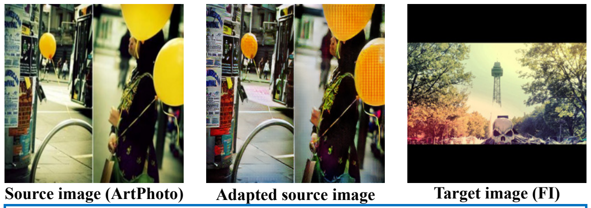
The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

CycleEmotionGAN: Emotional Semantic Consistency Preserved CycleGAN for Adapting Image Emotions

Sicheng Zhao,[†] Chuang Lin,[‡♯] Pengfei Xu,[♯] Sendong Zhao,[§]
Yuchen Guo,[♢] Ravi Krishna,[†] Guiguang Ding,[♢] Kurt Keutzer[†]
[†]University of California, Berkeley, USA, [‡]Harbin Institute of Technology, China
[♯]Didi Chuxing, China, [§]Cornell University, USA, [♢]Tsinghua University, China
schzhao@gmail.com, chuanglin.hit@outlook.com, xupengfeipf@didichuxing.com, zhaosendong@gmail.com,
yuchen.w.guo@gmail.com, ravi.krishna@berkeley.edu, dinggg@tsinghua.edu.cn, keutzer@berkeley.edu

Abstract

Deep neural networks excel at learning from large-scale labeled training data, but cannot well generalize the learned knowledge to new domains or datasets. Domain adaptation studies how to transfer models trained on one labeled source domain to another sparsely labeled or unlabeled target domain. In this paper, we investigate the unsupervised domain adaptation (UDA) problem in image emotion classification. Specifically, we develop a novel cycle-consistent adversarial model, termed CycleEmotionGAN, by enforcing emotional semantic consistency while adapting images cycle-consistently. By alternately optimizing the CycleGAN loss, the emotional semantic consistency loss, and the target classification loss, CycleEmotionGAN can adapt source domain images to have similar distributions to the target domain without using aligned image pairs. Simultaneously, the annotation information of the source images is preserved. Extensive experiments are conducted on the ArtPhoto and FI datasets, and the results demonstrate that CycleEmotionGAN significantly outperforms the state-of-the-art UDA approaches.



Classification accuracy (%) of the model tested on the target
Trained on the target: **66.81**, Trained on the source: **20.17**
Trained on the target after adaptation using CycleEmotionGAN: **25.20**

Figure 1: An example of *domain shift*. The overall accuracy of a state-of-the-art image emotion classification model (He et al. 2016) drops from 66.81% (trained on the target FI) to 20.17% (if trained only on the source ArtPhoto). We propose CycleEmotionGAN, a novel cycle-consistent adversarial model, to perform unsupervised domain adaptation. Our model achieves significant performance improvements over the source-trained model baselines.

Introduction

Psychological studies have revealed that visual content (*e.g.* images and videos) can evoke rich emotions for human viewers (Detenber, Simons, and Bennett Jr 1998). Nowadays, humans have become used to using images and videos appearing alongside text in social networks to record their activities, share their experiences, and express their opinions (Zhao et al. 2018b). Analyzing the implied emotions of this large volume of multimedia data can help us to understand humans' behaviors, which can benefit wide applications, such as blog recommendation (Borth et al. 2013).

Recognizing emotions induced by image content, referred to as image emotion recognition (IER) (Zhao et al. 2014a), is a non-trivial problem, because of two challenges: affective gap (Zhao et al. 2014a) and perception subjectivity (Peng et al. 2015; Zhao et al. 2016). Inspired by psychology and art theory, different hand-crafted features (e.g. color and texture (Machajdik and Hanbury 2010), shape (Lu et al. 2012), and principles of art (Zhao et al. 2014a)) have been designed to bridge the affective gap. These methods mainly classified the images into one dominant emotion category, or regressed the images with average dimension values. To tackle the subjectivity issue, we can predict personalized emotion perceptions for each viewer (Zhao et al. 2016), or learn the emotion distributions for each image (Yang, She, and Sun 2017; Zhao et al. 2017a; 2017b).

With the advent of deep neural networks, several end-to-end approaches have been proposed to classify image emotions (Rao, Xu, and Xu 2016; You et al. 2016; Zhu et al. 2017b; Yang et al. 2018a) or learn emotion distributions (Peng et al. 2015; Yang, She, and Sun 2017). Current IER methods, especially ones based on deep neural networks, perform well with large-scale labelled training data. However, due to *domain shift* or *dataset bias* (Torralba and Efros 2011), they cannot well generalize their learned knowledge to new domains or datasets, as shown in Figure 1. Even a slight departure from a network's training domain can cause it to make incorrect predictions and significantly reduce its performance (Tzeng et al. 2017). Domain adaptation (DA) is a machine learning paradigm that seeks to train a model on a source domain that can, in turn, perform well on a different, but related, target domain. Though DA has been widely studied in various computer vision tasks (Patel et al. 2015), it has rarely been applied to the IER problem.

In this paper, we study the unsupervised domain adaptation (UDA) problem of classifying image emotions in one source domain and adapting this to another target domain. A novel cycle-consistent adversarial model, termed CycleEmotionGAN, is developed for image emotion classification. Similar to Cycle-Consistent Generative Adversarial Networks (CycleGAN) (Zhu et al. 2017a), using an adversarial loss, a mapping $G_{ST} : \mathbf{X}_S \rightarrow \mathbf{X}_T$ is learned to adapt the source images $\mathbf{X}_S$ to the target images $\mathbf{X}_T$ so that the distribution of images from $G_{ST}(\mathbf{X}_S)$ is indistinguishable from the distribution $\mathbf{X}_T$. Because this mapping is highly under-constrained (Zhu et al. 2017a), an inverse mapping $G_{TS} : \mathbf{X}_T \rightarrow \mathbf{X}_S$ is coupled and a cycle-consistency loss is introduced to enforce $G_{TS}(G_{ST}(\mathbf{x}_S)) \approx \mathbf{x}_S$ (and vice versa). To preserve the annotation information of the source images, we complement the CycleGAN loss with an emotional semantic consistency (ESC) loss that penalizes large semantic changes between the adapted and source images. In this way, the CycleEmotionGAN model can adapt the source domain images to appear as if they were drawn from the target domain, while preserving the annotation information. Meanwhile, a classification network is trained to learn the mappings between image content and emotions. That is, we alternately optimize the CycleGAN loss, ESC loss, and classification loss. Extensive experimental results on the ArtPhoto (Machajdik and Hanbury 2010) and FI (You et al. 2016) datasets demonstrate the effectiveness of the proposed UDA method for classifying image emotions.

