INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Data Augmentation using GANs for Speech Emotion Recognition

Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, Shrikanth Narayanan

Behavioral Signal Technologies Inc.
Los Angeles, CA, USA

{aggelina, geopar, dimitris, gpantaz, malvina, thodoris, nassos, alex, shri}@behavioralsignals.com

Abstract

In this work, we address the problem of data imbalance for the task of Speech Emotion Recognition (SER). We investigate conditioned data augmentation using Generative Adversarial Networks (GANs), in order to generate samples for underrepresented emotions. We adapt and improve a conditional GAN architecture to generate synthetic spectrograms for the minority class. For comparison purposes, we implement a series of signal-based data augmentation methods. The proposed GAN-based approach is evaluated on two datasets, namely IEMOCAP and FEEL-25k, a large multi-domain dataset. Results demonstrate a 10% relative performance improvement in IEMOCAP and 5% in FEEL-25k, when augmenting the minority classes.
Index Terms: Generative Adversarial Networks, Speech Emotion Recognition, data augmentation, data imbalance

1. Introduction

In all types of human-human or human-computer interaction, the manner in which the words are spoken conveys important non-linguistic information, especially with regards to the underlying emotions. Therefore, it has become obvious that modern speech analysis systems should be able to analyze this emotion-related non-linguistic dimension, along with the message of the utterance itself. For that reason, during the last years, methods that automatically identify the emotional content of a spoken utterance draw a growing research attention. Speech Emotion Recognition (SER) is a supervised audio task, which maps low-level audio features to either high-level class labels of distinct emotions or scalar values of affective dimensions, such as valence and arousal. In any case, annotated datasets are of great importance in building and evaluating SER systems. In this work, we deal with the problem of data imbalance in SER, and we propose a Generative Adversarial Network (GAN) architecture that generates artificial spectrograms for the minority emotional classes.

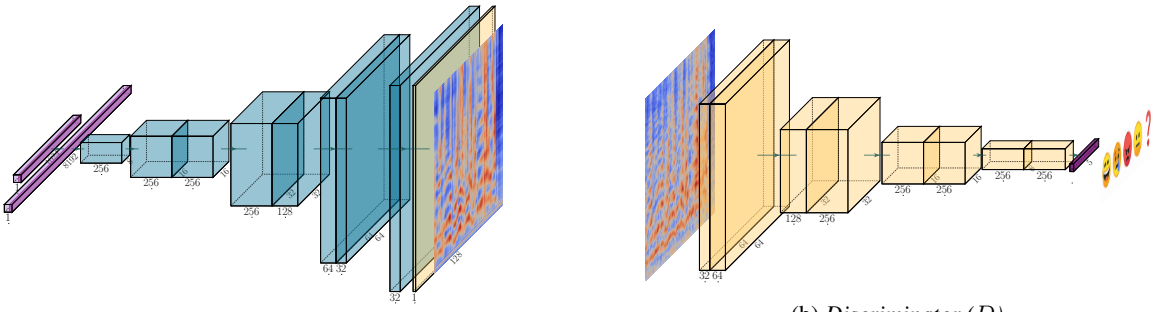
As with all classification problems, feature representation plays an important role in SER. Audio features need to efficiently characterize the emotional content, without depending on the speaker attributes or the background noise. Widely adopted hand-crafted audio representations include spectral-domain (e.g. spectral centroid and flux), cepstral-domain (e.g. MFCCs) and pitch-related features [1, 2]. Lately, spectrograms have also been used along with Convolutional Neural Networks (CNNs) as image classifiers [3, 4]. CNNs are able to deal with high-dimensional inputs and learn features that are invariant to small variations and distortions. Furthermore, it has been proved that Recurrent Neural Networks (RNNs), such as Long Short-Term Memory units (LSTM), are able to take into account the temporal information in speech, resulting in a more robust modeling of the speech signals [5, 6, 7, 4]. LSTMs can also be combined with CNNs [8, 9], in order to automatically learn the best signal representation. Spectrograms are extracted both from the speech and glottal flow signals in [10], while spectrogram encoding is performed by a stacked autoencoder and an RNN is trained to predict four primary emotions.

