Domain Adaptation for Emotion Detection from Face Expressions

Muhammad Sohaib Khalid
ITU
ASTP, Lahore
msds19096@itu.edu.pk

Amna Shahbaz
ITU
ASTP, Lahore
msds19060@itu.edu.pk

Jawad Tariq
ITU
ASTP, Lahore
msds19038@itu.edu.pk

Asif Ejaz
ITU
ASTP, Lahore
msds19010@itu.edu.pk

Muhammad Taimur Adil
ITU
ASTP, Lahore
msds19040@itu.edu.pk

Abstract

Facial expression recognition is a challenging task in the domain of deep learning. Ensuring good performance is the cornerstone of the research problem related to transfer learning and domain adaptation. In the facial expression recognition task, a well-performing deep neural network model trained on one dataset (source domain) typically underperforms when subjected to a different but relevant dataset (target domain). This under-performance is the result of facial expression variation across different domains, even in the case of same expression. In our case, we have Western faces as source domain and Pakistani faces as target domain. Further problem arises in using Pakistani faces is the lack of samples, and co-existence of different ethnicity in the region. In our approach, we used different approaches to address the problem of domain adaptation. WGAN failed to produce expected results, whereas Cycle-GAN and feature-based domain adaptive models performed reasonably good on our data. We were able to get very close to the baseline results using domain adaptation approaches used and even got slightly better accuracy for VGG16 in one of the approach.

1. Introduction

Facial expression analysis is a domain of interest in deep and transfer learning. The way in which human beings express their emotions has always been a topic of interest in psychological study. External factors have always proven to play a major role as different groups of people show different reaction to the same stimuli, and the way of their emotional expression sets them apart from any other group of interest. The common factors in this change of expressions are observed as gender, cultural influences, age, and the environment.

In this problem, our main focus is the facial expression recognition, as the reaction to a stimulus is first observed through face. As cultural factors influence the display of emotions, and we are dealing with the Indo-Pak ethnicity, we show how an ethnicity-specific classifier is built using the target domain data that is unlabeled. If the training data is an unbiased sample of an underlying distribution, then the learned classification function will make accurate predictions for new samples. However, if the training data is not an unbiased sample, then there will be differences between how the training data is distributed and how the test data is distributed. This becomes the major challenge in domain adaptation. The assumptions made by our proposed method will not be based on specific features used for emotional representation, so the proposed solution can be applied to a different kind of data also. We propose to use Cycle-GAN for the purpose of mapping source domain to target domain, and conduct the comparative analysis with results of conventional GAN model. Furthermore, we applied feature-space domain adaptation to our problem, which resulted in significant improvements.

2. Related Work

There were two commonly used techniques in the observed literature to achieve the similar results related to the problem being addressed. First is a state-of-the-art approach based on a frame-level technique, which is the facial expression analysis using single frame. For this approach, the pre-requisite is to perform image registration on facial images, and then drawing out the representative features of the image i.e. geometrical features. These extracted features are then given as an input to the classifier. This method is observed to show good results concerning frontal images, but it failed to capture the diversity of emotions.

