Classification-aware Semi-supervised Domain Adaptation

Gewen He[1†], Xiaofeng Liu[2,3†], Fangfang Fan[2], Jane You[4]
[1]Florida State University, [2]Harvard University,[3]Carnegie Mellon University, [4]HK PolyU
†Contribute equally

Abstract

Deep neural networks are usually data-starved, but manually annotation can be costly in many specific tasks. For instance, the emotion recognition from the audio. However, there is a large amount of public available labeled image-based facial expression recognition datasets. How could these images help for the audio emotion recognition with limited labeled data according to their inherent correlations can be a meaningful and challenging task. In this paper, we propose a semi-supervised adversarial network that allows the knowledge transfer from the labeled videos to the heterogeneous labeled audio domain hence enhancing the audio emotion recognition performance. Specifically, face image samples are translated to the spectrograms class-wisely. To harness the translated samples in a sparsely distributed area and construct a tighter decision boundary, we propose to precisely estimate the density on feature space and incorporate the reliable low-density sample with an annealing scheme. Moreover, the unlabeled audios are collected with the high-density path in a graph representation.

As a possible "recognition via generation" framework, we empirically demonstrated its effectiveness on several audio emotional recognition benchmarks. We also demonstrated its generality on recent large-scaled semi-supervised domain adaptation tasks.

1. Introduction

Sufficiently large-scale labeled data required by deep neural networks can be rarely available in many practical scenarios [39]. The advancement of emotion recognition with the modalities other than facial image is largely hindered by the available labeled data [1, 2].

In the meantime, the available image data for facial expression recognition (IFER) are relatively richer [37, 33]. Hence, a worthwhile research question is can we facilitate the audio emotion recognition (AER) with IFER data. Many recognitive psychology studies evidenced the correlation of a person's facial expression and the emotional state content in their voice [10, 52]. This could because the infants develop their visual and auditory perceptions by fusing visual sense with audio cues [18, 43]. Therefore, a mapping of these two heterogeneous domains can be potentially attained.

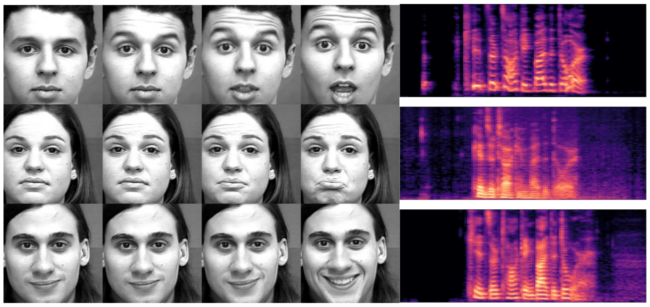

(a) IFER (b) Audio Spectrogram

Figure 1: Audio emotion recognition with IFER data by transferring the IFER in CK+ dataset (a) to the corresponding audio spectrograms (b).

From the generation perspective, many works have been proposed for visual-audio transfer. For example, [8] use the conditional generative adversarial networks (GAN) [16], and [19] propose to apply the Cycle GAN, which are typically style transfer methods. However, these methods target for generating realistic samples with good visual/auditory quality, and not specially designed from the recognition with data augmentation perspective.

Conventionally, the GAN frameworks are not well-matched to supervised/semi-supervised recognition tasks [39, 32, 35]. This is because of the GAN-generated results are expected to align with the central part of the real data distribution. For instance, the class-conditioned style transfer method tires to generate several samples with different appearance or background environment but keep the same class label. Those inner-class samples are generated according to the distribution of the training set and expected to be similar to them. This is achieved by decoding the high-density area points in the shared feature space that follows the real training data distribution [33]. However, the tight decision boundary highly relies on the reliable samples distributed in the low-density areas of the feature space [40, 36, 29, 26, 38, 25, 28, 27]. Thus, the generated samples usually cannot support the network to adjust the boundary.

Recently, [2] propose to generate AER training data with labeled paired visual-audio data. However, this setting is somewhat weird considering the number of labeled visual-audio pairs is even more limited than labeled audio data since the latter is a subset of the former. The deformation or variance of generated samples is originated from an additional random noise input. Without the noise injection, it can degrade to a simple autoencoder to regenerate the existed labeled audio data. Therefore, [2] does not utilize the information of additional large scale labeled image-only IFER data. Besides, the extended semi-supervised setting in [2] indicates the unavailable of some source visual (i.e., face image) label, and propose to predict the label of source sample using an IFER network pretrained with the labeled part of source visual image.

In this paper, we propose to augment the audio-based emotion recognition with the large scale labeled visual IFER data following an unpaired semi-supervised heterogeneous data augmentation manner. Specifically, we achieve the vicinal risk minimization using a semi-supervised classification-aware face-spectrogram translator with the GAN [16, 55, 34, 24] and variational autoencoder (VAE) [14, 7] as its backbone. The facial expression images and spectrograms are not necessary to be paired for our training of the translator, which enables us to resort widely available IFER data.

Our setting can also be regarded as the semi-supervised domain adaptation problem [57]. The typical adversarial method improves generalization on unlabeled part of target data with feature level GAN [48]. However, we play the adversarial game on high dimensional generation and produce more training data for the semi-supervised task. Hence it not only be able to improve the accuracy of AER in our experiments but also to expend or create a new AER dataset for all of the other advanced AER methods.

To summarize, our contributions are: 1) We evidenced that it is possible to facilitate audio emotion recognition with limited labeled data using a large amount of labeled IFER data by exploring the visual-audio correlation in an unpaired manner. 2) We propose a novel classification-aware semi-supervised translator that can well address the large gap of heterogeneous domains on pixel-level. 3) We give a more precisely density estimation to incorporate reliable low-density generation with an annealing scheme and explore the usability of unlabeled target samples following the high-density path on a graph.