In summary, the contributions of this paper are threefold:

1. We propose to adapt image emotions from one source domain to a target domain in an unsupervised manner. To the best of our knowledge, this is the first domain adaptation work on image emotion classification.

2. We develop a novel cycle-consistent adversarial model, CycleEmotionGAN, for image emotion classification, which alternately optimizes the CycleGAN loss, the ESC loss, and the target classification loss. Thanks to the emotional semantic consistency loss, the adapted images are indistinguishable from the target images, while preserving the annotation information of the source images.

3. We conduct extensive experiments on the ArtPhoto and FI datasets, and the results demonstrate the superiority of the proposed CycleEmotionGAN model.

Related Work

Image Emotion Recognition: Two models are typically employed by psychologists to represent emotions: categorical emotion states (CES), and dimensional emotion space (DES). CES models consider emotions to be one of a few basic categories, while DES models usually employ a 3D or 2D Cartesian space to represent emotions. CES is straightforward for users to understand and label, while DES is more descriptive. In this paper, we classify images into one of Mikels's eight emotions (positive ones are *amusement*, *awe*, *contentment*, *excitement*, and negative ones are *anger*, *disgust*, *fear*, and *sadness*) (Mikels et al. 2005).

In the early years, researchers mainly hand-crafted features at different levels for IER, such as low-level ones like color, texture (Machajdik and Hanbury 2010), and shape (Lu et al. 2012); mid-level ones such as principles-of-art (Zhao et al. 2014a) and composition (Machajdik and Hanbury 2010); and finally high-level ones such as adjective noun pairs (Borth et al. 2013). Some work also fused different levels of features (Zhao et al. 2014b; 2017a; 2018c).

Recently, with the success of convolutional neural networks (CNNs) on many computer vision tasks, CNNs have also been employed in IER. Peng et al. (2015) fine-tuned a pre-trained CNN to predict emotion distributions. You et al. (2015) designed a progressive CNN architecture to make use of noisily labeled data for binary sentiment classification (You et al. 2016). Rao, Xu, and Xu (2016) learned multi-level deep representations (MldrNet), based on which Zhu et al. (2017b) integrated the different levels of features with a Bidirectional GRU model to exploit their dependencies. Yang et al. (2018b) employed deep metric learning to optimize both the retrieval and classification tasks by jointly optimizing cross-entropy loss and a novel sentiment constraint. Different from improving global image representations, several methods (You, Jin, and Luo 2017; Yang et al. 2018a) consider the local information for IER.

All the above methods employ a supervised manner to learn the mapping between image content and emotions. Please refer to (Zhao et al. 2018a) for a more comprehensive survey of IER. In this paper, we study how to adapt the models from the labeled source domain to the unlabeled target domain for classifying image emotions.

Unsupervised Domain Adaptation: In computer vision, unsupervised domain adaptation (UDA) is an open theoretical and practical problem (Patel et al. 2015). Our literature review here primarily focuses on CNN methods due to their empirical superiority for the problem. Please refer to (Patel et al. 2015) for reviewing the non-deep UDA methods. Typically, deep UDA methods employ a conjoined architecture with two streams to represent the models for the source and target domains, respectively (Zhuo et al. 2017). In addition to the traditional task loss based on the labeled source data, deep UDA models are usually trained jointly with another loss, such as a discrepancy loss, adversarial loss, or reconstruction loss, to deal with the *domain shift*.

Discrepancy-based methods explicitly measure the discrepancy between the source and target domains on corresponding activation layers of the two streams, including the multiple kernel variant of maximum mean discrepancies on the fully connected (FL) layers (Long et al. 2015), correlation alignment (CORAL) on the last FL layer (Sun, Feng, and Saenko 2017), and CORAL on the last FL layer and the last convolutional (conv) layer (Zhuo et al. 2017).

Adversarial generative models combine the domain discriminative model with a generative component generally based on GANs (Goodfellow et al. 2014). The Coupled Generative Adversarial Networks (CoGAN) (Liu and Tuzel 2016) can learn a joint distribution of multi-domain images with a tuple of GANs. By enforcing a self-regularization loss, Shrivastava et al. (2017) proposed SimGAN to improve the realism of a simulator's output using unlabeled real data.

