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

CycleGAN-based Emotion Style Transfer as Data Augmentation for Speech Emotion Recognition

Fang Bao, Michael Neumann, Ngoc Thang Vu

Institute for Natural Language Processing (IMS), University of Stuttgart, Stuttgart, Germany
baofg@studi.informatik.uni-stuttgart.de, {michael.neumann|thang.vu}@ims.uni-stuttgart.de

Abstract

Cycle consistent adversarial networks (CycleGAN) have shown great success in image style transfer with unpaired datasets. Inspired by this, we investigate emotion style transfer to generate synthetic data, which aims at addressing the data scarcity problem in speech emotion recognition. Specifically, we propose a CycleGAN-based method to transfer feature vectors extracted from a large unlabeled speech corpus into synthetic features representing the given target emotions. We extend the CycleGAN framework with a classification loss which improves the discriminability of the generated data. To show the effectiveness of the proposed method, we present results for speech emotion recognition using the generated feature vectors as (i) augmentation of the training data, and (ii) as standalone training set. Our experimental results reveal that when utilizing synthetic feature vectors, the classification performance improves in within-corpus and cross-corpus evaluation.

Index Terms: speech emotion recognition, data augmentation, cycle-consistent generative adversarial networks

1. Introduction

Automatic speech emotion recognition (SER) is an important and rapidly growing research field with great potential to improve naturalistic voice-based human-computer interfaces. It has been acknowledged that data scarcity is one of the major challenges in this field [1], which is reflected not only in the lack of large, naturalistic labeled speech corpora, but also by the unbalanced distribution over emotions [2]. To approach both problems, we propose a method based on CycleGANs [3] to generate feature vectors representing a given target emotion. This way, the proportion of emotional categories can be controlled, thus building a large and balanced synthetic dataset.

Generative Adversarial Networks (GANs) [4] have successfully been applied to a variety of computer vision tasks as well as to speech-related applications, such as speech enhancement [5] and voice conversion [6]. Adversarial training schemes have also been used for SER recently. Sahu et al. deployed adversarial auto-encoders [7] to represent emotional speech in compressed feature space while maintaining the discriminability between emotion classes [8]. Chang and Scherer utilized a deep convolutional GAN to learn a discriminative representation of emotional speech in a semi-supervised way [9]. Han et al. proposed a conditional adversarial training framework consisting of two networks trained in an adversarial manner. One learns to predict dimensional representations of emotions, while the other aims at distinguishing between predictions and real labels from the dataset [10].

Moreover, GANs can also be used for synthetic data generation to improve classification performance. Sahu et al. have shown this by using the reconstructed samples from the adversarial auto-encoder as synthetic training data [8]. In a follow-up study, they investigated the use of vanilla GAN and conditional GAN for generating high-dimensional (1582-d) feature vectors from a low-dimensional (2-d) space [11]. It was shown that a vanilla GAN cannot achieve convergence and the conditional GAN only converges when it is initialized with pre-trained weights and the power of its discriminator is limited. The classification performance has been improved by augmenting the training dataset with synthetic feature vectors. Based on this work, we propose to generate synthetic feature vectors through emotion style transfer.

Previously, emotion style transfer has mainly been researched in the area of speech synthesis [12, 13]. To the best of our knowledge, this paper is the first to investigate emotion style transfer for improving classification performance of SER. The approach is inspired by remarkable recent advances in unsupervised image-to-image translation [3, 14, 15, 16, 17, 18]. All these approaches have in common that a mapping between source and target domain can be learned without paired training data. In this paper, we first introduce our model which adapts CycleGANs to generate synthetic feature vectors with transferred emotions from a large unlabeled speech corpus. Then, we present an analysis of the quality of generated samples, and finally, report results for SER using the generated samples to augment the training dataset. Our contributions are as follows:

- We introduce emotion style transfer for generating synthetic feature vectors for the purpose of classification.
- We propose a novel CycleGAN-based architecture that ensures similarity between real and generated samples on the one hand and provides discriminability among generated samples on the other hand.
- We show that a neural network classifier trained on the combination of real and synthetic feature vectors achieves better classification performance than the classifier trained only on the real feature vectors.