2. Related works

Semi-supervised domain adaptation Domain adaptation is a form of transfer learning, in which the task remains the same, but there is a domain shift or a distribution change between the source and the target [57]. In our setting, the visual face expression image is the source domain and the audio modality is the target domain.

With a problem-oriented taxonomy, we are targeting to the heterogeneous and semi-supervised domain adaptation [57]. Most recent researches focus on unsupervised domain adaptation where the target domain is totally unlabeled [59]. However, the limited labeled target data can be a more realistic scenario in many real-world tasks. The adversarial semi-supervised domain adaptation has not been fully explored [49]. Recently, [49] propose a feature-level game with entropy maximization to align the feature distribution. In contrast, we align two domains with the pixel-level game and explicitly generate the transferred data. Moreover, the utilization of the labeled target domain is based on the density that is essentially different from the entropy used in [49].

Semi-supervised Learning Our work is also closely related to the semi-supervised learning which makes use of unlabeled data for training, typically a small amount of labeled data with a large amount of unlabeled data. Noticing that the labeled and unlabeled data are independent identically distributed (i.i.d.). Generative model [11, 50], model-ensemble [22], and adversarial approaches [44] have contributed to the performance improvements of semi-supervised learning, but do not address domain shift in semi-supervised domain adaptation. We are exploring the usability of both the unlabeled data in the same domain and labeled data in the different domains.

Visual-Audio Correlation and Translation The related audio-visual studies have a large progress in recent years. An interesting study of cross-modal relationships of audio and visual cues was introduced in [8] where conditional GANs were applied to generate one modality while another modality was given as an input. To do so, the authors introduced two separate networks (image-to-sound and sound-to-image) to perform cross-modal generations in both ways. Inspired by this work, authors in [19], built a model called Cross-Modal Cycle Generative Adversarial Model to perform cross-modal mappings between image and audio. Authors in [9] introduced a system that performs audio-video synchronization between mouth and speech in a video. To facilitate the task, a two-stream network was implemented by having one network dedicated for audio and one for video and coupled together by using the constructive loss that is judging whether or not the embedding from the two streams belong to a synchronized video pair or not. Similarly in [21], an audio-visual study was performed to perform temporal synchronization. Likewise in

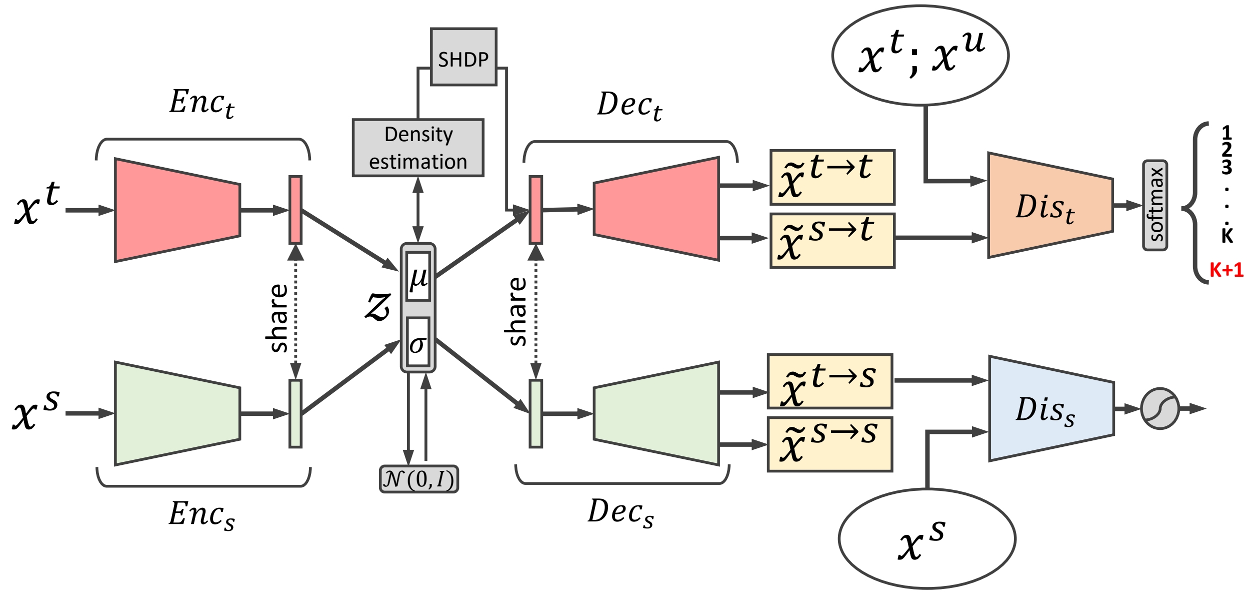


Figure 2: An overview of the model architecture. An example in the source domain is translated to the target domain in the translation unit. Meanwhile translated examples are categorized by density. Low-density samples are used in the adversarial setting. High-density samples are utilized as augmented examples.

[9] a two-stream network using constructive loss function was implemented.

3. Proposed methods

For the semi-supervised domain adaptation setting, there is a totally labeled source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{m_s}$ and a partially labeled target domain. We denote the labeled part as $\mathcal{D}_t = \{(x_i^t, y_i^t)\}_{i=1}^{m_t}$ while the unlabeled part as $\mathcal{D}_u = \{(x_i^u)\}_{i=1}^{m_u}$. m_s, m_t and m_u are the number of samples in each domains, and usually the avaliable m_s, m_u is larger than m_t. We have the shared K classes in all domains, for example the shared K expression in audio and visual datasets. Our objective is learning on $\mathcal{D}_s, \mathcal{D}_t$ as well as the training set of $\mathcal{D}_u$, and evaluate on the test set of $\mathcal{D}_u$.

