# Image2Audio: Facilitating Semi-supervised Audio Emotion Recognition with Facial Expression Image

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

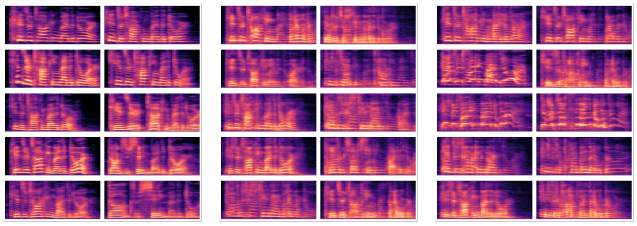
## Abstract

*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.*

## 1. Introduction

The advancement of emotion recognition with the modalities other than facial image is largely hindered by the available labeled data [1, 2]. However, the available image data for facial expression recognition (IFER) are relatively richer [17, 16]. Many recognitive psychology studies evidenced the correlation of a person's facial expression and the emotional state content in their voice [7, 24]. Therefore, a mapping of these two heterogeneous domains can be potentially attained.

Form the generation perspective, many works have been proposed for visual-audio transfer. For example, [6] use the conditional generative adversarial networks (GAN) [11], and [12] propose to apply the Cycle GAN. However, these methods target for generating realistic samples with good visual/auditory quality, and not specially designed from the recognition with data augmentation perspective.



(a) True Example (b) High Density (c) Low Density

Figure 1: Spectrogram representation of raw audio data. In each subfigure, the left column represents samples who belong to emotion Happy. The right column represents samples of emotion Sad. a) is the spectrograms of labeled audio modal. b) is the generated high-density spectrograms. c) is the generated low-density spectrograms.

Conventionally, the GAN frameworks are not well-matched to supervised/semi-supervised recognition tasks [18, 15]. This is because of the GAN-generated results are expected to align with the central part of the real data distribution [16]. However, the tight decision boundary highly relies on the reliable samples distributed in the low-density areas of the feature space [18]. Thus, the generated samples usually cannot support the network to adjust the boundary.

Recently, [2] propose to generate AER 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.

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 [11] and variational autoencoder (VAE) [10] 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 do-

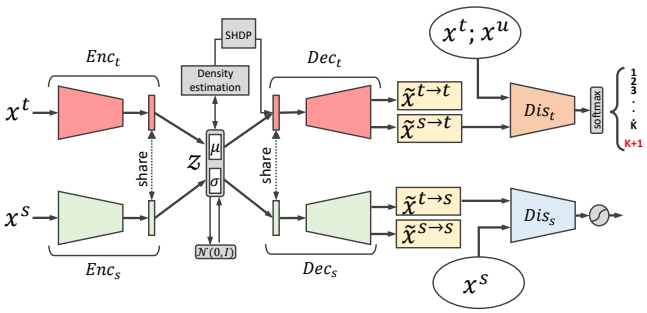



Figure 2: An overview of the model architecture. An example in the source domain is translated to the target domain in the translation unit. Meanwhile translated examples are categorized by density. Low-density samples are used in the adversarial setting. High-density samples are utilized as augmented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

main adaptation problem [27].

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. Proposed methods

For the unpaired semi-supervised domain adaptation setting, there is a totally labeled source domain $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{m_s}$ (e.g., IFER data) and a partially labeled target domain (e.g., audio emotion data). 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 $\mathcal{D}_u$.

### 2.1. Classification-Aware augmentation

In the unpaired semi-supervised domain adaptation setting, the AER classifier has relatively limited labeled examples in the audio domain. We propose to generate the new audio spectrogram that we are confident of its label.

We generate a new spectrogram 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 visual and audio domain and the latent component is also constrained to have the same semantic meanings in the two domains. The latent feature distribution of visual and audio data are expected to align with each other class-wisely.

We transfer the labeled visual data to its corresponding audio version while maintaining its class label. The generation is conditioned on the latent code of labeled IFER data and the generated spectrogram is assumed to preserve the emotional feature of the input IFER data. In the feature space, these data are expected to present the properties of a certain class of real spectrograms data points clustering and form a high-density audio spectrograms area together.

Based on the work of [14], our proposed translation unit has two VAEGANs, $(Enc_s, Dec_s, Dis_s)$ and $(Enc_t, Dec_t, Dis_t)$ for the audio spectrogram domain and the IFER domain respectively. 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 spectrogram and the IFER data share the latent space.

$Dis_t$ joint parameterizes the classifier and the true-fake discriminator [8]. The class $K+1$ refers to the new class representing generated data.

### 2.2. 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 audio spectrograms of higher density in the later batch. We propose to incorporate the reliable generated audio spectrograms $\tilde{x^t}$ in 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 $\epsilon$ that increases as the training proceeds. For every batch, the $\epsilon\%$ generated spectrograms with the highest density are added to $D_l$, the rest is 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 a spectrogram example $x$. Recall, in a variational auto-encoderas, the evidence lower bound is a lower bound of the density $\log p\,(\mathrm{x})$:$\log p\,(\mathrm{x}) \geq \mathbb{E}_{q(\mathrm{z}|\mathrm{x})}\left[\log \frac{p(\mathrm{x}|\mathrm{z})p(\mathrm{z})}{q(\mathrm{z}|\mathrm{x})}\right]$ where $z$ is the latent variable. We can approximate the density of a generated spectrogram example $x$ with importance sampling methods on the distribution of $q\,(z|x)$. In fact there are many well established methods to do more computationally efficient estimation [4][21][9].