2. Proposed method

2.1. CycleGAN as component

Given a labeled dataset X with N emotion classes, we generate synthetic samples for each emotion i using one CycleGAN. As shown in the left part of Fig. 1, the CycleGAN establishes a bijective mapping between a source domain S and a target domain T_i, where S is an external unlabeled dataset and T_i represents the samples of emotion i in the labeled dataset X. The two mapping functions G_i and F_i are used for translating from source to target and from target to source, respectively. The adversarial discriminator D_i^T encourages G_i to generate synthetic targets indistinguishable from real samples. The adversarial loss for

http://dx.doi.org/10.21437/Interspeech.2019-2293

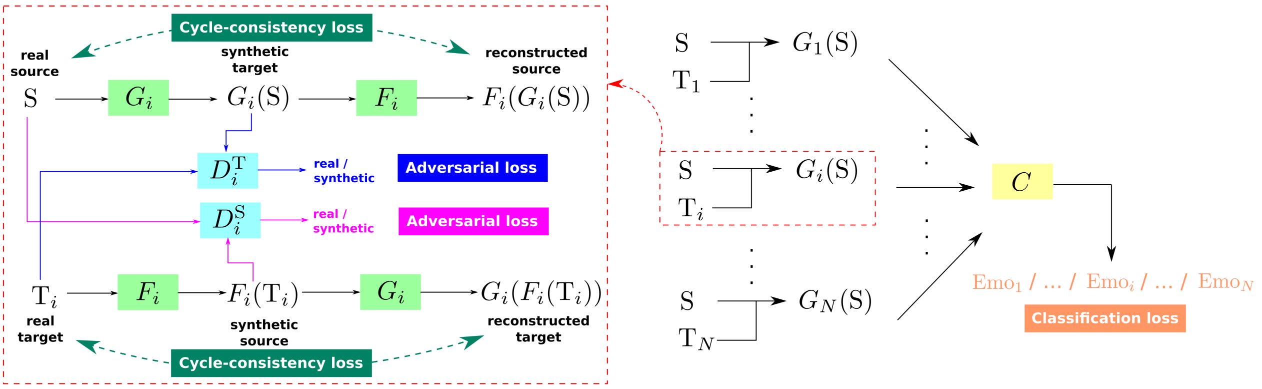


Figure 1: *Our model consists of N CycleGANs and a domain classifier, where N is the number of emotions to be classified. For each emotion i, we have a CycleGAN with two discriminators D_i^T, D_i^S and two mapping functions G_i, F_i as generators. The output of the mapping G_i from the source S is the desired synthetic samples $G_i(S)$. Cycle-consistency loss is built between real samples and their corresponding reconstructed samples. The domain classifier C learns to ensure the discriminability between the generated samples.*

G_i and D_i^{T} is defined as

$$\begin{aligned}\mathcal{L}_i^{\text{GAN}}(G_i, D_i^{\text{T}}, \text{S}, \text{T}_i) = & \mathop{\mathbb{E}}_{\text{t}\sim p_{\text{t}}}[\log D_i^{\text{T}}(\text{t})] \\ & + \mathop{\mathbb{E}}_{\text{s}\sim p_{\text{s}}}[\log(1 - D_i^{\text{T}}(G_i(\text{s})))]\end{aligned} \tag{1}$$

Similarly, for the generator F_i and the discriminator D_i^{S} we employ the adversarial loss $\mathcal{L}_i^{\text{GAN}}(F_i, D_i^{\text{S}}, \text{S}, \text{T}_i)$. The total adversarial loss is defined as

$$\begin{aligned}\mathcal{L}_i^{\text{GAN}}(G_i, F_i, D_i^{\text{T}}, D_i^{\text{S}}, \text{S}, \text{T}_i) = & \mathcal{L}_i^{\text{GAN}}(G_i, D_i^{\text{T}}, \text{S}, \text{T}_i) \\ & + \mathcal{L}_i^{\text{GAN}}(F_i, D_i^{\text{S}}, \text{S}, \text{T}_i)\end{aligned} \tag{2}$$

The generators G_i and F_i try to minimize it while the discriminators D_i^{T} and D_i^{S} try to maximize it.

In addition, a CycleGAN regularizes the adversarial training with a cycle consistency loss. It translates the synthetic target $G_i(\text{S})$ back to the source domain and computes the mean squared error (MSE) between the real source S and reconstruction $F_i(G_i(\text{S}))$. The same is done for T_i and the reconstructed target $G_i(F_i(\text{T}_i))$. Consequently, the total cycle consistency loss is defined as follows:

$$\begin{aligned}\mathcal{L}_i^{\text{cyc}}(G_i, F_i, \text{S}, \text{T}_i) = & \mathop{\mathbb{E}}_{\text{s}\sim p_{\text{s}}}[\|(F_i(G_i(\text{s})) - \text{s})\|_2^2] \\ & + \mathop{\mathbb{E}}_{\text{t}\sim p_{\text{t}}}[\|G_i(F_i(\text{t})) - \text{t}\|_2^2]\end{aligned} \tag{3}$$

