Emotional Semantics-Preserved and Feature-Aligned CycleGAN for Visual Emotion Adaptation

Sicheng Zhao, *Senior Member, IEEE*, Xuanbai Chen, Xiangyu Yue, Chuang Lin, Pengfei Xu, *Member, IEEE*, Ravi Krishna, Jufeng Yang, *Member, IEEE*, Guiguang Ding, Alberto L. Sangiovanni-Vincentelli, *Life Fellow, IEEE*, and Kurt Keutzer, *Life Fellow, IEEE*

***Abstract*—Thanks to large-scale labeled training data, deep neural networks (DNNs) have obtained remarkable success in many vision and multimedia tasks. However, because of the presence of domain shift, the learned knowledge of the well-trained DNNs cannot be well generalized to new domains or datasets that have few labels. Unsupervised domain adaptation (UDA) studies the problem of transferring models trained on one labeled source domain to another unlabeled target domain. In this article, we focus on UDA in visual emotion analysis for both emotion distribution learning and dominant emotion classification. Specifically, we design a novel end-to-end cycle-consistent adversarial model, called CycleEmotionGAN++. First, we generate an adapted domain to align the source and target domains on the pixel level by improving CycleGAN with a multiscale structured cycle-consistency loss. During the image translation, we propose a dynamic emotional semantic consistency loss to preserve the emotion labels of the source images. Second, we train a transferable task classifier on the adapted domain with feature-level alignment between the adapted and target domains. We conduct extensive UDA experiments on the Flickr-LDL and Twitter-LDL datasets for distribution learning and ArtPhoto and Flickr and Instagram datasets for emotion classification. The results demonstrate the significant improvements yielded by the proposed CycleEmotionGAN++ compared to state-of-the-art UDA approaches.**

***Index Terms*—Unsupervised domain adaptation (UDA), dynamic emotional semantic consistency (DESC), emotion distribution, generative adversarial networks (GANs), visual emotion analysis (VEA).**

I. Introduction

IT HAS been revealed that visual content, such as images and videos, can evoke strong emotions for human beings [1]. With the popularization of various mobile devices, cameras, and the Internet, people have become accustomed to recording their activities, sharing their experiences, and expressing their opinions by using images and videos appearing alongside text in social networks [2]. The generation of a large amount of multimedia data has made it convenient for researchers to process and analyze visual content. Understanding the implied emotions in the data is of great importance to behavioral sciences and enables various applications, including blog recommendation, decision making, and appreciation of art [3].

Recognizing the emotions induced by image content is often referred to as visual emotion analysis (VEA) [5]. This task mainly faces two challenges: 1) the affective gap [6] and 2) perception subjectivity [7], [8]. The former one reveals that the extracted feature-level information is inconsistent with the high-level emotions felt by human beings; while the latter indicates that due to different personal and contextual factors, such as education background, culture, and personality, different people may produce different emotions after viewing the same image [9], [10]. In order to bridge the affective gap, a variety of handcrafted features has been designed, such as color and texture [11], shape [12], principles-of-art [6], and adjective–noun pairs [3]. These methods mainly map the image content to one dominant emotion category (DEC). To deal with the subjectivity issue, either personalized emotion perception is predicted for each viewer [8] or an emotion distribution is learned for each image [13]–[15].

Recently, convolutional neural networks (CNNs) have been employed to deal with the issue of mapping the image content to emotions [5], [7], [13], [16], [17]. These CNN-based VEA methods can perform well on large-scale training datasets with labels. However, due to the presence of a *domain shift* or *dataset bias* [18], the performance of a model directly transferred from one labeled domain to another unlabeled domain drops significantly [19], as shown in Figs. 1 and 2.

Manuscript received 4 August 2020; revised 24 November 2020 and 8 February 2021; accepted 20 February 2021. Date of publication 24 March 2021; date of current version 16 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61701273, Grant U1936202, Grant 61876094, and Grant U1933114; in part by the Berkeley DeepDrive; and in part by the Natural Science Foundation of Tianjin, China, under Grant 20JCJQJC00020, Grant 18JCYBJC15400, and Grant 18ZXZNGX00110. This article was recommended by Associate Editor W. Hu. *(Sicheng Zhao and Xuanbai Chen contributed equally to this work.) (Corresponding author: Sicheng Zhao.)*

Sicheng Zhao and Guiguang Ding are with BNRist, Tsinghua University, Beijing 100084, China (e-mail: schzhao@gmail.com; dinggg@tsinghua.edu.cn).

Xuanbai Chen and Jufeng Yang are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: chenxuanbai@126.com; yangjufeng@nankai.edu.cn).

Xiangyu Yue, Ravi Krishna, Alberto L. Sangiovanni-Vincentelli, and Kurt Keutzer are with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94709 USA (e-mail: xyyue@berkeley.edu; ravi.krishna@berkeley.edu; alberto@berkeley.edu; keutzer@berkeley.edu).

Chuang Lin and Pengfei Xu are with Didi Chuxing, Beijing 100193, China (e-mail: chuanglin.hit@outlook.com; xupengfeipf@didiglobal.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3062750.

Digital Object Identifier 10.1109/TCYB.2021.3062750

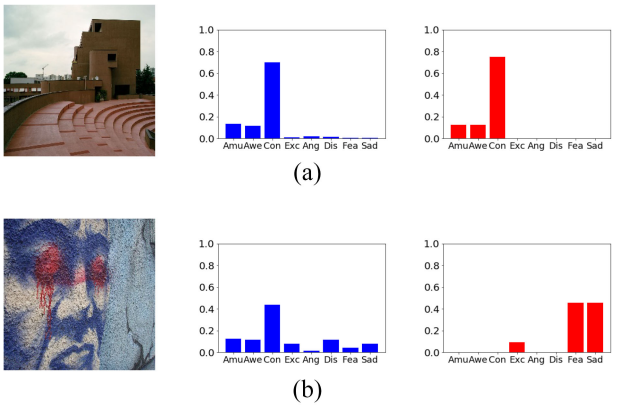

Fig. 1. Example of domain shift when performing the emotion distribution learning task. The classifier trained on Twitter-LDL is tested on the top image from Twitter-LDL and the bottom image from Flickr-LDL. The objects displayed, from left to right, are: the original image, the predicted emotion distribution, and the ground truth distribution. (a) Train on Twitter-LDL and test on Twitter-LDL. (b) Train on Twitter-LDL and test on Flickr-LDL.

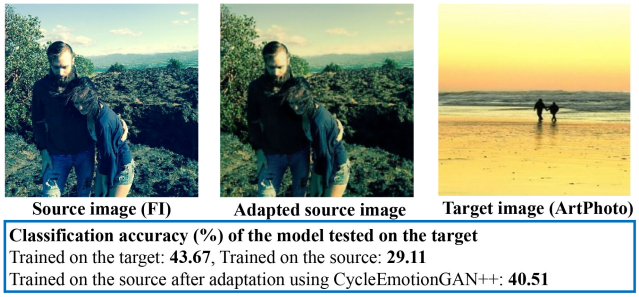

Fig. 2. Example of domain shift when performing dominant emotion classification. The overall accuracy of a state-of-the-art model (He *et al.* [4]) drops from 43.67% (trained on the target ArtPhoto) to 29.11% (trained only on the source FI). We propose CycleEmotionGAN++, which achieves significant performance improvements (11.40%) over the source-trained model baselines.

Domain adaptation (DA) is a machine-learning paradigm that tries to train a model on a source domain that can perform well on a different, but related, target domain. To the best of our knowledge, although DA has been used in various vision tasks [19], [20], it has rarely been used for VEA.

In this article, we study the unsupervised DA (UDA) problem of analyzing visual emotions in one labeled source domain and adapting it to another unlabeled target domain. A novel cycle-consistent adversarial UDA model, called CycleEmotionGAN++, is proposed for dominant emotion classification and emotion distribution learning. First, we generate an intermediate domain to align the source and target domains on the pixel level based on the generative adversarial network (GAN) [21]. Since this mapping from the source domain to the intermediate domain is highly underconstrained [22], we couple an inverse mapping and a cycle-consistency loss to enforce the reconstructed source to be as similar as possible to the original source. We add a multiscale structural similarity loss to the original cycle-consistency loss to better preserve the high-frequency-detailed information, and this combination is defined as the mixed cycle-consistency loss. In addition, we complement the mixed CycleGAN loss with a dynamic emotional semantic consistency (DESC) loss that penalizes large semantic changes between the adapted and source images. Two different classifiers are trained on the source domain and adapted domain, respectively, to dynamically preserve the semantic information. In order to make the adapted domain and target domain as similar as possible, we also add feature-level alignment by training a discriminator to maximize the probability of correctly classifying feature maps from adapted images and target images. In this way, the CycleEmotionGAN++ model can adapt the source-domain images to appear as if they were drawn from the target domain. Eventually, by optimizing the mixed CycleGAN loss, DESC loss, feature-level alignment loss, and task classification loss alternately, a transferable CycleEmotionGAN++ model is learned.

In summary, the contributions of this article are three-fold.

1) We propose to adapt visual emotions from one source domain to another target domain by using a novel end-to-end cycle-consistent adversarial model. To the best of our knowledge, this is the first work on UDA for both emotion distribution learning and dominant emotion classification tasks.
2) We develop a novel adversarial model, called CycleEmotionGAN++, by alternately optimizing the mixed CycleGAN loss, DESC loss, feature-level alignment loss, and task classification loss. The adapted images are indistinguishable from the target images, thanks to the mixed CycleGAN loss that can preserve the contrast and detailed information better by adding the multiscale structural similarity, the DESC loss that can preserve the annotation information of the source images, and the feature-level alignment loss that can align adapted and target images on the feature level.
3) We conduct extensive experiments on four datasets: a) Twitter-LDL and b) Flickr-LDL for emotion distribution learning and c) ArtPhoto and Flickr and d) Instagram (FI) for dominant emotion classification. The results demonstrate the significant improvements yielded by CycleEmotionGAN++.

CycleEmotionGAN++ is extended from CycleEmotionGAN, which was previously introduced in our AAAI 2019 paper [30]. The improvements include the following three aspects. First, the image translation is conducted with mixed CycleGAN by enforcing the multiscale structural similarity and with DESC; feature-level alignment is added to better align the source and target domains. Second, we conduct more UDA experiments for both emotion distribution learning and dominant emotion classification. Third, we provide a more comprehensive review to introduce the background and comparison.

