56. Zhang, W.; Gu, W.; Ma, F.; Ni, S.; Zhang, L.; Huang, S.L. Multimodal Emotion Recognition by extracting common and modality-specific information. In Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, Shenzhen, China, 4–7 November 2018; pp. 396–397.
57. Gu, W.; Zhang, Y.; Ma, F.; Mosalam, K.; Zhang, L.; Ni, S. Real-Time Emotion Detection via E-See. In Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, Shenzhen, China, 4–7 November 2018; pp. 420–421.
58. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [\[CrossRef\]](#) [\[PubMed\]](#)
59. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [\[CrossRef\]](#)
60. Zhu, W.; Wang, X.; Gao, W. Multimedia Intelligence: When Multimedia Meets Artificial Intelligence. *IEEE Trans. Multimed.* **2020**, *22*, 1823–1835. [\[CrossRef\]](#)
61. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394. [\[CrossRef\]](#)
62. Gao, J.; Li, P.; Chen, Z.; Zhang, J. A survey on deep learning for multimodal data fusion. *Neural Comput.* **2020**, *32*, 829–864. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 689–696.
64. Wu, Z.; Jiang, Y.G.; Wang, J.; Pu, J.; Xue, X. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In Proceedings of the 22nd ACM international Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 167–176.
65. Pan, Y.; Mei, T.; Yao, T.; Li, H.; Rui, Y. Jointly modeling embedding and translation to bridge video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4594–4602.
66. Xu, R.; Xiong, C.; Chen, W.; Corso, J. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.
67. Atrey, P.K.; Hossain, M.A.; El Saddik, A.; Kankanhalli, M.S. Multimodal fusion for multimedia analysis: A survey. *Multimed. Syst.* **2010**, *16*, 345–379. [\[CrossRef\]](#)
68. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **2017**, *37*, 98–125. [\[CrossRef\]](#)
69. Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. [\[CrossRef\]](#)
70. Jabbar, A.; Li, X.; Omar, B. A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–49. [\[CrossRef\]](#)
71. Pavan Kumar, M.; Jayagopal, P. Generative adversarial networks: A survey on applications and challenges. *Int. J. Multimed. Inf. Retr.* **2021**, *10*, 1–24. [\[CrossRef\]](#)
72. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
73. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
74. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
75. Vaccari, I.; Orani, V.; Paglialonga, A.; Cambiaso, E.; Mongelli, M. A Generative Adversarial Network (GAN) Technique for Internet of Medical Things Data. *Sensors* **2021**, *21*, 3726. [\[CrossRef\]](#)
76. Almezghwi, K.; Serte, S. Improved Classification of White Blood Cells with the Generative Adversarial Network and Deep Convolutional Neural Network. *Comput. Intell. Neurosci.* **2020**, *2020*, 6490479. [\[CrossRef\]](#) [\[PubMed\]](#)
77. Debie, E.; Moustafa, N.; Whitty, M.T. A privacy-preserving generative adversarial network method for securing eeg brain signals. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
78. Zhu, F.; He, M.; Zheng, Z. Data augmentation using improved cDCGAN for plant vigor rating. *Comput. Electron. Agric.* **2020**, *175*, 105603. [\[CrossRef\]](#)
79. Bi, L.; Hu, G. Improving Image-Based Plant Disease Classification with Generative Adversarial Network Under Limited Training Set. *Front. Plant Sci.* **2020**, *11*, 583438. [\[CrossRef\]](#)
80. Tan, X.; Sun, X.; Chen, W.; Du, B.; Ye, J.; Sun, L. Investigation on the data augmentation using machine learning algorithms in structural health monitoring information. *Struct. Health Monit.* **2021**, *2021*, 1475921721996238. [\[CrossRef\]](#)
81. Latif, S.; Asim, M.; Rana, R.; Khalifa, S.; Jurdak, R.; Schuller, B.W. Augmenting generative adversarial networks for speech emotion recognition. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Shanghai, China, 25–29 October 2020; Volume 2020, pp. 521–525.
82. Lai, Y.H.; Lai, S.H. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 263–270.

83. Yi, W.; Sun, Y.; He, S. Data augmentation using conditional GANs for facial emotion recognition. In Proceedings of the 2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama), Toyama, Japan, 1–4 August 2018; pp. 710–714.
84. Luo, Y.; Lu, B.L. EEG data augmentation for emotion recognition using a conditional Wasserstein GAN. In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 17–21 July 2018; pp. 2535–2538.