2.2. Discriminability between generated samples

The bijective mapping of CycleGAN ensures similarity between the distribution of real and synthetic data. However, to improve classification performance, we need to learn a generalized distribution from real data samples instead of merely reconstructing the exact same distribution. Therefore, we add a classifier C to discriminate between the generated data of each emotion class, which is displayed in the right part of Fig. 1. The classification loss can be defined as a softmax cross-entropy loss:

$$\mathcal{L}^{\text{cls}} = \sum_i y_i \log(C(G_i(\text{S}))) \tag{4}$$

where y_i is the label of the target emotion i. The overall loss for our method is defined as

$$\mathcal{L} = \sum_i \mathcal{L}_i^{\text{GAN}} + \lambda^{\text{cyc}} \sum_i \mathcal{L}_i^{\text{cyc}} + \lambda^{\text{cls}} \mathcal{L}^{\text{cls}} \tag{5}$$

The parameters λ^{cyc} and λ^{cls} are weights for cycle-consistency loss and classification loss, respectively. They affect the similarity of generated feature vectors to real data and the discriminability between emotions.

3. Data and features

3.1. Datasets

IEMOCAP. We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [19] database as labeled data. It contains five sessions of English dyadic conversations (one female and one male actor in each session). The entire corpus contains 10,039 utterances annotated with 10 emotion classes. To ensure comparability with related work, we use four of them: angry, happy, sad and neutral, where the happy class also incorporates the samples labeled as excited, resulting in 5,531 samples total [1] comprising approximately 7 hours of speech data.

MSP-IMPROV. For cross-corpus evaluation we utilize the MSP-Improv database [20] as testing set. It also contains English dyadic interactions between actors. There are 7,798 samples from 12 speakers across the same four emotion classes. [2]

TEDLIUM. We use the Tedlium corpus (release 2) [21] as unlabeled source data for generating synthetic feature vectors. It contains 1,495 Ted talks comprising 207 hours of English speech. The talks have been segmented according to the timing information in the transcripts, resulting in 92,973 segments.

3.2. Features

We use the openSMILE toolkit [22] for extracting the 'emobase2010' reference feature set for each utterance. It is based on the Interspeech 2010 Paralinguistic Challenge feature set [23] and consists of 1,582 features which are multiple functionals computed from a set of acoustic low level descriptors.

4. Experiments

4.1. Setup

Since there are four emotions to be classified, our model consists of four generators, four discriminators and one classifier.