II. Related Work

Emotion Representation: Two models are typically employed by psychologists to represent emotions: 1) categorical emotion states (CES) and 2) dimensional emotion space (DES) [10]. CES models usually consider classifying

emotions into several basic categories, such as Mikels' eight emotions (*amusement*, *anger*, *awe*, *contentment*, *disgust*, *excitement*, *fear*, and *sadness*). DES models usually employ a Cartesian space to represent emotions, such as the 3-D valence-arousal-dominance (VAD) space [31]. CES is intuitive for human beings to understand in labeling emotions for images, while DES is more abstract and fine grained. In this article, the classic Mikels' eight emotions are employed as our emotion model.

Visual Emotion Analysis: Similar to other machine learning and computer vision tasks, VEA also involves feature extraction and classifier training [10], [32]. While classifier training is mainly based on the existing machine-learning algorithms, the main focus in VEA is extracting discriminative features. In the early stage, researchers mainly handcrafted features on different levels [33], including low-level features, such as color and texture [11], shape [12]; mid-level features, such as principles-of-art [6], composition [11], and attributes [34]; and high-level features, such as skins [11] and adjective–noun pairs [3], [35]. Some other methods try to combine various features on different levels [14], [36]. A learning-based visual affective filtering framework has been proposed to synthesize user-specified emotions onto arbitrary input images or videos [37].

In recent years, CNNs have had great success on many machine-learning tasks including VEA. You *et al.* [38] built a large dataset for VEA and designed a progressive CNN architecture to make use of noisily labeled data for sentiment polarity classification. Various methods have been proposed to predict the probability distributions of image emotions [9], [13]–[15], [29], [39]. In order to predict emotion distributions more rapidly and accurately, some methods [7], [29] fine tune CNN models pretrained on ImageNet. You *et al.* [40] and Zhou *et al.* [30] also fine tuned a pretrained CNN to classify visual emotions on a new large-scale FI dataset [38], respectively. Yang *et al.* [17] proposed weakly supervised coupled networks (WSCNet) with two branches: 1) sentiment map detection and 2) coupled sentiment classification to improve the classification accuracy. Local information is also considered in [5] and [41]. Rao *et al.* [42] learned multilevel deep representations (MldrNet), including aesthetics CNN, AlexNet, and texture CNN. Yang *et al.* [43] optimized both retrieval and classification losses by using the sentiment constraints adapted from the triplet constraints, which is able to explore the hierarchical relation of emotion labels.

The above mapping methods between image content and emotions are all performed in a supervised manner. Refer to [10] for a comprehensive survey on VEA. In this article, we study how to adapt the models trained from one labeled source domain to another unlabeled target domain for VEA.

Unsupervised Domain Adaptation: In the early years, DA was introduced in a transform-based adaptation technique for object recognition [44]. Torralba and Efros [18] conducted a comparison study using a set of popular datasets and conducted a deep discussion regarding dataset bias. For UDA, it has been explored in [20] for the image classification task with extensive reviews of some nondeep approaches. These methods mainly focused on feature space alignment through minimizing the distance between the source domain and target domain, either by sample reweighting techniques [45], [46] or by constructing intermediate subspace representations [47], [48].

Recent efforts have shifted to employing deep models. In order to represent the source and target domains, Zhou *et al.* [49] proposed a conjoined architecture with two streams for UDA. Labeled source data are used for the supervised task loss and deep UDA models are usually trained jointly with another loss, such as a discrepancy loss, adversarial loss, or self-supervision loss, to deal with domain shift.

Discrepancy-based methods mainly measure the discrepancy directly between the source and target domains on corresponding activation layers, such as the multiple kernel variant of maximum mean discrepancies on the fully connected (FC) layers [50], correlation alignment (CORAL) [51], and geodesic CORAL [52] on the last FC layer, CORAL on both the last conv layer and FC layer [49], and contrastive domain discrepancy on multiple FC layers [53]. Adversarial discriminative models usually employ an adversarial objective with respect to a domain discriminator to encourage domain confusion. Representation discriminators include the feature discriminator [26], [54], output discriminator [26], [55], conditional discriminator [56], joint discriminator [57], prototypical discriminator [58], and gradient reversal layers [59]. Adversarial generative models combine the domain discriminative model with a generative component based on GAN [21] and its invariants, such as coupled GANs (CoGAN) [60], SimGAN [27], and CycleGAN [22], [28], [61]. Self-supervision-based methods incorporate auxiliary self-supervised learning tasks into the original task network to bring the source and target domains closer. The commonly employed self-supervision visual tasks include reconstruction [62]–[64], image rotation prediction [65], [66], and jigsaw prediction [67].

All these methods focus on objective tasks (i.e., with objective labels), such as digit recognition, gaze estimation, object classification, and scene segmentation. Zhao *et al.* [29] adapted a subjective variable image emotion to learn discrete distributions. Later, Zhao *et al.* [30] studied the UDA problem in image emotion classification. In this article, we study the UDA problem in both image emotion classification and emotion distribution learning tasks. The comparison between the proposed CycleEmotionGAN++ and the existing UDA methods is summarized in Table I, from which we can see the advantages of CycleEmotionGAN++ relative to other approaches.

Image Style Transfer: Image style transfer that aims to transfer visual appearance between images is closely related to DA and has achieved remarkable success recently [22]–[25], [68]–[71]. Reinhard *et al.* [68] proposed a method by using the $l\alpha\beta$ space to minimize correlation between channels to simplify the transfer process. Hwang *et al.* [70] proposed a scattered point interpolation scheme using moving least squares to deal with misalignments. Rabin *et al.* [71] proposed an image color transfer method based on the relaxed discrete optimal transport techniques. Zhu *et al.* [22] proposed CycleGAN by adding a constraint generator and corresponding loss to assure the transfer learning

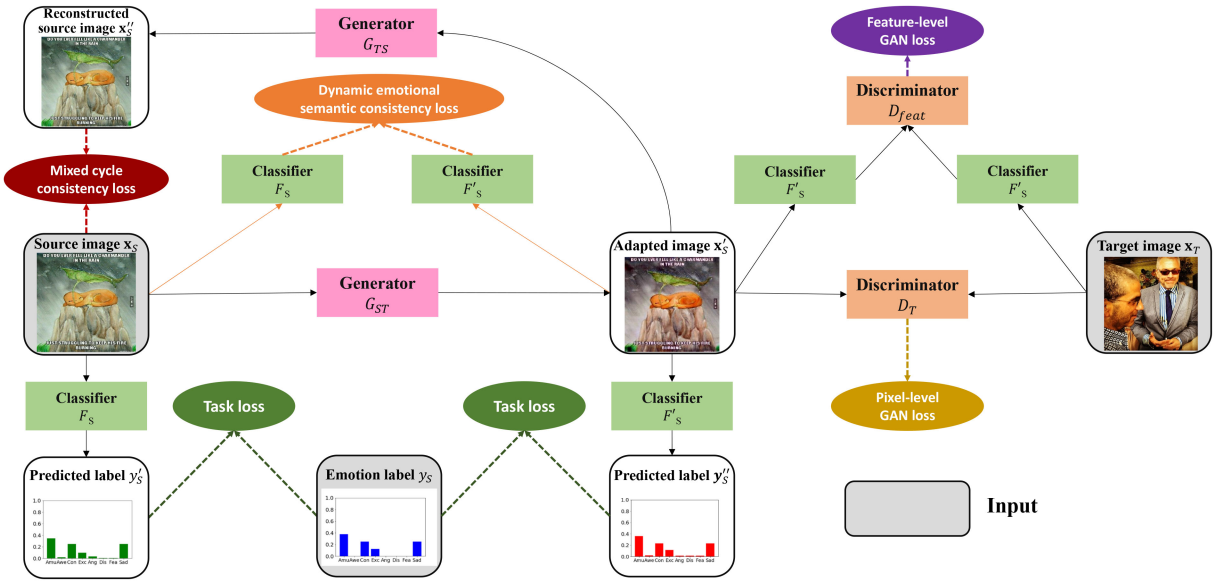

Fig. 3. Framework of the proposed CycleEmotionGAN++ model for visual emotion adaptation from one labeled source domain to another unlabeled target domain. The black solid lines with arrows indicate the operations in the training stage. The dashed lines with arrows correspond to different losses. For clarity, the target cycle is omitted.

TABLE I
COMPARISON OF THE PROPOSED CYCLEEMOTIONGAN++ MODEL WITH SEVERAL STATE-OF-THE-ART DA METHODS. THE FULL NAMES OF EACH ATTRIBUTE FROM THE SECOND TO THE LAST COLUMN ARE PIXEL-LEVEL ALIGNMENT, FEATURE-LEVEL ALIGNMENT, ESC, CYCLE CONSISTENCY, MULTISCALE STRUCTURAL SIMILARITY, EMOTION DISTRIBUTION LEARNING TASK, AND DOMINANT EMOTION CLASSIFICATION TASK, RESPECTIVELY

Method	pixel	feature	semantic	cycle	msssim	distrib	classifi
CycleGAN [22]	✓	✗	✗	✓	✗	✗	✗
SAPE [23]	✓	✓	✗	✗	✗	✗	✗
EICT [24]	✓	✓	✗	✗	✗	✗	✗
TAECT [25]	✗	✓	✗	✗	✗	✗	✗
ADDA [26]	✗	✓	static	✗	✗	✗	✗
SimGAN [27]	✓	✗	✗	✗	✗	✗	✗
CyCADA [28]	✓	✓	static	✓	✗	✗	✗
EmotionGAN [29]	✓	✗	static	✗	✗	✓	✗
CycleEmotionGAN (Ours)	✓	✗	static	✓	✗	✗	✓
CycleEmotionGAN++ (ours)	✓	✓	dynamic	✓	✓	✓	✓

process. Yan *et al.* [23] introduced an image descriptor to achieve semantics-aware photo enhancement (SAPE). Some methods specifically designed for emotion color transfer have emerged. Liu *et al.* [24] investigated emotional image color transfer (EICT) in a network with four modules to make the enhancement results meet the user's emotions. Liu and Pei [25] studied texture-aware emotional color transfer (TAECT) to adjust an image to a reference one by considering the texture information. These image style transfer methods might not perform well for DA since they do not explicitly align the distributions between different domains.