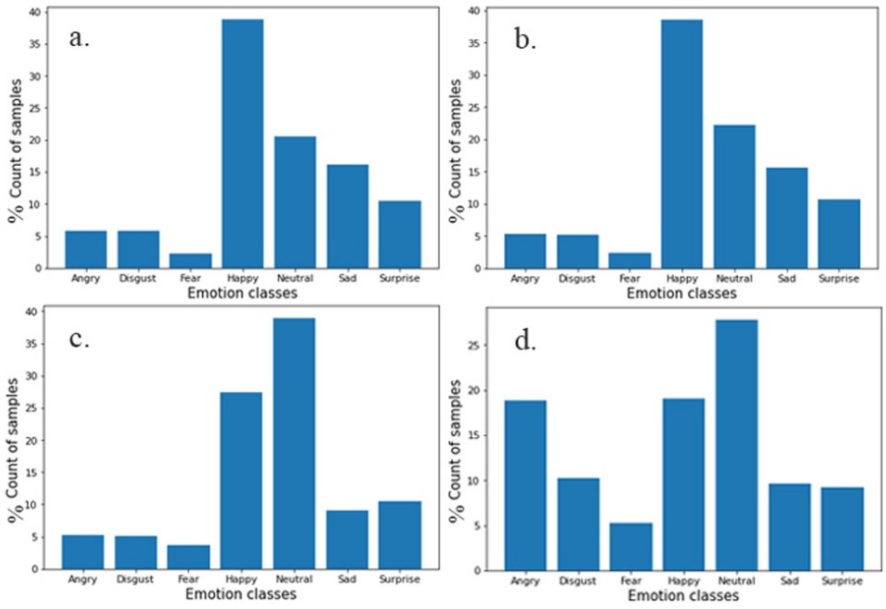

Figure 1. Percentage count of Samples in Source and Target datasets. **(a)** Percetange count of Samples in Source training set, **(b)** Percetange count of Samples in Source testing set, **(c)** Percetange count of Samples in Target Dataset 1, **(d)** Percetange count of Samples in Target Dataset 2

Figure 2. Images for different emotions from source and target domain

Second most commonly seen approach is of domain adaptation, along with other transfer learning techniques which depend on the type of information. These techniques include feature and parameter transfer. Former technique deals with the shared feature space representing both target domain and source domain, whereas the latter deals with finding the shared parameters between both domains.

In cases where the data representing the target domain is unlabeled, instance-based domain adaptation techniques are used. This is made possible through the observation of target domain's distribution. Using the information extracted from the distribution, certain weight is assigned to the samples of source domain while training, and certain samples are estimated.

In [5], an Identity-free conditional Generative Adversarial Network was proposed, known as IF-GAN. It was introduced for the purpose of reducing variations among the subjects for recognition of facial expression. A neutral face can be transformed to an expressive one using conditional generation, for a given input image. The neutral face subjected to transformation here takes the average expression of the images the model trained on.

[7] uses Cycle-GAN for image to image translation when there are no given paired samples for training. A mapping is learned from source domain to target domain, such that the distribution of the images being mapped could not be distinguished from that of target domain. This approach uses the concept of pix2pix framework [4] by Isola which learns the mapping from input to output image using conditional GAN. Conditional GANs also learn the loss function for the purpose of training the specific mapping.

In an approach suggested by Leon A. Gatys [3], neutral style transfer is using for the purpose of image translation which produces a new image by combining the style of one

