Identity-based Adversarial Training of Deep CNNs for Facial Action Unit Recognition

Zheng Zhang
zzhang27@cs.binghamton.edu
Shuangfei Zhai
szhai2@cs.binghamton.edu
Lijun Yin
lijun@cs.binghamton.edu

Department of Computer Science
State University of New York at
Binghamton
NY, USA.

Abstract

Automatic detection of facial action units (AU) is a key towards fully understanding spontaneous facial expressions in human emotion analysis. The most recent development shows encouraging results in utilizing deep learning models to recognize facial action units. However, existing AU-labeled spontaneous facial expression datasets are either in a small-scale due to labor-intensive annotations, or lack of sufficient variety in terms of amount, ethical background, age ranges, and facial appearance variations of subjects, thus limiting the learning effectiveness. To mitigate the issue of high redundancy and low level of variants existing among image frames of facial video sequences with respect to both subject identities (ID) and facial action units (AU), we propose a novel learning process with convolutional neural networks (CNNs), named Adversarial Training Framework (ATF). Architecture wise, ATF takes the form of a multi-task learning CNN, where a deep stack of convolutional layers, pooling layers and fully connected layers act as feature learners, and an AU loss together with an ID loss is computed with the shared deep representations. The crucial property of ATF is that during training, the underlying feature layers are trained to *minimize* the AU loss while *maximizing* the ID loss. This adversarial training w.r.t. the ID loss essentially makes the learned features effective for AU detection, while invariant to subject identities, thus alleviating the affection of personal identity to a great extent. We have conducted experiments on the public databases BP4D and DISFA, showing significant performance improvements over the peer methods.

1 Introduction

Research on facial expression recognition (FER) has been intensified in recent decades [6, 21]. Ekman proposed the Facial Action Coding System (FACS) [7], as a comprehensive, anatomically based system, to measure all visually discernible facial movement. Thus, detecting facial action units has become an essential task towards realizing an accurate and reliable FER system. Recently, deep network based algorithms [11, 23, 33] have been developed for extracting robust facial features. As a data-driven approach, the quality of the data with ground truth labels is crucial for affecting the performance of deep networks.

From existing facial action unit databases [18, 19, 30, 32], there are a large amount of 2D images that have been AU annotated. For example, in BP4D [30], over $140,000$ frames from 41 subjects have FACS coded while the numbers have been increased to $190,000$ with 140 subjects for BP4D+ [32]. Similar statistics can also be found in DISFA [19], with about $120,000$ frames from 27 subjects. Although there exists a great number of training samples, the number of class instances, *i.e.* subjects × AUs, is relatively small when AU occurrence needs be determined. This turns to cause the over-fitting problem for deep network based classifiers.

It is arguable whether the subjects' identities could help or jeopardize the task of AU recognition. Previous works [3, 4] have learned person-specific models for facial expression and AU recognition. By transferring the informative knowledge from a small set of person-specific data, it allows us to learn an accurate model for a new subject. Such inductive/transductive learning technique is also employed by [1] on AUs annotation. All of these work indicate that the person-specific information can help to improve the performance of AU related classification tasks. However, these works are typically built on hand-crafted features and the good AU detection system must be person-invariant, thus requiring the system to tackle the personal information in an automatic way. In order to regularize deep features, we need to find the relationship between subject identification and AU recognition. In [5], Chu et al. visualized the FC7 features from AlexNet [14] by a t-SNE embedding. t-SNE can project the high dimensional data into low dimensional space while keeping the distribution of data. This work discovered that through the feature learning process of deep neutral network, the feature representation will be increased as the individual differences are minimized. Inspired by this finding, here we present a compositional framework to add an adversarial training layer by maximizing the loss from human subject identification, with an attempt to improve the feature generality in the deep neutral network.