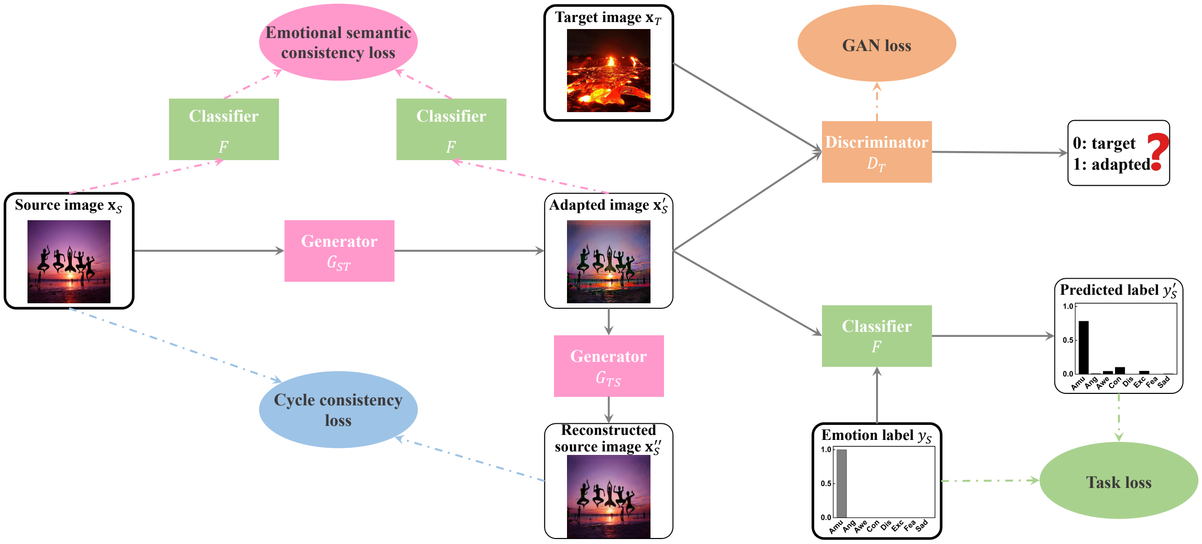


Figure 2: The framework of the proposed CycleEmotionGAN model for adapting image emotions from the source domain to the target domain. The black solid lines with arrows indicate the operations in the training stage. The dot dash lines with arrows correspond to different losses. For clarity the target cycle is omitted.

To penalize large low-level differences between the source and generated images for foreground pixels only, a masked Pairwise Mean Squared Error is minimized in Bousmalis et al. (2017). Hoffman et al. (2018) proposed to adapt representations at both the pixel-level and feature-level, while enforcing cycle-consistency and leveraging a task loss.

Adversarial discriminative models usually employ an adversarial objective with respect to a domain discriminator to encourage domain confusion. The domain-adversarial neural network (Ganin et al. 2016) optimizes the mapping to minimize the discriminator loss directly. Tzeng et al. (2017) proposed to use an inverted label GAN loss to split the optimization process into two independent objectives for generator and discriminator respectively.

Reconstruction based methods incorporate a reconstruction loss to minimize the difference between the input and the reconstructed input. Ghifary et al. (2015) designed a three-layer multi-task autoencoder with multiple output layers, each of which corresponds to one domain. Deep reconstruction classification networks (Ghifary et al. 2016) combine a traditional convolutional supervised network for source label prediction with a de-convolutional unsupervised network for target data reconstruction.

All the adapted targets of these methods are objective tasks, such as digit recognition, gaze estimation, and scene segmentation. Zhao et al. (2018d) adapted a subjective variable, image emotion, to learn discrete distributions. In this paper, we study the unsupervised domain adaptation problem in image emotion classification.

The CycleEmotionGAN Model

In this paper, we focus on one-source, homogeneous, and unsupervised domain adaptation, i.e. with only one source domain, labels from the source and target domains being observed in the same space, and the target domain being fully unlabeled. Suppose the source images and corresponding emotion labels drawn from the source distribution $P_S(\mathbf{x}, y)$ are $\mathbf{X}_S$ and $\mathbf{Y}_S$ respectively, and the target images drawn from the target distribution $P_T(\mathbf{x})$ are $\mathbf{X}_T$. Our goal is to learn a model that can correctly classify an image from the target domain into one of the L ($L = 8$ in this paper) emotion categories based on $\{\mathbf{X}_S, \mathbf{Y}_S\}$ and $\mathbf{X}_T$.

The main idea of CycleEmotionGAN is to learn a mapping $G_{ST} : \mathbf{X}_S \rightarrow \mathbf{X}_T$ to adapt the source images $\mathbf{X}_S$ to the target images $\mathbf{X}_T$. The requirement for G_{ST} is that the adapted images $\mathbf{X}'_S$ cannot be distinguished from the target images $\mathbf{X}_T$ by a discriminator D_T and that the emotion labels of $\mathbf{X}_S$ are preserved. Because the mapping G_{ST} is unstable and prone to failure (Zhu et al. 2017a), an inverse mapping $G_{TS} : \mathbf{X}_T \rightarrow \mathbf{X}_S$ is employed with a cycle-consistency loss to enforce $G_{TS}(G_{ST}(\mathbf{x}_S)) \approx \mathbf{x}_S$ (and vice versa). To preserve the emotion labels of the source images, we complement the CycleGAN loss with an emotional semantic consistency (ESC) loss which penalizes large semantic differences between the adapted and source images. In this way, the CycleEmotionGAN model can adapt the source domain images to be indistinguishable from the target domain, while preserving the annotation information. Finally, we can train a classifier F on the adapted dataset $\{\mathbf{X}'_S, \mathbf{Y}_S\}$ as if the training images $\mathbf{X}'_S$ and test images $\mathbf{X}_T$ were from the same distribution. The framework is shown in Figure 2.

CycleGAN Loss

CycleGAN aims to learn two mappings $G_{ST} : \mathbf{X}_S \rightarrow \mathbf{X}_T$ and $G_{TS} : \mathbf{X}_T \rightarrow \mathbf{X}_S$ between two domains S and T given training samples $\mathbf{X}_S$ and $\mathbf{X}_T$. Meanwhile, two discriminators D_T and D_S are trained, where D_T aims to distinguish between images $\mathbf{X}_T$ and $G_{ST}(\mathbf{X}_S)$, and D_S aims to distinguish between images $\mathbf{X}_S$ and $G_{TS}(\mathbf{X}_T)$. As in (Zhu et al. 2017a), the CycleGAN Loss contains two terms. One is the adversarial loss (Goodfellow et al. 2014) that matches the distribution of generated images to the data distribution in the target domain:

$$\begin{aligned}\mathcal{L}_{GAN}(G_{ST}, D_T) =& \mathbb{E}_{\mathbf{x}_S \sim P_S} \log D_T(G_{ST}(\mathbf{x}_S)) + \\ & \mathbb{E}_{\mathbf{x}_T \sim P_T} \log[1 - D_T(\mathbf{x}_T)],\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{GAN}(G_{TS}, D_S) =& \mathbb{E}_{\mathbf{x}_S \sim P_S} \log[1 - D_S(\mathbf{x}_S)] + \\ & \mathbb{E}_{\mathbf{x}_T \sim P_T} \log D_S(G_{TS}(\mathbf{x}_T)).\end{aligned} \tag{2}$$

