Pattern Recognition 100 (2020) 107111

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

ExprADA: Adversarial domain adaptation for facial expression analysis

Behzad Bozorgtabar[a,*], Dwarikanath Mahapatra[b], Jean-Philippe Thiran[a]

[a] Department of Radiology- Center of Biomedical Imaging, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland
[b] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

ARTICLE INFO

Article history:
Received 3 December 2018
Revised 3 October 2019
Accepted 13 November 2019
Available online 14 November 2019

Keywords:
Visual domain adaptation
Facial expression recognition
Adversarial learning

ABSTRACT

We propose a deep neural network based image-to-image translation for domain adaptation, which aims at finding translations between image domains. Despite recent GAN based methods showing promising results in image-to-image translation, they are prone to fail at preserving semantic information and maintaining image details during translation, which reduces their practicality on tasks such as facial expression synthesis. In this paper, we learn a framework with two training objectives: first, we propose a multi-domain image synthesis model, yielding a better recognition performance compared to other GAN based methods, with a focus on the data augmentation process; second, we explore the use of domain adaptation to transform the visual appearance of the images from different domains, with the detail of face characteristics (e.g., identity) well preserved. Doing so, the expression recognition model learned from the source domain can be generalized to the translated images from target domain, without the need for re-training a model for new target domain. Extensive experiments demonstrate that ExprADA shows significant improvements in facial expression recognition accuracy compared to state-of-the-art domain adaptation methods.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

The expression recognition accuracy of face images has improved significantly with the advent of deep convolutional neural networks (CNN) which utilize numerous annotated data. However, collecting numerous annotated samples in various domains is time-consuming and expensive. In addition, due to domain shift, CNNs would suffer from performance degradation when being applied to new datasets with different distribution of samples. For example, neural networks trained on labeled source images collected from RGB modality may not recognize target images collected from a near-infrared camera correctly as target images can have different characteristics from the source images, such as intensity distributions in which the face image is captured. To adopt CNNs trained on a source domain to a target domain, recently, unsupervised domain adaptation has been widely investigated, where we have access to labeled samples from source domain and only unlabeled target samples. In particular, Generative Adversarial Network (GAN) [1] variants have made great achievements in image-to-image translation task and mapping distributions for unsupervised domain adaptation. These GAN models could be trained in both with paired training data [2] and unpaired training data [3,4]. More Recently, the method [5] adopted GAN with cycle-consistency constraints to perform mapping images between domains. However, neither category information nor semantic structure can be preserved in the synthetic images. Moreover, these methods require training a new model for every new domain, which will limit their applications.

1.1. Motivation

Most of the existing face expression recognition (FER) methods are still concentrating on the recognition of frontal or near-frontal facial images. However, the performance of these methods will drop significantly in an unconstrained real-world environment, especially when there are large head pose variations. One possible solution would be rendering simulated faces in frontal view using a 3D Morphable Model [6,7] (see Fig. 1). However, learning directly from simulated face images would be challenging due to a domain gap between simulation and reality.

1.2. Contributions

The rationale behind our contributions is using visual domain adaptation for simulated face images to reduce the reality gap between simulation and reality. Doing so, we propose adversarial domain adaptation approach in the form of image-to-image transla-

[*] Corresponding author.
E-mail addresses: behzad.bozorgtabar@epfl.ch (B. Bozorgtabar), dwarikanath.mahapatra@inceptioniai.org (D. Mahapatra), jean-philippe.thiran@epfl.ch (J.-P. Thiran).

https://doi.org/10.1016/j.patcog.2019.107111
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

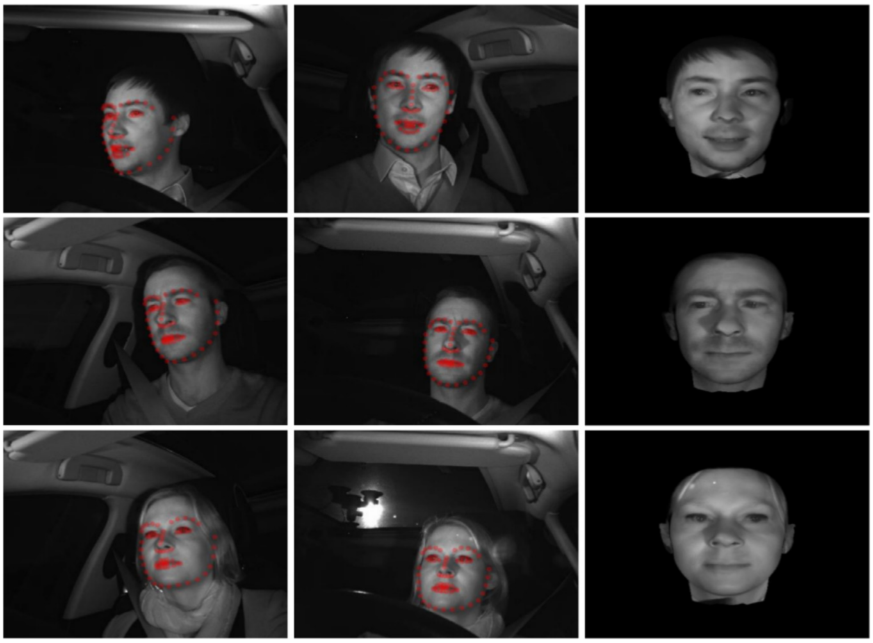


Fig. 1. Examples of simulated face images generated by [6]. **Left to right**: captured images from left and right cameras and simulated faces, respectively.

tion task for the facial expression analysis. In particular, we are interested in the problem of synthesizing photo-realistic face images in two different scenarios: first, we focus on the data augmentation process and generate face images with a desired expression category to alleviate the common problem of class imbalance and introduce more variations into training data thus resulting in more robust classification system; second, we investigate the use of domain adaptation to transform the visual appearance of the images from the target domain (simulated faces) into source domain (real face images) without affecting the face details such as identity or expression properties. Therefore, the expression recognition model learned from the labeled source domain containing real face images with arbitrary head poses can be generalized to the translated images from unlabeled target domain containing frontal simulated face images, without the need for re-training a model for target domain. Further, compared to other variants of adversarial domain adaptation methods [4,8,9], we demonstrate that a better performance can be achieved through a proposed method using in-the-wild data for emotion recognition.