3.1. Recap: Good Semi-supervised Learning that Requires a Bad GAN

[50] proposed a method to share the knowledge of the unlabeled data modeled by GAN with the classifier. [50] argues that the classifier should be able to play the role of the discriminator of a GAN in addition to the classification. [12] explores what does the classifier learns from the adversarial training in the setting of [50]. The first finding is if the generator distribution p_G exactly matches the true data distribution p, then for any optimal solution Dis of the supervised objective:

$$\max_{Dis} \mathop{\mathbb{E}}_{x,y\sim p} logP(y|x) \tag{1}$$

there exists an optimal solution Dis^* of the GAN based semi-supervised objective:

$$\max_{Dis} \mathop{\mathbb{E}}_{x\sim p_g} logP(fake|x) + \mathop{\mathbb{E}}_{x,y\sim p} logP(y|x) + \mathop{\mathbb{E}}_{x\sim p} logP(true|x) \tag{2}$$

such that Dis and Dis^* share the same generalization error. In another word, the joint parameterzied discriminator-classifier does not learn from the high density examples generated by GAN.

The second finding is, under mild assumptions, the GAN based semi-supervised classifier can strongly correctly classify the high-density subsets of the examples by learning to discriminate the data points generated in the low-density areas.If classifier can label training data and true-fake data correctly, under one extra mild assumption that the true-fake belief is also strong, [12] theoretically guaranteed that the joint classifier can encourage to place the K classes decision boundaries at the low-density region.

3.2. Classification-Aware augmentation

In the SSDA setting, target labeled examples are relatively limited. We propose to generate the new target example that we are confident of its label. We generate a new

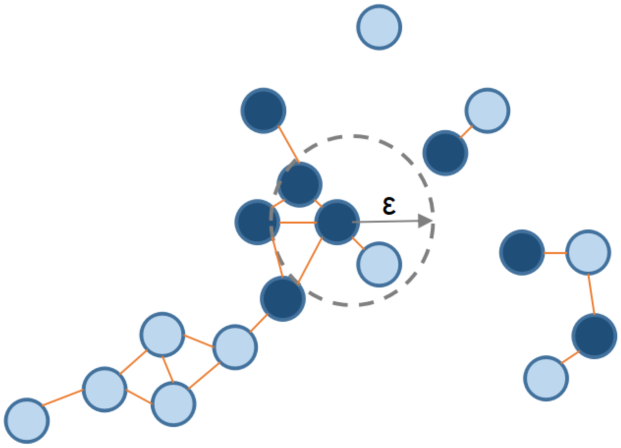


Figure 3: Graph-based manifold representation. Dark blue represents labeled, light blue represents unlabeled.

target example according to the learned conditional distribution $p(x|z)$ based on a latent code z. The latent code space of z is constrained to be shared among the source and target domain and the latent component is also constrained to have the same semantic meanings in the two domains.

The latent feature distribution of data points from the source and the target domain are expected to align with each other class-wisely. In the feature space, these data are expected to present the properties of a certain class of real source domain data points clustering and form a high-density region.

Based on the work of [23], our proposed translation unit has two VAEGANs, (Enc_s, Dec_s, Dis_s) and (Enc_t, Dec_t, Dis_t) for the target domain and the source domain respectively [39, 33, 32, 7]. The two autoencoders (Enc_s, Dec_s) and (Enc_t, Dec_t) share parameters weight at a few layers near the latent vector so that the source domain and the target domain share the latent space [56, 55, 30, 31].

Dis_t joint parameterizes the classifier and the true-fake discriminator [12]. The class $K + 1$ refers to the new class representing generated data.

3.3. Low density sample annealing

The translation unit is co-trained with the classifier Dis_t in a round-based training manner. During the training process, the translation unit generates target data points of higher density in the later batch. We propose to incorporate only the reliable generated target data points $\tilde{x^t}$ to $\mathcal{D}_t$ at each batch. The low-density portion is used in the adversarial training in the translation unit. The set of all the generated low-density examples is denoted as D_g hereinafter. The high-density portion is used as new labeled training data in D_t.

In the proposed annealing scheme, we have a hyper-parameter ϵ that increases as the training proceeds. For every batch, the $\epsilon\%$ generated target data points with the highest density are added to D_l, the rest are added to D_g. At the beginning of the training, all the generations are set to D_g. When the training converges, $\epsilon\%$ increase to 0.8.

Density estimation. Because the $\tilde{x^t}$ is generated by the decoder Dec_t in a trained VAE_t. We are able to estimate a pretty tight bound of the density of $\tilde{x^t}$. Recall, in a variational auto-encoderas, the evidence lower bound is a lower bound of the density $\log p\left(\mathrm{x}\right)$:$\log p\left(\mathrm{x}\right) \geq \mathbb{E}_{q(\mathrm{z}|\mathrm{x})}\left[\log \frac{p(\mathrm{x}|\mathrm{z})p(\mathrm{z})}{q(\mathrm{z}|\mathrm{x})}\right]$where z is the latent variable. We can approximate the density of a generated target example x with importance sampling methods on the distribution of $q\left(z|x\right)$. In fact there are many well established methods to do more computationally efficient estimation [5][45][13].

3.4. Reliability path

We assume, in the feature space, similar points are likely to share the same label and we adopt a regularizer to enforce this assumption. We propose to approximate the manifold by constructing a graph representation of all the examples in D_t and D_u in the feature space. We first construct a reliability path on the graph representation that all the nodes on it share the same label whenever the node's label is known. Then, the unlabeled examples on a reliability path is assigned with the path's label.

