A CONDITIONAL CYCLE EMOTION GAN FOR CROSS CORPUS SPEECH EMOTION RECOGNITION

Bo-Hao Su[1,2], *Chi-Chun Lee*[1,2]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[2]MOST Joint Research Center for AI Technology and All Vista Healthcare, Taiwan

2021 IEEE Spoken Language Technology Workshop (SLT) | 978-1-7281-7066-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/SLT48900.2021.9383512

ABSTRACT

Speech emotion recognition (SER) is important in enabling personalized services and multimedia applications in our life. It also becomes a prevalent topic of research with its potential in creating a better user experience across many modern technologies. However, the highly contextualized scenario and expensive emotion labeling required cause a severe mismatch between already limited-in-scale speech emotional corpora; this hinders the wide adoption of SER. In this work, instead of conventionally learning a common feature space between corpora, we take a novel approach in enhancing the variability of the source (labeled) corpus that is target (unlabeled) data-aware by generating synthetic source domain data using a conditional cycle emotion generative adversarial network (CCEmoGAN). Note that no target samples with label are used during whole training process. We evaluate our framework in cross corpus emotion recognition tasks and obtain a three classes valence recognition accuracy of 47.56%, 50.11% and activation accuracy of 51.13%, 65.7% when transferring from the IEMOCAP to the CIT dataset, and the IEMOCAP to the MSP-IMPROV dataset respectively. The benefit of increasing target domain-aware variability in the source domain to improve emotion discriminability in cross corpus emotion recognition is further visualized in our augmented data space.

Index Terms— speech emotion recognition, conditional cycle GAN, cross corpus, data augmentation, transfer learning

1. INTRODUCTION

Rapid progress in deep learning algorithms is key in driving advancement of technologies for human-centered services, e.g., vision-based behavior detection [1], sound-based multimedia applications [2], and natural language understanding [3]. As these services become more integrated into our daily life, the ability to automatically sense emotional states has become a critical component. For example, speech emotion recognition (SER) technology has enabled many applications, such as homecare platforms or devices [4], and intelligent vehicle assistance [5], to become more ubiquitous and pervasive. However, the existing emotional speech corpora are often highly contextualized for specific scenarios or interaction settings, which create severe idiosyncratic variations hindering the generalization of current SER algorithms across corpora. Developing sophisticated learning algorithms to perform unsupervised domain adaptation from existing labeled emotional corpora to unlabeled databases is crucial to guarantee the robustness and ease of wide applicability of SER systems in real world in-the-wild applications.

In fact, more and more computational studies have started to investigate different unsupervised learning strategies in handling the mismatch between emotional corpora to alleviate the issue of corpus-specific discrepancy to generalize SER. A variety of techniques have already been proposed in the literature, e.g., from simply eliminating cross corpus acoustic feature value differences through normalization scheme [6] to more sophisticated approaches in deriving common feature space through domain adaptation [7] or matrix factorization [8]. This problem has also been cast as an unsupervised transfer learning problem [9, 10] and also as a multi-task optimization [11]. The current state-of-art approach is to learn a domain-invariant acoustic representation for task of cross corpus SER with an adversarial strategy [12].

While many of these past studies have demonstrated improved cross corpus SER by deriving a generic context-invariant feature representation, this strategy still suffers from robustness issues due to its lack of ability to enhance the data variability in a manner that also achieves source emotion label consistency jointly. Not only these speech emotion corpora are unique in its own way, most of them are usually limited in-scale; hence, by directly learning a hidden layer that is domain invariant between these corpora, its representational power is likely to suffer from inadequate variability naturally limited by the amount of available emotional speech data. Furthermore, by simply ensuring two databases acoustic representations align in an overlapping space, it does not guarantee that similar samples between corpus would have similar emotion labels; this phenomenon is termed as label distortion that is especially common and also detrimental when performing cross corpus unsupervised emotion recognition. In this work, instead of seeking a common acoustic representation space, we use target-and-source bidirectional data augmentation as an strategy to increase the variability