The other is a cycle-consistency loss that ensures the learned mappings G_{ST} and G_{TS} are cycle-consistent, preventing them from contradicting each other. In this way, the image translation cycle is able to bring the reconstructed image back to the original image. That is $G_{TS}(G_{ST}(\mathbf{x}_S)) \approx \mathbf{x}_S$ and $G_{ST}(G_{TS}(\mathbf{x}_T)) \approx \mathbf{x}_T$. According to (Zhu et al. 2017a), the cycle-consistency loss is defined as:

$$\begin{aligned}\mathcal{L}_{cyc}(G_{ST}, G_{TS}) = & \mathbb{E}_{\mathbf{x}_S \sim P_S} \parallel G_{TS}(G_{ST}(\mathbf{x}_S)) - \mathbf{x}_S \parallel_1 \\ & + \mathbb{E}_{\mathbf{x}_T \sim P_T} \parallel G_{ST}(G_{TS}(\mathbf{x}_T)) - \mathbf{x}_T \parallel_1 .\end{aligned} \tag{3}$$

The objective of the CycleGAN loss is

$$\begin{aligned}\mathcal{L}_{CycleGAN}(G_{TS}, G_{TS}, D_T, D_S) = \mathcal{L}_{GAN}(G_{ST}, D_T) + \\ \mathcal{L}_{GAN}(G_{TS}, D_S) + \alpha \mathcal{L}_{cyc}(G_{TS}, G_{TS}),\end{aligned} \tag{4}$$

where α controls the relative importance of the GAN loss with respect to the cycle-consistency loss.

Emotional Semantic Consistency Loss

In addition to generating adapted images from source images, the generator G_{ST} should also preserve the emotion labels of the source images. This is an essential ingredient which enables training a classifier that uses the adapted images together with the emotion labels corresponding to the source images. For this purpose, we propose the use of an emotional semantic consistency loss to minimize the difference between the predicted emotions of source images and adapted images:

$$\mathcal{L}_{ESC}(G_{ST}) = \mathbb{E}_{\mathbf{x}_S \sim P_S} d(F(\mathbf{x}_S), F(G_{ST}(\mathbf{x}_S))), \tag{5}$$

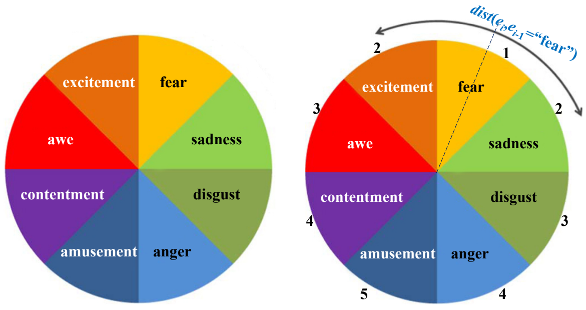
where $d(\cdot, \cdot)$ is a function that measures the distance between two emotion labels. Therefore, the augmented CycleGAN loss with the ESC loss is

$$\begin{aligned}\mathcal{L}_{aCycleGAN}(G_{TS}, G_{TS}, D_T, D_S) = \mathcal{L}_{GAN}(G_{ST}, D_T) + \\ \mathcal{L}_{GAN}(G_{TS}, D_S) + \alpha \mathcal{L}_{cyc}(G_{TS}, G_{TS}) + \lambda \mathcal{L}_{ESC}(G_{ST}),\end{aligned} \tag{6}$$

where λ controls the relative importance of the CycleGAN loss with respect to the ESC loss.

Here, we adopt two strategies for defining $d(\cdot, \cdot)$. First, since the output of network F is a probability distribution with each element representing the probability of corresponding emotion, we employ the symmetrized Kullback–Leibler divergence (SKL) to measure the distance between two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$:

$$\begin{aligned}SKL(\boldsymbol{p} \parallel \boldsymbol{q}) &= KL(\boldsymbol{p} || \boldsymbol{q}) + KL(\boldsymbol{q} || \boldsymbol{p}), \\ KL(\boldsymbol{p} || \boldsymbol{q}) &= \sum_{l=1}^{L} \big(p_l \ln p_l - p_l \ln q_l\big).\end{aligned} \tag{7}$$



(a) Mikels' wheel (b) Mikels' emotion distance

Figure 3: Mikels' emotion wheel and an example of Mikels' emotion distance for the emotion category *fear* (Zhao et al. 2016; 2018c).

Second, inspired by the research on emotion theory, we employ the Mikels' Wheel (Zhao et al. 2016) (which determines the relationship between two emotions) to measure the similarity between two emotions, from similar to complete opposites. Pairwise emotion distance is defined as 1+"the number of steps required to reach one emotion from another", as shown in Figure 3. Pairwise emotion similarity is defined as the reciprocal of pairwise emotion distance. $d(\cdot, \cdot)$ equals 1-pairwise emotion similarity.

Classification Loss

Generally, a separate task model is learned based on the adapted source images $\mathbf{X}'_S$ and the corresponding emotion labels $\mathbf{Y}_S$ after CycleGAN (Zhu et al. 2017a) to perform the final adaptation task. Contrary to this, the proposed CycleEmotionGAN is augmented with a classifier $F(\mathbf{x}'_S) \rightarrow y'_S$, which assigns emotion y'_S to the adapted image $\mathbf{x}'_S$. Similar to the CNN-based emotion classification method (You et al. 2016), the classifier F is optimized by minimizing the following cross-entropy loss:

$$\mathcal{L}_{task}(F) = \mathbb{E}_{(\mathbf{x}_S, y_S) \sim P_S} \sum_{l=1}^{L} \mathbb{1}_{[l = y_S]} \log(\sigma(F^{(l)}(G_{ST}(\mathbf{x}_S)))), \tag{8}$$

where σ is the softmax function, and $\mathbb{1}$ is an indicator function.