Graph representation. [3] Given n points $x_1, ..., x_n$ in $\mathbb{R}^l$, we construct a weighted graph with N nodes, one for each point in the D_t and D_u, and a set of edges connecting neighboring nodes. Nodes i and j are connected by an edge if $\|f(x_i) - f(x_j)\|^2 \leq \beta$, parameter $\beta \in \mathbb{R}$. Weight the edge with a Gaussian radial basis function:$W_{ij} = e^{-\gamma\|f(x_i)-f(x_j)\|^2}$. An example of graph representation construction is shown in Fig. 3.

We define the smoothness of a graph representation$S = \frac{1}{2}\sum_{ij}\left(y_i - y_j\right)^2 W_{ij}$where y_i, y_j are the labels of node

i, j, they are either known or predicted with Dis_t. S measures the smoothness. The lesser S is, the more smoothy the graph is. S can be computed with the Laplacian Eigenmaps $S = y^T Ly$where y is the labels on the graph who depends on Dis_t and L is graph laplacian [3]. According to [54], we add S to the objective function as a regularization term.

3.5. Training objective and its interpretation

There are three sources of data augmentation in our method: the generated target data points with high density, they are the new supplement to D_t; the generated low density data points are added to D_g and help dis_t learn the low-density separation, and those examples in D_u who are provided with label via the reliability path.

In the translation unit, let $GAN_{s\to t}$ denote the GAN consists of the encoder Enc_s, generator Dec_t and the discriminator Dis_t. $GAN_{s\to t}$ converts data points from source domain to target domain. In this generation path, $P_{Dis_t}(K+1|x)$ is the true or fake signal for the adversarial training. Similarly we denote $GAN_{t\to s}$ as the GAN consists of Enc_t, Dec_s and the discriminator Dis_s. Dis_s is a regular discriminator. VAE_s, VAE_t, $GAN_{s\to t}$, $GAN_{t\to s}$ together are the translation unit. The learning objective includes three components: the examples in source and target domain can be reconstructed in VAE_s and VAE_t respectively; minimization of the GAN loss of the translation in both directions between the source domain and the target domain; the cycle-reconstruction loss of the two direction of translation $\mathcal{L}_{s\to t}$, $\mathcal{L}_{t\to s}$:

$$\min_{(Enc_s,Enc_t,Dec_s,Dec_t,Dis_s,Dis_t)} \max_{(Dis_s,Dis_t)} \mathop{\mathbb{E}}_{(D_s,D_t,D_u)} [$$
$$\mathcal{L}_{VAE_s}(Enc_s, Dec_s) + GAN_{t\to s}(Enc_t, Dec_s, Dis_s) +$$
$$\mathcal{L}_{t\to s}(Enc_t, Dec_s, Enc_s, Dec_t) + \mathcal{L}_{VAE_t}(Enc_t, Dec_t) +$$
$$GAN_{s\to t}(Enc_s, Dec_t, Dis_t) +$$
$$\mathcal{L}_{s\to t}(Enc_s, Dec_t, Enc_t, Dec_s)]$$

For the paired training data from source and target domain are available or, in another word, we know the ground truth translation of the input example, we follow the philosophy of fix point learning [51] to replace the objective function $GAN_{t\to s}(Enc_t, Dec_s, Dis_s)$ and $GAN_{s\to t}(Enc_s, Dec_t, Dis_t)$ in Equation (8) to the $L1$ loss between the ground truth translation and the generated one:

$$GAN_{t\to s}(Enc_t, Dec_s, Dis_s) : \mathcal{L}_{l_1}(G(Enc_t, Dec_s), GT_S)$$
$$GAN_{s\to t}(Enc_s, Dec_t, Dis_t) : \mathcal{L}_{l_1}(G(Enc_s, Dec_t), GT_T)$$

where $\mathcal{L}_{l_1}$ indicates the l_1 loss, $G(Enc_s, Dec_t)$ denotes the generation of the target dexample. $G(Enc_t, Dec_s)$ means similarly. GT_T and GT_S represent the ground truth examples in both domain respectively.

Lastly, the smoothness regularizer S encourages the examples of the same class from D_t and D_u clustering in the features space. The overall objective function for Dis_t is:

$$\max_{Dis_t} \mathop{\mathbb{E}}_{x\in D_g} logP_{Dis_t}(K+1|x) + \mathop{\mathbb{E}}_{x,y\in D_t} logP_{Dis_t}(y|x)$$
$$+ \mathop{\mathbb{E}}_{x\in D_u} [logP_{Dis_t}(y < K+1|x)$$
$$+ \sum_{k=1}^{K} P_{Dis_t}(k|x) logP_{Dis_t}(k|x)] - \lambda S$$

where $S = y^T Ly$, λ is a hyper-parameter to control a trade-off between smoothness term and classification.

4. Experiments

4.1. Augmenting Audio from Visual

We evaluate the proposed classification-aware visual to audio data augmentation in this section. To evaluate the effectiveness of the proposed method, extensive experiments have been conducted on two publicly available multimodal emotion expression datasets.