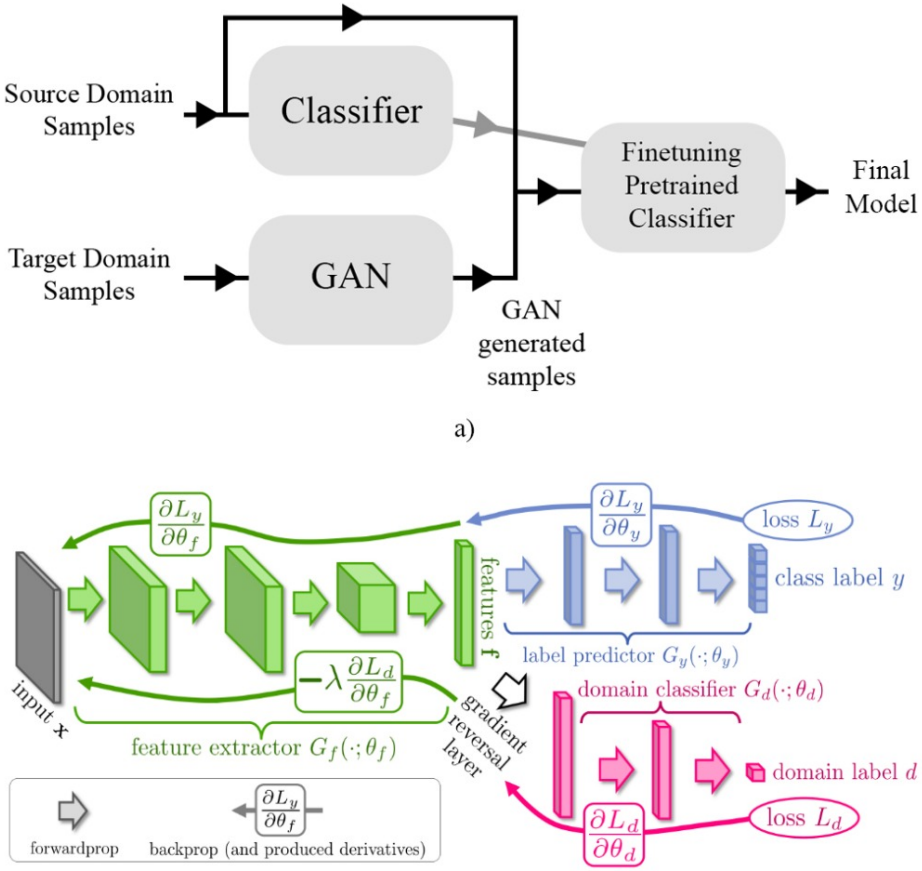

Figure 3. Different Domain Adaptation Process. **(a)** Input Space Domain Adaptation, **(b)** Feature Space Domain Adaptation (Followed from Domain-Adversarial Training of Neural Networks published in Journal of Machine Learning Research 17 (2016)

image by the contents of another.

Our approach does not solely rely on learning the mapping between two specific images but between the two domains, and the constraint of being given paired images will not restrict us in learning. Three base papers were concerned before finalizing the approach proposed by us. [6] introduces an unsupervised domain adaptation method, which is especially suitable for unlabeled small target dataset. They train a GAN on the target dataset and use the GAN generated samples to fine-tune the model pretrained on the source dataset, whereas [2] proposes an domain adaptation approach that promotes the emergence of features that are (i) discriminative for the main learning task on the source domain and (ii) indiscriminate with respect to the shift between the domains. On the other hand, in [1] the authors present an approach that learns, in an unsupervised manner, a transformation in the pixel space from one domain to the other. Their GAN-based method adapts source-domain images to appear as if drawn from the target domain.

3. Methods

WGAN lies in the category of input space domain adaptation where the classifier trained on source domain is fine-tuned using the generated samples so that it can adapt to target domain. WGAN in the first approach basically generates new target domain like samples from random noise input.

What makes WGAN different from GAN is 1-Wasserstein distance, which is used to measure the difference between the source and target distributions. In WGAN, we can train the discriminator to convergence. It would also remove needing to balance generator updates with discriminator updates, which feels like one of the big sources for training GANs. When learning generative models, we assume the data we have comes from some unknown distribution P. We want to learn a distribution that approxi-

mates P. This approach was later discontinued.

Proceeding to CycleGAN, the main objective us to learn the mapping between the source and target domains. If we consider our source domain as X and target as Y, two mappings are learned i.e. from X to Y and from Y to X. There are also the adversarial discriminators called DY and DX. The purpose of DX is to differentiate between the translated images and those from the domain being mapped. Adversarial loss is used here for the purpose of complementing the generated images distribution to the distribution of data.

A loss termed as cycle consistency loss is also used for the purpose of intercepting the contradiction between both the learned mappings.

Specific case of adversarial autoencoders are seen in this method where an image is mapped to itself. This mapping is seen as the image's translation in the other domain, after learning the common distribution between domains.