One way to understand the proposed ATF is by drawing the analogy between AU recognition and domain adaption [8, 17, 20]. Domain adaption deals with the problem when the training data distribution differs slightly from the testing data distribution, a model trained on the training data might not perform well during the test. And one commonly used approach is to learn a transformation of the input data, such that the distribution of transformed data points are about the same on the training set and the test set. In the setting of AU recognition, frames taken from the test subjects differ significantly from those of the training subjects, due to the unique biases attached to individual subjects, much like the problem in domain adaption. Our approach can thus be considered as a deep learning based domain adaptation method, with a specific focus on AU recognition, where the goal is to learn subject/domain invariant representations that can help generalization. To our knowledge, this is the first work to apply an adversarial training process on AU related classification tasks for facial expression analysis. The experimental results show that such an adversarial training is an effective mechanism to facilitate the feature generality on AU occurrence detection.

2 Related work

The existing works have been focused on the investigation of facial action information and learning methodologies.

Deep Model Based AU Recognition: With the extensive use of deep learning networks, it is needed to address how the deep features are correlated to perceived facial actions. To address this issue, Khorrami [13] set zero to all neurons of specific convolutional layers and deconvolved to visualize the reconstructed images. Jaiswal [12] concatenated the Convolutional and Bi-directional Long Short-Term Memory Neural Networks (CNN-BLSTM) to jointly learn the shapes and appearances from dynamic sequences. The developed system has achieved a good performance on the FERA-2015 [26] Challenge dataset. A similar architecture has also be developed by Chu [5]. With a consideration of the sparse AU distributions, Zhao [34] proposed a Deep Region and Multi-label Learning (DRML) method to guide the deep network by focusing on important facial regions under certain AU correlations. Inspired by this work, Li [15, 16] has further added two layers by enhancing and cropping nets to a pre-trained CNN model and improved AU detection performance significantly.

Multi-task Learning (MTL): MTL means different tasks are learning in parallel while using a shared representation. What is learned for each task can help other tasks to learn better [2]. For instance, Wu [27] applied the MTL to learn the relationships among AUs and face shapes while updating feature points, resulting in a boosted performance of both facial AUs recognition and landmarks localization. Zhang [31] has also optimized facial landmark detection successfully along with the other tasks, *i.e.*, estimation of facial appearance attribute, expression, demographic, and head pose. Our approach differs significantly from standard multi-task learning in that, instead of jointly minimizing several losses each of which corresponds to a task, we only minimize the loss of the task of interest (i.e., AU recognition) while adversarially maximize the loss of an auxiliary task (i.e., subject ID classification). On the other hand, our method points to a promising direction of effectively leveraging tasks which do not necessarily benefit from each other, and thus can be considered as a novel way of performing multi-task learning in this regard.

Deep Domain Adaptation and Adversarial Learning: Domain adaption work, based on their objects, can be categorized into discrepancy-based [17, 28] and adversarial-based approaches [24, 25, 29]. Our model falls into the second group and is inspired by the domain adversarial networks by Ganin et al. [8], where deep neural networks are trained to yield domain invariant features. Our contribution is to draw a previously undiscovered connection between the deep domain adaptation problem with the AU recognition problem. In contrast to [8] traditional domain adaption conventions, where source and target domains are usually the data distributions from training and testing sets, we consider different subjects as domains during the training phase. By treating images sampled from the same subject as a domain, we are able to readily borrow the tools developed in the domain adaptation literature. We also believe that this work could inspire further studies which apply different methodologies than ours. The adversarial module is also related to the recent advances in adversarial training of deep neural networks [9, 10], where instead of training a model by minimizing a single objective, an adversarial module is engaged which concurrently maximizes certain objectives. Our model is thus an instantiation of such techniques in the specific context of AU recognition.

3 Methodology

3.1 Deep CNN Architecture

VGG-19 [22] is the base model in this work, which has been widely used in many computer vision tasks, *e.g.*, object detection, images classification, etc.. It contains 16 convolution layers (grouped by 5 max-pooling layers) and 3 fully connected (FC) layers with dropout rate of 0.5 applied on the first two FC layers. We keep all the layers of VGG-19 except for the last FC layer (fc8) which is originally designed for classification with 1,000 possible categories. We initialize all the weights and biases of VGG-19 from a pre-trained model on ImageNet, which is freely available at Caffe model zoo. All the experiments in this paper follow the fine-tuning protocol, unless otherwise mentioned.