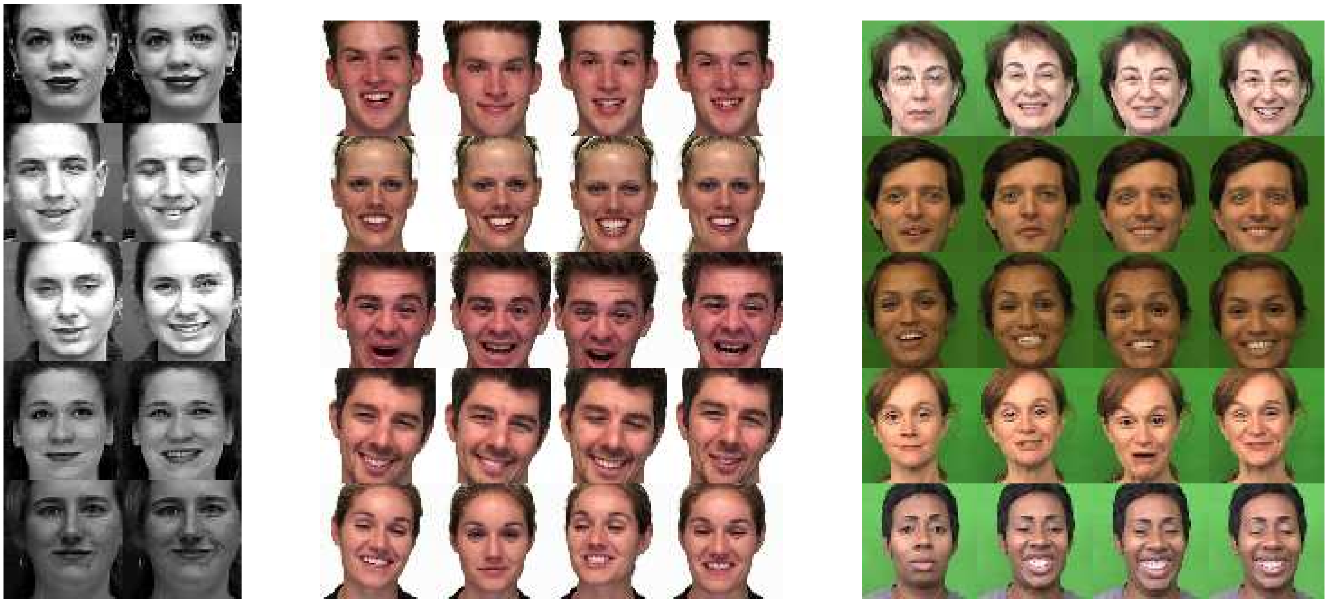
CREMA-D [6] is a multi-modal emotion data set with both facial and audio expressions. 91 actors and actresses are participated to generate the six universal emotions: Happy, Sad, Anger, Fear, Disgust and Neutral in 7442 clips.

RAVDESS [41] includes 24 gender-balanced professional actors vocalizing two statements in Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprised emotions. There are a totally of 2452 trials as Audio-Visual files.

[2] separate both CREMA-D and RAVDESS to four parts, i.e., S1 for classifier training, S2 and S3 for the additional network structure's training, and S4 for testing. We follow their setting and use S1 for thee labeled training set, S2 and S3 as the unlabeled training data, while leave S4 as testing set.

The large scale audio clips are hard to collect, especially the number of the actor is very limited. To augment the audio recognition, we propose to utilize the facial image in both of these multi-modal datasets and the large scale IFER datasets: CMU Multi-PIE, CK + [42], MMI Dataset [53], Oulu-CASIA VIS Dataset [58].

For these IFER datasets, we only use the data with shared emotions with CREMA-D or RAVDESS datasets. All of these IFER datasets are merged into a large one. We do not use the video-based facial expression recognition version of IFER datasets is because the expression development (from neutral to the apex of expression) of these datasets is essentially different from the AER which has the same emotion from the start to the end. Moreover, the correlation of paired facial expression image and audio data has been evidenced by many prior works.



(a) CK+

(b) RAVDESS

(c) CREMA-D

Figure 4: Sample preprocessed IFER data.

Data Preparation For the visual modal datasets we follow the [33] to locate the 49 facial landmarks. Then, face alignment is done to reduce in-plane rotation and crop the region of interest based on the coordinates of these landmarks to a size of 64×64. Augmentation procedures such as flipping and grayscale are employed to increase the number of training images and alleviate the chance of over-fitting.

All the images are processed with the standard histogram equalization and linear plane fitting to remove unbalanced illumination. Finally, we normalize them to have zero mean and unit variance. In the testing phase, a single center crop with a size of 60×60 is used as input data.

For the audio modal of CREMA-D, RAVDESS, we make use of spectrogram representation of the raw audio signals. We resize the spectrograms to 156 × 64 in 2-D array.

Quantitative evaluations We measure the classification accuracy gains from data augmentation. The supposed high-quality generation (which are those samples of high density in our setting) should contain features not presented in the labeled training examples thus improve the accuracy on the testing set. The classifier should also learn from the unlikely examples to place the decision boundaries better. The main results are shown in Table 1. We have four sets of experiments.

The first experiments with IFER datasets and RAVESS. In Experiment 1 we do not utilize the pairing information between the visual and audio modal of each RAVDESS clip. The second experiment works with the same dataset as Experiment 1, but we consider the pairing information in this case which means we calculate the l_1 loss of the visual to audio and audio to visual translation in the training objective. Experiments 3 and 4 are based on IFER datasets and CREMA-D with a similar setting as Experiments 1 and 2.

First, all three data augmentation techniques are validated to effectively improve the performance of the classifier on the testing set. The high-density samples of generated spectrograms are convincingly useful new labeled training examples in the spectrogram domain. The pairing information in visual-audio examples do improve classification performance but is not significant. This may be because the translation unit already learns to guess likely example in the distribution of audio spectrograms that keeps the emotion feature of the input IFER.