2. Related work

Deep neural networks have brought stunning progresses in domain adaptation in recent years, enabling learning invariant representations across image domains. One of the interesting lines for domain adaptation is using Maximum Mean Discrepancy (MMD) as a metric to measure the domain discrepancy. Tzeng et al. [10] proposed a Deep Domain Confusion (DDC) by using domain confusion objective to ensure that domains are indistinguishable in the learned representation. They also used MMD loss as a regularizer during fine-tuning of their network. This idea has been extended by Long et al. [11,12] via embedding the joint distributions of the network's activations of multiple task-specific layers in a tensor-product Hilbert space and then aligning the joint distribution based on the MMD criterion.

2.1. Adversarial domain adaptation

Adversarial domain adaptation methods [13–15] have become an increasingly popular deep learning approach to domain adaptation, which can learn representations to overcome the distributional variations between source and target domains. Pei et al. [16] proposed the multiple adversarial networks, each for a specific class to exploit the complex multimode structures, yielding effective deep transfer learning. Motiian et al. [17] introduced a deep learning model to extend adversarial learning to exploit the label information of target samples in few-shot learning regime.

Adversarial domain adaptation approaches can be categorized into either feature-level [18,19] or image pixel-level [8,15] adaptation methods. Feature-level adversarial domain adaptation methods seek to match the feature distributions from trained network across the source and target domains. However, shortcoming of these approaches would be to ignore semantic consistency during domains distribution alignment. In this paper, we focus on the second category to perform similar distribution matching in the raw pixel space. Recently, GAN based models [1] have achieved promising results in many image synthesis applications, including image-to-image translation (pix2pix) [2] and CycleGAN [4]. For example, CycleGAN [4] can learn transformations between image domains without one-to-one mapping between domains' training data. Li et al. [20] proposed a Deep CNN model for Identity-Aware Transfer (DIAT) of facial attributes, which can be used for several facial manipulation tasks. However, for each reference attribute label, these methods train a separate CNN model to transfer the input image to the desired attribute. Unlike these approaches, we build a multi-class image-to-image translation model to generate photo-realistic face images, each having specific expression class.

Similar to our approach, Zhao et al. [21] proposed a Dual-Agent Generative Adversarial Network (DA-GAN) to synthesize realistic profile faces by augmenting samples with extreme face poses. In their framework, the simulator produces synthesis faces with arbitrary poses, which are fed to DA-GAN for realism refinement. The discriminator used in DA-GAN focuses on distinguishing the realism of synthetic profile face images from a simulator using unlabeled real data while perceiving the face identity information. Unlike this approach, we present a category-guided image translation to generate a new face image using a desired expression with the focus on the data augmentation.

More recently, StarGAN [22] was proposed for multiple image-to-image translation task. Unlike [22], we aim to improves the realism of simulated face images, while preserving face image details such as facial identity using the proposed loss functions. Perarnau et al. [9] and Lample et al. [23] proposed image manipulation methods by imposing constraints on the latent space to enforce it to be independent from the image attributes, which result in loss of image details during image synthesis. Liu et al. [24] introduced few-shot unsupervised image-to-image translation framework. In addition, they showed that their approach can be used to the few-shot image classification task. Seminal work [38] proposed a method to generate synthetic face images for data augmentation to train face expression classier. This method has been further developed in [39] by using additional face parsing loss to generate high-quality face images conditioned on the attributes of interest. However, these approaches cannot directly be used for domain adaptation from one face dataset into another.

In continue, we first introduce our proposed approach in Section 3. Then, we discuss our implementation details and experimental results in Section 4 and Section 5, respectively.

3. Proposed approach

In this work, we study two scenarios for visual domain adaptation. These scenarios include image-to-image translation from simulated face image x^t (belongs to target image space $\mathcal{X}^t$) to its counterpart realistic image (belongs to the source image space $\mathcal{X}^s$). For the first scenario, we focus on the data augmentation process and use *category-guided image translation* to generate a new face image with desired expression while preserving other face details (Fig. 2 - left). In the second scenario, we use an *unsupervised image translation* to generate realistic face images given simulated face image x^t, so that the established classifier C^s trained on the source images can be directly generalized to the generated images (Fig. 2 - right). In the following, we introduce the objectives for the proposed model optimization.

3.1. Category-guided image translation

We aim to generate photo-realistic face images with a desired expression category from simulated face images to introduce more variations into training data through data augmentation, thus resulting in more robust classification system (see Fig. 3). Having said that, there is also domain gap between simulated face images (target domain) and real face images (source domain), which needs to be addressed. Doing so, we employ a variant of AC-GAN [25] to map the target face images drawn from target input space $\mathcal{X}^t$ towards the source domain space, conditioned on the face expression categorical vector y from the set of possible facial expression label space $\mathcal{Y}$. The generated new image $x^{t\to s}$ appears to be drawn from $\mathcal{X}^s$ with the desired expression category while the face image content remains unchanged. Our domain adaptation model consists of a generator and discriminator networks. First, we train a single generator $G^{t\to s}$ with the encoder $G^{t\to s}_{enc}$ – decoder $G^{t\to s}_{dec}$ networks to produce realistic transformed image $x^{t\to s} = G^{t\to s}_{dec}\left(G^{t\to s}_{enc}(x), y\right)$ conditioned on the expression classes. During training, we randomly use a set of expression labels y to make the generator more flexible in generating images. We also train a discriminator D^s using an adversarial formulation to not only distinguish between real and fake generated images, but also to classify the image to its corresponding expression categories.

3.1.1. Adversarial loss

We adopt an adversarial loss:

$$\mathcal{L}_{adv}^{t\to s} = \mathbb{E}_{x^s}\left[\log D^s_{domain}(x^s)\right] + \mathbb{E}_{x^t,y}\left[\log\left(1 - D^s_{domain}\left(G^{t\to s}_{dec}\left(G^{t\to s}_{enc}(x), y\right)\right)\right)\right] \quad (1)$$

The term $D^s_{domain}(\cdot)$ denotes a probability distribution over source domain images. A generator used in our model is trained to maximally fool the discriminator in a *min-max* game. On the other hand, the discriminator simultaneously seeks to identify the fake source samples for each expression category.

