Semi-supervised Multimodal Emotion Recognition With Improved Wasserstein GANs

Jingjun Liang, Shizhe Chen, Qin Jin
School of Information, Renmin University of China, Beijing, China

***Abstract*—Automatic emotion recognition has faced the challenge of lacking large-scale human labeled dataset for model learning due to the expensive data annotation cost and inevitable label ambiguity. To tackle such challenge, previous works have explored to transfer emotion label from one modality to the other modality assuming that the supervised annotation does exist in one modality or explored semi-supervised learning strategies to take advantage of large amount of unlabeled data with the focus on a single modality. In this work, we address the multi-modal emotion recognition problem with the acoustic and visual modalities and propose a multi-modal network structure of the semi-supervised learning approach with an improved generative adversarial network CT-GAN. Extensive experiments conducted on a multi-modal emotion recognition corpus demonstrate the effectiveness of the proposed approach and prove that utilizing unlabeled data via GANs and combining multi-modalities both benefit the classification performance. We also carry out some detailed analysis experiments such as influence of unlabeled data quantity on recognition performance and impact of different normalization strategies for semi-supervised learning etc.**

I. Introduction

Automatic emotion recognition empowers machines with the capability to communicate naturally with humans, which plays an essential role in maintaining long-term human-machine interactions. It has a wide range of applications in modern dyadic interaction scenarios involving various human relationships such as therapist-patient, teacher-student, agent-customer, and employer-employee interactions etc. [1] In recent years, there have been growing interests in exploring automatic technologies to recognize emotional states of individuals in various scenarios especially with the rapid development of deep neural networks.

We humans express emotions in a number of different ways including both verbal and nonverbal communications, such as emotional speech, facial expressions, and body languages etc. Therefore, emotional signals from multi-modalities could be used to predict a subject's emotional state. In previous research, combining different modalities including speech, facial expression, and texts have been investigated. For example, early fusion at the feature level has been widely applied for multi-modality fusion, which involves feature embedding via feed-forward layers in each modality and then fusing multi-modalities via simple concatenation. How to effectively combine multiple modalities matters for automation multi-modality emotion recognition, and has been an active research topic.

Besides multi-modality, limitations on available supervised emotion data is another big challenge for automatic emotion recognition due to the fact that it is expensive to collect large-scale of emotion data and it usually involves inevitable label ambiguity. There have been some research efforts to address the data shortage problem, such as applying transfer learning to take advantage of emotion annotations from another modality [2] or applying unsupervised or semi-supervised learning to take advantage of abundant multimedia corpus and alleviate the problem of label ambiguity [3].

Using generative adversarial networks (GANs) for semi-supervised learning has been proposed by Salimans et al. [4] to combine the supervised and unsupervised data. Such approach has made great success on image representation learning [5]. Chang et al. [6] apply similar GAN-based semi-supervised learning on the emotion recognition task and propose the semi-supervised acoustic representations for valance prediction. However, the training process of generative adversarial networks is often unstable without good heuristics. Several works have proposed improved methods to handle this issue [4, 7, 8]. Recently, a new optimization scheme for generative adversarial network training CT-GAN [9] is proposed and it can be seamlessly embedded into semi-supervised learning framework.

In this work, we investigate the effectiveness of applying adversarial network on semi-supervised emotional recognition. Additionally, unlike the previous work [6] which is limited in acoustic uni-modality, we extend it and investigate the capacity of semi-supervised emotion recognition in visual modality and multi-modalities as well. We propose a multi-modality network structure of the semi-supervised learning approach with an improved generative adversarial network CT-GAN. The results demonstrate that utilizing unlabeled data via GANs and combining multi-modalities both benefit the classification performance.

To summarize, the main contributions of this work are listed as follows:

- We design a multi-modality semi-supervised learning network structure via GANs on emotion recognition.
- We carry out extensive comparison experiments on a widely used multi-modality emotion recognition corpus to evaluate the effectiveness of proposed approach.
- We also conduct detailed investigation of different impact factors on recognition performance including unlabeled data quantity, normalization strategies etc.

II. Related Work

A. Emotion Recognition

Previous works have explored a variety of multi-modal features for emotion recognition tasks. Brady et al. [10] derive high-level acoustic, visual and physiological features from the low-level descriptors using sparse coding and deep learning. Viktor et al. [11] use early fusion to concatenate mult"i-modal features as the input for the prediction models and improves performance successfully. Huang et al. [12] use a late fusion strategy that combines the output of different modalities and trains a second level model to predict categories of multi-modal samples. Chang et al. [6] use semi-supervised GANs for emotion recognition on acoustic modality and the result is competitive to state-of-the-art performance. Similar to previous works, we explore emotion features from multiple modalities including the acoustic modality and facial modality and implement multi-modal fusion in this work.

B. GANs

Generative Adversarial Networks [13] was proposed using game theory for generative tasks. In order to train the generator, GANs build up a discriminator to play an adversarial game with the generator, rather than setting a normal loss function for the generator. The discriminator is designed to distinguish between samples from the generator and samples from the real data while the generator learns to output samples that can fool the discriminator.

However, the training progress of original Generative Adversarial Networks is often unstable and hard to converge. **WGAN** [7] proposed to use Wasserstein distance and weight clipping method to alleviate vanishing gradient problem. Furthermore, **WGAN-GP** [8] illustrated the risk of the weight clipping method that it pushes weights towards the bound of clipping range. They proposed an alternative to impose Lipschitz continuity: penalize the norm of gradient of the critic with respect to its input. Recently, **WGAN-CT** [9] illustrated that the critic function can freely violate the 1-Lipschitz continuity at the beginning of training stage since the generated samples could be distant from real samples. The 1-Lipschitz continuity is not enforced until the output of generative model become close enough to real distribution. So they add a consistent regularization term into loss function and benefits the training of generative model.