In addition to the metric above. We adopt the evaluation metric for generated samples proposed by [50], the Inception Score (IS). We quantify the quality of generated spectrograms with $exp(\mathbb{E}_x KL(p(y|x)||p(y)))$ and make use of an Inception network pre-trained on performing emotion recognition in real spectrogram datasets, e.g., the learned classifier in our framework as [2]. The higher the IS is the better the quality of the generated samples. Another applied qualitative metric is the Frechet Inception Distance (FID) [20]. It compares the statistics of generated samples to the real ones, instead of only evaluating generated ones. Lower FID values mean better image quality and diversity.

The translation quality metric is reported in Table 1 lower part. To reflect the comparative goodness of the generated samples, we use the spectrogram representa-

	UP IFER CRE	P IFER CRE	UP IFER RAV	P IFER RAV
Base	30.81%	-	30.65%	-
- Low - Rel	49.2%	51.17%	50.34%	53.12%
- Low	51.83%	54.68%	52.74%	53.55%
- High - Rel	41.1%	43.82 %	42.9%	42.71%
All	54.53%	58.71%	53.34%	56.12%
IS BaseScore	3.12	-	3.24	-
IS Low	2.65	2.63	2.77	2.80
IS High	2.72	2.84	2.87	2.89
FID Low	64.2	63.7	59.1	57.5
FID High	61.3	60.4	57.5	56.2

Table 1: Classification accuracy and generation quality metric. UP denotes not using pairing information in bi-modal datasets. P means using pairing information. CRE, RAV mean the two multi-modal datasets. IFER means the merged large IFER dataset. Base refers to learn to classify the spectrograms only with labeled examples and there is no knowledge transferring from IFER data sets. - Low - Rel refers to learn to classify with labeled target examples D_t that are supplemented with the new spectrograms generated from IFER data. - Low refers to we further supplement D_t by assigning labels to data in D_u with the reliability path. - High - Rel means we do not supplement D_t with data augmentation. All means adopting all the proposed techniques.

tions of real audio in the comparison which are denoted as BaseScore.

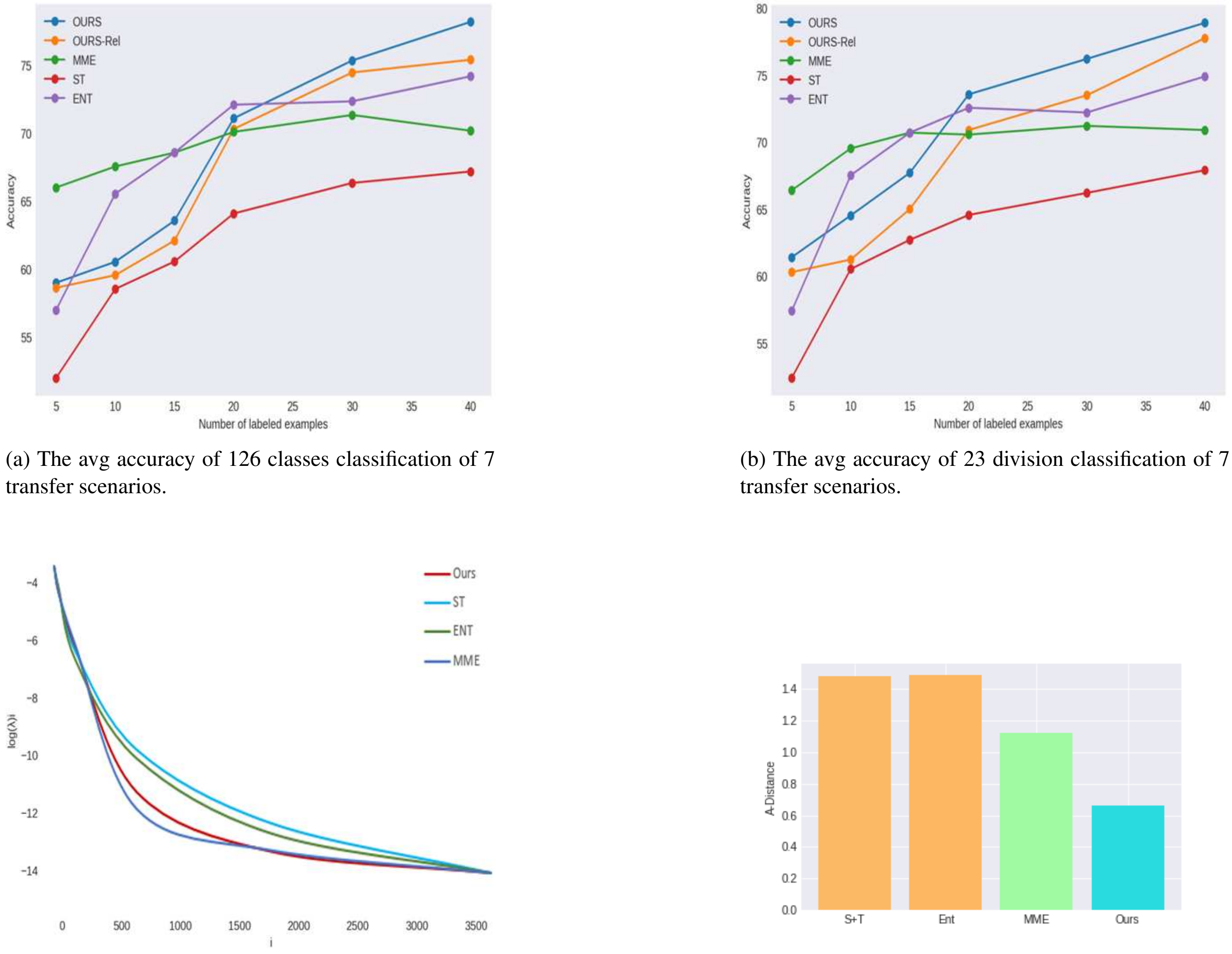
4.2. Bench marking with other Semi-supervised Domain Adaption Methods

Our framework is also a general-purpose semi-supervised domain adaptation solution. The benchmark focus on the accuracy comparison as the number of labeled target example changes. The comparison is based on DomainNet [46], we randomly choose some labeled target domain examples per class as labeled training target examples. [46] has 345 classes in 24 divisions. We conduct the study on 4 domains: $Real$ (R), $Clipart$ (C), $Painting$ (P), $Sketch$ (S). We adopt the same experiment setting as [49] for fair comparison: we use 126 classes that have largest number of examples and construct 7 adaptation scenarios: R→C, R→P, P→C, C→S, S→P, R→S and P→R. We also provide testing results of classification by the 23 divisions (excluding the ’others’ division).