III. PROPOSED CYCLEEMOTIONGAN++ MODEL

In this article, we focus on UDA for VEA from one source domain with emotion labels to another target domain without any labels. Suppose the source images and corresponding emotion labels drawn from the source-domain distribution $P_S(\mathbf{x}, \mathbf{y})$ are $\mathbf{x}_S$ and $\mathbf{y}_S$, respectively, and target images drawn from the target-domain distribution $P_T(\mathbf{x})$ are $\mathbf{x}_T$. Our objective is to train a model that can map an image from the target domain to L (8 in our setting) classes of emotion categories.

The framework of the proposed CycleEmotionGAN++ model is shown in Fig. 3. The main idea is to train a mapping network $G_{ST} : \mathbf{x}_S \rightarrow \mathbf{x}_T$, which is used to generate adapted images $\mathbf{x}'_S$ from source images, with the requirement that the adapted images $\mathbf{x}'_S$ are indistinguishable from the target images $\mathbf{x}_T$ by the discriminator D_T. Because the generator mapping G_{ST} is underconstrained and unstable [22], we impose some constraints. That is, an inverse mapping G_{TS} is employed to reconstruct the source images from the adapted images. A cycle-consistency loss is used to enforce that the reconstructed images $\mathbf{x}''_S$ and the source images $\mathbf{x}_S$ are as close as possible. In order to overcome the drawbacks of a traditional cycle-consistency loss, the multiscale structural similarity is added to the loss. There is a similar cycle from the target to the source. In order to make the adapted images and target images similar, we should ensure that they are similar not only on a pixel level but also on a feature level. So, we train another discriminator D_{feat} to perform a feature-level alignment. To preserve the emotion labels of the source images, we propose DESC loss with two different classifiers to penalize 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 train the task classifier F'_S on the adapted dataset $\{\mathbf{x}'_S, \mathbf{y}_S\}$ by considering that the adapted images $\mathbf{x}'_S$ and target images $\mathbf{x}_T$ are from the same distribution.

A. Mixed CycleGAN Loss

CycleGAN [22] 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 tries to maximize the probability of correctly classifying target images $\mathbf{x}_T$ and adapted images $\mathbf{x}'_S$, while the generator G_{ST} tries to generate images to fool D_T. D_S and G_{TS} perform similar operations. As in [22], the CycleGAN loss contains two terms. One is the adversarial loss [21] that matches the distribution of generated images to the data distribution in the target domain

$$L_{\text{GAN}}(G_{ST}, D_T, \mathbf{x}_S, \mathbf{x}_T) = E_{\mathbf{x}_T \sim P_T}\big[\log D_T(\mathbf{x}_T)\big] + E_{\mathbf{x}_S \sim P_S}\big[\log(1 - D_T(G_{ST}(\mathbf{x}_S)))\big] \tag{1}$$

$$L_{\text{GAN}}(G_{TS}, D_S, \mathbf{x}_T, \mathbf{x}_S) = E_{\mathbf{x}_S \sim P_S}\big[\log D_S(\mathbf{x}_S)\big] + E_{\mathbf{x}_T \sim P_T}\big[\log(1 - D_S(G_{TS}(\mathbf{x}_T)))\big]. \tag{2}$$

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 so that the reconstructed image is close to the original image, which means $G_{TS}(G_{ST}(\mathbf{x}_S)) \approx \mathbf{x}_S$ and $G_{ST}(G_{TS}(\mathbf{x}_T)) \approx \mathbf{x}_T$. The difference is penalized by using the $L1$ norm and according to [22], the cycle-consistency loss is defined as

$$L_{\text{cyc}}(G_{ST}, G_{TS}) = E_{\mathbf{x}_S \sim P_S}\|G_{TS}(G_{ST}(\mathbf{x}_S)) - \mathbf{x}_S\|_1 + E_{\mathbf{x}_T \sim P_T}\|G_{ST}(G_{TS}(\mathbf{x}_T)) - \mathbf{x}_T\|_1. \tag{3}$$

According to [72], the $L1$ norm loss can preserve the luminance and color of the images. But it does not perform well in preserving the high-frequency-detailed information. Based on a top-down assumption [72] that the human visual system (HVS) is highly adapted for extracting structural information from the scene, a measure of structural similarity (SSIM) that considers luminance, contrast, and structure information is a good approximation of the perceived image quality. By considering HVS, SSIM can obtain more high-frequency-detailed information that preserves the contrast better. In addition, multiscale structural similarity (MS-SSIM) can overcome the drawback of SSIM in facing different viewing conditions and scales. However, MS-SSIM is not particularly sensitive to uniform biases, which can cause changes in brightness or shifts of colors. So, we can use MS-SSIM combined with the $L1$ norm as the *mixed cycle-consistency loss* to preserve both the detailed information and brightness of images

$$L_{\text{mixed}-\text{cyc}}(G_{ST}, G_{TS}) = \alpha \cdot (L_{\text{MS}}(G_{TS}(G_{ST}(\mathbf{x}_S)), \mathbf{x}_S) + L_{\text{MS}}(G_{ST}(G_{TS}(\mathbf{x}_T)), \mathbf{x}_T)) + (1 - \alpha) \cdot L_{\text{cyc}}. \tag{4}$$

According to [73], MS-SSIM is defined as

$$L_{\text{MS}}(x, y) = \big[l_M(x, y)\big]^{\alpha_M} \cdot \prod_{j=1}^{M}\big[c_j(x, y)\big]^{\beta_j}\big[s_j(x, y)\big]^{\gamma_j} \tag{5}$$

where the exponents α_M, β_j, and γ_j are used to adjust the relative importance of different components and M is the scale number [73]. Several parameters in this equation are preserved: luminance, contrast, and structure comparison that are defined as

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{6}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{7}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{8}$$

where μ_x and μ_y are the means of x and y, σ_x^2 and σ_y^2 are the variance of x and y, σ_{xy} is the covariance of x and y, and C_1, C_2, and C_3 are small constants given by $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$, and $C_3 = C_2/2$, respectively. L is the dynamic range of the pixel values, and $K_1 \ll 1$ and $K_2 \ll 1$ are two scalar constants.

Therefore, the objective of the mixed CycleGAN loss is

$$L_{m\text{CycleGAN}} = L_{\text{GAN}}(G_{ST}, D_T, \mathbf{x}_S, \mathbf{x}_T) + L_{\text{GAN}}(G_{TS}, D_S, \mathbf{x}_T, \mathbf{x}_S) + \beta L_{\text{mixed}-\text{cyc}}(G_{ST}, G_{TS}) \tag{9}$$

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

B. Dynamic Emotional Semantic Consistency Loss

The classifier F'_S is trained based on the adapted images and the emotion labels of corresponding source images with the assumption that the emotion labels do not change during the adaptation process. However, this assumption is not always valid. In order to preserve the emotion labels of the source images, we add a DESC loss. That is, we try to enforce the predicted emotions of the source images $\mathbf{x}_S$ and adapted images $\mathbf{x}'_S$ to be as close as possible. Since we have already known that source images and adapted images have different styles, we use two different classifiers to compute the loss. For source images, we use F_S that is trained on the source domain; for adapted images, we use F'_S that is trained on the adapted domain. The DESC loss is defined as

$$L_{\text{DESC}}(G_{ST}) = E_{\mathbf{x}_S \sim P_S} d\big(F_S(\mathbf{x}_S), F'_S(G_{ST}(\mathbf{x}_S))\big) \tag{10}$$

$$L_{\text{DESC}}(G_{TS}) = E_{\mathbf{x}_T \sim P_T} d\big(F'_S(\mathbf{x}_T), F_S(G_{TS}(\mathbf{x}_T))\big) \tag{11}$$

where $d(\cdot, \cdot)$ is a function that measures the distance between two emotion labels. In this article, we define d in two ways. The first one is using the symmetric Kullback–Leibler (SKL) divergence to measure the difference of two distributions $\mathbf{p}$ and $\mathbf{q}$

$$\text{SKL}(\mathbf{p}\|\mathbf{q}) = \text{KL}(\mathbf{p}\|\mathbf{q}) + \text{KL}(\mathbf{q}\|\mathbf{p}) \tag{12}$$

$$\text{KL}(\mathbf{p}\|\mathbf{q}) = \sum_{l=1}^{L}\big(\mathbf{p}_l \ln \mathbf{p}_l - \mathbf{p}_l \ln \mathbf{q}_l\big). \tag{13}$$

Second, we employ Mikels' wheel [8] to calculate the distance between emotions. As one of our overall goals is dominant emotion classification, we select the emotion category with the largest probability as our emotion label. Pairwise emotion distance is defined as 1+ "the number of steps required to reach one emotion from another," as shown in Fig. 4. Pairwise emotion similarity is defined as the reciprocal of the pairwise emotion distance. $d(\cdot, \cdot)$ equals 1-pairwise emotion similarity. From the definitions of these two methods, for the emotion distribution learning task, we can only use the first

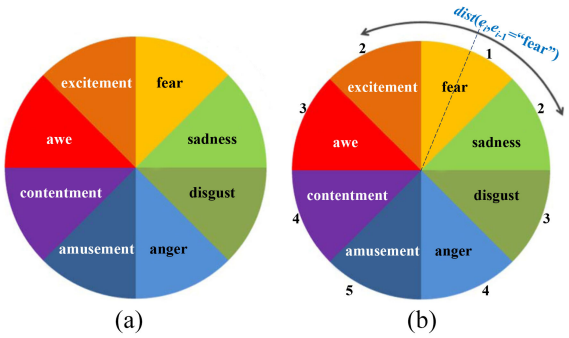

Fig. 4. Mikels' emotion wheel and an example of Mikels' emotion distances for the emotion category *fear* [8]. (a) Mikels' wheel. (b) Mikels' emotion distance.

method as the DESC loss, and for dominant emotion classification, we can use both methods. We name the models using these two methods as CycleEmotionGAN++-SKL and CycleEmotionGAN++-Mikels, respectively.