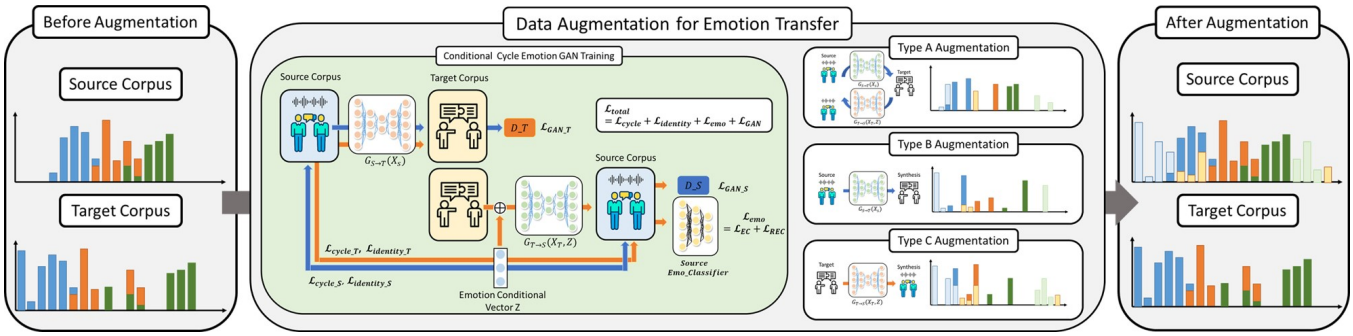

Fig. 1: Architecture of cross corpus speech emotion recognition using our proposed conditional cycle emotion GAN data augmentation.

in an emotionally consistent manner to improve the emotion transferability from source (labeled) to target (unlabeled) dataset.

Specifically, we learn a condition cycle generative adversarial network inspired by works of [13, 14], which learns a bi-directional mapping function between source and target data samples with an additional emotion conditional vector to constraint the generative adversarial training. We then generate synthesized source domain samples that are *target* domain-aware using the proposed conditional cycle emotion GAN (CCEmoGAN). Further, each of the CCEmoGAN-generated source domain data sample is assigned with original emotion labels on the real source sample from which it is generated from. After carrying out this data generation process, we can then easily train a emotion classifier on the *augmented* source dataset and directly use this recognition network to perform unsupervised target domain emotion classification. We evaluate our framework using three public speech emotion corpora, i.e, the IEMOCAP, the MSP-IMPROV, and the CreativeIT (CIT). We regard the IEMOCAP as the source domain and perform emotion recognition on the target domains, i.e., the MSP-IMPROV and the CIT. This data augmentation strategy surpasses the current state-of-the-art cross corpora method of DANN [12] by 4.85%, 2.94% in valence for the task of using the IEMOCAP model on the CIT and also using the IEMOCAP model on the MSP-IMPROV respectively; similar results have been observed in activation, i.e., 7.03%, 4.8%.

2. RESEARCH METHODOLOGY

In this work, we define source corpus as the dataset with emotion labels (the IEMOCAP), and target corpus as the dataset without labeling (the MSP-IMPROV and the CIT).

2.1. Databases and Acoustic Features

2.1.1. Acoustic Features:

We extract 1582 dimensional utterance level functional features using the openSMILE toolkit [15] with Emo-base config file. This set has been used as an effective acoustic feature set in conducting cross corpus SER experiments [16]. Max-min normalization is applied on all datasets to ensure efficient training of our conditional cycle GAN data augmentation network and the emotion recognition network.

2.1.2. The USC IEMOCAP Corpus:

The USC IEMOCAP database is an audio-visual English database [17]. It consists of 5 dyadic sessions with a total of 10 actors (5 males and 5 females). In each session, these actors are requested to perform both scripted and spontaneous dialog interactions. There is approximately 12 hours of data manually segmented into utterances, where each utterance is rated by at least 3 annotators on both categorical emotion labels and dimensional attributes. We use a total of 10039 utterances in this work, and the label of activation and valence are divided into three classes using the boundary of [0, 2], (2, 4), [4, 5].