Baseline. S+T [47] is the distance-based classification that has been extensively used in few-shot learning. The model is trained with labeled examples in the source and target domain. Unlabeled examples in the target domain is not used. ENT [17] is a model trained with labeled source and target and unlabeled target using standard entropy minimization. Entropy is calculated on unlabeled target examples and the entire network is trained to minimize it. MME [49] assumes domain invariant prototypes per class. The unlabeled target examples are used to align prototypes dominated by source examples to target domain feature distribution and to search for more discriminative feature extractor.

Results We report the results of the benchmark study of the 126 class classification in DomianNet in Figure. 5 (a). Even if the labeled target examples are very limited our method still able to achieve similar accuracy as compared to other methods. This finding may evidence that the generated target domain examples do posses class-specific features. As the number of target examples increases, our method tends to perform better than the other methods, this is an expected result for two reasons: some other methods are designed for prototype-based few-shot learning, they can not fully utilize the additional target examples; our method learns a better decision boundary with the help of the low-density examples. The next finding is the proposed reliability path effectively guess the label of the unlabeled target example. The performance gain is lesser initially when the number of labeled examples is small and tends to increase as we have more target labeled examples. Figure. 5 (b) shows a similar result in 23 divisions classification. When the number of labeled target examples is small, our method achieves a better result than the 126 class classification case.

Quantitative Feature Analysis We quantitatively estimate the features representation of the target domain in each method. Following the analysis of [15], the co-variance matrix of features in the target domain can measure if the features representation is discriminative. Each eigenvector of the covariance matrix corresponds to a component of the feature and the associated eigenvalue represents the contri-



(a) The avg accuracy of 126 classes classification of 7 transfer scenarios.

(b) The avg accuracy of 23 division classification of 7 transfer scenarios.

(c) Eigenvalues

(d) $\mathcal{A}$-distance

Figure 5: Accuracy vs number of target labeled examples: a) and b). Quantitative feature analysis: c) and d).

bution. If the feature is discriminative, we will need few components to summarize the feature hence the first few eigenvalues are expected to be large. Figure 5. (c) shows that our feature representation's eigenvalues decreases more quickly. Next, following [4], we computed the $\mathcal{A}$-distance of feature distribution in the source and target domain. In the benchmark, only MME is designed for the feature alignment. Figure 5. (d) shows the result.

5. Conclusions

We proposed a novel unpaired semi-supervised data augmentation method which can also be regard as a image-level heterogeneous semi-supervised domain adaptation framework. It is based on a GAN and VAE backbone with joint parameterized discriminator and classifier. The modules are optimized with a serials of semi-supervised objective. Other than explicitly class-aware conditional alignment, we also propose to give a tighter support of decision boundary in semi-supervised setting by exploring the low-density area. We encourage the generation of low-density sample with precisely density estimation while selecting the reliable samples following the high density-path in a graph. We empirically demonstrated the superiority of our method over many baselines and shown its generality on semi-supervised domain adaptation benchmarks.

References

[1] 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. 1

[2] Christos Athanasiadis, Enrique Hortal, and Stylianos Asteriadis. Audio-visual domain adaptation using conditional semi-supervised generative adversarial networks. *Neurocomputing*, 2019. 1, 2, 5, 6

[3] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. 4, 5

[4] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007. 8

[5] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015. 4

[6] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 5

[7] Tong Che, Xiaofeng Liu, Site Li, Yubin Ge, Ruixiang Zhang, Caiming Xiong, and Yoshua Bengio. Deep verifier networks: Verification of deep discriminative models with deep generative models. *arXiv preprint arXiv:1911.07421*, 2019. 2, 4

[8] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM, 2017. 1, 2

[9] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 2, 3

[10] Erin Cvejic and Chris Kim, Davis. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication*, 52(6):555–564, 2010. 1

[11] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan R Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *Advances in neural information processing systems*, pages 6510–6520, 2017. 2

[12] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Ruslan R Salakhutdinov. Good semi-supervised learning that requires a bad gan. In *Advances in neural information processing systems*, pages 6510–6520, 2017. 3, 4

[13] Xinqiang Ding and David J Freedman. Improving importance weighted auto-encoders with annealed importance sampling. *arXiv preprint arXiv:1906.04904*, 2019. 4

[14] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 2

[15] Abhimanyu Dubey, Otkrist Gupta, Ramesh Raskar, and Nikhil Naik. Maximum-entropy fine grained classification. In *Advances in Neural Information Processing Systems*, pages 637–647, 2018. 7

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2

[17] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005. 7

[18] Tobias Grossmann. The development of emotion perception in face and voice during infancy. *Restorative neurology and neuroscience*, 28(2):219–236, 2010. 1

[19] Wangli Hao, Zhaoxiang Zhang, and He Guan. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In *Thirty-Second AAAI Conference*, 2018. 1, 2