Data with non-uniform or highly skewed distributions among classes is a common issue in SER. During the processes of data collection and annotation, neutral speech samples are much more frequent than the emotionally-charged ones, leading to highly imbalanced datasets. A common way to address data imbalance is through data augmentation techniques. In [11] authors divide data augmentation techniques into feature-space through oversampling and data-space synthetic sample generation through transformations. Their experiments favor data-space augmentation for digit classification. In [12] seven data-space augmentation techniques are evaluated for singing voice detection on spectrogram data, with pitch shifting and random frequency filters being the most effective. Pitch augmentation has also proven to be beneficial for environmental sound classification [13] and for music genre classification [14]. For the task of SER, the work in [15] applies speed pertubations on the raw signal, while [16] proposes a combination of oversampling and vocal tract length pertubation. Recent approaches focus on learned augmentation strategies [17, 18] or GANs [19] to generate training samples.

GANs are powerful generative models that try to approximate the data distribution by training simultaneously two competing networks, a generator and a discriminator [19]. A lot of research has focused on improving the quality of generated samples and stabilizing GAN training [20, 21]. Recently, the GAN ability to generate realistic in-distribution samples has been leveraged for data augmentation. Specifically, in [22] authors train a GAN that generates in-class samples. In [23] the CycleGAN architecture [21] is adapted for emotion classification from facial expressions. As for the speech domain, in [24] synthetic feature vectors are used to improve the classifier's performance on an emotion task. In [25], a conditional GAN architecture is proposed to address data imbalance.

In this work, we extend the methodology of [25] in the SER domain, focusing on spectrogram generation for the minority emotional classes. We propose modifications in the original network architecture and the training process to improve the quality of the generated spectrograms. Extensive experimental results on the aforementioned approach and a series of other audio data augmentation techniques prove that the proposed method addresses data imbalance more effectively. To the best of our knowledge, this is the first time GANs are used to address the problem of data imbalance through data augmentation in the context of SER or other audio classification task.

http://dx.doi.org/10.21437/Interspeech.2019-2561



(a) *Generator (G)* (b) *Discriminator (D)*

Figure 1: *Architecture of the proposed GAN*

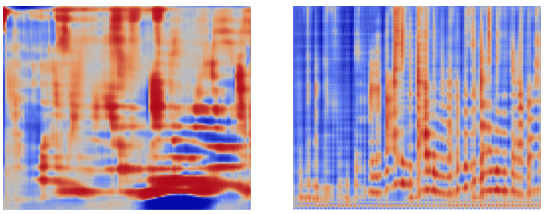
2. Proposed Method

2.1. Motivation

Real-world emotion recognition datasets suffer from data imbalance, as non-neutral emotions are usually very sparse in the initial mined data sources. However, only few works have focused on the problem of data augmentation for SER to address the imbalance issue. Most of them are limited to signal-based transformations, such as time stretching, pitch shifting and noise addition, which have been adopted for generic audio classification and music information retrieval tasks as well. We apply these approaches for comparison purposes (see Sec. 2.5 for further details). Instead, in this work, we propose using GANs to generate artificial samples for the minority classes.

2.2. Method Description

In this work, we adapt the Balancing GAN (BAGAN) methodology proposed in [25], which addresses the imbalance issue in various image classification tasks. The basic concept behind this approach is the training of a GAN to generate realistic samples for the minority class. The generator contains a series of transposed convolutional and upsampling layers, while the discriminator consists of a series of convolutional layers. However, the original architecture did not generate high-quality spectrograms as shown in Fig. 2a (Note: compare with the spectrogram in Fig. 2b generated by our proposed approach discussed next).