2.1.3. The MSP-IMPROV Corpus:

The MSP-IMPROV database is an acted audiovisual corpus in English [18], which is composed of six sessions with each session includes two actors (1 male and 1 female). In each session, actors are set in a context with a designated sentence that would elicit a target emotion. The database includes the entire improvisation along with the designated sentence. There is around 9 hours of data, and all utterances are manually annotated by at least five annotators following a crowdsourcing approach. In this paper, we use 8438 utterances labeled with three classes of activation and valence with the boundary cutoff [0, 2], (2, 4), [4, 5].

2.1.4. The USC CreativeIT (CIT) Corpus:

The USC CreativeIT database is a multimodal affective interaction database in English, which consists of dyadic theatrical paraphrase improvisations and 2-sentence scripted plays [19]. This emotion corpus features the use of Stanislavsky Active Analysis to elicit naturalistic affective behaviors and interactions. It includes a total of 16 actors (8 males, 8 females) forming 8 pairs to engage in 3 to 5 minutes long interactions. Each interaction is rated by 3 raters using continuous-

in-time annotation scheme on emotion dimensional attributes (the scale ranges between 1 to -1). There are total 2163 utterances used in this work, which are separated into three classes using the boundary of [-1, -0.33], (-0.33, 0.33), [0.33, 1].

2.2. Conditional Cycle Emotion GAN Model

In this paper, we propose to use a conditional cycle emotion GAN to learn a generative mapping function between source and target to perform source data augmentation. A brief description of our conditional cycle emotion GAN augmentation network is below:

2.2.1. Cycle GAN:

Our framework considers a bi-directional mapping between source and target corpus, two generators are used here. $G_{S\to T}$ generates the synthetic instances from source to target domain, and $G_{T\to S}$ generates from target to source corpus. The standard Cycle GAN loss $\mathcal{L}(G_{S\to T}, D_T)$ is defined as below:

$$\mathcal{L}(G_{S\to T}, D_T) = \mathbb{E}_{T\sim P_{data}(T)}[\log D_T(T)] + \mathbb{E}_{S\sim P_{data}(S)}[\log(1 - D_T(G_{S\to T}(S)))] \quad (1)$$

Similar to the $G_{S\to T}$, we form an additional target to source loss $\mathcal{L}(G_{T\to S}, D_S)$, that GAN loss function:

$$\mathcal{L}_{GAN}(G_{T\to S}, G_{S\to T}, D_S, D_T) = \mathcal{L}(G_{T\to S}, D_S) + \mathcal{L}(G_{S\to T}, D_T) \quad (2)$$

In order to stabilize the training process, we add an extra identity loss and cycle loss as regularization terms. Identity loss constraints the transformation of a source instance through $G_{T\to S}$ to be identical to that original source sample. In addition, cycle consistency loss enforces that after bidirectional transformation, samples should be identical, which means the result for each source sample feeding forward $G_{S\to T}$ then through $G_{T\to S}$ should be consistent to the original sample. All the criterion is measured in mean-square error (MSE), and the losses are defined as:

$$\mathcal{L}_{identity} = \mathbb{E}_{S\sim P_S}[||G_{T\to S}(S) - S||^2] + \mathbb{E}_{T\sim P_T}[||G_{S\to T}(T) - T||^2] \quad (3)$$

$$\mathcal{L}_{cycle} = \mathbb{E}_{S\sim P_S}[||G_{T\to S}(G_{S\to T}(S)) - S||^2] + \mathbb{E}_{T\sim P_T}[||G_{S\to T}(G_{T\to S}(T)) - T||^2] \quad (4)$$

Table 1: Train and test within each dataset (no transfer), and the star symbol (*) means training with balanced distribution by random up-sampling

	Activation		Valence	
	DNN*	SVM	DNN*	SVM
MSP-IMPROV	**66.42**	61.38	**53.12**	52.24
CreativeIT	**48.56**	43.54	**49.79**	38.84