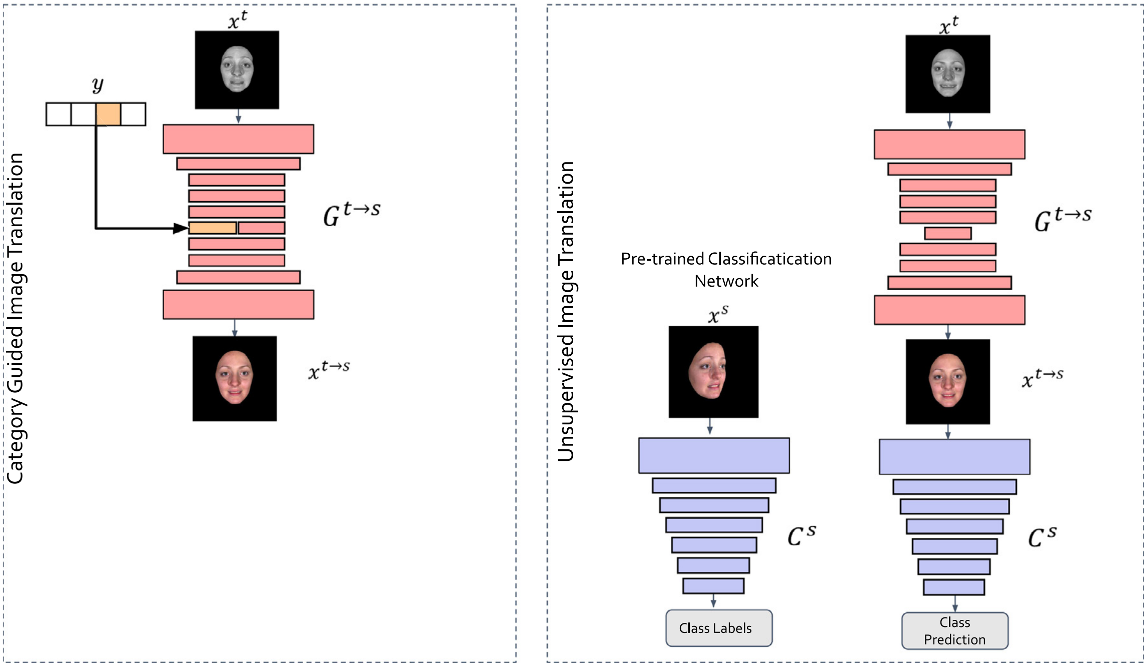


Fig. 2. Overview of proposed scenarios for domain adaptation at test time. **Left:** a category-guided image translation to generate a new face image using a desired expression with the focus on the data augmentation; **Right:** an unsupervised image translation, where the classifier C^s trained on the source images can be directly generalized to the generated images.

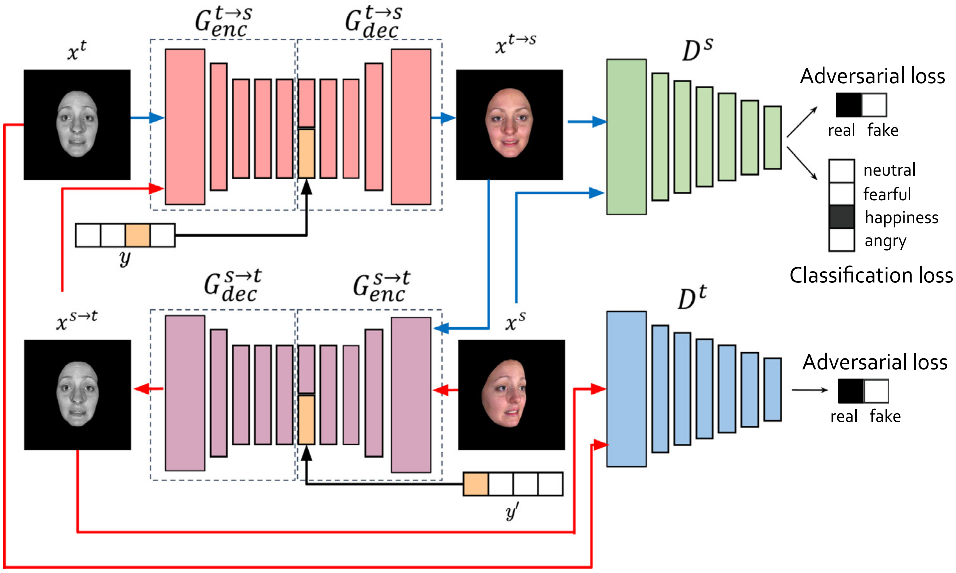


Fig. 3. Illustration of ExprADA (training stage). The proposed framework contains the paired generator and discriminator, where images of two domains can be translated bidirectionally. The blue and red arrows illustrate the data flows of target and source data, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1.2. Classification loss

We deploy a classifier by returning additional output from the discriminator to perform an auxiliary task of classifying the fake and the real source data into their respective expression categories. A classification loss of real source images $\mathcal{L}_{cls_r}$ to optimize the discriminator, is defined as follow:

$$\mathcal{L}_{cls_r}(D^s) = \mathbb{E}_{x^s,y'}\left[-\log D^s_{cls}\left(y'|x^s\right)\right] \tag{2}$$

where the term $D^s_{cls}(\cdot)$ represents a probability distribution over source domain classes computed by the discriminator and y' denotes original input labels. On the other side, a domain classification loss of fake images $\mathcal{L}_{cls_f}$ used to optimize the generator, formulated as follow:

$$\mathcal{L}_{cls_f}\left(G^{t\to s}\right) = \mathbb{E}_{x^t,y}\left[-\log D^s_{cls}\left(y|x^{t\to s}\right)\right] \tag{3}$$

where y is the source domain categories. Here, the generator is trained to minimize this objective to ensure that class-consistent images are generated.