(a) *original BAGAN* (b) *proposed approach*

Figure 2: *Sample spectrograms generated by original BAGAN as in [25] (a) and the proposed approach (b)*

We propose the fully convolutional architecture illustrated in Fig. 1. In Fig. 1a, we show the generator G architecture, where we use 2 dense layers to project the input state to a higher dimensionality and 8 transposed convolutional layers to produce a spectrogram image. We double the stride in every second deconvolution layer to increase the height and width of the intermediate tensors. The final layer of G is a convolutional layer that converts the input tensor to a spectrogram image. In Fig. 1b, the discriminator D uses 8 convolutional layers and a softmax classification layer to discriminate between fake spectrograms and spectrograms of a specific emotion class [1].

In brief, the main steps of the proposed methodology are: (a) Autoencoder training (b) GAN initialization and (c) GAN fine-tuning.

Autoencoder Training: For faster convergence, the GAN is initialized using a pre-trained autoencoder. The autoencoder consists of the encoder which corresponds to the D architecture, replacing the last softmax layer with a dense layer of size 100, and the decoder which has the same architecture as G. The autoencoder is trained using the whole imbalanced dataset, without any explicit class knowledge. In this step, the model learns weights close to a good solution, avoiding the issue of mode collapse [19, 26, 27] during adversarial training.

GAN Initialization: The learned weights are transferred to the GAN modules - the encoder weights are transferred to D and the decoder to G respectively. For class conditioning, we calculate the mean and covariance matrix of the learned latent vectors of the autoencoder that correspond to the images of each class. In this way, we model each class with a multivariate normal distribution. Then, we sample at random a latent vector from the distribution of a specific class and provide it as input to G, which outputs a realistic spectrogram for this class. Contrary to the autoencoder, GAN has explicit class knowledge.

GAN Fine-tuning: The proposed GAN is fine-tuned using both the minority and majority classes of the training data. In this way, it learns features that are shared between classes, e.g. dominant frequencies in the spectrogram. Such features contribute to a more qualitative image generation for the minority class. During fine-tuning, G takes as input the aforementioned latent vectors, that are extracted from the class-conditional latent vector generator. The latter takes as input uniformly distributed class labels. Then, the batches of real and generated images are forwarded to D. The goal of each one of the two competing networks, G and D, is to optimize its loss function, for which sparse categorical cross-entropy is used. D is optimized to match the real images with the correct class labels and the generated ones with the fake label. As for G, it is optimized to match the labels selected by D with the labels used to generate the images. Following the GAN fine-tuning, we use G to generate artificial spectrograms for each class separately to reach the majority class population.

[1] The reason D does not perform a binary classification between real and fake samples, is that minority class samples would be misclassified as fake, due to their rarity [25].

2.3. Implementation Details

The spectrograms are normalized in the [-1, 1] range applying min-max normalization, so we use *tanh* activation at the decoder output. In both modules, batch normalization, dropout with $p = 0.2$ and leaky ReLU activations are added after each (de)convolutional layer. Real and fake samples are fed to D separately in successive batches, mainly due to the use of batch normalization. In addition, we use Adam optimizer with learning rate 5×10^{-5} when training the autoencoder and decrease it to 10^{-6} when fine-tuning the GAN.

2.4. Architecture Modifications

The proposed GAN-based augmentation method is basically differentiated from the one proposed in [25] in the following ways: (a) we replace any upsampling layer with transposed convolutions, (b) we use leaky ReLU for all the intermediate activation layers, (c) we have added batch normalization and dropout and (d) we feed the discriminator with separate batches of real and fake images. The proposed fully convolutional architecture avoids extreme values in the generated images, i.e. regions with zero and one values, as demonstrated in Fig. 2. More examples of real and synthesized spectrograms for common distinct emotion classes are demonstrated in Fig. 3. Assuming a real-world imbalanced emotion dataset, the proposed approach can generate high-quality spectrograms for underrepresented classes.