2.2.2. Conditional Cycle Emotion GAN:

First of all, we learn a pre-trained emotion classifier F_s using the original source corpus to guide the learning process of the conditional cycle GAN. In order to control how the synthetic samples are being generated from the source-and-target cycle GAN, we add another hidden conditional vector Z acting as an emotion conditional input for the generator $G_{T\to S}$. The source and target data are randomly paired in the cycle GAN training stage with each source sample corresponds to a specific target sample and vice versa. It means that for that particular target sample, we could assign the corresponding source sample's emotion label as the hidden condition for the generator $G_{T\to S}$. We define this as the conditional emotion consistency loss as below:

$$\mathcal{L}_{CE} = \sum_i y_i \log(F_s(G_{T\to S}(G_{S\to T}(S_i), Z_i))) + \sum_i y_i \log(F_s(G_{T\to S}(S_i, Z_i))) \quad (5)$$

where i represents the sample index, and Z_i is the one-hot encoded vector corresponding to the annotation of instance S_i from source corpus. To strengthen the conditional hidden vector Z, we additionally impose a strict constraint that by giving a random emotion condition to each target samples, after transforming to source database, it would be mapped to the same category as dictated by F_s. This random conditional emotion loss is defined as:

$$\mathcal{L}_{RCE} = \sum_i y_r \log(F_s(G_{T\to S}(T_i, Z_r))) \quad (6)$$

where Z_r, y_r represent a random one-hot encoded vector indicating the label for each of the emotion category r. Further, to handle the weight clipping and gradient explosion or vanish issue in training GAN, we add an extra gradient penalty loss that limits the gradient to a specific range. After aggregating all the aforementioned structures together, the total conditional cycle emotion GAN loss is listed below:

$$\mathcal{L}_{CCEmoGAN}(G_{S\to T}, G_{T\to S}, D_S, D_T, S, T) = \mathcal{L}_{GAN} + \lambda_1\mathcal{L}_{identity} + \lambda_2\mathcal{L}_{cycle} + \lambda_3(\mathcal{L}_{CE} + \mathcal{L}_{RCE}) \quad (7)$$

where λ_1, λ_2, λ_3 are the weights of different losses. We set λ_1 as 5, λ_2 as 10, and λ_3 as 10. All the generators and discriminators are optimized during training process as:

$$G^*_{S\to T}, G^*_{T\to S} = arg \min_{G_{S\to T}, G_{T\to S}} \max_{D_S, D_T} \{\mathcal{L}_{CCEmoGAN}(G_{S\to T}, G_{T\to S}, D_S, D_T, S, T)\} \quad (8)$$

2.2.3. Cross Corpus Emotion Recognition with Data Augmentation:

After training CCEmoGANs, we use the generator of CCEmoGAN to generate synthetic source domain samples to conduct data augmentation. Note that we first balance the source

Table 2: Activation and valence result of baseline models and proposed model of data augmentation. All the results are presented in UAR metric. The abbreviation TC_A stands for the model training with only type A synthetic instances generated from traditional cycle GAN, and TC_OA means training by original source corpus with type A synthetic instances generated from traditional cycle GAN.

		Baseline Model							CCEmo-GAN Augmentation					
	Corp.	DANN*	CyCADA	CyEmoGAN	TC_A	TC_OA	TC_B	TC_OB	A	B	C	OA	OB	OC
Act.	I2C	44.1	42.17	40.63	41.4	39.45	**45.78**	43.72	46.55	51.4	40.25	40.95	**51.13**	40.46
	I2M	60.9	61.46	60.74	62.92	62.87	**65.46**	65.11	63.38	62.73	64.74	62.2	**65.7**	62.89
Val.	I2C	42.71	31.67	40.97	36.2	36.35	43.52	**44.51**	44.24	**47.75**	45.5	43.38	47.56	43.32
	I2M	47.17	41.75	43.36	42.49	48.46	43.45	**48.54**	44.51	39.43	37.03	48.59	**50.11**	49.4

domain emotion class distribution by random up-sampling the minority class to have the same number of samples as the majority class before training CCEmoGANs. There are a total of three types of samples that could be generated from the learned CCEmoGAN:

- **Type A : Variational Source Instance**

  In this manner, samples of source corpus belong to each emotion label are fed to $G_{S\to T}$ and then to $G_{T\to S}$. With this procedure, we augment the source samples by including synthetic data that has been transformed to the target domain and back to the source domain. This increases the variability of source data and include information about the target distribution at the same time.