C. Semi-supervised GANs

Salimans et al. [4] proposed a semi-supervised learning framework using Generative Adversarial Networks. It simply add the samples from generator and labeled with a new "generated" class into standard classifier to perform semi-supervised learning progress. Laine et al. presented a method that forming ensemble predictions during semi-supervised training and take advantage of the effect of temporal ensembling [14]. Based on their work, Xiang et al. [9] further embedded the consistent regularization term into discriminator's loss and performed over state-of-the-art semi-supervised learning results.

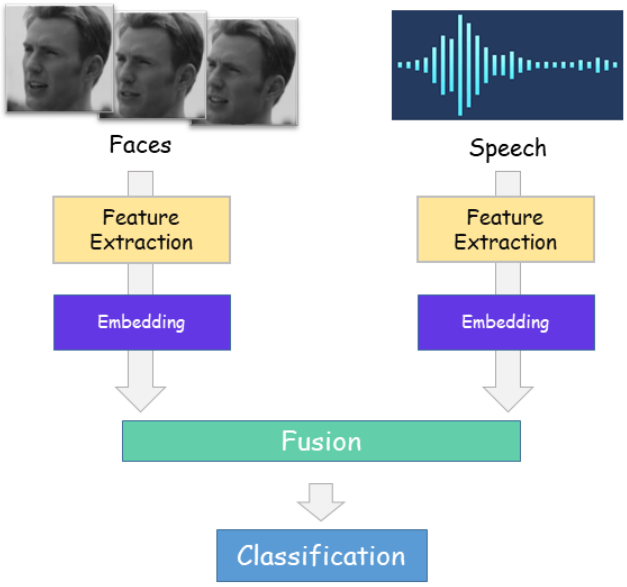


Fig. 1. Illustration of multi-modality emotion recognition framework. Different modalities are embedded and fused before fed into emotional classifier.

III. The Proposed Approach

In this section, we present the proposed multi-modality semi-supervised emotion recognition approach. We first explain the general multi-modality emotion recognition framework and the algorithm of semi-supervised learning with CT-GAN, then we present the multi-modality network architecture.

A. Multi-modality Emotion Recognition

The main tasks of emotion recognition can be generally divided into two categories, which are discrete emotion recognition (e.g. happy, angry, sad etc.) and continuous emotion recognition (e.g. in arousal/activation, valence dimensions etc.). As it is nearly impossible to achieve consistent annotation across corpus, it is common to transform the arousal and valence continuous emotion prediction task into the 3-point scale or 5-point scale discrete classification tasks. In a common interaction scenario, both the acoustic and facial modality signals are available for emotion prediction. Figure 1 illustrates a general multi-modality emotion recognition framework. Hand-crafted or deep learned feature representations are first extracted from different modalities. The raw feature representations are then embedded by a feed-forward network into a fixed dimensional embedding space and then fused for example by simple concatenation. The emotion classification is then executed on the fused multi-modality features. The classification network is normally trained with supervised multi-modality emotion data to maximize the probability of predicting the ground truth emotion labels. The general loss function can be expressed as:

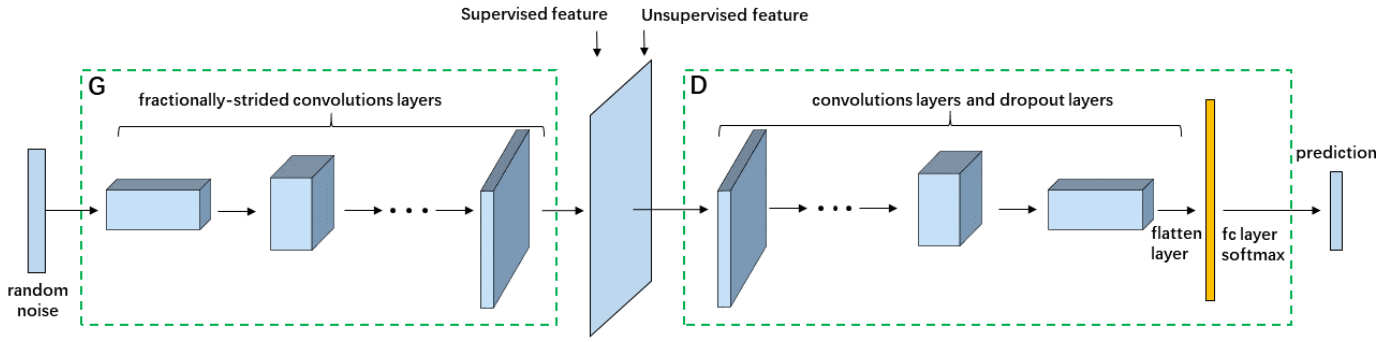


Fig. 2. Multi-modality Semi-supervised Network Structure: After each convolutional layer in the discriminator, one dropout layer is built to compute perturbed discriminator's output. The last layer of the discriminator is a standard classifier with a fully connected layer and a softmax layer