3.1.3. Bidirectional loss

Using adversarial losses alone cannot guarantee that the trained generator can maintain the detailed image contents. Inversely, we build a source-to-target generator $G^{s\to t}$ and a target discriminator D^t, so that images of different domains can be translated bidirectionally. This pair of models are trained with a same-way adversarial loss $\mathcal{L}^{s\to t}_{adv}$ following the Eq. (1). It should be noted that for target data, since class labels are not reliable, we do not use domain classification loss. Inspired by Zhu et al. [4], we propose a bidirectional loss between transformed images, formulated using l_1 loss as follow:

$$\begin{aligned}\mathcal{L}_{bi}\left(G^{s\to t}, G^{t\to s}\right) &= \mathbb{E}_{x^s}\left[\left\|\hat{x}^s - x^s\right\|_1\right] + \mathbb{E}_{x^t}\left[\left\|\hat{x}^t - x^t\right\|_1\right]\\ \hat{x}^s &= G^{t\to s}\left(G^{s\to t}(x^s)\right),\\ \hat{x}^t &= G^{s\to t}\left(G^{t\to s}\left(x^t\right)\right)\end{aligned} \tag{4}$$

3.1.4. Reconstruction loss

Using this loss, we aim to preserve class labels-excluding face details such as facial identity before and after image translation. By doing so, we use a pixel-wise l_1 loss to enforce the facial details consistency:

$$\begin{aligned}\mathcal{L}_{re}\left(G^{s\to t}, G^{t\to s}\right) = {}& \mathbb{E}_{x^t,y}\left[\left\|x^t - G^{t\to s}_{dec}\left(G^{t\to s}_{enc}\left(x^t\right), y\right)\right\|_1\right]\\ &+ \mathbb{E}_{x^s,y'}\left[\left\|x^s - G^{s\to t}_{dec}\left(G^{s\to t}_{enc}(x^s), y'\right)\right\|_1\right]\end{aligned} \tag{5}$$

3.1.5. Overall objective

Finally, the proposed training loss for the generator G joins all the losses. Meanwhile, the discriminator D is optimized using an adversarial loss and domain classification loss for the real images:

$$\begin{aligned}\mathcal{L}\left(G^{s\to t}, G^{t\to s}\right) = {}& \mathcal{L}^{t\to s}_{adv} + \mathcal{L}^{s\to t}_{adv} + \lambda_{bi}\mathcal{L}_{bi}\left(G^{s\to t}, G^{t\to s}\right) + \lambda_{cls}\mathcal{L}_{cls_f}\left(G^{t\to s}\right)\\ &+ \lambda_{re}\mathcal{L}_{re}\left(G^{s\to t}, G^{t\to s}\right),\\ \mathcal{L}\left(D^s, D^t\right) = {}& -\mathcal{L}^{t\to s}_{adv} - \mathcal{L}^{s\to t}_{adv} + \lambda_{cls}\mathcal{L}_{cls_r}(D^s),\end{aligned} \tag{6}$$

where λ_{bi}, λ_{re} and λ_{cls} are hyper-parameters, which tune the importance of bidirectional loss, reconstruction loss and domain classification loss, respectively.

3.2. Unsupervised image translation

In this training phase, our pipeline is identical to Section 3.1, except that we remove classification stream and related losses. Therefore, the generated face images are not conditioned on the face expression categorical information. Our goal is to make the trained CNN model on the source domain generally applicable to target dataset including simulated face images, where the label information is not available.

3.2.1. Classification network established on source domain

We train an expression classifier to classify the source face images correctly. The classification loss is the negative log-likelihood of the classifier prediction C^s, given the ground truth labels of face images from the source dataset, $\mathcal{L}_{SClass} = -\sum_{i=0}^{N_s} y^s_i \log \hat{y}^s_i$, where y^s_i and $\hat{y}^s_i$ are the label and prediction for the image $x^s_i \in \mathcal{X}^s$, respectively. Our expression recognition network is detached from the learning of our domain adaptation model. Compared with the integrated approaches, an independent classifier enables much more flexibility and allows us to take advantage of well-known pre-trained models. In particular, we use VGG-face model [26], pre-trained on the large face recognition dataset and can positively improves the expression recognition accuracy. Doing so, we can directly apply the established classification model C^s on $x^{t\rightarrow s}$ without re-training.

4. Implementation details

All networks are trained using Adam optimizer [27] $(\beta_1 = 0.5, \beta_2 = 0.999)$ and with a base learning rate of 0.0001. We linearly decay learning rate after the first 100 epochs. We use a horizontal flipping for data augmentation. The input image size and the batch size are set to 128 × 128 and 8 for all experiments, respectively. In practice, to stabilize the training of the GAN, the negative log likelihood in $\mathcal{L}_{adv}$ was replaced by a least-square loss. The hyper-parameters are set as: $\lambda_{bi} = 10$ and $\lambda_{re} = 10$ and $\lambda_{cls} = 1$, respectively. The whole model is implemented using PyTorch on a single NVIDIA GeForce GTX 1080 Ti.

4.1. Networks architectures

Tables 1 and 2 demonstrate the detailed network architectures of our proposed ExprADA. In particular, both generators have the same architecture. Regarding the generator's decoder, we use sub-pixel convolution instead of transposed convolution followed by instance normalization [28].

5. Experimental results

5.1. Datasets

5.1.1. Driver emotion dataset

We present the driver emotion dataset; a dataset of images that captured the driver's emotion using a near-infrared (NIR) camera in a real driving environment. The drivers show six basic facial expressions including anger, disgust, fear, happiness, sadness, surprise plus neutral faces. In our experiments, we use video frames (peak expressions) of 20 subjects for training and validation, and 6 subjects for the test, respectively.

5.1.2. BU-3DFE

The Binghamton University 3D Facial Expression Database (BU-3DFE) [29] contains 3D models from 100 subjects. Each subject has 13 different poses ranging from −90 to 90° in 15° steps. The subjects show a neutral face as well as six basic facial expressions and at four different intensity levels.

5.1.3. KDEF

The KDEF dataset [30] is a multi-view emotion dataset that contains 35 females and 35 males displaying seven discrete expressions (anger, fear, disgust, happiness, sadness, surprise and neutral) and each expression comes with 5 different yaw angles.

5.1.4. MMI

The MMI dataset [31] consists of 236 image sequences from 31 subjects, from which 208 sequences captured in frontal view were selected in our experiments. Each sequence is annotated as one of

Table 1
The generator architecture. There are some notations; n_y denotes the dimension of one-hot vector. IN and RB denote instance normalization and residual block, respectively.