For feature space domain adaptation, the architecture followed by us consists of label predictor and a deep feature extractor. The two functionally combine to represent a feed-forward structure. Domain classifier is also attached for the purpose of attaining unsupervised domain adaptation. This extractor is attached to the feature extractor, and a negative constant is multiplied with the gradient during training based on back propagation. If this condition is ignored, then there is standard training where the label prediction loss is minimized. The technique of gradient reversal makes sure that the distribution of features is undisguisable between both the domains.

4. Data

We used RAF_DB dataset as source dataset in experimentation. There are other datasets available on internet for public use like FER-2013, CK+ etc. but almost all of them have grayscale images. Our goal was adapting classifier trained on one domain to another. RAF_DB dataset has both grayscale and colored Western Face expression images. Target domain is of Pakistani faces and in-order to properly adopt classifier trained on Western faces to Pakistani faces, color information was one of the key feature. There are 12271 images in source training dataset and 3067 images in source testing dataset. The percentage count of samples in each emotion class in RAF-DB training set and testing set is given in Figure 1(a) and Figure 1(b) respectively.

One of the major challenge faced while working on this problem was unavailability of any public Pakistani facial expressions dataset. We collected Pakistan facial expression dataset on our own. We were able to collect in total 4.2k+ images with different facial expressions from different internet sources. These images were further divided into two datasets for use different use in experimentation. The data was fractionated into seven classes, having high expression variability. Noise removal was ensured by using frontal images with clarity of expressions, and discarding any sample which could affect the model's performance in a negative manner. The sample images were also resized to 224x224, as required by implemented model for training purposes. The percentage count of samples for each emotion in Target dataset 1 is provided in Figure 1(c). The percentage count of samples for each emotion in Target dataset 2 is provided in Figure 1(d).

It can be seen in fig 1 that there exists a great imbalance of classes in source and target dataset. This will affect our results significantly. In this project, we didn't deal with this class imbalance as it was not our goal. But this can be done to improve accuracy for domain adaptation approaches used. Apart from this, there is one another thing that made domain adaptation task difficult for us. It is the presence of different groups present in source dataset while they are just some particular age group present in target dataset.

5. Experiments and Results

We performed a series of experiments. We have used two target datasets in our experimentation. The first dataset is used in domain adaptation process and second dataset is kept unseen in all the ways for testing purposes. This was done to ensure model performance consistency on target domain. We used two classifiers in our experimentation. One is VGG16 pre-trained on ImageNet Dataset and second is ResNET18 pre-trained on ImageNet Dataset. These classifiers were trained on source domain and their accuracies on source domain are given in Table 1.

Classifier	Source Domain Training Accuracy	Source Domain Testing Accuracy
VGG16	94.8	79.45
ResNET18	92.3	80.17

Table 1. Source dataset accuracy results

In our experimentation, we first evaluated our classifiers (VGG16 and ResNET18) on target domain without doing any kind of domain adaptation. The baseline results for the classifiers used are provided in Table 2.

Classifier	Target Dataset 1 Accuracy (Unseen)	Target Dataset 2 Accuracy (Unseen)
VGG16	50.92	37.51
ResNET18	50.75	33.51

Table 2. Baseline results

Then in our next experiment, we fine-tuned our classi-

fiers directly on target domain to get an upper bound of accuracies on target domain for each classifier. These accuracies are given in Table 3.

Classifier	Target Dataset 1 Accuracy (Used in Fine-tuning)	Target Dataset 2 Accuracy (Unseen)
VGG16	92.23	42.75
ResNET18	96.47	43.03

Table 3. Results obtained by direct fine-tuning on target dataset

We used three different approaches for domain adaptation purpose.

1. Unsupervised Domain Adaptation using WGAN — WGAN results were not useable. So this approach was discontinued. This approach was presented in [1].