[1] Distribution: 1,103 angry, 1,636 happy, 1,708 neutral, 1,084 sad

[2] Distribution: 792 angry, 2,644 happy, 3,477 neutral, 885 sad

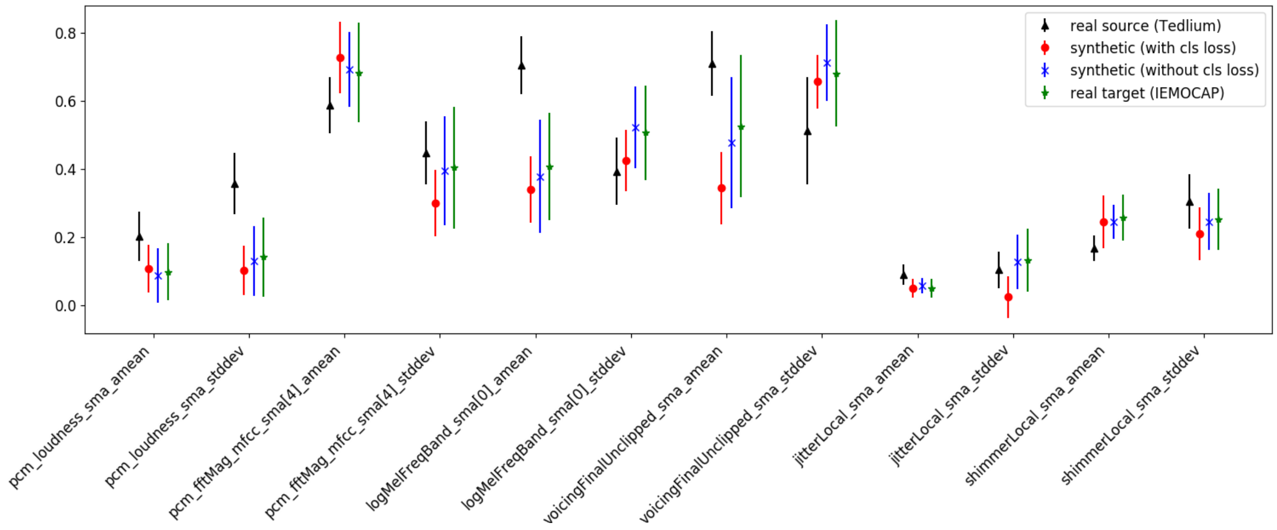


Figure 2: *Comparison of statistical distribution of the source, synthetic and target features. Due to space limitations, we show average value of the four emotion classes for only 12 features (representing different feature categories). It can be seen that the synthetic features are close to the desired target distribution, especially when no classification loss is involved (blue data points).*

They are all implemented by feed-forward neural networks. Each generator has three hidden layers with 1000, 500, 1000 neurons, respectively. Each discriminator has two hidden layers with 1000 neurons each. The classifier has two hidden layers with 100 neurons each. For all hidden layers Leaky Rectified Linear Units (leaky ReLUs) are used as activation function.

Due to the difficulty for generators to learn a high-dimensional distribution, we pre-train each pair of the generators G_i and F_i on their corresponding source and target data for 10k epochs with a learning rate of 0.0002 and a dropout of 0.2. The source data S consists of the full Tedlium corpus in each case, and the target data T_i consists of the particular portion of IEMOCAP annotated with emotion class i.

Initialized with the pre-trained weights for generators, our model is trained for 2k epochs with four parallel CycleGANs which transfer the unlabeled data to each of the target emotions individually. To reduce loss oscillation, the initial learning rate is set to 0.0002 and is linearly decayed every 50 epochs by a factor of 0.8. To balance the generators and discriminators, we update the generators twice and the discriminators once at each iteration. Besides that, we use one-sided label smoothing as introduced by [24]. For both training and pre-training we use Adam optimization [25] and a batch size of 64.

Our experiments are implemented with TensorFlow (v 1.10.0) [26]. In terms of preprocessing, min-max normalization is used for synthetic features generation. For classification we scale the features on each dataset with z-normalization separately, because Zhang et. al. [27] have shown that z-normalization yields an improvement over min-max normalization for cross-corpus classification.

4.2. Emotion transfer

In this experiment, we test the feasibility of adapting CycleGANs to emotion style transfer in feature space. The main objective is to generate feature vectors that preserve the distribution of the real target samples. We train our CycleGAN framework in two different configurations: (a) *without* classification loss (we set $\lambda^{\text{cls}} = 0$ in equation 5), and (b) *with* classification loss, setting $\lambda^{\text{cls}} = 2$. For both setups we set $\lambda^{\text{cyc}} = 5$.

We compare the distribution of the unlabeled source data, the emotional synthetic data and the target data in feature space. Fig. 2 shows for a subset of features that the synthetic and target feature vectors are similar in both mean and standard deviation, which means the source data are transferred to the four emotional categories successfully. To verify what is exemplified in Fig. 2, we visually inspected the distribution of all 1,582 features and calculated the mean of absolute differences between all feature values from different datasets, presented in Table 1. For this investigation all features have been normalized using min-max normalization.

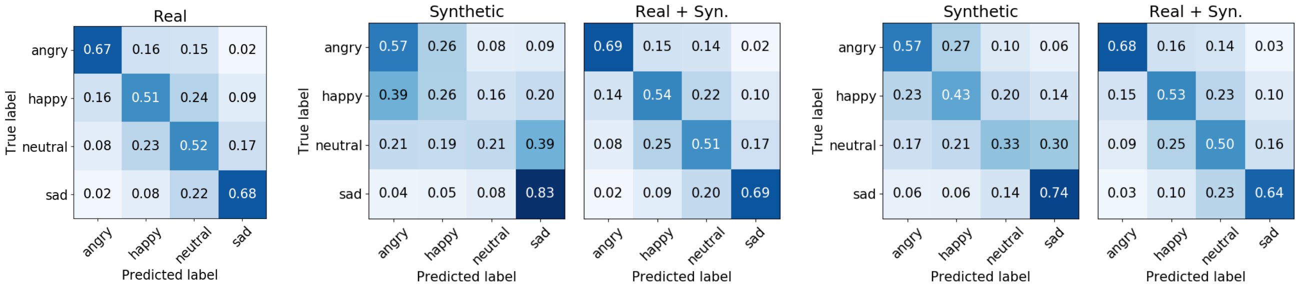
Table 1: *Mean of absolute differences between all feature values from the source, synthetic and target dataset.*