CycleEmotionGAN Learning

In our implementation, the generators G_{ST} and G_{TS} are convolutional neural networks with residual connections that maintain the resolution of the original image as illustrated in Figure 2. The discriminators D_T, D_S and the classifier F are also convolutional neural networks. The optimization of the proposed CycleEmotionGAN model is achieved by alternating between two stochastic gradient descent (SGD) steps. During the first step, we update D_T, D_S and F using SGD, while keeping G_{ST} and G_{TS} fixed. During the second step, we fix D_T, D_S and F, and update G_{ST} and G_{TS} using SGD. The detailed training procedure is shown in Algorithm 1, where $\boldsymbol{\theta}_{ST}$, $\boldsymbol{\theta}_{TS}$, $\boldsymbol{\phi}_S$, $\boldsymbol{\phi}_T$, and $\boldsymbol{\varphi}$ are the parameters of G_{ST}, G_{TS}, D_S, D_T, and F, respectively.

Algorithm 1: Adversarial training procedure of the proposed CycleEmotionGAN model

```
Input: Sets of source images $\mathbf{x}_S \in \mathbf{X}_S$ with emotion
       labels $y_S \in \mathbf{Y}_S$, and target images $\mathbf{x}_T \in \mathbf{X}_T$, the
       maximum number of steps $T$
Output: Predicted emotion label of target image $\mathbf{x}_T$
for $t \leftarrow 1$ to $T$ do
    Sample a mini-batch of source images $\mathbf{x}_S$, and
      target images $\mathbf{x}_T$;
    /* Updating $\boldsymbol{\theta}_{ST}$ and $\boldsymbol{\theta}_{TS}$ when fixing
         $\boldsymbol{\phi}_T$ , $\boldsymbol{\phi}_S$ and $\boldsymbol{\varphi}$                */
    Update $\boldsymbol{\theta}_{ST}$ and $\boldsymbol{\theta}_{TS}$ by taking an SGD step on
      mini-batch loss $\mathcal{L}_{aCycleGAN}$, $\mathcal{L}_{ESC}$ in Eq. (4),
      Eq. (5);
    /* Updating $\boldsymbol{\phi}_T$ , $\boldsymbol{\phi}_S$ when fixing
         $\boldsymbol{\theta}_{ST}$, $\boldsymbol{\theta}_{TS}$, and $\boldsymbol{\varphi}$                 */
    Compute $G_{ST}(\mathbf{x}_S; \boldsymbol{\theta}_{ST})$ with current $\boldsymbol{\theta}_{ST}$;
    Compute $G_{TS}(\mathbf{x}_T; \boldsymbol{\theta}_{TS})$ with current $\boldsymbol{\theta}_{TS}$;
    Update $\boldsymbol{\phi}_T$ by taking an SGD step on mini-batch
      loss $\mathcal{L}_{GAN}$ in Eq. (1);
    Update $\boldsymbol{\phi}_S$ by taking an SGD step on mini-batch
      loss $\mathcal{L}_{GAN}$ in Eq. (2);
    /* Updating $\boldsymbol{\varphi}$ when fixing $\boldsymbol{\phi}_T$ , $\boldsymbol{\phi}_S$,
         $\boldsymbol{\theta}_{ST}$, and $\boldsymbol{\theta}_{TS}$                  */
    Compute $G_{ST}(\mathbf{x}_S; \boldsymbol{\theta}_{ST})$ with current $\boldsymbol{\theta}_{ST}$;
    Update $\boldsymbol{\varphi}$ by taking an SGD step on mini-batch loss
      $\mathcal{L}_{task}(F)$ in Eq. (8);
end
return $F(\mathbf{x}_T; \boldsymbol{\varphi})$.
```

Experiments

In this section, we first introduce the detailed experimental setup, including the datasets, baselines, evaluation metrics, and implementation details. We then evaluate the performance of the proposed model and report and analyze the results as compared to the state-of-the-art approaches.

Datasets

The Artistic (**ArtPhoto**) dataset (Machajdik and Hanbury 2010) consists of 806 artistic photographs from a photo sharing site organized by emotion categories. The artists take the photos, upload them to the website, and determine the emotion categories of the photos. The artists try to evoke a certain emotion in the viewers through the photo with conscious manipulation of the emotional objects, lighting, colors, *etc*. In this dataset, each image is assigned to one of the eight Mikels' emotion categories.

The Flickr and Instagram (**FI**) dataset (You et al. 2016) is collected from 3 million weakly labeled web images in Flickr and Instagram by labeling with Mikels' eight emotion categories. A group of 225 Amazon Mechanical Turk (AMT) workers were employed to label the images. Each image is assigned to 5 AMT workers. In total, 23,308 images receiving at least three agreements between workers are included in the FI dataset.

Evaluation Metrics

Similar to (You et al. 2016; Yang et al. 2018a), we employ the classification accuracy (Acc) to evaluate our results. Acc is defined as the proportion of all predictions which are correct. $0 \leq Acc \leq 1$, with larger values representing better performances.

Baselines

To the best of our knowledge, CycleEmotionGAN is the first work on unsupervised domain adaptation for classifying image emotion. To demonstrate its effectiveness, we compare it to the following baselines: (1) source-only i.e. train on the source domain and test on the target domain directly; and (2) CycleGAN (Zhu et al. 2017a) i.e. first adapt the source images to the adapted ones cycle-consistently, and then train the classifier on the adapted source images with the emotion labels from corresponding source images. For comparison, we also report the results of an oracle setting, where the regressor is both trained and tested on the target domain.

Implementation Details

The generators G_{ST} and G_{TS} employ the CycleGAN architecture (Zhu et al. 2017a), which has shown impressive results for style domain transfer. This network contains two stride-2 convolutions, several residual blocks and two fractionally-strided convolutions with stride $\frac{1}{2}$. We use 9 blocks for 256 × 256 and higher-resolution training images. Similar to (Johnson, Alahi, and Fei-Fei 2016), we use instance normalization.

The discriminators D_T and D_S use 70 × 70 PatchGANs (Johnson, Alahi, and Fei-Fei 2016), which aim to classify whether 70 × 70 overlapping image patches are real or fake. Such a patch-level discriminator architecture has fewer parameters than a full-image discriminator, and can be applied to arbitrarily-sized images in a fully convolutional fashion.