C. Feature-Level Alignment Loss

Since we want the adapted images $\mathbf{x}'_S$ and target images $\mathbf{x}_T$ to be similar, we should ensure that they are similar not only on a pixel level but also on a feature level. Our model trains a discriminator D_{feat} that tries to maximize the probability of correctly classifying adapted images $\mathbf{x}'_S$ and target images $\mathbf{x}_T$. The feature-level information we leverage is the output of the last layer in F'_S. So, the information is an L-dimension vector. We assume that the adapted images drawn from the distribution of $F'_S(G_{ST}(\mathbf{x}_S))$ are $\mathbf{x}'''_S$ and rename the distribution as P'_S

$$\begin{aligned} &L_{\text{GAN}}\big(F'_S, D_{\text{feat}}, \mathbf{x}_T, F'_S(G_{ST}(\mathbf{x}_S))\big) \\ &\quad = E_{\mathbf{x}'''_S \sim P'_S}\big[\log D_{\text{feat}}\big(F'_S(G_{ST}(\mathbf{x}_S))\big)\big] \\ &\qquad + E_{\mathbf{x}_T \sim P_T}\big[\log\big(1 - D_{\text{feat}}\big(F'_S(\mathbf{x}_T)\big)\big)\big]. \end{aligned} \tag{14}$$

D. Task Classification Loss

Under the assumption that the emotion labels of the adapted images do not change during the adaptation process, we can train a transferable task classifier F'_S based on the adapted images $\mathbf{x}'_S$ and corresponding source emotion labels $\mathbf{y}_S$. Besides $F'_S(\mathbf{x}'_s) \rightarrow \mathbf{y}''_s$, which assigns emotion $\mathbf{y}''_S$ to the adapted image $\mathbf{x}'_s$, the proposed CycleEmotionGAN++ is augmented with another classifier $F_S(\mathbf{x}_s) \rightarrow \mathbf{y}'_s$ assigning emotion $\mathbf{y}'_S$ to the source image $\mathbf{x}_s$ for DESC. For emotion distribution learning, we use the KL-Divergence as the task loss

$$L_{\text{task}}(F_S) = E_{(\mathbf{x}_S, \mathbf{y}_S)\sim P_S} \text{KL}\big(\mathbf{y}_S \| F_S(\mathbf{x}_S)\big) \tag{15}$$

$$L_{\text{task}}\big(F'_S\big) = E_{(\mathbf{x}_S, \mathbf{y}_S)\sim P_S} \text{KL}\big(\mathbf{y}_S \| F'_S(G_{ST}(\mathbf{x}_S))\big). \tag{16}$$

For dominant emotion classification, following [38], the two classifiers F_S and F'_S are optimized by minimizing the standard cross-entropy loss:

Algorithm 1: Adversarial Training Procedure of Our CycleEmotionGAN++ Model

Input: Sets of source images $\mathbf{x}_s \in \mathbf{x}_S$ with emotion labels $\mathbf{y}_s \in \mathbf{y}_S$, target images $\mathbf{x}_t \in \mathbf{x}_T$, the maximum number of steps of the first and second parts T_1, T_2, respectively, a threshold *thres*.

Output: Predicted emotion label distributions of target domain image $\mathbf{x}_T$.

for $i \leftarrow 1$ ***to*** T_1 **do**
  Sample a mini-batch of source images $\mathbf{x}_s$ and target images $\mathbf{x}_t$.
  /* Updating θ_{ST} and θ_{TS} when fixing ϕ_S, ϕ_T, δ_S and δ'_S */
  Update θ_{ST} and θ_{TS} by taking an SGD step on mini-batch loss $L_{mCycleGAN}$ plus $L_{DESC}(G_{ST})$ and $L_{DESC}(G_{TS})$ in Eq. (9), Eq. (10) and Eq. (11).
  /* Updating ϕ_S, ϕ_T when fixing θ_{ST}, θ_{TS}, δ_S and δ'_S */
  Compute $G_{ST}(\mathbf{x}_s, \theta_{ST})$ with current θ_{ST}.
  Compute $G_{TS}(\mathbf{x}_t, \theta_{TS})$ with current θ_{TS}.
  Update ϕ_T and ϕ_S by taking an SGD step on mini-batch loss L_{GAN} in Eq. (1), Eq. (2).
  /* Updating δ_S and δ'_S when fixing θ_{ST}, θ_{TS}, ϕ_S and ϕ_T */
  Compute $G_{ST}(\mathbf{x}_s, \theta_{ST})$ with current θ_{ST}.
  Update δ_S, δ'_S by taking an SGD step on mini-batch loss $L_{task}(F_S)$ and $L_{task}(F'_S)$ in Eq. (15) / Eq. (17) and Eq. (16) / Eq. (18).
end
Compute $G_{ST}(\mathbf{x}_s, \theta_{ST})$ with current θ_{ST}.
for $j \leftarrow 1$ ***to*** T_2 **do**
  Update ϕ by taking an SGD step on mini-batch loss L_{GAN} and in Eq. (14).
  if $Accuracy(D_{feat}) > thres$ **then**
    Update δ'_S by taking an SGD step on mini-batch loss $L_{task}(F'_S)$ in Eq. (16) / Eq. (18).
  end
end
return $F'_S(\mathbf{x}_t, \delta'_S)$;

$$L_{\text{task}}(F_S) = E_{(\mathbf{x}_S, \mathbf{y}_S)\sim P_S} \sum_{l=1}^{L} \mathbb{1}_{[l=\mathbf{y}_S]} \log\Big(\sigma\Big(F_S^{(l)}(\mathbf{x}_S)\Big)\Big) \tag{17}$$

$$L_{\text{task}}\big(F'_S\big) = E_{(\mathbf{x}_S, \mathbf{y}_S)\sim P_S} \sum_{l=1}^{L} \mathbb{1}_{[l=\mathbf{y}_S]} \log\Big(\sigma\Big(F_S^{'(l)}(G_{ST}(\mathbf{x}_S))\Big)\Big) \tag{18}$$

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

E. CycleEmotionGAN++ Learning

Our model objective loss combines the CycleGAN loss, DESC loss, and the feature-level alignment loss

$$\begin{aligned} L_{\text{Model}} &= L_{m\text{CycleGAN}} + \gamma L_{\text{DESC}}(G_{ST}) + \gamma L_{\text{DESC}}(G_{TS}) \\ &\quad + L_{\text{GAN}}\big(F'_S, D_{\text{feat}}, \mathbf{x}_T, F'_S(G_{ST}(\mathbf{x}_S))\big) \end{aligned} \tag{19}$$

where γ controls the relative importance of the DESC loss to the overall loss.

In our implementation, the generators G_{ST} and G_{TS} are CNNs with residual connections that maintain the resolution of the original image as illustrated in Fig. 3. The discriminators D_T, D_S, and D_{feat} and the classifiers F_S and F'_S are also

CNNs. The optimization of the our model is divided into two parts. In the first part, we optimize G_{ST}, G_{TS}, D_T, D_S, F_S, and F'_S. Those networks are optimized by alternating between three stochastic gradient descent (SGD) steps. During the first step, we fix D_T, D_S, F_S, and F'_S and update G_{ST} and G_{TS}. During the second step, we update D_T and D_S, while keeping G_{ST}, G_{TS}, F_S, and F'_S fixed. During the third step, we update F_S and F'_S, while keeping G_{ST}, G_{TS}, D_T, and D_S fixed. After the first part, we use G_{ST} to generate the adapted domain $\mathbf{x}'_S$ with $G_{ST}(\mathbf{x}_S)$. In the second part, we also optimize D_{feat} and F'_S by using SGD steps. F'_S is fixed when D_{feat}'s accuracy is lower than 0.8. The detailed training procedure is summarized in Algorithm 1, where θ_{ST}, θ_{TS}, ϕ_S, ϕ_T, δ_S, δ'_S, and ϕ are the parameters of G_{ST}, G_{TS}, D_S, D_T, F_S, F'_S, and D_{feat}, respectively.

IV. Experiments

In this section, we introduce the experimental settings, evaluate the performance of the proposed model, and report and analyze the results as compared to state-of-the-art approaches.

A. Datasets

Flickr-LDL [9] is a subset of FlickrCC [3] and contains 11 500 images that are labeled by 11 viewers based on Mikels' eight emotion categories. *Twitter-LDL* [9] contains 10 045 images obtained by searching from Twitter using emotions keywords. The images are labeled by 8 viewers also based on Mikels' emotion categories. The original labels of each image in Flickr-LDL and Twitter-LDL datasets are the number of votes on each emotion category. To obtain the emotion distribution labels, we divide the votes of each category by the total number of voters.

ArtPhoto [11] contains 806 artistic photographs organized by Mikels' emotion categories. The photographers took the photos, uploaded them to the website, and determined which one of the eight emotion categories each photo belongs to. The artists try to evoke a certain emotion in the viewers through the photos with conscious manipulation of the emotional objects, lighting, colors, etc. The FI dataset [38] contains images from the FI websites, which are labeled into one of Mikels' emotion categories by a group of 225 Amazon Mechanical Turk (AMT) workers. 23 308 images that received at least three agreements among workers are included in the FI dataset.

B. Evaluation Metrics

1) Distribution Learning on Twitter-LDL and Flickr-LDL: We use different metrics to evaluate the performance of our model: the sum of squared difference (SSD) [14], Kullback–Leibler (KL) divergence,[1] Bhattacharyya coefficient (BC),[2] Canberra distance (Canbe),[3] Chebyshev distance (Cheb),[4] and Cosine similarity (Cos)[5]. For BC and Cos, larger values indicate better results and for SSD, KL, Canbe, and Cheb, smaller values indicate better results. KL is leveraged as the main metric.

2) Emotion Classification on ArtPhoto and FI: Similar to [38], we employ the emotion classification accuracy (Acc) as the evaluation metric. Acc is defined as the proportion of correct predictions out of total predictions. We compute the accuracy for each category and employ the average accuracy as the main metric.

C. Baselines

To the best of our knowledge, CycleEmotionGAN++ is the first work on UDA for both emotion distribution learning and dominant emotion classification. To demonstrate its effectiveness, we compare it with the following baselines.

1) *Source-Only:* A lower bound that trains a model on the source domain and tests it directly on the target domain.
2) *Color Style Transfer Methods:* CycleGAN [22]: first, translating the source images into adapted images via CycleGAN and then training a task classifier on the adapted images; SAPE [23]: first, using the descriptor to generate semantics-aware photo enhanced style images and then training a task classifier on the enhanced images; EICT [24]: first, translating source images' style and then training a task classifier on the adapted images; and TAECT [25]: first, transferring source images to the color in the database extracted from reference images and then training a task classifier on the adapted images.
3) *UDA Methods:* ADDA [26]: first, training the task classifier on the source images and then aligning the feature-level information of the source and target domains; SimGAN [27]: first, translating the source images into the target style using the generator augmented with a self-regularization loss, and then training on the adapted images with corresponding source labels; and CyCADA [28]: first, translating source images into the target style with cycle-consistency loss and semantic-consistency loss, and then training the task classifier on the adapted images with feature-level alignment.
4) *Oracle:* An upper bound, where the classifier is both trained and tested on the target domain.