2. Semi-supervised Domain Adaptation using CycleGAN — Approach based on the [2].

3. Feature Space Unsupervised Domain Adaptation — Approach based on the [3].

The first two approaches basically used Input Space domain adaptation. The third approach as the name suggests, used Feature Space domain adaptation. In input space domain adaptation we basically use GAN generated samples to fine-tune our pre-trained classifier on source domain so that it can adapt to target domain. WGAN in the first approach basically generates new target domain like samples from random noise input while CycleGAN translates source samples to target domain like samples. WGAN results were not satisfactory even though we tried 3 different architectures for it and trained for 2 weeks. Thus this approach was discontinued. On the other hand, CycleGAN produced fairly good results and fine-tuning was performed for its translated samples. In feature space, we basically try to make feature extracted by our model from input images independent of any domain information. We do this using a domain classifier network trained in parallel to our classifier and loss generated by this network is fed to the model as well so that it becomes invariant to any domain information. Feature space domain adaptation techniques don't require any GAN to help out in its purpose and thus very less expensive than input space domain adaptation. It took us 4 days to train classifiers using third approach.

WGAN generated samples for all three architectures used are given in Figure 4. As it can be that results for WGAN are not satisfactory and usable. Henceforth, we discontinued this approach. Training specifications for WGAN models are given in Table 4.

The next series of experiments are for second approach "Semi-supervised Domain Adaptation using CycleGAN".

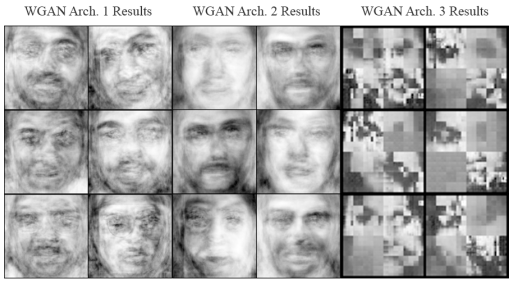

Figure 4. WGAN results for different architecture.

Specifications	WGAN Arch. 1	WGAN Arch. 2	WGAN Arch.3
Model Type	Linear	Linear	Convolutional
Training Epochs	10k	1k	7.5k
Training Time	7 Days	3 Days	7 Days

Table 4. Training specifications of different WGAN models

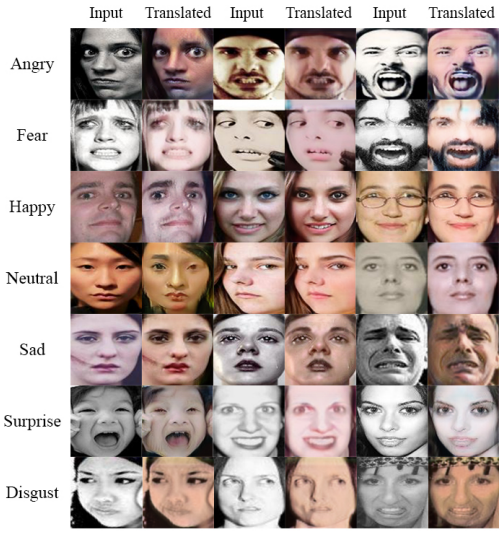

Figure 5. CycleGAN translated images.

The results for some of the CycleGAN translated images are given in fig 4. We trained 7 different CycleGAN for each emotion class so that CycleGAN don't mess up the emotion of the source image while translation. It took us 2 weeks to train these 7 CycleGAN and translate image through them. Some of the input and translated images are shown in Figure 5. As we can see that CycleGAN mostly tries to adapt color features from target domain to source domain input. In order to get better results from CycleGAN, it can be trained more so that it can capture more feature information from target domain and map them to source domain input images.

Using these translated Images, we fine-tuned our classi-

fiers and Table 5 provides the accuracy score on both target datasets.

Classifier	Target Dataset 1 Accuracy (Used in Fine-tuning)	Target Dataset 2 Accuracy (Unseen)
ResNET18	48.13	33.42

Table 5. CycleGAN based approach results after fine-tuning

The final series of experimentation is for approach "Feature Space Unsupervised Domain Adaptation". Here we retrained both the classifier with an additional domain classifier network in them. This domain classifier network help in making the features used in classifier independent of any domain information. Table 6 represents the results for both classifiers on source and both target datasets.