The classifier F is based on the ResNet101 (He et al. 2016) architecture, which is initialized with the weights trained for ImageNet classification. The output of the last FL layer is changed to L, which can produce a probability distribution over the L emotion categories. The original loss layer is replaced with the cross-entropy loss from Eq. (8).

Similar to CycleGAN (Zhu et al. 2017a), we follow Shrivastava's strategy (Shrivastava et al. 2017) and update the discriminators using a history of generated images rather than the ones produced by the latest generative networks. We keep an image buffer that stores the 50 previously generated images. α in Eq. (4) is set to 10 in all experiments as in (Zhu et al. 2017a). λ in Eq. (6) is set to 10 and 5 for SKL and Midels' wheel definitions of $d(\cdot, \cdot)$, respectively. We also conduct an empirical analysis on the sensitivity of results to λ. We use the Adam solver with a batch size of 1. All generator and discriminator networks are trained from scratch with a learning rate of 0.0002, and the classifier F is trained with a learning rate of 0.0001.

Results and Analysis

The performance comparisons between the proposed CycleEmotionGAN model and the state-of-the-art approaches

Table 1: Classification accuracy (%) comparison between the proposed CycleEmotionGAN model and state-of-the-art approaches from the source ArtPhoto to the target FI. The best method trained on the source domain is emphasized in bold.

	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sadness	Average
source-only	2.06	6.50	9.51	4.16	**47.83**	**66.67**	18.18	37.37	20.17
CycleGAN (Zhu et al. 2017a)	29.63	**19.51**	21.97	7.37	36.02	17.02	**24.24**	**39.15**	22.68
CycleEmotionGAN - SKL	28.19	17.89	23.93	**9.26**	25.47	38.30	22.22	38.08	24.67
CycleEmotionGAN - Mikels	**35.39**	16.26	**33.44**	6.81	35.40	23.05	20.20	35.23	**25.20**
oracle (train on target)	63.69	59.57	40.65	84.16	70.32	34.34	64.77	64.60	66.81

Table 2: Classification accuracy (%) comparison between the proposed CycleEmotionGAN model and the state-of-the-art approaches from the source FI to the target ArtPhoto. The best method trained on the source domain is emphasized in bold.

	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sadness	Average
source-only	0.00	14.29	40.00	**57.14**	14.29	20.00	**45.45**	37.50	29.49
CycleGAN (Zhu et al. 2017a)	**20.00**	14.29	40.00	28.57	**42.86**	30.00	18.18	**50.00**	32.05
CycleEmotionGAN - SKL	**20.00**	**28.57**	**70.00**	28.57	**42.86**	20.00	36.36	43.75	**37.18**
CycleEmotionGAN - Mikels	**20.00**	**28.57**	60.00	42.86	28.57	**40.00**	27.27	43.75	**37.18**
oracle (train on target)	30.00	28.57	40.00	14.29	71.43	60.00	63.64	43.75	44.87

as measured by classification accuracy are shown in Table 1 (from the source ArtPhoto to the target FI) and Table 2 (from the source FI to the target ArtPhoto).

From the results, we have the following observations. **(1)** The source-only method *i.e.* directly transferring the models trained on the source domain to the target domain performs the worst in both adaptation settings. Due to the influence of *domain shift* or *dataset bias*, the joint probability distributions of observed images and emotion labels greatly differ in the two domains. This results in the model's low transferability from the source domain to the target domain. **(2)** Both adaptation methods, CycleGAN and CycleEmotionGAN, outperform the source-only methods, with CycleEmotionGAN performing better. This demonstrates the effectiveness of CycleEmotionGAN for unsupervised domain adaptation in classifying image emotions. Specifically, the relative performance improvements of CycleEmotionGAN over source-only and CycleGAN measured by classification accuracy are 24.94% and 11.11% from the source ArtPhoto to the target FI, and 26.08% and 16.01% from the source FI to the target ArtPhoto, respectively. These results demonstrate that the proposed CycleEmotionGAN model can achieve superior performance relative to state-of-the-art approaches. The performance improvements benefit from the alternate exploration of CycleGAN loss, emotional semantic consistency loss, and classification loss in CycleEmotionGAN. **(3)** Both SKL and Mikels' wheel distance measures work in the CycleEmotioGAN model. Mikels' wheel performs marginally better than SKL from the source ArtPhoto to the target FI. **(4)** For the relatively similar emotion categories, such as *amusement*, *contentment*, and *excitement*, the source-only model is extremely unbalanced, misclassifying *amusement* as *excitement* when going from ArtPhoto to FI. The CycleGAN and CycleEmotionGAN models can better distinguish these emotion categories. **(5)** For the emotion categories such as *amusement*, the images significantly differ between ArtPhoto and FI. When testing on ArtPhoto using the models directly trained on FI, the classification accuracy is 0. After image style transfer using CycleGAN, the performance is significantly improved. **(6)** The oracle method, i.e. testing on the target domain using the model trained on the same domain, achieves the best performance. However, this model is trained using the ground truth emotion labels from the target domain, which are of course unavailable in unsupervised domain adaptation. **(7)** There is still an obvious performance gap between all adaptation methods and the oracle method, especially when adapting from the small-scale ArtPhoto to the large-scale FI. Due to the complexity and subjectivity of emotions (Yang et al. 2018b), effectively adapting image emotions is still a challenging problem.

We visualize the results of image-space adaptation between ArtPhoto and FI in Figure 4. We can see that both CycleGAN and CycleEmotionGAN can adapt the source images to be more visually similar to the target ones, with CycleEmotionGAN performing better. For example, the hue of the adapted image (c) in the last column by CycleEmotionGAN is more similar to that of the target one than the hue in the original source image and the hue in the image adapted by CycleGAN. This further demonstrates the effectiveness of the proposed CycleEmotionGAN model.

Parameter Sensitivity. We investigate the impact of the hyperparameter λ in Eq. (6) on performance. Figure 5 and Figure 6 give an illustration of the variation of emotion classification performance. We can observe that generally the performance first increases and then decreases as λ varies. This confirms the validity of alternately optimizing the ESC loss and CycleGAN loss, since a good trade-off between the two can enhance the transferability.