3.2 Baseline model

In this paper, we focus on the AU occurrence detection which we cast as a binary classification problem. Each input image is associated with a binary FACS code, where at each dimension 1 indicates the presence of a corresponding action unit is presented and 0 otherwise. An RGB facial image is passed through the entire network which outputs a 2048-D feature vector at FC7. Two more (2048, M) FC layers G_{AU} with sigmoid units is placed on top of FC7 for action unit predictions with parameters denoted as θ_{AU}. As shown in Figure 1(a), the output of G_{AU} contains M neurons corresponding to M Action Units. Then ground truth labels are compared for computing the loss (cross entropy), whose gradient is further back-propagated for parameter updating.

Action Unit Loss: AUs recognition (or detection) can be naturally regarded as a multi-label classification problem. We first define the action unit loss L_{AU}, which is computed by the mean value across all M AUs that have been considered:

$$L_{AU} = \frac{1}{M}\sum_{j=1}^{M} H_j(p_j^{AU}(x), y_j^{AU}), \tag{1}$$

where x denotes the input image, $p^{AU}(x) \in R^{11}$ is the output of the last FC layer for input x, with each dimension representing the predicted score of an AU, $y_j^{AU} \in \{0,1\}$ and $H_j(\cdot,\cdot)$ denotes the label and loss for the jth AU, respectively. The loss for the jth AU is computed with cross entropy, as shown in following equation:

$$H_j(x, y_j^{AU}) = -y_j^{AU} \log p_j^{AU}(x) - (1 - y_j^{AU}) \log(1 - p_j^{AU}(x)). \tag{2}$$

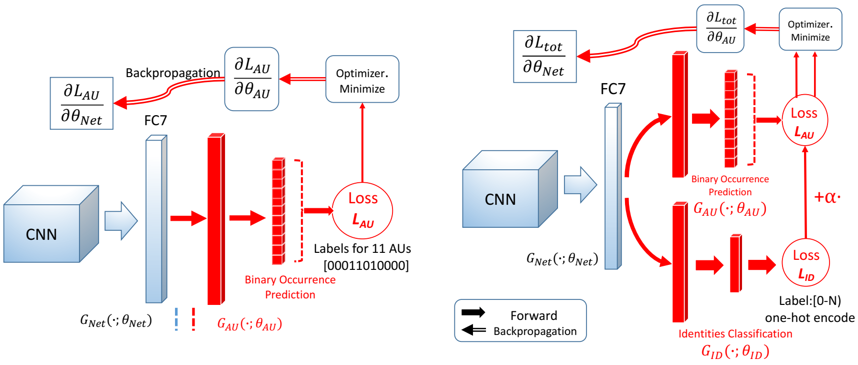
The same setting can also be found in many related works for AU recognition, e.g., [5, 34]. We use this model as the baseline for this work.

3.3 Jointly Minimizing the ID Loss

It is common that the training images of many facial expression databases with AU annotations (e.g., [30]) are organized by subject identities. Such a property can potentially serve as an additional source of information to benefit the learning of AU recognition models. Similar to the multi-task learning neural nets, we utilize the ID information to train an ID classifier along with the AU classifiers. We now proceed to formally define this joint minimization model (JM).