Classifier	Target Dataset 1 Accuracy (Used in Fine-tuning)	Target Dataset 2 Accuracy (Unseen)
VGG16	51.36	37.37
ResNET18	46.72	32.41

Table 6. Results after Feature space domain adaptation

For better understanding of performance of different approaches, confusion matrices for each experiment can be found in figures section. In Figure 6, confusion matrices for baseline results for each classifier on both target dataset can be found. In Figure 7, confusion matrices for classifiers fine-tuned on target dataset directly can be found. In Figure 8, confusion matrices for classifiers fine-tuned on CycleGAN translated samples can be found. In Figure 9, confusion matrices for classifiers trained using feature space domain adaptation approach can be found.

6. Conclusion

WGAN didn't produced promising results, hence the approach involving WGAN was discontinued. By using "Semi-supervised Domain Adaptation using CycleGAN", we are able to produce reasonably good results which were very close to baseline results. For "feature Space unsupervised domain adaptation", we were able to achieve slightly better accuracy than baseline results for one of the target datasets. It was for VGG16 classifier. CycleGAN based approach can make a significant change in results if we train CycleGANs more. Same goes for Feature space based approach. There was an imbalance of classes in source dataset and target datasets which limited the performance of the approaches used. Along this, there several age groups present in source dataset as compared to target dataset. If these things are also dealt carefully, then we can achieve way better results than reported.

References

[1] Nathan Silberman David Dohan Dumitru Erhan Bousmalis, Konstantinos and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 12(1):3722–3731, 2017.

[2] Evgeniya Ustinova Hana Ajakan Pascal Germain Hugo Larochelle François Laviolette Mario Marchand Ganin, Yaroslav and Victor Lempitsky. "domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[3] Alexander S. Ecker Gatys, Leon A. and Matthias Bethge. Image style transfer using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[4] Jun-Yan Zhu Tinghui Zhou Isola, Phillip and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] Ying-Hsiu Lai and Shang-Hong Lai. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 13(1):263–270, 2018.

[6] Xiaoqing Xiangjun Wang Wang and Yubo Ni. Unsupervised domain adaptation for facial expression recognition using generative adversarial networks. *Computational intelligence and neuroscience*, 2018.

[7] Taesung Park Phillip Isola Zhu, Jun-Yan and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