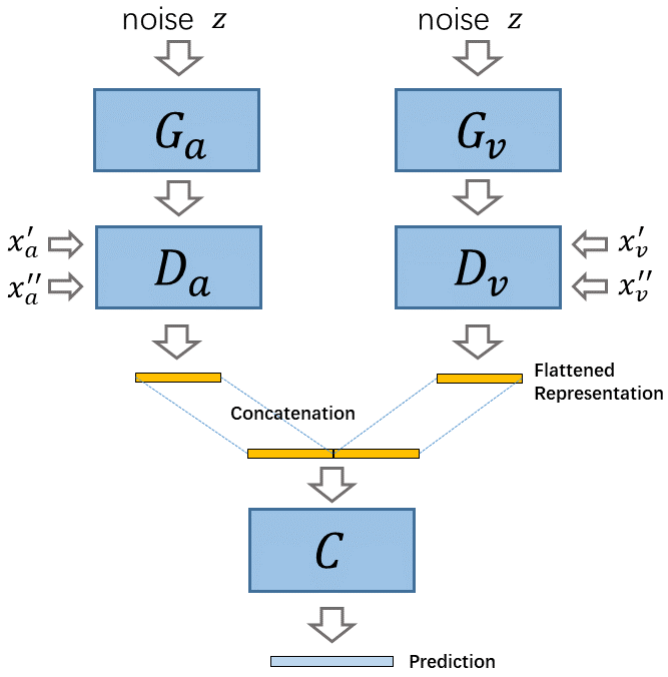


Fig. 3. Components of multi-modality semi-supervised network structure: acoustic generator G_a, acoustic discriminator D_a, visual generator G_v, visual discriminator D_v, multi-modality classifier C, random noise z, acoustic and visual feature pairs of supervised data $(x_a^{'}, x_v^{'})$, acoustic and visual feature pairs of unsupervised data $(x_a^{''}, x_v^{''})$

$$Loss = -\mathbb{E}_{x,y\sim\mathbb{P}_{x,y}}[logC(y|(P_a(x_a) \oplus P_v(x_v)))] \quad (1)$$

where x, y is the supervised data, x_a, x_v are acoustic and visual features of sample x, y is the corresponding emotion label. $\oplus$ is fusion operation.

B. Semi-supervised Learning with CT-GAN

Salimans et al. [4] proposed the semi-supervised learning strategy using Generative Adversarial Networks. Based on the standard $K - classes$ classifier, they add an additional $K+1$ class which is called "generative" class and modify the adversarial objective function which is shown as below.

$$\begin{aligned} L_{dis} = &-\mathbb{E}_{x',y\sim\mathbb{P}_{x',y}}[logD(y|x^{'})] \\ &-\mathbb{E}_{z\sim\mathbb{P}_{noise}}[logD(K+1|G(z))] \\ &-\mathbb{E}_{x\sim\mathbb{P}_r}[log(1-D(K+1|x))] \end{aligned} \quad (2)$$

$$L_{gen} = -\mathbb{E}_{z\sim\mathbb{P}_{noise}}[log(1-D(K+1|G(z)))] \quad (3)$$

where G and D are the generator and discriminator respectively. $(x^{'}, y)$ is the supervised instance-label data pair, r is the union of supervised and unsupervised data, $noise$ is the random noise fed into the generator.

For the discriminator, it aims to classify the supervised data into their corresponding categories correctly and distinguish the real data (both supervised and unsupervised one) from fake data at the same time. For the generator, it aims to synthesize the fake sample which is similar to real sample and confuse the discriminator.

Following the improved technique of semi-supervised learning proposed by [9, 14], we modify the loss function by replacing the cross-entropy value function with Wasserstein distance and imposing the Lipschitz continuity over the discriminator. A consistency regularization was added and it essentially builds up a temporal self-ensembling structure to benefit the learning progress. The modified loss function is shown as below.

$$\begin{aligned} L_{dis} = &-\mathbb{E}_{x',y\sim\mathbb{P}_{x',y}}[logD(y|x^{'})] \\ &-\mathbb{E}_{z\sim\mathbb{P}_{noise}}[logD(K+1|G(z))] \\ &-\mathbb{E}_{x\sim\mathbb{P}_r}[log(1-D(K+1|x))] + \lambda CT \end{aligned} \quad (4)$$

$$L_{gen} = \|\mathbb{E}_{z\sim\mathbb{P}_{noise}}[D\_(G(z))] - \mathbb{E}_{x\sim\mathbb{P}_r}[D\_(x))]\|_2^2 \quad (5)$$

where $D\_(x)$ is the output of second-to-last layer of the discriminator

C. Multi-modality Model Architecture

In our multi-modality semi-supervised learning model, the structure of generator (G) and discriminator (D) are built upon a set of symmetric deep convolutional layers as in DCGAN [5], shown in Figure 2. A 100-dimensional random noise vector is transformed into a feature map with several fractionally-strided convolutional layers. And then it is fed into the discriminator with the real samples. On the contrary, the discriminator consists of several convolutional layers and transforms the feature map into a flattened high level representation. After each convolutional layer in the discriminator, one dropout layer is built to compute perturbed discriminator's output. The last layer of the discriminator is a standard classifier with a fully connected layer and a softmax layer. When the semi-supervised training finishes, we take the discriminator as a more effective classifier compared to the fully-supervised model.

For multi-modality fusion experiments, we train two Deep Convolutional Generative Adversarial Networks for acoustic and visual modalities respectively. Different from uni-modal network structure, we remove the last fully connected layer and softmax layer and take the flattened representation as the outputs of the discriminator. To apply multi-modality fusion, we concatenate the representation from acoustic and visual modalities and feed them into a new fully connected layer and softmax layer as a multi-modality classifier (C). The multi-modality model architecture is shown in Figure 3 and the loss function is modified to Eq. 6 and Eq. 7.

$$\begin{aligned} L_{dis} = & -\mathbb{E}_{x^{'},y\sim\mathbb{P}_{x^{'},y}}[logC(y|(D_a(x_a^{'})\oplus D_v(x_v^{'})))] \\ & -\mathbb{E}_{z\sim\mathbb{P}_{noise}}[logC(K+1|(D_a(G_a(z))\oplus D_v(G_v(z))))] \\ & -\mathbb{E}_{x\sim\mathbb{P}_r}[log(1-C(K+1|((D_a(x_a)\oplus D_v(x_v)))))] \\ & +\lambda CT \end{aligned} \tag{6}$$

$$\begin{aligned} L_{gen} = & \|\mathbb{E}_{z\sim\mathbb{P}_{noise}}[D_a(G_a(z))\oplus D_v(G_v(z))] \\ & -\mathbb{E}_{x\sim\mathbb{P}_r}[D_a(G_a(x))\oplus D_v(G_v(x))]\|_2^2 \end{aligned} \tag{7}$$

where $\oplus$ means the concatenation operation.