- **Type B : Transform Source to Target**

  We directly map the source corpus sample to the target domain distribution using $G_{S\to T}$ and assign the corresponding categorical emotion label from the source sample for the transformed instance to augment the source training data.

- **Type C : Transform Target to Source**

  Final augmentation method is directly mapping all the target corpus samples to source domain using $G_{T\to S}$ and annotating these synthetic data with a hidden vector Z. In this strategy, we map each of the target samples with three different categorical one-hot encode vectors Z through $G_{T\to S}$ to generate the same amount of data for each of the three emotion categories.

After applying CCEmoGAN data augmentation, we train an network of three dense layers as our recognition network, and then directly evaluate the network on the target corpus.

3. EXPERIMENTAL SETUP AND RESULTS

Detailed settings of our model are listed below: in order to achieve emotion consistency in our CCEmoGAN, we pretrain an emotion classifier using 3 dense layers of size 500, 100, 3 neurons on the source corpus (the IEMOCAP) using a leave one person out (LOPO) scheme to decide the optimum epoch. Here, the number of epoch for activation and valence is 15 and 20 respectively. All the label distributions are balanced using random up-sampling before training the classifier. $G_{S\to T}$ generator consists of three dense layers with 1000, 500, 1000 neurons, and $G_{T\to S}$ contains three dense layers with 1000, 500, 1000 neurons due to the one-hot encoded conditional emotion hidden vector Z, the input dimension of $G_{T\to S}$ is 1585 (1582+3). We pretrain the generator for 50 epochs with identity loss on our dataset as initialization. The learning rate of generator, discriminator are 2e-5 and emotion classifier is 2e-4 without decaying, and activation functions are LeakyReLU for generators and ReLU for emotion classifier, batch normalization and early stopping are also utilized. Note that we take extra caution to ensure no information about speaker and emotion in the testing fold is included in the training of CCEmoGAN.

3.1. Baseline Models

We compare with the following baseline models. Due to the imbalance of the original emotion classes in the three datasets, we also present results after balancing the label (indicated with a star sign in the Table).

- **Domain Adaptation Neural Network (DANN)**

  This architecture is first proposed in [12] to overcome the issue of domain discrepancy when training on cross corpus emotion datasets. DANN aims to find a common space that could perform well on different datasets. It has become the state-of-the-art model in performing transfer learning for SER. In this work, the encoder is built using two dense layers with 512, 128 neurons and emotion, and classifier portion includes two dense layers each with 128, 32 neurons.

- **Traditional Cycle GAN (TC)**

  Cycle consistent GAN is proposed in [14], which has the advantage of using unpaired samples in training GAN for two datasets. Traditional cycle GAN aims at learning both the distribution of two datasets with a bidirectional mapping through generative model. We take this as a comparable baseline model, i.e., essentially it is a cycle GAN without any emotion constraint. All the parameters associated with generators, discrim-

Table 3: Data augmentation result of activation and valence implemented by different variants of cycle GAN models. All the results are presented in UAR metric. The abbreviation meaning is the same as above tables.

Data Augmentation									
		CyCADA				CyEmoGAN			
Emo	Corp.	A	B	OA	OB	A	B	OA	OB
Act.	I2C	47.42	45.79	41.04	44.54	42.37	41.02	38.85	42.52
	I2M	59.63	60.02	64.05	63.01	61.35	61.83	58.24	57.91
Val.	I2C	31.77	32.54	37.71	35.83	40.38	43.15	36.73	40.01
	I2M	45.64	48.08	50.17	48.19	33.39	32.7	35.23	33.57

inators of source and target are set the same as our proposed model.