ID Loss: Followed by FC7, the ID classifier is implemented as another two (2048, C) FC layers (see the lower branch of Figure 1(b)) with softmax units G_{ID} which predicts the subject identities from the facial images. C denotes the number of subjects appearing in the training set. Different from AU recognition, this is a multi-class classification problem with ground truth label of $y^{ID} \in \{1,...,C\}$. Here the ID loss, L_{ID}, is defined as follows:

$$L_{ID} = -\sum_{i=1}^{C} y_i^{ID} log(p_i^{ID}(x)), \tag{3}$$


Backpropagation
$\frac{\partial L_{AU}}{\partial \theta_{AU}}$
Optimizer. Minimize
$\frac{\partial L_{AU}}{\partial \theta_{Net}}$
FC7
CNN
Loss L_{AU}
Labels for 11 AUs [00011010000]
Binary Occurrence Prediction
$G_{Net}(\cdot;\theta_{Net})$
$G_{AU}(\cdot;\theta_{AU})$
$\frac{\partial L_{tot}}{\partial \theta_{AU}}$
Optimizer. Minimize
$\frac{\partial L_{tot}}{\partial \theta_{Net}}$
FC7
CNN
Loss L_{AU}
Binary Occurrence Prediction
$G_{AU}(\cdot;\theta_{AU})$
+α·
Loss L_{ID}
$G_{Net}(\cdot;\theta_{Net})$
Label:[0-N] one-hot encode
Forward
Backpropagation
Identities Classification
$G_{ID}(\cdot;\theta_{ID})$

(a) Baseline Model

(b) Joint Minimization Model

Figure 1: Alternated way to consider identity

where $p^{ID}(x) \in R^C$ denotes the softmax prediction for the input x, with the ith dimension denoting the probability of x belong to the ith subject.

The introduction of ID loss layer gives rise to a question of how to combine it with the loss function from the AU recognition layer. Given that our objective is to minimize L_{AU} which in turn increases the accuracy of prediction, a more intuitive way from MTL is to add these two losses together in the optimization process in pursuit of a faster convergence rate or a better recognition performance [31]. Therefore, the loss function of our second baseline model is a combination of the two losses:

$$L_{tot} = L_{AU} + \alpha L_{ID}, \tag{4}$$

where α in equation 4 is a hyperparameter controlling the weight of the ID loss. Due to the fact that the AU classifier and ID classifier share all the layers below and including FC7 in our VGG-19 model, the joint minimization model essentially encourages the model to learn representations (encoded by FC7) that can both distinguish AU as well as be sensitive to the ID of an image.

3.4 Adversarial Training Framework (ATF)

Although the joint minimization model is by nature based on utilizing side class information, we argue that the learned representations to remember the subject ID can *hurt* the generalization ability of the AU classifier. A successful AU classifier should work well in a person-independent fashion. Encouraging a model to memorize the ID information of training subjects (typically in the number of dozens) actually only makes the AU classifier easier to overfit on the training set. This leads us to developing the core idea, in which we propose to adversarially train the feature learning layers (layers including and below FC7) to maximize the ID classification loss, meanwhile minimizing the AU classification loss. The parameters of G_{AU} and G_{ID} on the other hand, are updated to minimize their respective loss, same as in the joint minimization setting. We name this training strategy along with the model architecture the adversarial training framework (ATF). Note that in ATF, we are no longer updating all the parameters by minimizing a single objective function, instead, different set of parameters have their dedicated objectives, which interact with each other during