The parameters of classifier θ_C are updated with the parameters of the two discriminators θ_D and separated from the parameters in the two generators θ_G. We first train the discriminators and the classifier for S_D iterations and then fix the discriminators as well as the classifier to train the generator for S_G iterations. The adversarial training procedures of the two modules are presented in Algorithm 1.

IV. Experiments

In this section, we present detailed comparison experiments on the arousal and valence emotion recognition tasks with the semi-supervised settings. Specifically we want to find answers for:

- If our semi-supervised strategy benefits from large amount of unlabeled data to boost the emotion recognition accuracy.

Algorithm 1 Multi-modality Semi-supervised Training Procedure

Require: acoustic generator G_a, acoustic discriminator D_a, visual generator G_v, visual discriminator D_v, multi-modality classifier C;
Input: random noise z, feature and label triplets of supervised data $(x_a^{'}, x_v^{'}, y)$, acoustic and visual feature pairs of unsupervised data $(x_a^{''}, x_v^{''})$;
for epoch = 0, ..., N **do**
  for batch = 0, ..., M **do**
    Select batch sample in $(x_a^{'}, x_v^{'}, y)$ and $(x_a^{''}, x_v^{''})$
    for iter = 0, ..., S_D **do**
      Compute discriminative loss L_{dis} using Eq. 6
      Adam update θ_D and θ_C with L_{dis}
    end for
    for iter = 0, ..., S_G **do**
      Compute generative loss L_{gen} using Eq. 7
      Adam update θ_G with L_{gen}
    end for
  end for
end for

TABLE I
3-CLASSES DATA DISTRIBUTION IN AROUSAL AND VALENCE ON THE IEMOCAP DATASET

label	1	2	3	Total
arousal	1112	7235	1692	10039
valence	3223	4869	1947	10039

- If the multi-modality fusion in the GAN structure can improve performance over the uni-modality baseline.

A. Data Description

We utilize both labeled and unlabeled data for semi-supervised experiments. For the labeled data, we use the IEMOCAP dataset [15], which consists of around 12 hours of video recordings of situational dialogues between two speakers. There are in total 10 speakers in the dataset and each utterance in the dialog comes with arousal and valence labels, both measured on a 5-point scale from at least three distinct annotators. Following the same processing, we convert it into a 3-point scale ("low" level contains values in the range [1, 2], "medium" level contains values in the range (2, 4), and "high" level contains values in the range [4, 5]). For the unlabeled data, we use the AMI [16] dataset which consists of about 100 hours of unlabeled meeting recordings. It provides video recordings for each speaker and transcripts of their speech. The 3-class data distribution of IEMOCAP is presented in Table I. The distribution across 3 classes is unbalanced especially for the arousal.

B. Implementation Details

Samples Extraction: Due to various video length in AMI and IEMOCAP datasets, we pre-define the crop width in advance. We implement face detection with the open-source toolkit Seetaface [17]. Each face image is transformed into the gray scale with height and width of 64 pixels.

For unlabeled AMI corpus, to ensure that the selected crop region is not entirely of silence, we randomly select three consecutive words in the transcript and look up the time region in the video. We then take the middle time as center of the crop and extract the audio and video segment defined by the crop width. Furthermore, we apply frame-level face detection in the cropped video segments. We drop the samples which no detected faces.

For IEMOCAP, we split the video into utterance-level according to the transcript. Each video contains two actors and they perform improvisation by turns. For each frame we only extract face images of the actor who are performing improvisation. So, we finally get the facial expression and audio of each utterance. For those frames when faces cannot be detected, we use the detection result of the previous frame as the result of the current frame.

Multi-modality Features: We extract features from the acoustic modality and visual modality respectively. **Acoustic Modality:** Following the acoustic feature extraction procedure in [6], we extract the spectrograms, denoted as **FFT**, which are computed using a short time Fourier transform with frame size of 25ms and window shift of 10ms. Each frame of spectrogram is a 128 dimensional vector. Additionally, we extract 128-dimensional logMel filter-banks frame-wise feature with the same time step which denoted as **Fbank**. **Visual Modality:** Then we utilize the state-of-the-art Dense Convolutional Neural Networks [18] (DenseNet) to extract facial features. The DenseNet is pre-trained on the FERPlus [19] dataset for facial expression recognition following the setup in [20]. We extract the activation from the last pooling layer of DenseNet for each face image, and get 342-dimensional frame-level feature, which is referred as the **Dface** feature.

Labels: According to the guidance of IEMOCAP, each utterance is annotated by at least two annotators and the average score is used as the ground-truth label. However, the label distribution is unbalanced, as shown in Table I. We check each annotator's source annotation and observe that the mean pooling labels are quiet different from source annotation in all the sessions. For example, if two annotators labeled an utterance as 2 and 4 respectively (such case is very common in IEMOCAP), the mean pooling would treat it as 3, the "medium" level, which is very different from the original label. To handle such situation, fuzzy label strategy has been designed to improve classification performance in previous works [6, 21, 22]. Similar to their label generation strategy, we use one-hot vector to represent each annotator's annotation and then compute the mean of these vector as fuzzy label. For example, if three annotators label the utterance 2, 4 and 5, we represent their annotation as vectors [1,0,0], [0,0,1] and [0,0,1]. The final ground-truth label would be [0.33,0,0.67]. These fuzzy labels will not be transformed back into one-hot vectors in training, but in validation and testing process, we still take the dimension of maximum value as the correct label (level "high" in the example). If only two annotators label the utterance and they hold different opinions, when the fuzzy label would has a tie like [0.5,0.5,0], we will regard both two levels as the correct class.