- **Cycle-Consistent Adversarial Domain Adaptation**

  Cycle-Consistent Adversarial Domain Adaptation (CyCADA) proposed in [20] aims to mitigate the issue of domain shift between training and testing corpus through generative model as well. Different from DANN, CyCADA jointly considers the label consistency loss in their structure. In order to compare fairly with our proposed model, the setting of cycle GAN generators and discriminators are the same as ours, and the pre-trained source emotion classifier is also applied in this architecture.

- **Cycle Emotion GAN (CyEmoGAN)**

  CycleEmotionGAN [21] is the latest generative model used in image-based emotion recognition that is also based on cycle GAN. Similar to the setting of CyCADA, this architecture consists of a cycle GAN which jointly considers the consistency loss between source and target. The major difference is that the classifier of main task is jointly updated when learning the cycle GAN. All the settings of cycle GAN and classifier are the same as ours for fair comparison.

3.2. Results and Analysis

3.2.1. Baseline Model Comparison:

We first run a within corpus experiment (LOPO) for each of the three datasets to assess performance of the recognition network in a *non-transfer* setting (results are in Table 1). This also indicates the upper bound of the within-dataset performance. The performance of baseline models and our proposed methods are listed in Table 2, and all of them are measured in the percentage of unweighted average recall (UAR). Different types of combination for data augmentations are also presented.

By examining the baseline models, we observe that the discrepancy between the IEMOCAP and the MSP-IMPROV is less severe compared to the IEMOCAP and the CIT, especially evident from the activation dimension. This phenomenon may potentially due to the similarity in the collection environment and interaction settings of both the IEMOCAP and the MSP-IMPROV. Recognition performance of activation is generally better than valence, which is inituitive as activation is known to be an easier construct to recognize using speech directly. Valence is known to be challenging to compute from speech, and the discrepancy between corpora creates a much larger degradation of performance on this dimension, i.e, we observe methods of baseline domain adaption is also generally more useful for valence.

The results of using our method are listed in Table 2. Our method consistently outperforms all of the current state-of-the-art especially in the scenario of tremendous discrepancy such as the setting in using model trained in the IEMOCAP and test in the CIT for both dimensions (Table 2). Specifically, when transferring from the IEMOCAP and the CIT, the UAR of our proposed methods are 51.13% and 47.75% in activation and valence respectively, which are 5.35% and 3.24% better than state-of-the-art traditional cycle GAN (TC) model. Similar results are obtained when transferring from the IEMOCAP to the MSP-IMPROV, where the improvement are 0.24% and 1.57% respectively in activation and valence as compared to TC model. Due to much lesser discrepancy in the activation dimension, the improvement is limited when transferring from the IEMOCAP to the MSP-IMPROV. However, we observe a consistent improvement in the more challenging task such as valence recognition in the cross corpus experiments.

3.2.2. Comparison of Data Augmentation using Different Generative Models:

Comparing to the baseline models, our proposed use of data augmentation method achieves a better performance in general by increasing the quantities and variability of training data. Instead of directly using a pre-trained or a jointly-trained classifier to predict emotion, we first generate an augmented dataset and re-train the source classifier jointly with synthetic and original data, which is then evaluated directly on the target test samples.

For different types of generated samples, only type A and type B synthetic samples can be generated by Cycle GAN based generative model such as CyCADA and CyEmoGAN. For type A, the variational source instance's label is annotated as the original source label, and for type B, the fake target sample from generator $G_{S\to T}$ is assigned to its corresponding (paired) source sample's class label. However, when taking target samples to be transformed using $G_{T\to S}$, no label information could be provided. However, in contrast to these previous cycle GAN models, our CCEmoGAN model employs the use of condition hidden vector Z, which is capable of assigning labels to type C synthetic samples. All of the cross corpus recognition results are presented in Table 2, 3.