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Gunter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv: Learning*, 2017. 6

[21] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 2

[22] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2

[23] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017. 4

[24] Xiaofeng Liu. *Research on the technology of deep learning based face image recognition*. PhD thesis, Thesis, 2019. 1. 2

[25] Xiaofeng Liu, Fangfang Fan, Lingsheng Kong, Zhihui Diao, Wanqing Xie, Jun Lu, and Jane You. Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing*, 2020. 2

[26] Xiaofeng Liu, Zhenhua Guo, Site Li, Lingsheng Kong, Ping Jia, Jane You, and BVK Kumar. Permutation-invariant feature restructuring for correlation-aware image set-based recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4986–4996, 2019. 2

[27] Xiaofeng Liu, Zhenhua Guo, Jane You, and BV Kumar. Attention control with metric learning alignment for image set-based recognition. *arXiv preprint arXiv:1908.01872*, 2019. 2

[28] Xiaofeng Liu, Zhenhua Guo, Jane You, and BVK Vijaya Kumar. Dependency-aware attention control for image set-based face recognition. *IEEE Transactions on Information Forensics and Security*, 15:1501–1512, 2019. 2

[29] Xiaofeng Liu, Xu Han, Yukai Qiao, Yi Ge, Site Li, and Jun Lu. Unimodal-uniform constrained wasserstein training for medical diagnosis. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[30] Xiaofeng Liu, Yuzhuo Han, Song Bai, Yi Ge, Tianxing Wang, Xu Han, Site Li, Jane You, and Jun Lu. Importance-aware semantic segmentation in self-driving with discrete wasserstein training. *AAAI*, 2020. 4

[31] Xiaofeng Liu, Wenxuan Ji, Jane You, Georges El Fakhri, and Jonghye Woo. Severity-aware semantic segmentation with reinforced wasserstein training. *CVPR*, 2020. 4

[32] Xiaofeng Liu, BVK Vijaya Kumar, Yubin Ge, Chao Yang, Jane You, and Ping Jia. Normalized face image generation with perceptron generative adversarial networks. In *2018*

IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA), pages 1–8. IEEE, 2018. 1, 4

[33] Xiaofeng Liu, B V K Vijaya Kumar, and Jane Jia. Hard negative generation for identity-disentangled facial expression recognition. *Pattern Recognition*, 88:1–12, 2019. 1, 4, 6

[34] Xiaofeng Liu, Site Li, Lingsheng Kong, Wanqing Xie, Ping Jia, Jane You, and BVK Kumar. Feature-level frankenstein: Eliminating variations for discriminative recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 637–646, 2019. 2

[35] Xiaofeng Liu, Zhaofeng Li, Lingsheng Kong, Zhihui Diao, Junliang Yan, Yang Zou, Chao Yang, Ping Jia, and Jane You. A joint optimization framework of low-dimensional projection and collaborative representation for discriminative classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1493–1498. IEEE, 2018. 1

[36] Xiaofeng Liu, BVK Vijaya Kumar, Chao Yang, Qingming Tang, and Jane You. Dependency-aware attention control for unconstrained face recognition with image sets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 548–565, 2018. 2

[37] Xiaofeng Liu, BVK Vijaya Kumar, Jane You, and Ping Jia. Adaptive deep metric learning for identity-aware facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–29, 2017. 1

[38] Xiaofeng Liu, Yang Zou, Tong Che, Peng Ding, Ping Jia, Jane You, and BVK Kumar. Conservative wasserstein training for pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8262–8272, 2019. 2

[39] Xiaofeng Liu, Yang Zou, Lingsheng Kong, Zhihui Diao, Junliang Yan, Jun Wang, and Jane Li. Data augmentation via latent space interpolation for image classification. In *2018 24th ICPR*, pages 728–733. IEEE, 2018. 1, 4

[40] Xiaofeng Liu, Yang Zou, Yuhang Song, Chao Yang, Jane You, and BV K Vijaya Kumar. Ordinal regression with neuron stick-breaking for medical diagnosis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 2

[41] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song. *PloS one*, 13(5):e0196391, 2018. 5

[42] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 94–101. IEEE, 2010. 5

[43] Dominic W Massaro and Michael M Cohen. Perceiving talking faces. *Current Directions in Psychological Science*, 4(4):104–109, 1995. 1

[44] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training. *arXiv preprint arXiv:1507.00677*, 2015. 2

[45] Sebastian Nowozin. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. 2018. 4

[46] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019. 7

[47] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 7

[48] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. *arXiv preprint arXiv:1904.06487*, 2019. 2

[49] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. *arXiv preprint arXiv:1904.06487*, 2019. 2, 7

[50] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 2, 3, 6

[51] Siddiquee, Md Mahfuzur Rahman, and et al. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE ICCV*, pages 191–200, 2019. 5

[52] Marc Swerts and Emiel Krahmer. Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, 36(2):219–238, 2008. 1

[53] Michel Valstar and Maja Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, page 65. Paris, France, 2010. 5

[54] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012. 5

[55] Chao Yang, Xiaofeng Liu, Qingming Tang, and C-C Jay Kuo. Towards disentangled representations for human retargeting by multi-view learning. *arXiv preprint arXiv:1912.06265*, 2019. 2, 4

[56] Chao Yang, Yuhang Song, Xiaofeng Liu, Qingming Tang, and C-C Jay Kuo. Image inpainting using block-wise procedural training with annealed adversarial counterpart. *arXiv preprint arXiv:1803.08943*, 2018. 4

[57] Jing Zhang and et al. Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. *ACM CSUR*, 52(1):7, 2019. 2

[58] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z Li, and Matti PietikäInen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011. 5

[59] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019. 2