85. Nie, W.; Ren, M.; Nie, J.; Zhao, S. C-GCN: Correlation based graph convolutional network for audio-video emotion recognition. *IEEE Trans. Multimed.* **2020**, *23*, 3793–3804. [\[CrossRef\]](#)
86. Douzas, G.; Bacao, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Syst. Appl.* **2018**, *91*, 464–471. [\[CrossRef\]](#)
87. Ali-Gombe, A.; Elyan, E. MFC-GAN: Class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing* **2019**, *361*, 212–221. [\[CrossRef\]](#)
88. Zhang, W.; Li, X.; Jia, X.D.; Ma, H.; Luo, Z.; Li, X. Machinery fault diagnosis with imbalanced data using deep generative adversarial networks. *Measurement* **2020**, *152*, 107377. [\[CrossRef\]](#)
89. Gao, Y.; Zhai, P.; Mosalam, K.M. Balanced semisupervised generative adversarial network for damage assessment from low-data imbalanced-class regime. *Comput.-Aided Civ. Infrastruct. Eng.* **2021**, *36*, 1094–1113. [\[CrossRef\]](#)
90. Luo, J.; Bouazizi, M.; Ohtsuki, T. Data Augmentation for Sentiment Analysis Using Sentence Compression-Based SeqGAN With Data Screening. *IEEE Access* **2021**, *9*, 99922–99931. [\[CrossRef\]](#)
91. Ma, F.; Xu, X.; Huang, S.L.; Zhang, L. Maximum Likelihood Estimation for Multimodal Learning with Missing Modality. *arXiv* **2021**, arXiv:2108.10513.
92. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [\[CrossRef\]](#)
93. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
94. Khorrami, P.; Le Paine, T.; Brady, K.; Dagli, C.; Huang, T.S. How deep neural networks can improve emotion recognition on video data. In Proceedings of the 2016 IEEE International Conference On Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 619–623.
95. Zhang, C.; Cui, Y.; Han, Z.; Zhou, J.T.; Fu, H.; Hu, Q. Deep partial multi-view learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [\[CrossRef\]](#) [\[PubMed\]](#)
96. Qian, Y.; Chen, Z.; Wang, S. Audio-Visual Deep Neural Network for Robust Person Verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1079–1092. [\[CrossRef\]](#)
97. Xi, C.; Lu, G.; Yan, J. Multimodal sentiment analysis based on multi-head attention mechanism. In Proceedings of the 4th International Conference on Machine Learning and Soft Computing, Haiphong City, Viet Nam, 17–19 January 2020; pp. 34–39.
98. Zhang, K.; Li, Y.; Wang, J.; Wang, Z.; Li, X. Feature Fusion for Multimodal Emotion Recognition Based on Deep Canonical Correlation Analysis. *IEEE Signal Process. Lett.* **2021**, *28*, 1898–1902. [\[CrossRef\]](#)
99. Yan, X.; Xue, H.; Jiang, S.; Liu, Z. Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling. *Appl. Artif. Intell.* **2021**, *2021*, 1–16. [\[CrossRef\]](#)
100. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980
101. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
102. Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks. *arXiv* **2016**, arXiv:1701.00160.
103. Saxena, D.; Cao, J. Generative Adversarial Networks (GANs) Challenges, Solutions, and Future Directions. *ACM Comput. Surv.* (CSUR) **2021**, *54*, 1–42. [\[CrossRef\]](#)
104. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026–8037. Available online: <https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html> (accessed on 4 December 2021).
105. Gaur, A.; Nsaka, P.; de Raveschoot, V.W.P.; Zhuang, L. Evaluating the Efficacy of Data Augmentation Using Generative Adversarial Networks For Identification of Leukemia Cells. Available online: http://cs230.stanford.edu/projects_spring_2021/reports/13.pdf (accessed on 4 December 2021).
106. Leevy, J.L.; Khoshgoftaar, T.M.; Bauder, R.A.; Seliya, N. A survey on addressing high-class imbalance in big data. *J. Big Data* **2018**, *5*, 1–30. [\[CrossRef\]](#)
107. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference On Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 439–448.
108. Goshvarpour, A.; Abbasi, A.; Goshvarpour, A. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomed. J.* **2017**, *40*, 355–368. [\[CrossRef\]](#)
109. Li, P.; Liu, H.; Si, Y.; Li, C.; Li, F.; Zhu, X.; Huang, X.; Zeng, Y.; Yao, D.; Zhang, Y.; et al. EEG based emotion recognition by combining functional connectivity network and local activations. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 2869–2881. [\[CrossRef\]](#) [\[PubMed\]](#)
110. Frénay, B.; Verleysen, M. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2013**, *25*, 845–869. [\[CrossRef\]](#)

-
111. Frénay, B.; Kaban, A. A Comprehensive Introduction to Label Noise. In Proceedings of the 2014 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2014), Bruges, Belgium, 23–25 April 2014.
 112. Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1944–1952.