From Table 3, we observe the accuracy obtained using our proposed model clearly outperforms all the others, es-

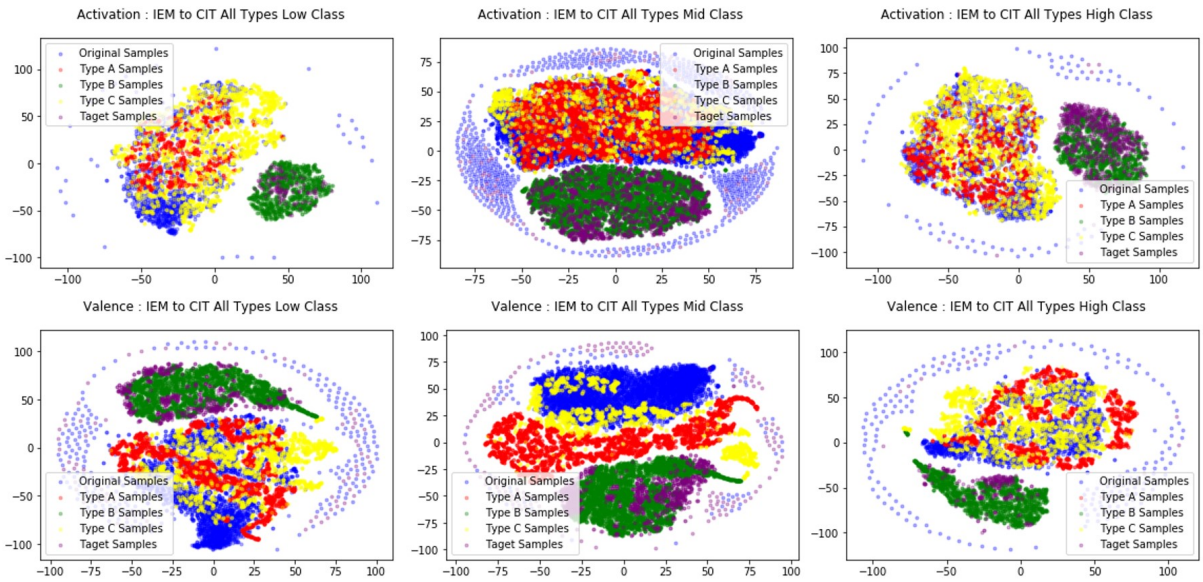

Fig. 2: Visualization of three different types of synthetic data in the setting from the IEMOCAP to the CIT. The top figure shows a visualization of activation class low, mid and high respectively, and below is for valence dimension. Blue, red, green, yellow, purple stand for samples of original source corpus, type A, type B, type C and target corpus respectively.

pecially in setting where the mismatch is more severe, i.e., from the IEMOCAP to the CIT. Specifically, in the setting of IEM2CIT, we obtain a 6.59% and 4.6% improvement in UAR for activation and valence respectively. For IEM2MSP, the improvement is limited to 2.69% in the activation dimension due to less mismatch (further activation tends to be an easier task), but the UAR improves 1.92% in the more challenging task of valence recognition.

We also observe that the best case usually comes from using type B synthetic samples, and the result is quite intuitive. The type B synthetic sample aims at recreating the target domain distribution, hence, by augmenting source corpus with type B samples when training the emotion recognition network naturally helps in improving the recognition accuracy in the target domain for both activation or valence dimensions. We further visualize the distribution of samples in each emotion class (high, mid, low) for the source, synthesized, and target domain samples to investigate the effect of our use of CCEmoGAN as the sample generator. All the figures are shown in Figure 2. We observe that almost all of the intended characteristics for each type of synthetic instances are well captured and generated, e.g., type B samples for each emotion class should be well aligned with the target domain samples, and type A samples should be well aligned with source domain samples for each emotion class, and so on. Except for some mid class synthetic samples, the well-aligned data distributions that are also emotionally consistent are key in utilizing our proposed CCEmoGAN to achieve the improved accuracy, and we observe this phenomenon even in severely mismatched scenario, i.e., from the IEMOCAP to the CIT. The bidirectional mapping between source and target corpus with consistent emotion constraint to control synthetic samples using an additional condition vector Z provides a significant boost in accuracy through generating target-domain aware synthetic samples to be used in source corpus emotion recognition network training.