D. Implementation Details

The generators G_{ST} and G_{TS} use the model in [22], which shows impressive results for neural style transfer and super-resolution. G_{ST} and G_{TS} contain two stride-2 convolutions, several residual blocks, and two fractionally strided convolutions with stride $(1/2)$. Our model uses nine blocks for 256×256 training images. For normalization, we choose instance normalization [74]. The discriminators D_T and D_S are deployed by using 70×70 PatchGAN [75], which aims 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 work on the arbitrarily sized images in a fully convolutional fashion with good performance. The task classifier F_S and F'_S for the baselines and our method all use the ResNet-101 [4] architecture pretrained on ImageNet. We fine tune the model and

[1] https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
[2] https://en.wikipedia.org/wiki/Bhattacharyya_distance
[3] https://en.wikipedia.org/wiki/Canberra_distance
[4] https://en.wikipedia.org/wiki/Chebyshev_distance
[5] https://en.wikipedia.org/wiki/Cosine_similarity

TABLE II
CLASSIFICATION ACCURACY (%) COMPARISON BETWEEN CYCLEEMOTIONGAN++ AND STATE-OF-THE-ART APPROACHES WHEN ADAPTING FROM ARTPHOTO TO FI. THE BEST ACCURACY OF EACH EMOTION CATEGORY AND THE AVERAGE ACCURACY ARE EMPHASIZED IN BOLD

Method	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
Source-only	47.63	2.83	25.86	6.33	5.57	8.67	16.50	51.15	23.86
CycleGAN [22]	28.19	20.24	13.58	29.87	30.90	21.59	25.00	32.86	25.99
SAPE [23]	30.89	17.32	21.26	25.99	9.82	33.68	17.00	31.45	26.04
EICT [24]	30.08	8.66	16.28	33.51	2.45	24.56	15.00	24.73	24.00
TAECT [25]	36.69	7.87	22.92	21.68	14.11	**34.39**	19.00	20.49	25.09
ADDA [26]	39.51	1.62	32.90	8.13	4.33	7.96	7.50	**79.22**	26.33
SimGAN [27]	23.58	3.94	21.54	**34.95**	30.06	29.82	10.00	27.91	26.13
CyCADA [28]	33.74	18.90	22.83	44.27	23.31	10.18	14.00	34.63	29.62
CycleEmotionGAN-Mikels (ours)	42.90	4.45	**41.41**	5.10	6.81	15.93	3.00	74.25	28.00
CycleEmotionGAN-SKL (ours)	55.14	12.96	36.50	4.73	6.19	4.96	1.00	63.40	27.50
CycleEmotionGAN++-Mikels (ours)	**66.26**	16.54	21.54	27.78	3.68	14.39	6.00	40.28	31.74
CycleEmotionGAN++-SKL (ours)	44.86	**40.49**	18.33	32.99	**30.96**	17.88	**50.00**	27.53	**32.01**
Oracle (train on target)	77.24	44.88	72.03	65.59	60.12	61.40	48.00	65.37	66.11

update the classification loss into (15)–(18). The discriminator D_{feat} uses the architecture as [28], which contains three linear layers mapping an L-dimension vector to a 2-D one. The GAN loss keeps the same as standard ones. As shown in Algorithm 1, D_{feat} is trained after acquiring fine tuned F'_S and x'_S generated by G_{ST}.

Similar to LSGAN [76], our model trains a GAN loss $L_{\text{GAN}}(G_{ST}, D_T, \mathbf{x}_S, \mathbf{x}_T)$ by minimizing G_{ST} in $E_{\mathbf{x}_S \sim P_S}[(D_T(G_{ST}(\mathbf{x}_S)) - 1)^2]$ and minimizing D_T in $E_{\mathbf{x}_T \sim P_T}[(D_T(\mathbf{x}_T) - 1)^2] + E_{\mathbf{x}_S \sim P_S}[D_T(G_{ST}(\mathbf{x}_S))^2]$. It is more stable to train GANs on high-resolution images. $L_{\text{GAN}}(G_{TS}, D_S, \mathbf{x}_T, \mathbf{x}_S)$ is also optimized by using LSGAN. We follow [27] to reduce the model oscillation by updating the discriminators using a history of generated images rather than the ones produced by the latest generators. We leverage an image pool that stores the 50 recently created images.

α, β, and γ in (4), (9), and (19) are empirically set to 0.5, 10, and 50, respectively. Similar to [73], in order to simplify parameter selection, we set $\alpha_j = \beta_j = \gamma_j$ for all j's in (5), and normalize the cross-scale settings such that $\sum_{j=1}^{M} \gamma_j = 1$, which makes different parameter settings comparable. M is set to 5 and K_1 and K_2 are set to 0.01 and 0.03, respectively. We also set $\beta_1 = \gamma_1 = 0.0448, \beta_2 = \gamma_2 = 0.2856, \beta_3 = \gamma_3 = 0.3001, \beta_4 = \gamma_4 = 0.2363$, and $\alpha_5 = \beta_5 = \gamma_5 = 0.1333$, respectively. For the first part, we use the Adam optimizer for generators with a learning rate of 0.0002 and the SGD optimizer for classifiers with a learning rate of 0.0001 and a batch size of 1. We train the first part for 200 epochs, keeping the same learning rate for the first 100 epochs and linearly decaying it to 0 over the next 100 epochs. Then, we choose the classifier F'_S and G_{ST} with the best validation performance during the first stage (image translation training) and use the classifier and the adapted images x'_S generated by G_{ST} during the second stage. For the second part, we use the Adam optimizer with a batch size of 64 and a learning rate of 0.0001. We train D_{feat} and F'_S for 200 epochs and F'_S is updated only when D_{feat}'s accuracy is larger than 0.8. All our experiments are conducted on a machine with 4 NVIDIA TITAN *V* GPUs, each with 12-GB memory.

E. Results and Analysis

Comparison With State of the Art: The performance comparisons between the proposed CycleEmotionGAN++ model and state-of-the-art approaches are shown in Tables II–V. From the results, we have several observations.

1) The source-only method directly transferring the models trained on the source domain to the target domain performs the worst in all adaptation settings. Due to the influence of *domain shift*, the style of images and distribution of labels are totally different in the two different domains, which results in the model's low transferability from one domain to another.
2) All the style transfer and DA methods outperform the source-only method, with CycleEmotionGAN++ performing the best since these methods can overcome the *domain shift* to some extent. Specifically, the performance improvements of our model over source-only, CycleGAN, SAPE, EICT, TAECT, ADDA, SimGAN, and CyCADA measured by KL are 15.67%, 10.39%, 9.46%, 12.73%, 12.40%, 11.05%, 11.27%, and 2.40% when adapting from the source Twitter-LDL to the target Flickr-LDL, respectively. The performance improvements of our model over these methods measured by average classification accuracy are 34.16%, 23.16%, 22.93%, 33.38%, 27.58%, 21.57%, 22.50%, and 8.07% when adapting from the source ArtPhoto to the target FI, respectively. The improvements imply that our model can achieve superior performance relative to these approaches.
3) For the Twitter-LDL and Flickr-LDL datasets, we observe that our model obtains the best performance in most of the evaluation metrics except BC in the Twitter-LDL→Flickr-LDL process and Canbe in the Flickr-LDL→Twitter-LDL process. For ArtPhoto and FI datasets, a better model cannot ensure better accuracy on every single category, but only for the average accuracy. The two methods of (dynamic) emotional semantic consistency (ESC) loss obtain similar performance. For the original CycleEmotionGAN [30], Mikels' wheel obtains better performance, while SKL outperforms Mikels' wheel for CycleEmotionGAN++.
4) The oracle method achieves the best performance on both emotion distribution learning and dominant emotion classification tasks. However, this model is trained using the ground-truth emotion labels from the target domain $\mathbf{x}_T$, which are actually unavailable in the UDA setting.

TABLE III
CLASSIFICATION ACCURACY (%) COMPARISON BETWEEN CYCLEEMOTIONGAN++ AND STATE-OF-THE-ART APPROACHES WHEN ADAPTING FROM FI TO ARTPHOTO

Method	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
Source-only	**30.00**	28.57	**55.00**	35.71	14.29	20.00	18.18	30.30	29.11
CycleGAN [22]	15.00	28.57	20.00	7.14	21.43	**60.00**	40.90	42.42	31.65
SAPE [23]	**30.00**	7.14	45.00	35.71	7.14	**60.00**	30.43	54.55	33.54
EICT [24]	25.00	28.57	30.00	21.43	14.29	50.00	26.09	42.42	31.64
TAECT [25]	20.00	14.29	40.00	21.43	28.57	55.00	30.43	30.30	31.01
ADDA [26]	25.00	**42.86**	40.00	7.14	14.29	55.00	40.90	30.30	32.91
SimGAN [27]	15.00	14.29	45.00	28.57	7.14	25.00	31.82	**63.64**	32.91
CyCADA [28]	20.00	21.43	45.00	0.00	35.71	40.00	59.09	57.58	38.61
CycleEmotionGAN-Mikels (ours)	**30.00**	**42.86**	45.00	21.43	21.43	50.00	55.00	33.33	37.97
CycleEmotionGAN-SKL (ours)	25.00	28.57	35.00	**42.86**	0.00	55.00	54.55	42.42	37.34
CycleEmotionGAN++-Mikels (ours)	25.00	28.57	30.00	**42.86**	**42.86**	55.00	72.73	27.27	39.87
CycleEmotionGAN++-SKL (ours)	**30.00**	36.00	40.00	21.43	14.29	40.00	**77.27**	45.45	**40.51**
Oracle (train on target)	55.00	35.71	30.00	14.29	42.86	55.00	59.09	45.45	43.67

5) For the ArtPhoto and FI datasets, 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 [43], we find the accuracies of all DA methods are not very high and effectively adapting image emotions is still a challenging problem.