Hyper-parameters: We set the crop width as 128, the same as [6]. The λ in Eq. 6 is set as 1. S_G and S_D is both set to 1. The quantity of utterances from AMI is 20000. The filter size of convolutional layers is set as 5 with stride as 2. The number of channels in the feature map generated in G is $\{512, 256, 128, 64\}$ and $\{128, 256, 512, 1024\}$ in D. The dropout rate in D is 0.2 and batch size is set as 64. We apply Adam algorithm with learning rate of 2e-4 to optimize the parameter.

C. Evaluation Metrics

For evaluation, we utilized a 5-fold leave-one-session-out validation. Each fold takes four of the five session as the training set and leaves one session for validation and test. This one left-out session contains both male and female speakers. So, we randomly split the session into validation set and test set according to gender. Two common evaluation metrics unweighted average recall (UAR) and Macro f1 [23] are used for performance measurement. We use UAR to select the best model on the validation set. The equation of UAR and Macro f1 is listed as follows:

$$\begin{aligned} UAR &= \frac{1}{n}\sum_{i=0}^{n-1}\frac{c_{ii}}{\sum_{j=0}^{n-1} c_{ij}} \\ MAP &= \frac{1}{n}\sum_{j=0}^{n-1}\frac{c_{ii}}{\sum_{i=0}^{n-1} c_{ij}} \\ Macro\ f1 &= \frac{1}{n} \times \frac{2 \times UAR \times MAP}{UAR + MAP} \end{aligned} \tag{8}$$

where c is an $n \times n$ confusion matrix of which rows represent real classification and columns represent predictions.

D. Compared Models

We design the following experimental setups and compare the classification performance under different settings.

1) **Fully-supervised Baseline**: it is a fully-supervised uni-modality classifier, which is equivalent to only using the discriminator in the networks as shown Figure 2. We refer to it as **FSBase**.

2) **Multi-modality Fully-supervised Baseline**: We combine the two uni-modality fully-supervised baseline and apply concatenation fusion to implement a multi-modality baseline. We refer it as **MFSBase**.

3) **Semi-supervised Baseline**: In this setup, we add more training samples from the unlabeled corpus by using the above FSBase classifier. We feed the unlabeled data into the pre-trained fully-supervised baseline model to annotate the unlabeled samples and then select those samples with high classification confidence. We then finetune the fully supervised baseline model with these new samples. We refer it as **SSBase**.

4) **Semi-supervised with GANs**: We use the generative adversarial networks with consistency term as shown in Figure 2 to build the semi-supervised model based on a single modality. We refer to it as **SSGAN**.

5) **Multi-modality Semi-supervised with GANs**: We build the multi-modality semi-supervised emotion recognition

TABLE II
AROUSAL CLASSIFICATION PERFORMANCE IN UNI-MODALITY SETTING ON THE TEST SET

Feature	Model	UAR	MAF1
Fbank	FSBase	65.16%	64.68%
	SSBase	65.43%	65.00%
	SSGAN	**65.61%**	**65.47%**
FFT	FSBase	58.72%	60.78%
	SSBase	57.03%	58.12%
	SSGAN	60.94%	62.31%
Dface	FSBase	54.11%	54.72%
	SSBase	54.23%	55.02%
	SSGAN	56.40%	55.85%

TABLE III
AROUSAL UAR SCORES ON THE VALIDATION SET AND THE TEST SET WITH VISUAL AND ACOUSTIC FEATURES

Modality	Model	Val	Test	Gap
Acoustic	FSBase	65.07%	65.16%	+0.09%
	SSGAN	66.16%	65.61%	-0.55%
Visual	FSBase	57.31%	54.11%	-3.20%
	SSGAN	58.23%	56.40%	-1.83%

model with GANs following the description in section III-C. We refer to it as **MSSGAN**.

E. Experimental Results

Uni-modality Settings: We first present the experiment results under the uni-modality setting. We compare the classification performance of different models with three types of features. For the arousal classification task, shown in Table II, the SSBase model gets a little higher UAR score compared to the FSBase model (e.g. 65.43% vs 64.16% with Fbank). It shows that adding high confident unsupervised data benefits classification model. Furthermore, It needs to be noted that the performance of visual features drops dramatically with all the model. It indicates that acoustic cues are more informative than visual cues for arousal. This result matches the phenomenon discovered in previous emotion research [24, 25]. Additionally, we discover that the gap of arousal UAR score between the validation set and the test set with visual features is more obvious than with acoustic features. In Table III, we report the validation result and test result of arousal in both acoustic and visual modalities. As we can see, the capacity of visual model loses much in the process transferring from validation set to test set. It may be caused by the fact that we split the validation set and test set by gender and arousal expression is quite different between male faces and female faces.

For the valence classification task, shown in Table IV, SSBase does not always get better performance than FSBase model. It demonstrates that simply adding unsupervised data and adjusting the pre-trained model may disturb the parameter and classification capacity. When we apply semi-supervised generative adversarial strategy, the unsupervised data is used in a correct way and boosts the model for a higher performance. In this task, the visual features, on the contrary, are more effective in classifying valence which agrees with the previous research [26]. Table V shows the valence gap between the validation set and the test set with visual features. The gap is much smaller compared to that with acoustic features.