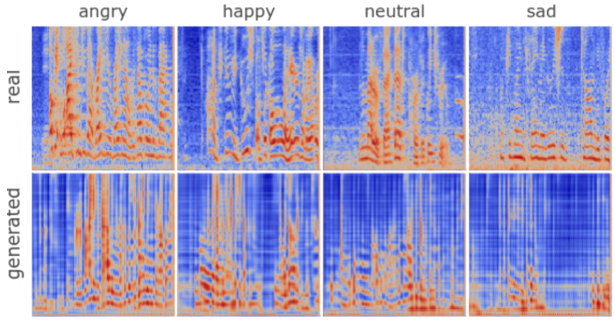


Figure 3: *Real and generated spectrograms for each emotion class, from IEMOCAP dataset.*

2.5. Baseline Methodologies

For comparison purposes, we implement a series of baseline methods to balance our initial dataset. A first approach would be the random removal of samples from the majority classes so that all classes are of equal size. This random selection can be applied with a number of ratios, considering the less populated to the most dominant class. Since this technique results in less data for training, maybe removing useful information as well, we additionally investigate various data augmentation methods. After the augmentation process, all classes are represented by the same number of samples as the majority class.

We focus on signal-based transformations, that are followed in the literature [12], [13], [14]. We apply time stretch (TS), that changes the audio signal duration without affecting its pitch, pitch shift (PS), that changes the pitch without affecting its duration and finally noise addition to the original speech utterance (either Gaussian noise, GN, or true background audio noise, BN). In the case of BN, background noise has been extracted from signals of the ESC-50 [28] and FMA [29] datasets. In addition, we try the simple technique of sample copying (CP), randomly adding identical copies of preexisting samples.

Combining all the aforementioned methods, we create a set of experimental augmentation strategies, described in Table 1: Signal-based Audio Augmentation (SA), SA with replacement (SAR), SA with replacement of the majority class only (SAR_M), SAR adding only Background Noise (SAR_B), SAR using only TS and PS (SAR_S). The replacement mentioned refers to the case of replacing audio samples with their noisy counterparts, instead of adding them. The number of samples chosen for replacement for each class is equal to the difference between the specific class population and the minority class. This method aims to balance the percentages of noise samples of every class, in an attempt to prevent any bias towards classes with unusually high or low noise distribution. It can be applied for either all the classes or only the majority.

Table 1: *Dataset augmentation strategies*

Method	CP	TS	PS	GN	BN	Replace
CP	✓	-	-	-	-	-
SA	-	✓	✓	✓	✓	-
SAR	-	✓	✓	✓	✓	All
SAR_M	-	✓	✓	✓	✓	Majority
SAR_B	-	-	-	-	✓	All
SAR_S	-	✓	✓	-	-	All

3. Experimental Evaluation

3.1. Datasets Description

IEMOCAP (interactive emotional dyadic motion capture database) is a widely adopted corpus for emotional data, collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC) [30]. It has been recorded from ten actors in dyadic sessions, including both emotional scripts and improvised hypothetical scenarios. The scenarios have been designed to elicit specific types of emotions, namely: happiness, anger, sadness, frustration and neutral state, while additional emotions (excitement, fear, disgust, surprise and other) are also included in the final annotations. It contains approximately 12 hours of speech and it is considered a standard in most of the SER publications during the last years. In this work, we use four emotion classes: $angry$, $happy$, sad and $neutral$, merging the $happy$ and $excited$ classes, which results in 5531 speech utterances of about 7 hours total duration.

Despite its wide adoption, the IEMOCAP dataset (a) contains limited number of speakers and (b) is quite balanced. On the contrary, in the real world, high imbalance can be noticed, as well as diversity of different domains. Therefore, part of our internal (not publicly available) dataset, FEEL-$25k$, is also used to evaluate the augmentation methods. In particular, FEEL-$25k$ contains almost $25k$ utterances from several domains, including films, TV series and podcasts. Its total duration is approximately 49 hours and the ratio of the less populated (sad) to the most dominant (neutral) class is $1/5$. The emotion classes are: $angry$, $happy$, $neutral$, sad and $ambiguous$. The latter contains speech samples for which the inter-annotator agreement was lower than a particular threshold. Each segment has

been labeled by 3 to 7 human annotators. A separate and large dataset, which is constructed similarly to FEEL-25k and consists of data drawn from the same broad domains, is used for testing. It is composed of almost 50k utterances of 100 hours of total duration.