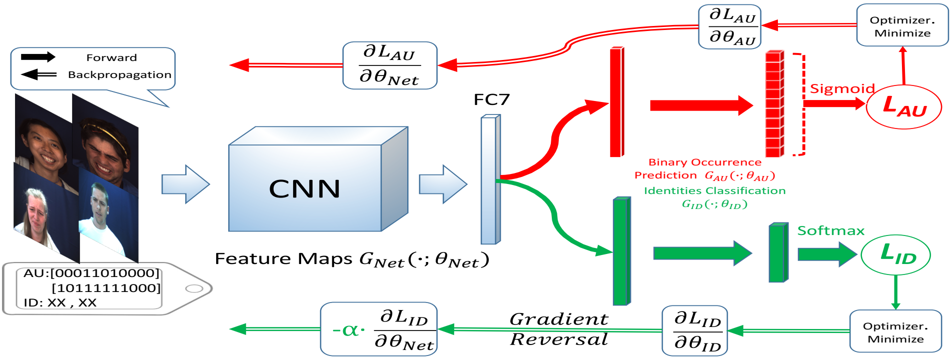

Forward
Backpropagation
AU:[00011010000]
[10111111000]
ID: XX , XX
CNN
Feature Maps $G_{Net}(\cdot;\theta_{Net})$
FC7
$\frac{\partial L_{AU}}{\partial \theta_{Net}}$
$\frac{\partial L_{AU}}{\partial \theta_{AU}}$
Optimizer. Minimize
Sigmoid
L_{AU}
Binary Occurrence Prediction $G_{AU}(\cdot;\theta_{AU})$
Identities Classification $G_{ID}(\cdot;\theta_{ID})$
Softmax
L_{ID}
Optimizer. Minimize
$-\alpha\cdot\frac{\partial L_{ID}}{\partial \theta_{Net}}$
Gradient Reversal
$\frac{\partial L_{ID}}{\partial \theta_{ID}}$

Figure 2: The architecture, along with the forward and backward propagation path of the proposed ATF.

the time of training.

We now formally describe the update rules for each set of parameters. Recall that we divide the set of parameters for ATF into three disjoint subsets: the feature layers parameters θ_{Net}, the AU classifier parameters θ_{AU}, and the ID classifier parameters θ_{ID}. Each of them is updated as follows respectively,

$$\begin{aligned} \theta_{Net} &= argmin\ L_{AU} - \alpha L_{ID}, \\ \theta_{AU} &= argmin\ L_{AU}, \\ \theta_{ID} &= argmin\ L_{ID}, \end{aligned} \tag{5}$$

where L_{AU}, L_{ID}, α are defined as in Eq. 4. In other words, the network parameters θ_{Net} are updated to improve the AU classification accuracy, as well as to confuse the best ID classifier (by increasing its classification loss), so as to learn subject invariant deep representations. As illustrated in Figure 2, the ATF includes both forward and backward propagations of the neural network.

4 Experiment

The recent work by Li et al [15] introduced enhancing layers (E-Net) and cropping layers (C-Net) on a basis of pre-trained $VGG19$ [22] and obtained state of art performance on AU occurrence detection. In this section, we first plug the proposed ATF into E-Net to illustrate the effectiveness of adversarial training on AU detection across two databases. Then we visualize the training process in detail with comparison to the peer models (*Baseline* and *JM*). The impact of ID layer on learning deep features is also demonstrated and discussed. It is worth noting that the C-Net is not suitable for ATF because it crops an entire face into small patches located by 20 landmarks and trains the shallow convolution network on individual face patch, thus losing the holistic identity information.

4.1 Datasets

BP4D: 41 subjects are included in BP4D with 23 females and 18 males. Spontaneous facial expressions were elicited through 8 well-designed tasks. FACS Coders selected most expressive 20s from all 328(41 × 8) sequences for occurrence coding.

DISFA: Another benchmark dataset comes from Denver Intensity of Spontaneous Facial Action (DISFA) which contains two-views videos of 27 adult subjects. Frames are labeled by 66 facial landmark points and 0-5 scale AU intensity codes. We extracted the data frames captured by the left camera and considered 8 AUs as positive if their intensities are larger than 0.

Similar to the settings in [15], subjects in these two databases are split into 3 folds for cross validation. In our experiment, the training and testing data of each fold are exactly same as the ones used in [15] to assure a fair comparison. For training, some samples are repeated 4 to 7 times in order to balance the AUs that are less occurring. Thus it alleviates the negative impact of data unbalance on ATF. Note that no face alignment and further data augmentation are required in our experiment.

4.2 Implementation Details

First of all, each image is resized into 224 × 224 for fit of $VGG19$ [22]. We subtract each image by its mean value and divide it by the variance of pixel intensities. Then the batched images with a size of 50 are input to the network directly. Second, for the first two groups of convolution layers in $VGG19$, parameters are left out during the fine-tuning process in order to keep the lower level generic features. The output logits of the AU layer indicate the presence or absence of each action unit. Third, an sigmoid function is applied on these outputs in order to normalize them in the range from 0 to 1.