$\lvert\text{target} - \text{source}\rvert$	0.084
$\lvert\text{target} - \text{syn}_{(\lambda^{\text{cls}}=0)}\rvert$	0.013
$\lvert\text{target} - \text{syn}_{(\lambda^{\text{cls}}=2)}\rvert$	0.049

4.3. Within-corpus evaluation

In this experiment, we build three feed-forward neural network classifiers which are trained on: (i) only real samples taken from IEMOCAP, (ii) only synthetic features and (iii) the combination of both. We perform leave-one-session-out cross-validation on IEMOCAP to ensure that results are speaker-independent.The hyper-parameters are: 2 hidden layers in each case, for (i): 100 neurons per layer, batch size of 64, dropout of 0.2, learning rate 1e-5, 70 training epochs, for (ii): 200 neurons, batch size 256, dropout 0.5, learning rate 1e-5, 5 epochs, and for (iii): 1000 neurons, batch size 256, dropout 0.5, learning rate 5e-6, 30 epochs. We report unweighted average recall (UAR) as performance measure. Since the neural networks are initialized with random weights, we repeat all experiments five times and report mean and standard deviation of the results.

Table 2 shows our results and, for comparison, the results reported by Sahu et. al. [11] for the three experimental settings. It can be seen that the baseline performance without using synthetic data is comparable to [11]. Using the combined dataset (real + syn.), we achieve an improvement over the baseline



Real

	angry	happy	neutral	sad
angry	0.67	0.16	0.15	0.02
happy	0.16	0.51	0.24	0.09
neutral	0.08	0.23	0.52	0.17
sad	0.02	0.08	0.22	0.68

(a) *Baseline without synthetic data.*

Synthetic

	angry	happy	neutral	sad
angry	0.57	0.26	0.08	0.09
happy	0.39	0.26	0.16	0.20
neutral	0.21	0.19	0.21	0.39
sad	0.04	0.05	0.08	0.83

Real + Syn.

	angry	happy	neutral	sad
angry	0.69	0.15	0.14	0.02
happy	0.14	0.54	0.22	0.10
neutral	0.08	0.25	0.51	0.17
sad	0.02	0.09	0.20	0.69

(b) *Synthetic feature vectors generated* ***with*** *classification loss ($\lambda^{cls} = 2$).*

Synthetic

	angry	happy	neutral	sad
angry	0.57	0.27	0.10	0.06
happy	0.23	0.43	0.20	0.14
neutral	0.17	0.21	0.33	0.30
sad	0.06	0.06	0.14	0.74

Real + Syn.

	angry	happy	neutral	sad
angry	0.68	0.16	0.14	0.03
happy	0.15	0.53	0.23	0.10
neutral	0.09	0.25	0.50	0.16
sad	0.03	0.10	0.23	0.64

(c) *Synthetic feature vectors generated* ***without*** *classification loss ($\lambda^{cls} = 0$).*

Figure 3: *Averaged results on IEMOCAP.*

when incorporating the classification loss into the CycleGAN ($\lambda^{\text{cls}} = 2$). Augmenting the dataset with synthetic features generated without this loss does not yield an improvement.

Table 2: *Results for cross-validation evaluation on IEMOCAP.*

	Real	**Syn.**	**Real + Syn.**
Sahu [11]	59.42	34.09	60.29
$\lambda^{\text{cls}} = 0$	**59.48 ± 0.71**	**51.57 ± 0.60**	58.79 ± 0.77
$\lambda^{\text{cls}} = 2$		46.59 ± 0.75	**60.37 ± 0.70**

Using only synthetically generated samples as training data, we observe a significantly higher performance on the test set (51.57%) than reported in [11], which implies that our CycleGAN approach generates feature vectors that are closer to the underlying distribution of real data. Interestingly, the UAR for the setting with $\lambda^{\text{cls}} = 2$ is notably lower than for $\lambda^{\text{cls}} = 0$. To gain a deeper understanding of the performance differences, we analyzed the prediction errors, shown in Fig. 3a- 3c.

It can be seen from the confusion matrices that the predictions and error patterns based on the augmented dataset (real + syn., right-hand sides of Fig. 3b, 3c) are similar to the baseline (Fig 3a). For the setting *with* classification loss (Fig. 3b), we observe improvements for the three classes angry, happy, sad – whereas the result for sad drops below the baseline in the setting *without* classification loss (Fig. 3c).