TABLE IV
VALENCE CLASSIFICATION PERFORMANCE IN UNI-MODALITY SETTING ON THE TEST SET

Feature	Model	UAR	MAF1
	Chang et al. [6]	49.80%	-
Fbank	FSBase	50.39%	49.76%
	SSBase	50.15%	49.54%
	SSGAN	51.15%	51.64%
FFT	FSBase	47.10%	47.46%
	SSBase	46.98%	47.45%
	SSGAN	47.27%	48.67%
Dface	FSBase	58.71%	60.92%
	SSBase	58.88%	61.22%
	SSGAN	**61.78%**	**62.97%**

TABLE V
VALENCE UAR SCORES ON THE VALIDATION SET AND THE TEST SET WITH VISUAL FEATURES

Modality	Model	Val	Test	Gap
Acoustic	FSBase	50.64%	50.39%	-0.25%
	SSGAN	50.41%	51.15%	-0.74%
Visual	FSBase	60.11%	60.92%	-0.81%
	SSGAN	61.94%	62.97%	+1.03%

Multi-modality Settings: In Table VI and Table VII, we present the multi-modality experiment results on arousal and valence. For the arousal task, the multi-modality model performances worse than the uni-modality model with acoustic features in FSBase or SSGAN setup. It could be due to the limited information in visual we mentioned above and it disturbs the overall classification capacity. For the valence task, the multi-modality model shows that two types of modality are complementary for valence classification and boosts the performance. Compared with the result under FSBase and SSGAN setups using different feature combination, the improvement is all significant with the proposed approach and it demonstrates the effectiveness of the semi-supervised learning strategy.

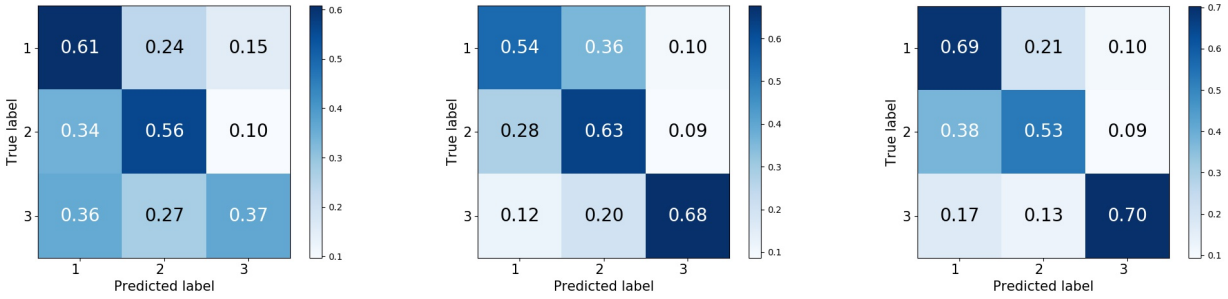
Confusion Metrics Analysis: In Figure 4, we present the confusion matrices comparison of valence prediction with

TABLE VI
AROUSAL CLASSIFICATION PERFORMANCE IN MULTI-MODALITY SETTING ON TEST SET

Feature	Model	UAR	MAF1
FFT+Dface	MFSBase	59.73%	60.92%
	MSSGAN	62.94%	62.49%
Fbank+Dface	MFSBase	62.00%	62.42%
	MSSGAN	**64.10%**	**63.87%**

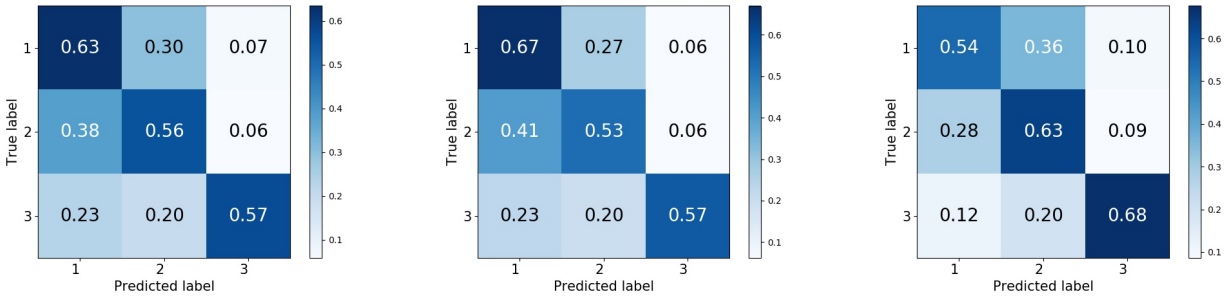
TABLE VII
VALENCE CLASSIFICATION PERFORMANCE IN MULTI-MODALITY SETTING ON TEST SET

Feature	Model	UAR	MAF1
FFT+Dface	MFSBase	58.88%	60.18%
	MSSGAN	63.21%	63.39%
Fbank+Dface	MFSBase	61.82%	62.96%
	MSSGAN	**63.98%**	**65.26%**



(a) Acoustic (b) Visual (c) Multi-modality

Fig. 4. Confusion matrices comparison with different modalities. We use Fbank and Dface feature to implement valence classification comparison experiment as example. It demonstrates that multi-modality fusion strategy benefits the classification performance.



(a) Fully-supervised Baseline (b) semi-supervised Baseline (c) semi-supervised with GANs

Fig. 5. Confusion matrices comparison under different training setup. We take the comparison experiment with Dface features as example. It shows the effectiveness of semi-supervised learning with GANs.

Fbank features and Dface features. All the results are based on semi-supervised learning experiments with GANs. As we can see, the prediction of acoustic model is very unbalanced that majority of its prediction falls in low level. This problem is alleviated with visual features and multi-modality features. In the multi-modality setup, the improved performance over that of the uni-modality systems is due to more accurate low and high valence predictions. It can also demonstrate that the two types of modalities are complementary for valence classification. In Figure 5, we present the confusion matrices comparison of valence prediction with Dface features under different training setup. We can see that the confusion matrices are almost the same between fully-supervised baseline model and semi-supervised baseline model. It shows that simply adding unlabeled data to finetune the fully-supervised model can not effectively improve the classification capacity. When we apply the algorithm of semi-supervised learning with GANs on it, the performance of medium and high valence classification improves significantly. Although the performance of low valence is decreased by the unlabeled data, the overall performance becomes better as shown in Table IV. The results show that semi-supervised learning with GANs is able to take advantage of unlabeled data under limited supervised data situation.