Ablation Study: First, we perform various experiments to evaluate how each component contributes to the adaptation performance, with results shown in Tables VI and VII. We can observe the following.

1) DESC loss boosts the performance by a large margin; after adding it, the performance improves significantly, proving that preserving the emotion label is of vital importance. Since we use the classifier trained on the adapted domain to test, we should make sure that it can use emotion labels from the source domain as its labels.
2) From the third and fourth rows of each DA setting, we can see the improvements that feature-level loss and multiscale structural similarity contribute. Each of these two components can improve the performance of the model trained in the source domain.
3) The last row of each setting, which contains all of the three components performs the best. For example, it obtains the best average classification accuracy on ArtPhoto and FI datasets.

Second, we compare the proposed dynamic semantic consistency (DESC) loss and the original ESC loss. The main difference between these two methods is that DESC uses two classifiers, one for each of the source domain and the adapted domain to dynamically preserve the emotion labels. The results are shown in Tables VIII and IX. For each process, we use CycleGAN and CycleGAN+Feat as baselines. For Table VIII, the latter of every two rows obtains better performance in all evaluation metrics except Canbe. For Table IX, the latter of every two rows obtains better average classification accuracy.

Visualization of Adapted Images: As illustrated from Fig. 5, we visualize the adapted images to demonstrate the effectiveness and necessity of image translation. We compare the adapted images generated by CycleGAN [22], SAPE [23], EICT [24], TAECT [25], CycleEmotionGAN-SKL, and

TABLE IV
COMPARISON OF CYCLEEMOTIONGAN++ WITH STATE OF-THE-ART METHODS WHEN ADAPTING FROM THE SOURCE-DOMAIN TWITTER-LDL TO THE TARGET-DOMAIN FLICKR-LDL. THE BEST METHOD TRAINED ON THE SOURCE DOMAIN IS EMPHASIZED IN BOLD

Method	$SSD\downarrow$	$KL\downarrow$	$BC\uparrow$	$Canbe\downarrow$	$Cheb\downarrow$	$Cos\uparrow$
Source-only	0.1864	0.6845	0.7974	5.8347	0.2964	0.7846
CycleGAN [22]	0.1781	0.6441	0.8000	5.8173	0.2905	0.7904
SAPE [23]	0.1768	0.6375	0.8045	5.8434	0.2875	0.7935
EICT [24]	0.1803	0.6614	0.8129	5.9373	0.2904	0.8102
TAECT [25]	0.1777	0.6589	0.8046	5.9027	0.2843	0.8001
ADDA [26]	0.1876	0.6489	**0.8304**	6.0150	0.2830	0.8026
SimGAN [27]	0.1787	0.6505	0.8073	5.8621	0.2891	0.7904
CyCADA [28]	**0.1589**	0.5914	0.8130	5.7576	0.2757	0.8126
CycleEmotionGAN-SKL (Ours)	0.1672	0.6123	0.8156	5.8374	0.2775	0.8089
CycleEmotionGAN++-SKL (ours)	**0.1589**	**0.5772**	0.8214	**5.7542**	**0.2718**	**0.8167**
Oracle (train on target)	0.1419	0.5306	0.8248	5.5078	0.2597	0.8335

TABLE V
COMPARISON OF CYCLEEMOTIONGAN++ WITH STATE-OF-ART METHODS WHEN ADAPTING FROM FLICKR-LDL TO TWITTER-LDL

Method	$SSD\downarrow$	$KL\downarrow$	$BC\uparrow$	$Canbe\downarrow$	$Cheb\downarrow$	$Cos\uparrow$
Source-only	0.1856	0.6650	0.7791	6.0716	0.3066	0.8004
CycleGAN [22]	0.1712	0.6392	0.7964	6.0234	0.2896	0.8132
SAPE [23]	0.1694	0.6101	0.8086	6.0700	0.2799	0.8263
EICT [24]	0.1793	0.6482	0.7819	6.0652	0.2865	0.8097
TAECT [25]	0.1735	0.6392	0.7916	6.0635	0.2803	0.8139
ADDA [26]	0.1693	0.6306	0.7924	6.0630	0.2883	0.8141
SimGAN [27]	0.1751	0.6338	0.7919	6.0560	0.2938	0.8088
CyCADA [28]	0.1493	0.5617	0.8120	**6.0178**	0.2712	0.8375
CycleEmotionGAN-SKL(ours)	0.1541	0.5800	0.8124	6.0512	0.2688	0.8327
CycleEmotionGAN++-SKL (ours)	**0.1410**	**0.5412**	**0.8273**	6.0336	**0.2529**	**0.8462**
Oracle (train on target)	0.1274	0.5003	0.8439	5.8543	0.2389	0.8629

CycleEmotionGAN++-SKL. We can observe that all these methods can adapt the source images to be more similar to the images of the target domain. However, CycleEmotionGAN++-SKL performs better than the other methods. For example, the hue of the adapted image (b) in the second but last line generated by CycleEmotionGAN++-SKL is more yellow which is closer to target domain FI. Therefore, the images generated by our model are more similar to the target images, as compared to the original images and the images generated by CycleGAN. This further demonstrates the effectiveness of the proposed model.

Similar to GAN [21] and CycleGAN [22] based image generation methods, the proposed CycleEmotionGAN++ also suffers from low quality problem. As our goal is to improve the accuracy of the task classifier, that is, F'_S, we did not employ any high-resolution or high-quality generation methods, such as MSG-GAN [77] and EventSR [78], which usually require

TABLE VI
ABLATION STUDY ON DIFFERENT COMPONENTS OF CYCLEEMOTIONGAN++ FOR EMOTION DISTRIBUTION LEARNING. BASELINE DENOTES PIXEL-LEVEL ALIGNMENT WITH CYCLE CONSISTENCY, +DESC DENOTES ADDING DESC LOSS, +FEAT DENOTES ADDING FEATURE-LEVEL ALIGNMENT, AND +MSSSIM DENOTES ADDING MULTISCALE STRUCTURE SIMILARITY

DA setting	Method	$SSD\downarrow$	$KL\downarrow$	$BC\uparrow$	$Canbe\downarrow$	$Cheb\downarrow$	$Cos\uparrow$
Twitter-LDL → Flickr-LDL	CycleGAN (Baseline)	0.1781	0.6441	0.8000	5.8173	0.2905	0.7904
	+DESC	0.1592	0.6027	0.8184	5.8149	0.2732	0.8156
	+DESC+Feat	**0.1558**	0.5901	**0.8245**	5.7745	0.2713	0.8145
	+DESC+msssim	0.1578	0.5949	0.8196	5.8100	**0.2707**	0.8164
	+DESC+msssim+Feat	0.1589	**0.5772**	0.8214	**5.7542**	0.2718	**0.8167**
Flickr-LDL → Twitter-LDL	CycleGAN (Baseline)	0.1712	0.6392	0.7964	**6.0234**	0.2896	0.8132
	+DESC	0.1478	0.5724	0.8176	6.0701	0.2627	0.8395
	+DESC+Feat	0.1477	0.5513	0.8256	6.0516	0.2593	0.8400
	+DESC+msssim	0.1498	0.5642	0.8228	6.0591	0.2641	0.8402
	+DESC+msssim+Feat	**0.1410**	**0.5412**	**0.8273**	6.0336	**0.2529**	**0.8462**

TABLE VII
ABLATION STUDY ON DIFFERENT COMPONENTS OF CYCLEEMOTIONGAN++ FOR DOMINANT EMOTION CLASSIFICATION

DA setting	Method	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
ArtPhoto → FI	CycleGAN (Baseline)	28.19	20.24	13.58	29.87	30.90	21.59	25.00	32.86	25.99
	+DESC	46.95	26.62	32.48	33.69	**38.65**	31.23	8.00	9.54	27.94
	+DESC+Feat	2.64	21.26	**38.54**	**37.10**	12.27	23.51	11.00	**47.35**	30.19
	+DESC+msssim	**51.95**	24.70	19.97	31.38	21.67	**35.75**	13.50	7.99	30.05
	+DESC+msssim+Feat	44.86	**40.49**	18.33	32.99	30.96	17.88	**50.00**	27.53	**32.01**
FI → ArtPhoto	CycleGAN (Baseline)	15.00	28.57	20.00	7.14	21.43	**60.00**	40.90	42.42	31.65
	+DESC	**30.00**	7.14	35.00	**35.71**	**35.71**	55.00	36.36	51.52	37.97
	+DESC+Feat	25.00	14.29	**50.00**	**35.71**	7.14	40.00	50.00	**57.58**	39.24
	+DESC+msssim	25.00	28.57	**50.00**	14.29	21.43	50.00	59.09	42.42	38.61
	+DESC+msssim+Feat	**30.00**	**36.00**	40.00	21.43	14.29	40.00	**77.27**	45.45	**40.51**

TABLE VIII
COMPARISON BETWEEN THE PROPOSED DESC LOSS AND THE ORIGINAL ESC LOSS IN [30] FOR EMOTION DISTRIBUTION LEARNING. WE USE CYCLEGAN AND CYCLEGAN+FEAT AS BASELINES

DA setting	Method	$SSD\downarrow$	$KL\downarrow$	$BC\uparrow$	$Canbe\downarrow$	$Cheb\downarrow$	$Cos\uparrow$
Twitter-LDL → Flickr-LDL	CycleGAN+ESC	0.1672	0.6123	0.8156	5.8374	0.2775	0.8089
	CycleGAN+DESC	0.1592	0.6027	0.8184	5.8149	0.2732	0.8156
	CycleGAN+Feat+ESC	0.1589	0.5914	0.8130	5.7576	0.2757	0.8126
	CycleGAN+Feat+DESC	0.1558	0.5901	0.8245	5.7745	0.2713	0.8145
Flickr-LDL → Twitter-LDL	CycleGAN+ESC	0.1541	0.5800	0.8124	6.0512	0.2688	0.8327
	CycleGAN+DESC	0.1478	0.5724	0.8176	6.0701	0.2627	0.8395
	CycleGAN+Feat+ESC	0.1493	0.5617	0.8120	6.0178	0.2712	0.8375
	CycleGAN+Feat+DESC	0.1477	0.5513	0.8256	6.0516	0.2593	0.8400

TABLE IX
COMPARISON BETWEEN THE PROPOSED DESC LOSS AND THE ORIGINAL ESC LOSS IN [30] FOR DOMINANT EMOTION CLASSIFICATION

DA setting	Method	Amu	Ang	Awe	Con	Dis	Exc	Fea	Sad	Avg
ArtPhoto → FI	CycleGAN+ESC	55.14	12.96	36.50	4.73	6.19	4.96	1.00	63.40	27.50
	CycleGAN+DESC	46.95	26.62	32.48	33.69	38.65	31.23	8.00	9.54	27.94
	CycleGAN+Feat+ESC	33.74	18.90	22.83	44.27	23.31	10.18	14.00	34.63	29.62
	CycleGAN+Feat+DESC	2.64	21.26	38.54	37.10	12.27	23.51	11.00	47.35	30.19
FI → ArtPhoto	CycleGAN+ESC	25.00	28.57	35.00	42.86	0.00	55.00	54.55	42.42	37.34
	CycleGAN+DESC	30.00	7.14	35.00	35.71	35.71	55.00	36.36	51.52	37.97
	CycleGAN+Feat+ESC	20.00	21.43	45.00	0.00	35.71	40.00	59.09	57.58	38.61
	CycleGAN+Feat+DESC	25.00	14.29	50.00	35.71	7.14	40.00	50.00	57.58	39.24

more computation cost. We leave generating high-quality images as our future work.