Part	Layers	Input Size → Output Size	Filter Size	Stride	Padding
Encoder	Conv+IN+ReLU	$(h, w, 6) \rightarrow (h, w, 64)$	7 × 7	1	3
	Conv+IN+ReLU	$(h, w, 64) \rightarrow \left(\frac{h}{2}, \frac{w}{2}, 128\right)$	4 × 4	2	1
	Conv+IN+ReLU	$\left(\frac{h}{2}, \frac{w}{2}, 128\right) \rightarrow \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	4 × 4	2	1
	Conv+IN+ReLU	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \rightarrow \left(\frac{h}{8}, \frac{w}{8}, 512\right)$	4 × 4	2	1
	Conv+IN+ReLU	$\left(\frac{h}{8}, \frac{w}{8}, 512\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 1024\right)$	4 × 4	2	1
Bottleneck	RB:Conv+IN+ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 1024\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 1024\right)$	3 × 3	1	1
	RB:Conv+IN+ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 1024\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 1024\right)$	3 × 3	1	1
	RB:Conv+IN+ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 1024\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 1024\right)$	3 × 3	1	1
	RB:Conv+IN+ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 1024\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 1024\right)$	3 × 3	1	1
	RB:Conv+IN+ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 1024\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 1024\right)$	3 × 3	1	1
	RB:Conv+IN+ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 1024\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 1024\right)$	3 × 3	1	1
Decoder	Sub-Pixel Conv+IN+ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 1024 + n_y\right) \rightarrow \left(\frac{h}{8}, \frac{w}{8}, 512\right)$	3 × 3	2	1
	Sub-Pixel Conv+IN+ReLU	$\left(\frac{h}{8}, \frac{w}{8}, 512\right) \rightarrow \left(\frac{h}{4}, \frac{w}{4}, 256\right)$	3 × 3	2	1
	Sub-Pixel Conv+IN+ReLU	$\left(\frac{h}{4}, \frac{w}{4}, 256\right) \rightarrow \left(\frac{h}{2}, \frac{w}{2}, 128\right)$	3 × 3	2	1
	Sub-Pixel Conv+IN+ReLU	$\left(\frac{h}{2}, \frac{w}{2}, 128\right) \rightarrow (h, w, 64)$	3 × 3	2	1
	Image output:Conv+Tanh	$(h, w, 64) \rightarrow (h, w, 3)$	7 × 7	1	3
	Side output:Conv+Tanh	$(h, w, 64) \rightarrow (h, w, 3)$	7 × 7	1	3

Table 2
The discriminator architecture. FC and m denote fully connected layer and the number of target classes, respectively.

Part	Layers	Input Size → Output Size	Filter Size	Stride	Padding
Discriminator	Conv+Leaky ReLU	$(h, w, 6) \rightarrow \left(\frac{h}{2}, \frac{w}{2}, 64\right)$	4 × 4	2	1
	Conv+Leaky ReLU	$\left(\frac{h}{2}, \frac{w}{2}, 64\right) \rightarrow \left(\frac{h}{4}, \frac{w}{4}, 128\right)$	4 × 4	2	1
	Conv+Leaky ReLU	$\left(\frac{h}{4}, \frac{w}{4}, 128\right) \rightarrow \left(\frac{h}{8}, \frac{w}{8}, 256\right)$	4 × 4	2	1
	Conv+Leaky ReLU	$\left(\frac{h}{8}, \frac{w}{8}, 256\right) \rightarrow \left(\frac{h}{16}, \frac{w}{16}, 512\right)$	4 × 4	2	1
	Conv+Leaky ReLU	$\left(\frac{h}{16}, \frac{w}{16}, 512\right) \rightarrow \left(\frac{h}{32}, \frac{w}{32}, 1024\right)$	4 × 4	2	1
	Conv+Leaky ReLU	$\left(\frac{h}{32}, \frac{w}{32}, 1024\right) \rightarrow \left(\frac{h}{64}, \frac{w}{64}, 2048\right)$	4 × 4	2	1
Outputs	**Output Layer**:Conv	$\left(\frac{h}{64}, \frac{w}{64}, 2048\right) \rightarrow \left(\frac{h}{64}, \frac{w}{64}, 1\right)$	3 × 3	1	1
	Output Layer:FC	$\left(\frac{h}{64}, \frac{w}{64}, 2048\right) \rightarrow FCm$	-	-	-

the six basic facial expressions. Each sequence starts from a neutral expression, reaches a peak phase in the middle of a sequence, and ends up with a neutral expression again. Since there is no annotation for the peak frames, we chose three frames in the middle of each sequence as the peak frames and labeled them with the provided labels for a related sequence, resulting in a total of 624 images for our experiments.

5.2. Baselines

We study the performance of our two baseline methods:

- ***ExprADA$_{CGIT}$*** : Our proposed model using category-guided image translation;
- ***ExprADA$_{UIT}$*** : Our proposed model using unsupervised image translation.

5.3. Qualitative evaluation

For the qualitative evaluation, we consider a visual domain adaptation experiment on the BU-3DFE dataset [29]. For this purpose, a simulated frontal face image is generated using standard rendering pipeline [7] from images of two camera views. We use a 3D Morphable Model using bilinear face model [7] to render a simulated face image. The results of our proposed *ExprADA$_{CGIT}$* have been compared with the SimGAN method [8] (see Fig. 4). From the results, it is obvious that our facial expression transfer results are more realistic and facial expression is well distinguishable demonstrating the importance of classification objective function.

5.3.1. Comparison with SimGAN

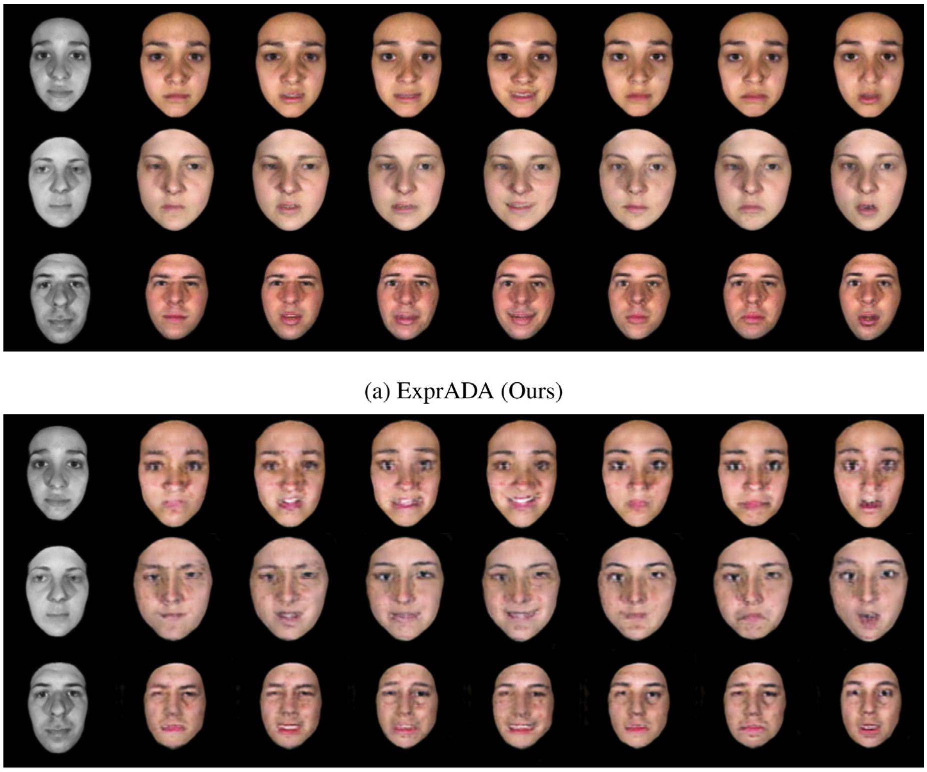
SimGAN [8] considers learning from simulated and unlabeled real images through adversarial training. However, we present a new objective to prevent the image content distortion during the image translation and to preserve face image details e.g., face pose and identity, whereas SimGAN was proposed for simpler scenarios e.g., eye image refinement.