Conclusion

In this paper, we make significant progress toward solving the unsupervised domain adaptation (UDA) problem of classifying image emotions. A novel cycle-consistent adversarial model, termed CycleEmotionGAN, is developed by complementing the CycleGAN loss with an emotional semantic consistency (ESC) loss. Two implementations for the ESC loss, the symmetrized KL divergence and Mikels' Wheel distance, are employed to preserve the emotion labels of the

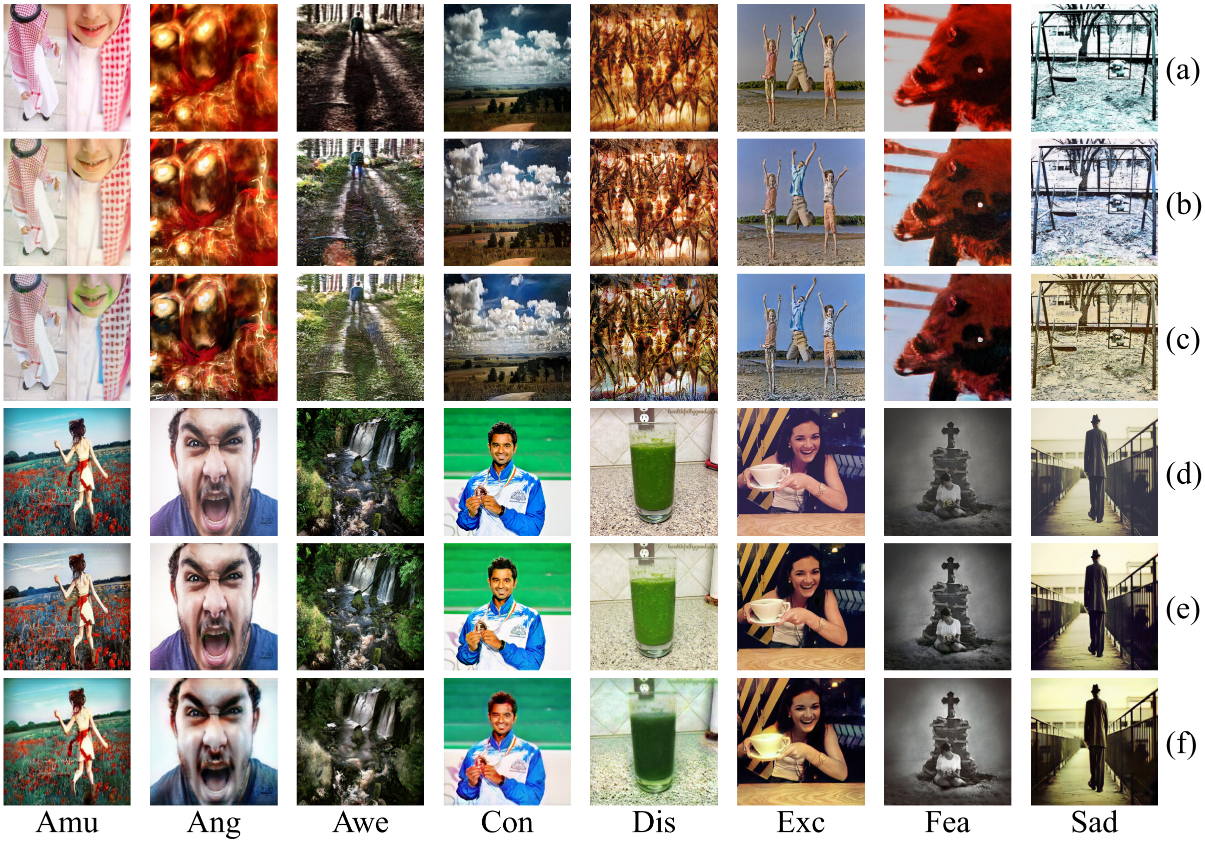
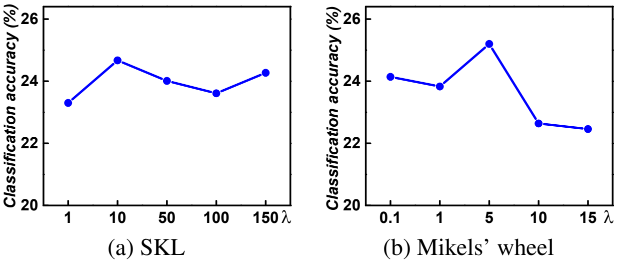


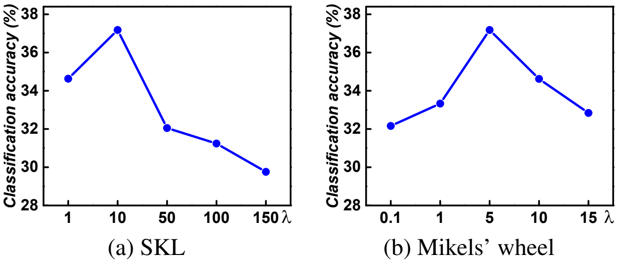
Figure 4: Visualization of the adapted results between ArtPhoto and FI. Example images from the ArtPhoto (a) and FI (d) datasets, alongside their image-space adaptations to the opposite domain by CycleGAN (b) (e) and CyleEmotionGAN-Mikels (c) (f), respectively. The adapted images look more visually similar to the target images than the source images.



(a) SKL

(b) Mikels' wheel

Figure 5: Sensitivity of the proposed CycleEmotionGAN model to λ in Eq. (6) when going from ArtPhoto to FI.



(a) SKL

(b) Mikels' wheel

Figure 6: Sensitivity of the proposed CycleEmotionGAN model to λ in Eq. (6) when going from FI to ArtPhoto.

source images. The alternate optimization of the CycleGAN loss, ESC loss, and classification loss enables CycleEmotionGAN to adapt the source domain images to have similar distributions to those of the target domain images. The extensive experiments conducted on the ArtPhoto and FI datasets demonstrate that CycleEmotionGAN significantly outperforms the state-of-the-art UDA approaches for image emotion classification.