### 2.3. 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

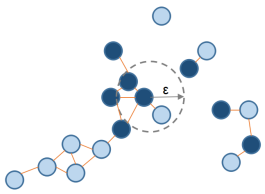



Figure 3: Graph-based manifold representation. Dark blue represents labeled, light blue represents unlabeled.

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 spectrogram 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$S = \frac{1}{2}\sum_{ij}(y_i - y_j)^2 W_{ij}$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 [26], we add $S$ to the objective function as a regularization term.

### 2.4. Training objective and its interpretation

There are three sources of data augmentation in our method: the high density generated spectrograms translated from IFER data are the new supplement to $D_t$; the low density generated spectrograms $D_g$ that 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 IFER data to spectrogram. 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 that translate IFER data to audio spectrogram. The learning objective includes three components: the IFER data and spectrogram data can be reconstructed in $VAE_s$ and $VAE_t$ respectively; minimization the GAN loss of the translation from IFER to spectrogram as well as the translation in the other way; 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 is available or, in another word, we know the ground truth translation of the input example, we follow the philosophy of fix point learning [23] 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:

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

where $\mathcal{L}_{l_1}$ indicates the $l_1$ loss, $G(Enc_s, Dec_t)$ denotes the spectrograms generation from IFER data. $G(Enc_t, Dec_s)$ means similarly. $GT_T$ and $GT_S$ represent the ground truth audio and image samples 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:

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

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

## 3. Experiments

To evaluate the effectiveness of the proposed method, extensive experiments have been conducted on two publicly available multimodal emotion expression datasets.

CREMA-D [5] 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 [19] 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.

[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, S4 for testing.

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 + [20], MMI Dataset [25], Oulu-CASIA VIS Dataset [28].

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.

For the audio modal of CREMA-D, RAVDESS, we make use of spectrogram representation of the raw audio signals. We resize the spectrograms to $156 \times 64$ in 2-D array. The samples of extracted audio representation in RAVDESS are shown in Fig. 1 (a).

We measure the classification accuracy gains from data augmentation. The main results are shown in Table 1. 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.

| | 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 bimodal 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.

In addition to the metric above. We adopt the evaluation metric for generated samples proposed by [22], 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) [13]. 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.

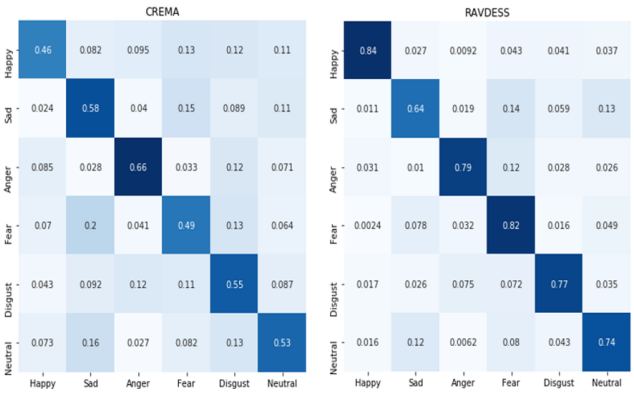



Figure 4: Confusion matrix for AER classification on CREMA and RAVDESS datasets using our methods.

The translation quality metric is reported in Table 1 lower part. To reflect the comparative goodness of the generated samples, we use the spectrogram representations of real audio in the comparison which are denoted as BaseScore. The samples of generated high-density examples and low-density examples are shown in Fig. 1. (b)(c).

## 4. 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, 3, 4

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

[4] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015. 2

[5] 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. 3

[6] 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

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

[8] 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

[9] Xinqiang Ding and David J Freedman. Improving importance weighted auto-encoders with annealed importance sampling. *arXiv preprint arXiv:1906.04904*, 2019. 2

[10] Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016. 1

[11] 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

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

[13] 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. 4

[14] 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. 2

[15] 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

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

[17] 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

[18] Xiaofeng Liu, Yang Zou, Lingsheng Kong, Zhihui Diao, Junliang Yan, Jun Wang, Site Li, Ping Jia, and Jane You. Data augmentation via latent space interpolation for image classification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 728–733. IEEE, 2018. 1

[19] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 3

[20] 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. 4

[21] Sebastian Nowozin. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. 2018. 2

[22] 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. 4

[23] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 191–200, 2019. 3

[24] 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

[25] 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. 4

[26] 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. 3

[27] Jing Zhang, Wanqing Li, Philip Ogunbona, and Dong Xu. Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. *ACM Computing Surveys (CSUR)*, 52(1):7, 2019. 2

[28] 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. 4