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a novel transfer learning strategy for SER by using conditional cycle emotion GAN (CCEmoGAN) data augmentation. Instead of learning a common invariant feature space, our idea is to increase the variability of the source data that is guided toward unlabeled target domain with an additional emotion consistency constraint in order to handle the issue of highly contextualized and limited in-scale emotion spoken corpora. The model could generate specific emotion from target to source corpus just by adapting the emotion conditional vector Z during training. By generating different variants of emotion-aware synthetic samples to derive an augmented source corpus that can be used to train a robust emotion recognizer, our experiments demonstrate a state-of-art speech emotion transfer recognition accuracy in both activation and valence when using the IEMOCAP as the source data on two different target data, the MSP-IMPROV and the CIT.

The current conditional cycle emotion GAN is learned on balanced emotion distribution through random upsampling of source dataset. We would like to further investigate the improvement in SER transfer as a function on the amount of synthetic samples, and examine whether the types of the original source data when learning the conditional cycle emotion GAN augmentation network would have an effect. Also, we would continue to explore the idea of *guided* data augmentation as an emotion transfer approach, e.g., training the augmentation network from a much larger but not necessary emotion corpora (ASR databases).

5. REFERENCES

[1] Gabriele Pasetti Monizza, Tammam Tillo, and Dominik T Matt, "Computer vision approach for indoor location recognition within an augmented reality mobile application," in *Cooperative Design, Visualization, and Engineering: 16th International Conference, CDVE 2019, Mallorca, Spain, October 6–9, 2019, Proceedings*. Springer Nature, 2019, vol. 11792, p. 45.

[2] Mahesh K Singh, AK Singh, and Narendra Singh, "Multimedia utilization of non-computerized disguised voice and acoustic similarity measurement," *Multimedia Tools and Applications*, pp. 1–16, 2019.

[3] Ruhi Sarikaya, Geoffrey E Hinton, and Anoop Deoras, "Application of deep belief networks for natural language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 778–784, 2014.

[4] A Menychtas, M Galliakis, P Tsanakas, and I Maglogiannis, "Real-time integration of emotion analysis into homecare platforms," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 3468–3471.

[5] Hans-Jörg Vögel, Christian Süß, Thomas Hubregtsen, Elisabeth André, Björn Schuller, Jérôme Härri, Jörg Conradt, Asaf Adi, Alexander Zadorojniy, Jacques Terken, et al., "Emotion-awareness for intelligent vehicle assistants: A research agenda," in *2018 IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIAS)*. IEEE, 2018, pp. 11–15.

[6] Zixing Zhang, Felix Weninger, Martin Wöllmer, and Björn Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 523–528.

[7] Zhengwei Huang, Wentao Xue, Qirong Mao, and Yongzhao Zhan, "Unsupervised domain adaptation for speech emotion recognition using pcanet," *Multimedia Tools and Applications*, vol. 76, no. 5, pp. 6785–6799, 2017.

[8] Peng Song, Wenming Zheng, Shifeng Ou, Xinran Zhang, Yun Jin, Jinglei Liu, and Yanwei Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[9] Peng Song, "Transfer linear subspace learning for cross-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, , no. 1, pp. 1–1, 2017.

[10] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," *arXiv preprint arXiv:1808.05561*, 2018.

[11] Biqiao Zhang, Emily Mower Provost, and Georg Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 85–99, 2017.

[12] Mohammed Abdelwahab and Carlos Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[13] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang, "Attribute-guided face generation using conditional cyclegan," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 282–297.

[14] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei 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.

[15] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[16] Cevahir Parlak, Banu Diri, and Fikret Gürgen, "A cross-corpus experiment in speech emotion recognition.," in *SLAM@ INTERSPEECH*, 2014, pp. 58–61.

[17] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[18] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[19] Angeliki Metallinou, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[20] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," *arXiv preprint arXiv:1711.03213*, 2017.

[21] Sicheng Zhao, Chuang Lin, Pengfei Xu, Sendong Zhao, Yuchen Guo, Ravi Krishna, Guiguang Ding, and Kurt Keutzer, "Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 2620–2627.