For further studies, we plan to extend the CycleEmotionGAN model to other image emotion recognition (IER) tasks, such as emotion distribution learning. We also aim to investigate methods that adapt well when the source domain and the target domain employ different emotion models and when there is more than one source domain.

Acknowledgments

This work was supported by Berkeley DeepDrive, the National Natural Science Foundation of China (Nos. 61701273, 61571269), the Project Funded by China Postdoctoral Science Foundation (Nos. 2018T110100, 2017M610897), the National Key R&D Program of China (No. 2017YFC011300), and National Postdoctoral Program for Innovative Talents (No. BX20180172). Corresponding authors: Chuang Lin, Pengfei Xu.

References

Borth, D.; Ji, R.; Chen, T.; Breuel, T.; and Chang, S.-F. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 223–232.

Bousmalis, K.; Silberman, N.; Dohan, D.; Erhan, D.; and Krishnan, D. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 3722–3731.

Detenber, B. H.; Simons, R. F.; and Bennett Jr, G. G. 1998. Roll 'em!: The effects of picture motion on emotional responses. *JBEM* 42(1):113–127.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR* 17(1):2096–2030.

Ghifary, M.; Bastiaan Kleijn, W.; Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, 2551–2559.

Ghifary, M.; Kleijn, W. B.; Zhang, M.; Balduzzi, D.; and Li, W. 2016. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, 597–613.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A. A.; and Darrell, T. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 1994–2003.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 694–711.

Liu, M.-Y., and Tuzel, O. 2016. Coupled generative adversarial networks. In *NIPS*, 469–477.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning transferable features with deep adaptation networks. In *ICML*, 97–105.

Lu, X.; Suryanarayan, P.; Adams Jr, R. B.; Li, J.; Newman, M. G.; and Wang, J. Z. 2012. On shape and the computability of emotions. In *ACM MM*, 229–238.

Machajdik, J., and Hanbury, A. 2010. Affective image classification using features inspired by psychology and art theory. In *ACM MM*, 83–92.

Mikels, J. A.; Fredrickson, B. L.; Larkin, G. R.; Lindberg, C. M.; Maglio, S. J.; and Reuter-Lorenz, P. A. 2005. Emotional category data on images from the international affective picture system. *BRM* 37(4):626–630.

Patel, V. M.; Gopalan, R.; Li, R.; and Chellappa, R. 2015. Visual domain adaptation: A survey of recent advances. *IEEE SPM* 32(3):53–69.

Peng, K.-C.; Sadovnik, A.; Gallagher, A.; and Chen, T. 2015. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *CVPR*, 860–868.

Rao, T.; Xu, M.; and Xu, D. 2016. Learning multi-level deep representations for image emotion classification. *arXiv:1611.07145*.

Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; and Webb, R. 2017. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2242–2251.

Sun, B.; Feng, J.; and Saenko, K. 2017. Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*. 153–171.

Torralba, A., and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR*, 1521–1528.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*, 2962–2971.

Yang, J.; She, D.; Lai, Y.-K.; Rosin, P. L.; and Yang, M.-H. 2018a. Weakly supervised coupled networks for visual sentiment analysis. In *CVPR*, 7584–7592.

Yang, J.; She, D.; Lai, Y.; and Yang, M.-H. 2018b. Retrieving and classifying affective images via deep metric learning. In *AAAI*.

Yang, J.; She, D.; and Sun, M. 2017. Joint image emotion classification and distribution learning via deep convolutional neural network. In *IJCAI*, 3266–3272.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 381–388.

You, Q.; Luo, J.; Jin, H.; and Yang, J. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *AAAI*, 308–314.

You, Q.; Jin, H.; and Luo, J. 2017. Visual sentiment analysis by attending on local image regions. In *AAAI*, 231–237.

Zhao, S.; Gao, Y.; Jiang, X.; Yao, H.; Chua, T.-S.; and Sun, X. 2014a. Exploring principles-of-art features for image emotion recognition. In *ACM MM*, 47–56.

Zhao, S.; Yao, H.; Yang, Y.; and Zhang, Y. 2014b. Affective image retrieval via multi-graph learning. In *ACM MM*, 1025–1028.

Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; Xie, W.; Jiang, X.; and Chua, T.-S. 2016. Predicting personalized emotion perceptions of social images. In *ACM MM*, 1385–1394.

Zhao, S.; Ding, G.; Gao, Y.; and Han, J. 2017a. Approximating discrete probability distribution of image emotions by multi-modal features fusion. In *IJCAI*, 4669–4675.

Zhao, S.; Yao, H.; Gao, Y.; Ji, R.; and Ding, G. 2017b. Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. *IEEE TMM* 19(3):632–645.

Zhao, S.; Ding, G.; Huang, Q.; Chua, T.-S.; Schuller, B. W.; and Keutzer, K. 2018a. Affective image content analysis: A comprehensive survey. In *IJCAI*, 5534–5541.

Zhao, S.; Gao, Y.; Ding, G.; and Chua, T.-S. 2018b. Real-time multimedia social event detection in microblog. *IEEE TCYB* 48(11):3218–3231.

Zhao, S.; Yao, H.; Gao, Y.; Ding, G.; and Chua, T.-S. 2018c. Predicting personalized image emotion perceptions in social networks. *IEEE TAFFC* 9(4):526–540.

Zhao, S.; Zhao, X.; Ding, G.; and Keutzer, K. 2018d. Emotiongan: Unsupervised domain adaptation for learning discrete probability distributions of image emotions. In *ACM MM*, 1319–1327.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2242–2251.

Zhu, X.; Li, L.; Zhang, W.; Rao, T.; Xu, M.; Huang, Q.; and Xu, D. 2017b. Dependency exploitation: a unified cnn-rnn approach for visual emotion recognition. In *IJCAI*, 3595–3601.

Zhuo, J.; Wang, S.; Zhang, W.; and Huang, Q. 2017. Deep unsupervised convolutional domain adaptation. In *ACM MM*, 261–269.