Visualization of Predicted Results: Some predicted emotion label distributions are visualized in Fig. 6 for the Twitter-LDL dataset. The first two examples in the blue frame show that our model's results are close to the ground-truth label distributions, which demonstrates the effectiveness of our proposed model. In other examples in the red frame, the results of our model, as well as the oracle, are not close enough to the ground truth. In these two examples, we can observe that, though the oracle performs better than all the adaptation methods, it is still very different from the ground truth, demonstrating the need for further research.

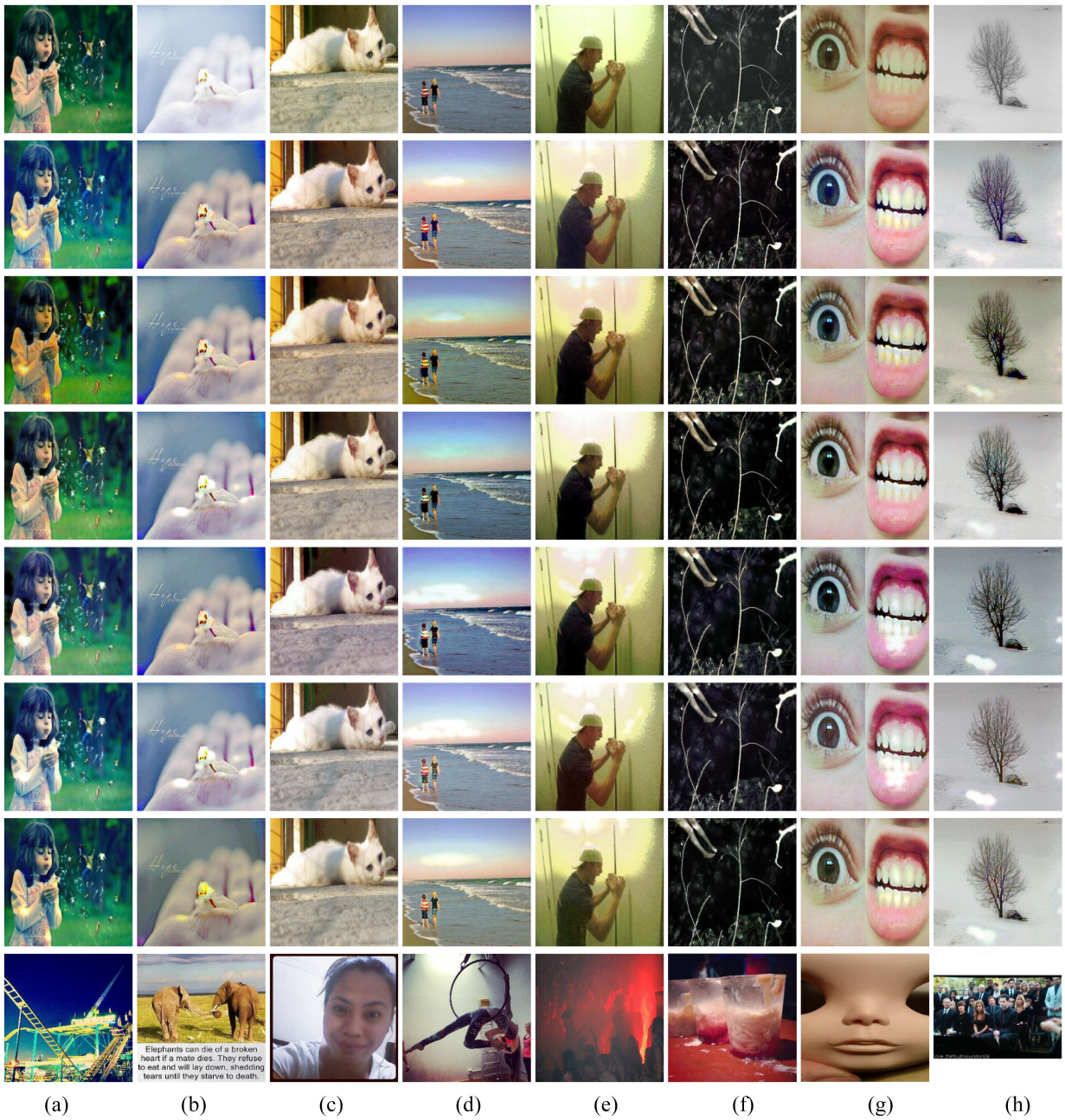

Fig. 5. Visualization of images across Mikels' emotion categories from ArtPhoto, adapted in order to make them have the style of FI. From top to bottom are: original ArtPhoto images, images generated by CycleGAN [22], SAPE [23], EICT [24], TAECT [25], CycleEmotionGAN-SKL [30], CycleEmotionGAN++-SKL, and original FI images. (a) Amu, (b) Awe, (c) Con, (d) Exc, (e) Ang, (f) Dis, (g) Fea, and (h) Sad.

Convergence: In order to display the training process more directly, we visualize some loss curves when adapting from the source-domain Twitter-LDL to the target-domain Flickr-LDL in Fig. 7. We can observe that during the first part training of generating an adapted domain $\{x'_S\}$, the best validation KL performance appears between 50 and 100 epochs, and then with the decrease of the training loss, the validation KL performance becomes unstable, showing overfitting in Fig. 7 (a). We use the networks with the best performance, that is, F'_S and G_{ST}, to generated adapted images x'_S for the second part of the training. From the other three figures, we observe that the losses decrease gradually with the increase of epoch number.

V. Conclusion

In this article, we studied the UDA problem for both emotion distribution learning and dominant emotion classification. We proposed an end-to-end cycle-consistent adversarial model, CycleEmotionGAN++, to bridge the gap between different domains. We generated an adapted domain to align the source and target domains on the pixel level by improving CycleGAN with a multiscale structured cycle-consistency loss. During the image translation, we proposed DESC loss to preserve the emotion labels of the source images. We trained a transferable task classifier on the adapted domain with feature-level alignment between the adapted

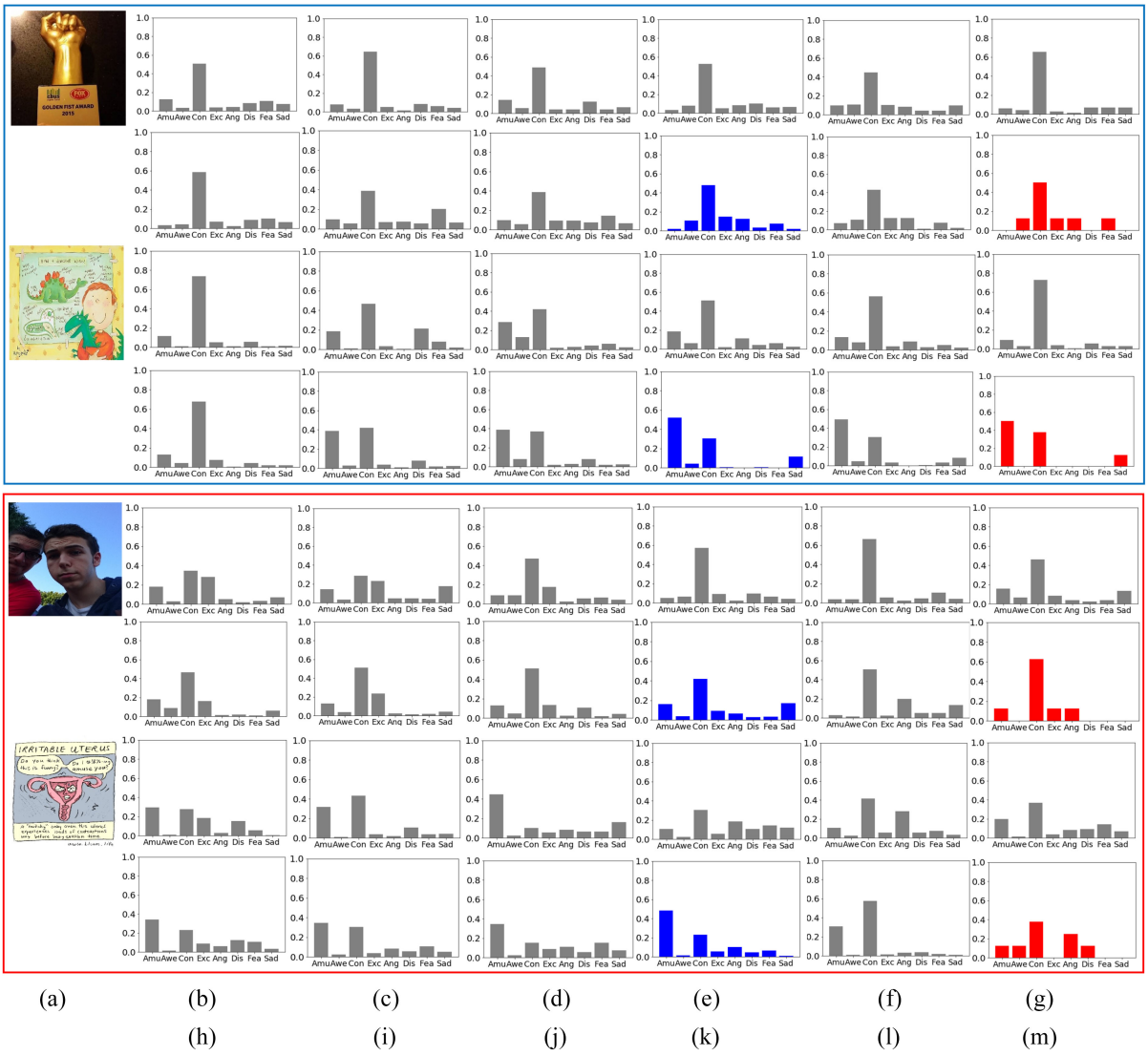

Fig. 6. Visualization of predicted emotion distributions on Twitter-LDL using the proposed CycleEmotionGAN++-SKL (CEGAN++-SKL), the original CycleEmotionGAN-SKL (CEGAN-SKL) [30] and several state-of-the-art approaches (source-only, CycleGAN [22], SAPE [23], EICT [24], TAECT [25], ADDA [26], SimGAN [27], CyCADA [28], and oracle). Original images and the corresponding ground-truth distributions are shown in the first column and last image of each group, respectively.