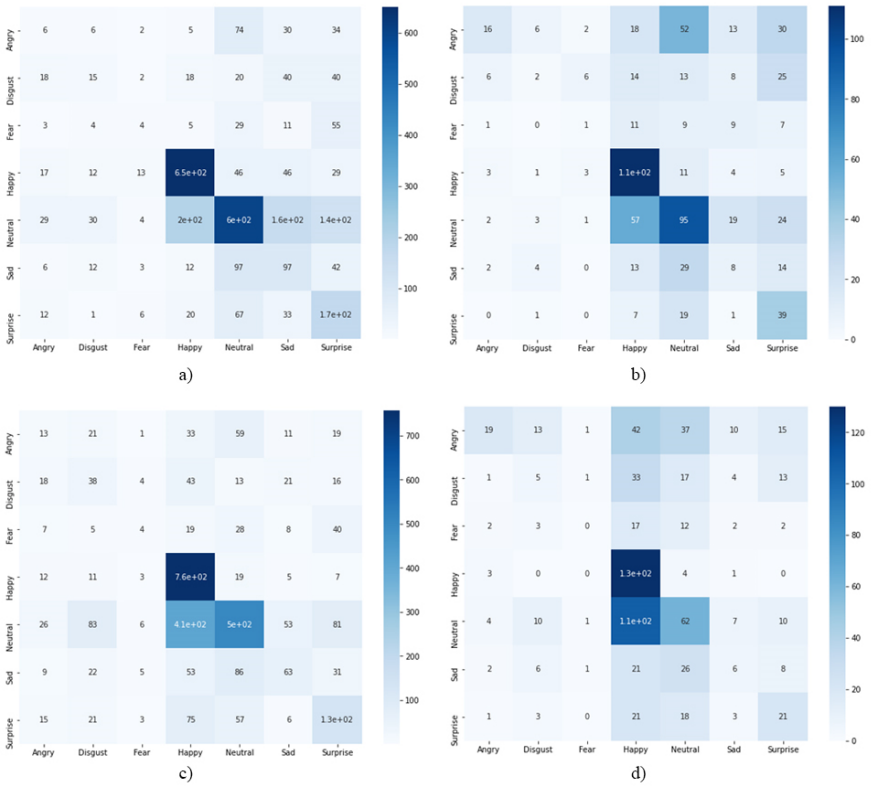

Figure 6. Confusion Matrices for baseline results. **(a)** VGG16 results on Target Dataset 1, **(b)** VGG16 results on Target Dataset 2, **(c)** ResNET18 results on Target Dataset 1, **(d)** ResNET18 results on Target Dataset 2

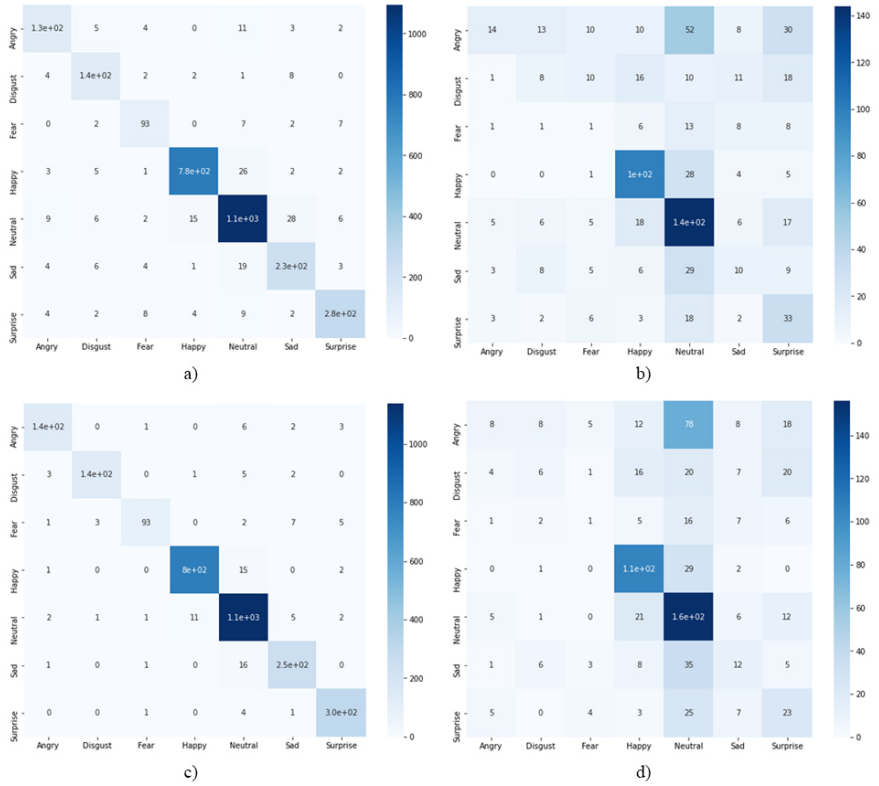

Figure 7. Classifiers fine-tuned on target dataset directly. **(a)** VGG16 results on Target Dataset 1, **(b)** VGG16 results on Target Dataset 2, **(c)** ResNET18 results on Target Dataset 1, **(d)** ResNET18 results on Target Dataset 2