5.4. Quantitative evaluation

We perform quantitative evaluation of our *ExprADA$_{UIT}$* baseline on generating images to tackle unsupervised domain adaptation task. Doing so, we used the VGG-Face model [26] as the backbone of our classifier, which was further fine-tuned on source domain samples from the driver emotion dataset. We trained our classifier with subject-independent subsets (20 subjects for training and validation and 6 subjects for the test). Then, we performed our unsupervised image translation scheme (*ExprADA$_{UIT}$*) by improving realism to the simulated faces (target domain) to generate transformed images $x^{t\to s}$ and obtain the expression recognition result. We compared our method with the state-of-the-art unsupervised domain adaptation schemes including CycleGAN [4], StarGAN [22], SimGAN [8], Adversarial Discriminative Domain Adaptation (ADDA) [19], Generate To Adapt (GTA) [32] and Cycle-Consistent Adversarial Domain Adaptation (CyCADA) [5], which are formulated as image-to-image translation task. As can be seen in Fig. 5, *ExprADA$_{UIT}$* achieves the highest classification accuracy, demonstrating that our method could generate the most realistic expressions among all the methods compared. In addition, we estimate the t-SNE components after using our domain adaptation model (Fig. 5). The different colored points represent the clusters in the embedding space, each corresponding to a different expression category.

5.4.1. Data augmentation

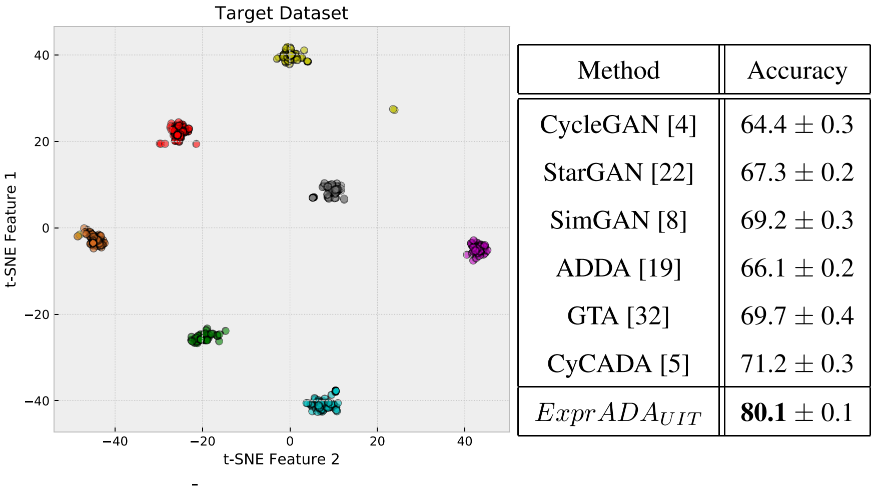
We also demonstrate quantitatively the usefulness our *ExprADA$_{CGIT}$* baseline in generating photo-realistic face images



(a) ExprADA (Ours)

(b) SimGAN [8]

Fig. 4. Comparison of facial expression transfer task by (a) our *ExprADA$_{CGIT}$* and (b) SimGAN [8] on the BU-3DFE dataset [29]. **Left to right**: input simulated face and seven exhibited expressions including *angry, disgusted, fearful, happiness, neutral, sadness* and *surprised*, respectively.



Method	Accuracy
CycleGAN [4]	64.4 ± 0.3
StarGAN [22]	67.3 ± 0.2
SimGAN [8]	69.2 ± 0.3
ADDA [19]	66.1 ± 0.2
GTA [32]	69.7 ± 0.4
CyCADA [5]	71.2 ± 0.3
$ExprADA_{UIT}$	**80.1** ± 0.1

Fig. 5. **Left:** The t-SNE components of the target embedding of the driver emotion dataset after domain adaptation. **Right:** Results (expression recognition accuracy %) on the driver emotion dataset for unsupervised domain adaptation. (For interpretation of the references to color in text, the reader is referred to the web version of this article.)

Table 3
Results (expression recognition accuracy %) on the driver emotion dataset using generated images for data augmentation.

Method	1K	2K	5K
CycleGAN [4]	66.8 ± 0.4	68.5 ± 0.3	71.2 ± 0.1
StarGAN [22]	69.2 ± 0.2	72.7 ± 0.2	75.4 ± 0.3
SimGAN [8]	72.5 ± 0.1	74.6 ± 0.3	76.8 ± 0.2
***ExprADA*$_{CGIT}$**	**82.2** ± 0.3	**84.5** ± 0.4	**86.9** ± 0.6

controlled by the expression category (see Table 3). Doing so, we augment real images from the driver emotion dataset with the images generated by *ExprADA*$_{CGIT}$ and then compare with other category-guided image translation methods to train our backbone expression classifier. The purpose of this experiment is to enrich source dataset further to improve the expression recognition performance. In particular, from each of the six expression categories, we generate 1K, 2K and 5K images, respectively. The performance of all methods starts to become saturated when more images (30K) are used. We achieved a higher recognition accuracy value using the images generated from *ExprADA*$_{CGIT}$ than other category-guided image translation methods e.g., CycleGAN. This shows that our model has learned to generate more diverse realistic images.