V. ABLATION STUDY

A. *Unlabeled Data Quantity*

To gain more insights about the impact of unlabeled data, we analyze the classification performance change with different quantity of unlabeled training data. We select the combination of Fbank and Dface features and conduct a multi-modality semi-supervised valence classification experiment on one of the 5-fold leave-one-session-out validation. To show the impact of unlabeled data, we keep the hyper-parameter unchanged except the number of unlabeled training data. We train a fully-supervised model with 7936 supervised utterances as baseline at first and then add 5000 unlabeled samples step by step till all the 25000 unlabeled samples from the AMI corpus are used.

The results are shown at Table 6. As we can see, the performance of semi-supervised model gradually improves with the increase of unlabeled samples at the beginning and it reaches the peak when the quantity of unsupervised data is nearly double the size of the supervised data. However, if the unlabeled data quantity is more than 20000, which is nearly triple the size of the supervised data, the performance starts to drop. The change of UAR score and Marco f1 is almost similar. This phenomenon suggests that the balance of the labeled and unlabeled data plays an important role in semi-supervised learning and the best quantity of unlabeled data in our experiment is around 20000. Based on this analysis, the

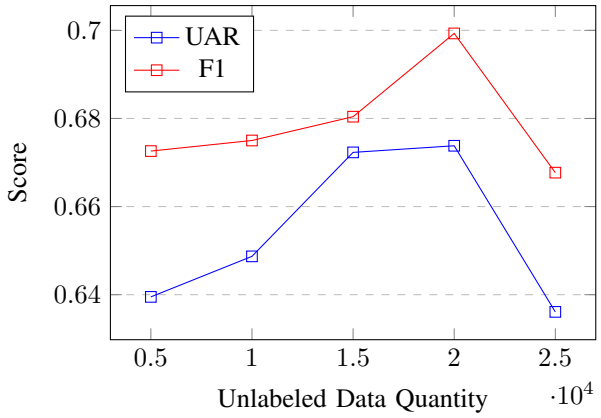


Fig. 6. Performance comparison between different quantity of unlabeled data. The best size of unlabeled data in our experiment is 20000 unlabeled data with around 8000 supervised training data. Too much unlabeled data harms the model capacity

quantity of unlabeled data we reported above are from the setup using 20000 unlabeled samples.

B. Unlabeled Data Normalization Strategy

In common semi-supervised learning scenario, supervised and unsupervised data is acquired in same configurations. But the annotator only annotate a little part of data and remain the rest part of it unlabeled because the corpus is too large to be covered (e.g. for a corpus contains more than 2 billion images, the annotators only label 20000 images to implement semi-supervised learning). In this situation, it's obvious that two types of data can be normalized with the same standard such as mean, variance. However in this work, supervised and unsupervised data is acquired from two different corpus and they are collected in totally different settings. Under this consideration, we want to know the best way for unlabeled dataset normalization. We design 3 types of Z-score normalization settings to investigate the influence on performance as follows.
1) **no normalization**: only compute the mean and variance of each feature dimension on the labeled training set, and then use them to normalize the labeled training, validation and testing sets, while no normalization is applied on the unlabeled data.
2) **self-normalization**: compute the mean and variance of each feature dimension on labeled and unlabeled data respectively and then normalize each feature dimension into $N(0, 1)$ Gaussian distribution respectively.
3) **shift-normalization**: only compute the mean and variance of each feature dimension on the labeled data and use it to normalize both labeled and unlabeled data.
4) **combine-normalization**: combine the labeled training set and unlabeled data and compute the mean and variance of each feature dimension and then use them to normalize both the labeled and unlabeled data.

We select the Dface feature and SSGAN setup to implement valence classification experiments. The results are shown in Figure 7. The bad performance of combine-normalization and shift-normalization demonstrates the difference is significant

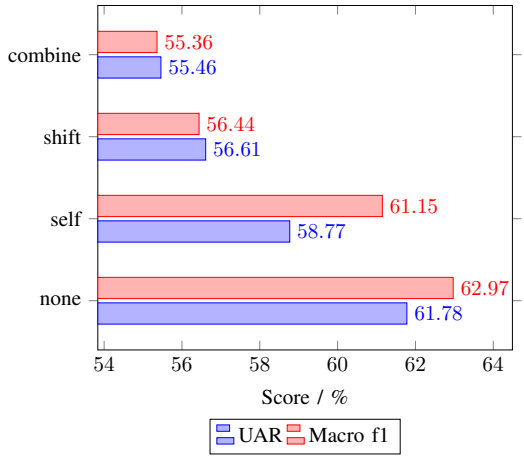
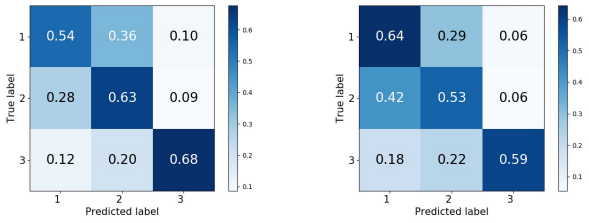


Fig. 7. Performance comparison of different normalization strategies.



(a) no-normalization (b) self-normalization

Fig. 8. Confusion matrices comparison with and without normalization on unlabeled data.

between IEMOCAP and AMI corpus. And to our surprise, the model with no normalization on unlabeled data performs the best. We further check their confusion matrices shown in Figure 8. As we can see, more prediction shifts to low level after normalization. It indicates that this processing narrows the difference between unlabeled data and confuses the classifier to output more medium level prediction. This investigation suggests that normalization is not always necessary and it depends on the real situation.