3.2. Experimental Setup

Feature Extraction and Classification: The data augmentation methods have been evaluated in terms of the classification performance of a CNN. In particular, we have chosen the VGG19 architecture [31], which results in state-of-the-art performance on IEMOCAP. The network takes as input mel-scaled spectrograms, that are extracted from fix-sized segments of 3 seconds, after breaking each spoken utterance. During the spectrogram extraction a short-term window of 50 mseconds with a 50% overlap ratio has been adopted, while the number of Mel coefficients is 128. This results in fix-sized spectrograms of 128×128. Logarithmic scale has been applied after the frequency power calculation.

Train - Test Data Split: For the evaluation experiments on IEMOCAP, we use 5 fold cross-validation, namely leave-one-session-out, using 4 sessions for training and 1 for testing. This setup is a common practice for IEMOCAP in related SER publications. As far as FEEL-25k is concerned, cross-validation is not needed due to the dataset's size and diversity. Instead, we have used a shuffle split of 80% - 20% for training and validation respectively. A separate dataset is used for testing, as explained in Sec. 3.1.

For both datasets, we perform spectrogram normalization (see Sec. 2.3), computing the parameters from the training set and applying them to the validation and test sets. We report the average performance on the test set, after calculating the majority voting of the segment-level labels for every utterance. When applying this classification scheme on the whole IEMOCAP dataset, we achieve an Unweighted Average Recall (UAR) of 56%, which shows a performance improvement of about 1.2% in comparison to the non pre-trained AlexNet and VGG16 [4].

Datasets Imbalance Strategy: Since IEMOCAP is almost balanced, we simulate the imbalance issue for each emotional class separately, i.e. $happy$, $angry$ and sad, except $neutral$. For every class, we remove 80% of the specific class from the training set, selected at random, in order to reproduce the difficulty of the classification task when this class is underrepresented. The validation set remains unmodified. In the case of FEEL-25k, which is gathered "in the wild" and as a result is imbalanced, we apply directly the audio data augmentation methodologies. The resulting training set in both datasets is then augmented using the aforementioned approaches.

3.3. Performance Results

In this section, we present the performance results for both datasets. In Table 2 we demonstrate the performance achieved on IEMOCAP. We use UAR metric to be comparable with other works in the literature. Each column named after an emotional class corresponds to the simulation described in Sec. 3.2, where we remove the 80% of the class samples in the training set and then augment it using one of the methodologies. In the final column, we compute the average scores of those simulations to assess the overall performance. The rows correspond to the different augmentation methods, as described in Sec. 2. For IEMOCAP, we did not try any random undersampling, since the minority class in the imbalanced training set contains a tiny amount of samples (approximately 180), making the CNN training almost impossible. We see that the proposed approach achieves almost 10% relative performance improvement.

Table 2: *IEMOCAP performance (UAR %)*

Dataset	Angry	Happy	Sad	Average
Imbalanced	47.8	52.2	46.9	49.0
CP	51.5	50.8	45.8	49.4
SA	49.7	49.6	47.6	49.0
Proposed approach	**53.5**	**55.2**	**52.1**	**53.6**

Extensive experimental results are presented in Table 3 for FEEL-25k for the various augmentation methods. We show both the UAR and F-score results, since F-score computation combines both recall and precision. It can be observed that all the attempts to balance the dataset give suboptimal results in comparison to the initial distribution, with the exception of data generation using the proposed approach, which achieves almost 5% relative improvement. In general, the signal-based transformations can lead to overfitting, due to the existence of similar samples in the training set, while random balance removes possibly useful information. On the contrary, the GAN-based augmentation method generates high-quality spectrograms. After the fine-tuning, it can be easily used to generate as many spectrograms as needed for the underrepresented emotion classes.