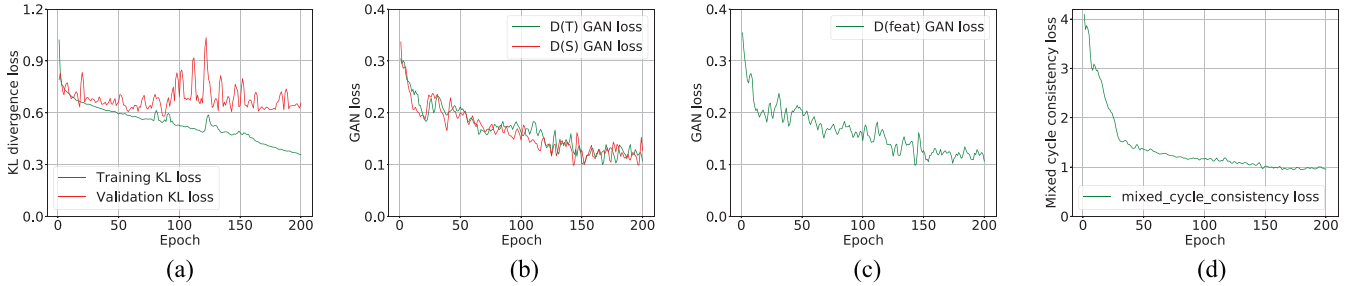

Fig. 7. Examples of loss curves when adapting from the source-domain Twitter-LDL to the target-domain Flickr-LDL. The figures from left to right: (a) Training and validation KL divergence loss, (b) GAN loss of D_T and D_S, (c) GAN loss of D_{feat}, and (d) mixed cycle consistency loss.

and target domains. We conducted extensive experiments on the Flickr-LDL and Twitter-LDL datasets for emotion distribution learning, and the ArtPhoto and FI datasets for dominant emotion classification. The results on these four datasets demonstrate the significant improvements yielded by the proposed method over state-of-the-art UDA approaches. For future work, we plan to extend the CycleEmotionGAN++ model to multimodal settings, such as audio-visual emotion recognition. We will also investigate domain generalization without accessing target data for VEA and study theoretical deduction to better understand the learning process.

REFERENCES

[1] P. J. Lang, "A bio-informational theory of emotional imagery," *Psychophysiology*, vol. 16, no. 6, pp. 495–512, 1979.

[2] S. Zhao, Y. Gao, G. Ding, and T.-S. Chua, "Real-time multimedia social event detection in microblog," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3218–3231, Nov. 2018.

[3] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2013, pp. 223–232.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[5] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer, "PDANet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2019, pp. 192–201.

[6] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 47–56.

[7] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 860–868.

[8] S. Zhao *et al.*, "Predicting personalized emotion perceptions of social images," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2016, pp. 1385–1394.

[9] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 224–230.

[10] S. Zhao, G. Ding, Q. Huang, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: A comprehensive survey," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 5534–5541.

[11] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2010, pp. 83–92.

[12] X. Lu, P. Suryanarayan, R. B. Adams Jr., J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2012, pp. 229–238.

[13] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 3266–3272.

[14] S. Zhao, G. Ding, Y. Gao, and J. Han, "Approximating discrete probability distribution of image emotions by multi-modal features fusion," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 4669–4675.

[15] S. Zhao, G. Ding, Y. Gao, and J. Han, "Learning visual emotion distributions via multi-modal features fusion," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2017, pp. 369–377.

[16] X. Zhu *et al.*, "Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 3595–3601.

[17] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7584–7592.

[18] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 1521–1528.

[19] S. Zhao *et al.*, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 23, 2020, doi: 10.1109/TNNLS.2020.3028503.

[20] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa, "Visual domain adaptation: A survey of recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 53–69, May 2015.

[21] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2223–2232.

[23] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM Trans. Graph.*, vol. 35, no. 2, pp. 1–15, 2016.

[24] D. Liu, Y. Jiang, M. Pei, and S. Liu, "Emotional image color transfer via deep learning," *Pattern Recognit. Lett.*, vol. 110, pp. 16–22, Jul. 2018.

[25] S. Liu and M. Pei, "Texture-aware emotional color transfer between images," *IEEE Access*, vol. 6, pp. 31375–31386, 2018.

[26] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput.Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2962–2971.

[27] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, "Learning from simulated and unsupervised images through adversarial training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2242–2251.

[28] J. Hoffman *et al.*, "CYCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1989–1998.

[29] S. Zhao, X. Zhao, G. Ding, and K. Keutzer, "EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2018, pp. 1319–1327.

[30] S. Zhao *et al.*, "Cycleemotiongan: Emotional semantic consistency preserved cyclegan for adapting image emotions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 2620–2627.

[31] P. J. Lang, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," Dept. Clinical Health Psychology, Univ. Florida, Gainesville, FL, USA, Rep.A-8, 2005.

[32] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, Mar. 2017.

[33] S. Zhao, H. Yao, and X. Jiang, "Predicting continuous probability distribution of image emotions in valence-arousal space," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2015, pp. 879–882.

[34] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: Image sentiment analysis from a mid-level perspective," in *Proc. Int. Workshop Issues Sentiment Discov. Opinion Mining (WISDOM)*, 2013, pp. 1–8.

[35] T. Chen, F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang, "Object-based visual sentiment concept analysis and application," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 367–376.

[36] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 526–540, Oct. 2018.

[37] T. Li, B. Ni, M. Xu, M. Wang, Q. Gao, and S. Yan, "Data-driven affective filtering for images and videos," *IEEE Trans. Cybern.*, vol. 45, no. 10, pp. 2336–2349, Oct. 2015.

[38] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2016, pp. 308–314.

[39] S. Zhao *et al.*, "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE Trans. Affect. Comput.*, vol. 11, no. 4, pp. 574–587, Oct./Dec. 2020.

[40] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2015, pp. 381–388.

[41] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 231–237.

[42] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Process. Lett.*, vol. 51, pp. 2043–2061, Jun. 2020.

[43] J. Yang, D. She, Y. Lai, and M.-H. Yang, "Retrieving and classifying affective images via deep metric learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 491–498.

[44] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 213–226.

[45] B. Gong, K. Grauman, and F. Sha, "Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 222–230.

[46] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola, "Correcting sample selection bias by unlabeled data," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2007, pp. 601–608.

[47] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 2960–2967.

[48] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 2066–2073.

[49] J. Zhuo, S. Wang, W. Zhang, and Q. Huang, "Deep unsupervised convolutional domain adaptation," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2017, pp. 261–269.

[50] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 97–105.
[51] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*. Cham, Switzerland: Springer, 2017, pp. 153–171.
[52] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 4376–4382.
[53] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4893–4902.
[54] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun, "No more discrimination: Cross city adaptation of road scene segmenters," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1992–2001.
[55] Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7472–7481.
[56] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 1640–1650.
[57] S. Cicek and S. Soatto, "Unsupervised domain adaptation via regularized conditional alignment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1416–1425.
[58] D. Hu *et al.*, "Semantic domain adversarial networks for unsupervised domain adaptation," 2020. [Online]. Available: arXiv:2003.13274
[59] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
[60] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016, pp. 469–477.
[61] X. Yue, Y. Zhang, S. Zhao, A. Sangiovanni-Vincentelli, K. Keutzer, and B. Gong, "Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2100–2110.
[62] M. Ghifary, W. Bastiaan Kleijn, M. Zhang, and D. Balduzzi, "Domain generalization for object recognition with multi-task autoencoders," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2551–2559.
[63] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 597–613.
[64] X. Chen, H. Li, C. Zhou, X. Liu, D. Wu, and G. Dudek, "FiDo: Ubiquitous fine-grained wifi-based localization for unlabelled users via domain adaptation," in *Proc. World Wide Web Conf. (WWW)*, 2020, pp. 23–33.
[65] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," 2019. [Online]. Available: arXiv:1909.11825
[66] J. Xu, L. Xiao, and A. M. López, "Self-supervised domain adaptation for computer vision tasks," *IEEE Access*, vol. 7, pp. 156694–156706, 2019.
[67] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 2229–2238.
[68] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 5, pp. 34–41, Jul./Aug. 2001.
[69] C.-Y. Wei, N. Dimitrova, and S.-F. Chang, "Color-mood analysis of films based on syntactic and psychological models," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, vol. 2. 2004, pp. 831–834.
[70] Y. Hwang, J.-Y. Lee, I. So Kweon, and S. Joo Kim, "Color transfer using probabilistic moving least squares," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 3342–3349.
[71] J. Rabin, S. Ferradans, and N. Papadakis, "Adaptive color transfer with relaxed optimal transport," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2014, pp. 4852–4856.
[72] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imag.*, vol. 3, no. 1, pp. 47–57, Mar. 2017.
[73] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals Syst. Comput. (ACSSC)*, vol. 2, 2003, pp. 1398–1402.
[74] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.
[75] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 702–716.
[76] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2794–2802.
[77] A. Karnewar and O. Wang, "MSG-GAN: Multi-scale gradients for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 7799–7808.
[78] L. Wang, T.-K. Kim, and K.-J. Yoon, "EventSR: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8315–8325.