5.4.2. Pose normalization & improving realism

We tested our unsupervised image translation method on the BU-3DFE dataset [29] and compared with two GAN-based domain adaptation techniques for realism refinement of the simulated face images. For the first experiment (first row in Table 4), we only evaluated the performance of the trained classifier for non-frontal face images without pose normalization or rendering simulated faces. We performed 5-fold cross validation using 100 subjects. Training data includes images of 80 subjects, while test data includes images of 20 subjects with varying poses. For our classifier, we fine-tuned VGG-Face model on the real frontal face images from BU-3DFE dataset. Then, we rendered frontal face images from test images of non-frontal faces and conducted experiments using different image translation methods. It can be observed from Table 4 that image translation on the simulated frontal faces contributes significantly to the expression recognition performance for the non-frontal faces (ranging from 15 to 45° in 15° steps) and brings additional gains.

5.4.3. Facial expression transfer

To further evaluate the generality of our proposed method, we tested transferring a subject's facial expression to different expressions using our proposed category-guided image translation *ExprADA*$_{CGIT}$. Facial expression transfer between two unpaired images is a challenging task due to the individual variances. For this experiment, we do not use simulated face images, but rather the source and target domains represent different facial expressions. We evaluate our expression transfer framework on two facial expression datasets: KDEF [30] and MMI [31]. For both datasets, we followed the same settings and trained a facial expression classifier with (90%/10%) splitting for training and test sets, respectively. We opt a ResNet-50 [33] as our expression classifier and reported the average accuracy of both datasets and for all expressions in Tables 5 and 6. We then trained each of baseline expression transfer models including CycleGAN, IcGAN and StarGAN using the same training set and performed image-to-image translation on the same unseen test set. Finally, we classified the expression of these generated images using the above-mentioned classifier. Our ExprADA method, showing a close result to real images, generates photo-realistic expression transfer results while preserving identity information.

Table 4
Results (expression recognition accuracy %) on the BU-3DFE dataset for unsupervised domain adaptation and varying face pose angles.

Method	±15	±30	±45
Test Set (without domain adaptation)	70.9 ± 0.2	65.9 ± 0.2	59.3 ± 0.2
CycleGAN [4]	71.6 ± 0.3	67.3 ± 0.3	61.5 ± 0.2
SimGAN [8]	71.8 ± 0.2	67.7 ± 0.2	62.3 ± 0.2
***ExprADA*$_{UIT}$**	**73.2** ± 0.1	**69.4** ± 0.1	**65.1** ± 0.1

Table 5
Quantitative results of expression transfer (recognition accuracy %) on the KDEF dataset.

Method	Accuracy
Real Images	87.1 ± 0.1
CycleGAN	82.4 ± 0.4
IcGAN	78.3 ± 0.3
StarGAN	84.5 ± 0.2
ExprADA$_{CGIT}$ w/o $\mathcal{L}_{bi}$	82.8 ± 0.2
ExprADA$_{CGIT}$ w/o $\mathcal{L}_{re}$	85.3 ± 0.3
ExprADA$_{CGIT}$	86.9 ± 0.2

Table 6
Quantitative results of expression transfer (recognition accuracy %) on the MMI dataset.

Method	Accuracy
Real Images	72.25 ± 0.1
CycleGAN	63.5 ± 0.2
IcGAN	60.1 ± 0.2
StarGAN	66.8 ± 0.3
ExprADA$_{CGIT}$ w/o $\mathcal{L}_{bi}$	67.1 ± 0.1
ExprADA$_{CGIT}$ w/o $\mathcal{L}_{re}$	68.5 ± 0.2
ExprADA$_{CGIT}$	70.7 ± 0.1

5.4.4. Ablation study

We investigated the sensitivity of the results for each component of our objective function including bidirectional loss and reconstruction loss, respectively. For each ablation experiment, we turn-off one of the loss terms in the final objective function and then we generate the related results for evaluation. Our ExprADA trained with each of the proposed loss terms, resulting in a performance gain in expression recognition accuracy for the translated images (Tables 5 and 6).

5.4.5. Comparison with other image-to-image translation methods

Compared to other adversarial image-to-image translation methods such as CycleGAN [4] and DIAT [20], which can be used for facial manipulation task, our method is more efficient as it formulated a multi-class image-to-image translation task in a single model while other approaches need to train a separate model for each reference attribute to transfer the input image to the desired attribute.

More recently, StarGAN [22] and STGAN [34] were proposed to manipulate multiple attributes for image-to-image translation task. STGAN [34] incorporated difference attribute vector and selective transfer units (STUs) to perform arbitrary image attribute editing. Unlike these approaches, our objective is to improve the realism of simulated face images using domain adaptation. We propose a new objective to prevent the content distortion during the image translation and to preserve face image details. In addition, our method involves latent representation using an encoder-decoder architecture and models the relation between the latent representation and the facial expressions.

Relevant recent work [35–37] proposed facial attribute transfer methods with the objective of generating visually more pleasing results. For example, Yin et al. [35] proposed GeoGAN, a geometry-aware flow representation to address the problem of instance-level facial attribute transfer. Tang et al. [36] presented a Multi-Channel Attention Selection GAN (SelectionGAN) to address image synthesizing task by conditioning on a reference image and a target segmentation map. Chen et al. [37] presented a face attribute manipulation method, where they decompose semantic components from high-level attributes to control the face attribute transfer through the user. In comparison to these methods, our method is capable to model the transformation between simulated face images and real images, despite of the large gap between the source and target face images without using additional information e.g., segmentation map.

6. Conclusion

In this paper, we propose a method to simultaneously achieve image-to-image translation, discriminative modeling, and adversarial domain adaptation. More importantly, we present a new objective to prevent the content distortion during the image translation and to preserve face image details. We present two models within our framework, where one uses multi-domain image-to-image mapping with a focus on the data augmentation process to alleviate the common problem of class imbalance, and the other transforms the visual appearance of the images between domains in an unsupervised way. We have shown the superiority of our approach over existing visual domain adaptation methods using experiments on face datasets recorded in real world conditions. The results demonstrate the generality of our domain adaptation model. Furthermore, during evaluation, our approach does not need re-training a model (initially trained on the source data) for a new target dataset, which is a necessary aspect when deploying such models in practice.

6.1. Limitations

While our adversarial domain adaptation method could transfer appearance changes across face image domains, it requires access to many images in both source and target image domains at training time. The generalization ability from a few samples of a new image domain based on prior knowledge is entirely beyond the scope of this paper. We argue that this issue limits the use of ExprADA for the real scenario, where we need an unsupervised image-to-image translation algorithm that works on previously unseen target image domains that are available, at test time, only by a few example images. A possible future extension of this work would be to support few-shot generalization by leveraging few images of the target domain given at test time.