Table 3: *FEEL-25k performance (UAR & F-score %)*

Category	Dataset	UAR	F-score
-	Initial Dataset	52.3	52.7
Random Selection	0.4 Balanced	50.0	50.3
	0.6 Balanced	48.5	48.1
	0.8 Balanced	49.6	49.5
	Fully Balanced	49.4	48.9
Signal-based Augmentation	CP	51.1	49.8
	SA	50.7	49.2
	SAR	51.2	50.0
	SAR_M	51.0	50.5
	SAR_B	51.1	49.7
	SAR_S	49.3	48.0
Generation	Proposed approach	**54.6**	**55.0**

4. Conclusion

In this work, we propose a GAN architecture for in-class spectrogram generation to address the data imbalance issue for the task of SER, by augmenting the underrepresented classes. Through extensive experimentation, we provide conclusive evidence that the proposed approach is more effective in comparison to standard augmentation techniques. We provide experimental results both on IEMOCAP and FEEL-25k, showcasing the applicability of our approach, which boosts the performance with a relative improvement of 5% to 10%. In the future, we plan to combine LSTMs with our CNN classifier to take into account temporal information [8]. Additionally, we will try more sophisticated conditioning techniques and incorporate ideas from GANs for raw audio synthesis (e.g. WaveGAN [32]) to directly generate audio samples.

5. References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 1263–1267.

[3] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60 – 68, 2017, advances in Cognitive Engineering Using Neural Networks.

[4] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Nov 2018.

[5] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *IEEE Workshop on Spoken Language Technology*, Dec. 2018, pp. 126–131.

[6] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, Oct 2018.

[7] E. Tzinis, G. Paraskevopoulos, C. Baziotis, and A. Potamianos, "Integrating recurrence dynamics for speech emotion recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 927–931.

[8] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5200–5204.

[9] C. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, July 2017, pp. 583–588.

[10] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation Learning for Speech Emotion Recognition," in *Interspeech 2016*, September 2016, pp. 3603–3607.

[11] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2016, pp. 1–6.

[12] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *ISMIR*, January 2015.

[13] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, March 2017.

[14] L. Rafael Aguiar, M. G. Yandre Costa, and N. Carlos Silla, "Exploring data augmentation to improve music genre classification with convnets," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.

[15] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 2741–2745.

[16] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "Cnn+lstm architecture for speech emotion recognition with data augmentation," in *Workshop on Speech, Music and Mind 2018*, September 2018, pp. 21–25.

[17] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *CoRR*, vol. abs/1805.09501, 2018.

[18] J. Lemley, S. Bazrafkan, and P. Corcoran, "Smart augmentation learning an optimal data augmentation strategy," *IEEE Access*, vol. 5, pp. 5858–5869, 2017.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[21] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[22] A. Antoniou, A. J. Storkey, and H. A. Edwards, "Data augmentation generative adversarial networks," *CoRR*, vol. abs/1711.04340, 2018.

[23] X. Zhu, Y. Liu, Z. Qin, and J. Li, "Data augmentation in emotion classification using generative adversarial networks," *arXiv preprint arXiv:1711.00648*, 2017.

[24] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," *Interspeech 2018*, September 2018.

[25] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "Bagan: Data augmentation with balancing gan," *arXiv preprint arXiv:1803.09655*, 2018.

[26] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Advances in Neural Information Processing Systems*, December 2017.

[27] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," in *Advances in Neural Information Processing Systems*, December 2017, pp. 3308–3318.

[28] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, 2015, pp. 1015–1018.

[29] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," in *18th International Society for Music Information Retrieval Conference*, 2017.

[30] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[32] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations*, 2019.