Substantial differences between those configurations are observed in the predictions when using only synthetic data as train set (left-hand sides of Fig. 3b, 3c). For $\lambda^{\text{cls}} = 2$, the model appears to have a strong bias towards the classes angry and sad, given the high proportions of incorrect predictions of those two classes. For $\lambda^{\text{cls}} = 0$, the proportions of samples wrongly predicted as sad and angry, respectively, are also high, but Fig. 3c presents a more balanced confusion matrix for synthetic samples overall. The total UAR of 51.57% is higher than for $\lambda^{\text{cls}} = 2$ and the bias towards sad and angry not as severe.

These findings show that the proposed classification loss in our CycleGAN framework can in fact improve classification results (for real + syn.), but could potentially introduce a bias towards certain categories. In addition, we have recognized a strong overfitting problem when training *only* on synthetically generated feature vectors.

4.4. Cross-corpus evaluation

To investigate whether the generated synthetic samples are useful when applying a model to another, unseen dataset, we perform cross-corpus evaluation in the same three setups as described in section 4.3, using MSP-IMPROV as test set. We take 30% of the samples as development set for hyper-parameter tuning and the remaining 70% as test set, keeping class proportions equal in both sets. For (ii) *Syn.* and (iii) *Real + syn.*, the following hyper-parameters differ from the within-corpus setup: 200 neurons per layer, dropout of 0.8, learning rate of 1e-5 and 20 training epochs for both setups. The high dropout rate appears to be necessary because of the overfitting problem with synthetic samples.

Table 3: *Results for cross-corpus evaluation on MSP-IMRPOV.*

	Real	**Syn.**	**Real + Syn.**
Sahu [11]	45.14	33.96	45.40
$\lambda^{\text{cls}} = 0$	**45.58 ± 0.40**	39.35 ± 0.33	42.61 ± 0.34
$\lambda^{\text{cls}} = 2$		**41.58 ± 1.29**	**46.52 ± 0.43**

The results in Table 3 exhibit similar characteristics as the results for within-corpus evaluation. They show that the addition of synthetically generated training samples can improve classification performance for data which was not involved in CycleGAN training. When using only synthetic training samples, the UAR for $\lambda^{\text{cls}} = 2$ is higher than for $\lambda^{\text{cls}} = 0$, suggesting that the introduced classification loss is beneficial for cross-domain scenarios.

5. Conclusion

Data augmentation via generative adversarial networks is a relevant topic for SER. In contrast to previous methods which generate synthetic feature vectors from a low-dimensional space, we propose a CycleGAN-based method to transfer unlabeled data into different target emotions. Our experiments have shown a considerable similarity between the distribution of synthetic and target feature vectors. Furthermore, we introduced a classification loss to the network architecture to enable the synthetic samples to be distinguishable. Experiments on IEMOCAP and MSP-IMPROV have shown improvements in classification performance over previous methods when training on synthetic features as well as on the combination of real and synthetic samples. However, we also see a bias towards certain categories in the synthetic data and overfitting when training only on these samples. These problems need to be solved in future work, possible directions are varying the weight λ^{cls} to find the optimal balance between similarity and discriminability as well as utilizing several speech emotion corpora as target data for CycleGAN training.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and languagestate-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] 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.

[3] 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.

[4] 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.

[5] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *Proc. of Interspeech*, 2017.

[6] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," in *26th European Signal Processing Conference*, Rome, Italy, 2018.

[7] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," in *Proc. of International Conference on Learning Representations (ICLR)*, 2016.

[8] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," in *Proc. of Interspeech*, 2017.

[9] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2746–2750.

[10] J. Han, Z. Zhang, Z. Ren, F. Ringeval, and B. Schuller, "Towards conditional adversarial training for predicting emotions from speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6822–6826.

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

[12] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1145–1154, 2006.

[13] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken english," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[14] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1857–1865.

[15] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training." in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, no. 4, 2017.

[16] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *International Conference for Learning Representations (ICLR)*, 2017.

[17] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 700–708.

[18] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2868–2876.

[19] 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, 2008.

[20] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, 2017.

[21] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the tedlium corpus with selected data for language modeling and more ted talks." in *Proc. of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014.

[22] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM international conference on Multimedia*. ACM, 2013.

[23] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, S. S. Narayanan *et al.*, "The interspeech 2010 paralinguistic challenge." in *InterSpeech*, vol. 2010, 2010.

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems*, 2016.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations (ICLR)*, 2015.

[26] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016.

[27] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 523–528.