Acknowledgment

This work is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 688900 (ADAS&ME project - http://www.adasandme.com).

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[2] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[3] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: International Conference on Machine Learning, 2017, pp. 1857–1865.

[4] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2223–2232.

[5] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, Cycada: cycle-consistent adversarial domain adaptation, in: Proceedings of the 35th International Conference on Machine Learning, 2018.

[6] D. Engin, C. Ecabert, H.K. Ekenel, J.-P. Thiran, Face frontalization for cross-pose facial expression recognition, in: 2018 26th European Signal Processing Conference (EUSIPCO), IEEE, 2018, pp. 1795–1799.

[7] D. Vlasic, M. Brand, H. Pfister, J. Popović, Face transfer with multilinear models, ACM Trans. Graph. 24 (3) (2005) 426–433.

[8] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 3, 2017, p. 6.

[9] G. Perarnau, J. van de Weijer, B. Raducanu, J.M. Álvarez, Invertible conditional GANs for image editing, NIPS Workshop on Adversarial Training, 2016.

[10] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: maximizing for domain invariance, arXiv:1412.3474 (2014).

[11] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, 2015, pp. 97–105.

[12] M. Long, H. Zhu, J. Wang, M.I. Jordan, Deep transfer learning with joint adaptation networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2208–2217.

[13] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, J. Mach. Learn. Res. 17 (1) (2016). 2096–2030.

[14] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37, JMLR. org, 2015, pp. 1180–1189.

[15] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, D. Krishnan, Unsupervised pixel-level domain adaptation with generative adversarial networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2017, p. 7.

[16] Z. Pei, Z. Cao, M. Long, J. Wang, Multi-adversarial domain adaptation, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[17] S. Motiian, Q. Jones, S. Iranmanesh, G. Doretto, Few-shot adversarial domain adaptation, in: Advances in Neural Information Processing Systems, 2017, pp. 6670–6680.

[18] R. Volpi, P. Morerio, S. Savarese, V. Murino, Adversarial feature augmentation for unsupervised domain adaptation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5495–5504.

[19] E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in: Computer Vision and Pattern Recognition (CVPR), vol. 1, 2017, p. 4.

[20] M. Li, W. Zuo, D. Zhang, Deep identity-aware transfer of facial attributes, arXiv:1610.05586 (2016).

[21] J. Zhao, L. Xiong, P.K. Jayashree, J. Li, F. Zhao, Z. Wang, P.S. Pranata, P.S. Shen, S. Yan, J. Feng, Dual-agent GANs for photorealistic and identity preserving profile face synthesis, in: Advances in Neural Information Processing Systems, 2017, pp. 66–76.

[22] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.

[23] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, et al., Fader networks: manipulating images by sliding attributes, in: Advances in Neural Information Processing Systems, 2017, pp. 5969–5978.

[24] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, J. Kautz, Few-shot unsupervised image-to-image translation, arXiv:1905.01723 (2019).

[25] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2642–2651.

[26] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition., in: BMVC, vol. 1, 2015, p. 6.

[27] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980 (2014).

[28] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, arXiv:1607.06450 (2016).

[29] L. Yin, X. Wei, Y. Sun, J. Wang, M.J. Rosato, A 3d facial expression database for facial behavior research, in: Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on, IEEE, 2006, pp. 211–216.

[30] D. Lundqvist, A. Flykt, A. Öhman, The Karolinska Directed Emotional Faces (KDEF), 91, 1998, p. 630. CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet.

[31] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: 2005 IEEE International Conference on Multimedia and Expo, IEEE, 2005, pp. 5–pp.

[32] S. Sankaranarayanan, Y. Balaji, C.D. Castillo, R. Chellappa, Generate to adapt: aligning domains using generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8503–8512.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[34] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, S. Wen, Stgan: a unified selective transfer network for arbitrary image attribute editing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3673–3682.

[35] W. Yin, Z. Liu, C.C. Loy, Instance-level facial attributes transfer with geometry-aware flow, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9111–9118.

[36] H. Tang, D. Xu, N. Sebe, Y. Wang, J.J. Corso, Y. Yan, Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2417–2426.

[37] Y.-C. Chen, X. Shen, Z. Lin, X. Lu, I. Pao, J. Jia, et al., Semantic component decomposition for face attribute manipulation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9859–9867.

[38] B. Bozorgtabar, M.S. Rad, H.K. Ekenel, J. Thiran, Using photorealistic face synthesis and domain adaptation to improve facial expression analysis, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 2019, pp. 1–8.

[39] B. Bozorgtabar, M.S. Rad, H.K. Ekenel, J. Thiran, Learn to synthesize and synthesize to learn, Comput. Vis. Image Und. (2019).

Dr. Behzad Bozorgtabar is a scientist at Signal Processing Lab (LTS5) at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. He is leading computer vision team of LTS5 for a number of projects focusing on machine learning and computer vision. His previous role as a postdoctoral Researcher at IBM has seen him lead development of novel deep learning based methods in medical image analysis leading to peer-reviewed papers and patents. His research interests lie in the general area of machine learning, medical image analysis and computer vision, particularly in deep representation learning.

Dwarikanath Mahapatra is currently a Senior Scientist at the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. Earlier, he has served as a Research Staff Member at IBM Research Australia. He obtained his Ph.D. from the National University of Singapore in 2011, and worked as a post-doctoral research fellow at the ETH Zurich, Switzerland from 2011–15. Dwarikanath's research interests are mainly in medical image analysis, machine learning, deep learning, decision support systems and computer aided diagnosis. He also explores other aspects of computer vision such as object detection and tracking, surveillance and image classification using deep neural networks.

Prof. Jean-Philippe Thiran is the head of the Signal Processing Lab (LTS5) at the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. His main research focus is image analysis and computer vision, with contributions in the image and video analysis for facial expression recognition. More specifically, over the last couple of years the group of Prof. Thiran has developed a specific expertise in developing advanced models of deep learning and computer vision, for affective computing and image analysis. Important projects have been initiated in this domain, including emotion recognition in video sequences using deep learning (EU H2020 ADAS&ME project), automatic detection and recognition of objects in video sequences (two Swiss KTI projects and one industrial research mandate) and medical image classification (skin cancer lesion detection and classification Swiss KTI project), etc. He also served as the project leader in collaboration with PSA Peugeot CitroŁn, where LTS5 adapted a facial emotion detection device for use in a car based on the analysis of facial expressions.