C. Faces Deformation

During the face extraction process for IEMOCAP mentioned in section IV-B, we noticed the phenomenon that we can capture the frontal view of the left person but only get the side face of the right person in most cases because of the camera position. We present an example in Figure 9. To gain more insight of the impact of this factor, we compute the accuracy of two types of face respectively based on result of visual semi-supervised experiment with GANs. Shown in Table VIII, the performance on frontal faces is significantly better than the side one. The analysis also indicates that the performance of emotion recognition model trained on standard frontal faces can be better.

VI. Conclusion

We investigate effectiveness of the improved semi-supervised learning method with GANs on emotion recognition. Extending the uni-modality approach in previous work,

Fig. 9. Taking a frame from one of the videos as an example. We can capture the frontal view of the left person but only get the side face of the right person at most of time because of the camera position.

TABLE VIII
PERFORMANCE COMPARISON OF TWO TYPES OF FACES. THE PREDICTION ON FRONTAL FACES IS SIGNIFICANTLY BETTER THAN SIDE FACES.

Type	correct	wrong	Accuracy
Frontal face	1686	877	65.78%
Side face	1441	924	60.93%

we propose a multi-modality network structure to implement semi-supervised emotion recognition with acoustic and visual modalities. Extensive experiments demonstrate that unlabeled data and multi-modality fusion strategy both benefit the classification performance. We further conduct several comparison experiments to analyze the influence of unlabeled data quantity and normalization on the recognition performance. In the future, we will further explore other modalities to optimize the semi-supervised training process on emotion recognition.

VII. ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (No. 61772535), Beijing Natural Science Foundation (No. 4192028), and National Key Research and Development Plan (No. 2016YFB1001202).

REFERENCES

[1] Morena Danieli, Giuseppe Riccardi, and Firoj Alam. 2015. Emotion Unfolding and Affective Scenes:A Case Study in Spoken Conversations. In *Icmi-workshop on Emotion Representations Modelling for Companion Technologies.* 5–11.

[2] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018.* 292301. https://doi.org/10.1145/3240508.3240578

[3] Hao-Chen Dong, Yu-Feng Li, and Zhi-Hua Zhou. 2018. Learning From Semi-Supervised Weak-Label Data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.* 2926–2933.

[4] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford,and Xi Chen. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain.*

[5] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *Computer Science*(2015).

[6] Jonathan Chang and Stefan Scherer. 2017. Learning representations of emotional speech with deep convolutional generative adversarial networks. In *IEEE International Conference on Acoustics.*

[7] Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR* abs/1701.07875 (2017). arXiv:1701.07875 http://arxiv.org/abs/1701.07875

[8] Ishaan Gulrajani, Faruk Ahmed, Martn Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved Training of Wasserstein GANs. *CoRR* abs/1704.00028 (2017). arXiv:1704.00028 http://arxiv.org/abs/1704.00028

[9] Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. 2018. Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its Dual Effect. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings.*

[10] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S. Huang. 2016. Multi-Modal Audio, Video and Physiological Sensor Learning for Continuous Emotion Prediction. In *International Workshop on Audio/visual Emotion Challenge.* 97–104.

[11] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shiri Saleem, Rohi Kumar, Vembu Aravind Namandi, and Prasad Rohit. 2012. Emotion recognition using acoustic and lexical feature. In *INTERSPEECH 2012.* 366–369.

[12] Zhaocheng Huang, Ting Dang, Nicholas Cummins, Brian Stasak, and Julien Epps. 2015. An Investigation of Annotation Delay Compensation and Output-Associative Fusion for Multimodal Continuous Emotion Prediction. In *International Workshop on Audio/visual Emotion Challenge.*

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David WardeFarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *International Conference on Neural Information Processing Systems.* 2672–2680.

[14] Samuli Laine and Timo Aila. 2016. Temporal Ensembling for Semi-Supervised Learning. *CoRR* abs/1610.02242 (2016). arXiv:1610.02242 http://arxiv.org/abs/1610.02242

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP:interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (2008), 335–359.

[16] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, and Melissa Kronenthal. 2006. The AMI meeting corpus:a pre-announcement. In *International Workshop on Machine Learning for Multimodal Interaction*. 28–39.

[17] Shuzhe Wu, Meina Kan, Zhenliang He, Shiguang Shan, and Xilin Chen. 2017. Funnel-Structured Cascade for Multi-View Face Detection with Alignment-Awareness. *Neurocomputing* 221, C (2017), 138–145.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition.* 2261-2269.

[19] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. In *ACM International Conference on Multimodal Interaction(ICMI).*

[20] Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition. In *The Workshop on Audio/visual Emotion Challenge*. 19–26.

[21] Zakaria Aldeneh, Soheil Khorram, Dimitrios Dimitriadis, and Emily Mower Provost. 2017. Pooling acoustic and lexical features for the prediction of valence. In *Acm International Conference on Multimodal Interaction.*

[22] Stefan Scherer, John Kane, Christer Gobl, and Friedhelm Schwenker. 2013. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech Language* 27, 1 (2013), 263–287.

[23] Yutaka Sasaki et al.2007. The truth of the F-measure. *Teach Tutor mater* 1, 5 (2007), 1–5.

[24] J. A. Russell, J. A. Bachorowski, and J. M. Fernandez-Dols. 2003. Facial and vocal expressions of emotion. *Annual Review of Psychology* 54, 54 (2003), 329.

[25] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2010. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *IEEE International Conference on Acoustics Speech Signal Processing*.

[26] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2362–2365.