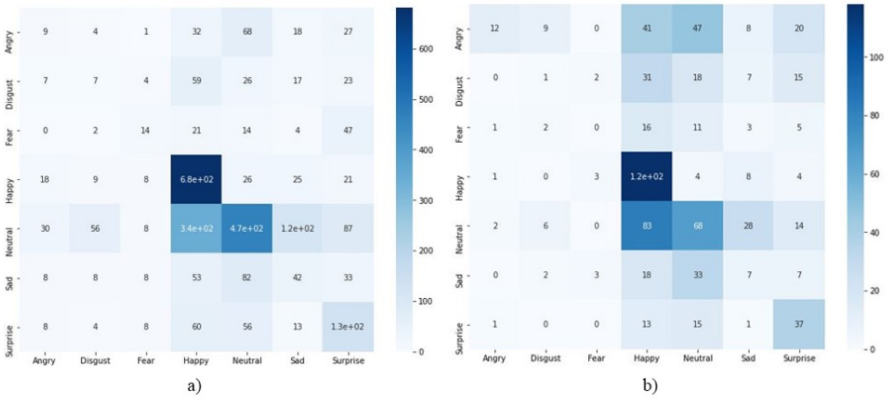

Figure 8. Classifiers fine-tuned on CycleGAN translated samples. **(a)** ResNET18 results on Target Dataset 1, **(b)** ResNET18 results on Target Dataset 2

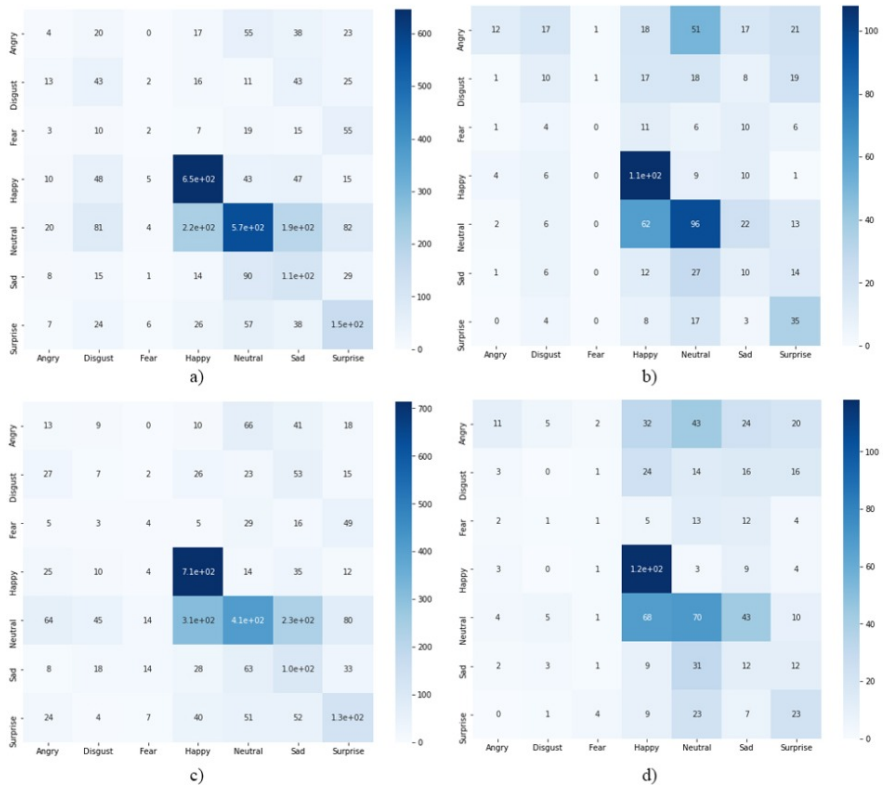

Figure 9. Classifiers trained using feature space domain adaptation approach. **(a)** VGG16 results on Target Dataset 1, **(b)** VGG16 results on Target Dataset 2t, **(c)** ResNET18 results on Target Dataset 1, **(d)** ResNET18 results on Target Dataset 2