The whole framework is implemented on TensorFlow with an initial learning rate of 0.0001 for both AU and Identity. The parameters are then optimized by Nesterov method with a momentum of 0.9. To further balance the optimization process of the two components, AU and Identity, we use a weight decay of 0.0001 and decrease the learning rate by 4% on every 2000 iterations. Note that an early-stop strategy is adopted before reaching the model over-fitting.

4.3 Evaluation and Results

To be comparable to the existing models, two common metrics, *i.e.*, F-score and Accuracy are used to evaluate the performance of the proposed framework. F-score is the harmonic mean of the precision and recall which is defined as $F1 = \frac{2*precision*recall}{precision+recall}$. For AU detection, the datasets are highly unbalanced in terms of both labels and classes (0 or 1). Even though we try to balance them by up-sampling as mentioned, there still exist a large number of 0s in labels. Therefore, F-score is more favorable than Accuracy for our performance evaluation.

4.3.1 Comparison with the State of Art

Experiment on BP4D: Table 1 shows a performance increase from 52.1% to 55.4% on F-score. Note that recognition rates on AU1 (Inner Brow Raiser), AU4 (Brow Lowerer), AU10 (Upper Lip Raiser), AU17 (Chin Raiser), and AU23 (Lip Tightener) have achieved the best as compared to the peer state of the art approaches. Since those AUs are subtle in spontaneous expressions of the BP4D database, it is very challenging to increase the recognition performance. The improvement of those AUs detection attributes to the use of the proposed ATF, by which the identity influence has been mitigated effectively.

Table 1: F-score on BP4D in terms of 12 AUs. (α=0.02)

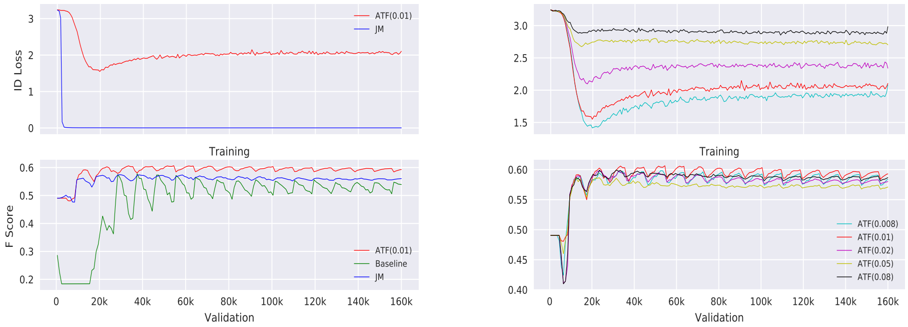
AU	JPML [33]	DRML [34]	FVGG [15]	E-Net [15]	**E-Net+ATF**
1	32.6	36.4	27.8	37.6	**39.2**
2	25.6	**41.8**	27.6	32.1	35.2
4	37.4	43.0	18.3	44.2	**45.9**
6	42.3	55.0	69.7	**75.6**	71.6
7	50.5	67.0	69.1	**74.5**	71.9
10	72.2	66.3	78.1	80.8	**79.0**
12	74.1	65.8	63.2	**85.1**	83.7
14	**65.7**	54.1	36.4	56.8	65.5
15	**38.1**	33.2	26.1	31.6	33.8
17	40.0	48.0	50.7	55.6	**60.0**
23	30.4	31.7	22.8	21.9	**37.3**
24	**42.3**	30.0	35.9	29.1	41.8
Avg.	45.9	48.3	43.8	52.1	**55.4**

Experiment on DISFA: Following the same protocols in [15], we use the best model trained on BP4D as a feature extractor to map the input images of DISFA database into the 2048-dimension vectors. Then a linear regressor is trained from scratch for inference of 8 AU labels. Table 2 shows the F-score of our model and the compared works. As we can see, our proposed ATF outperforms the peer approaches on average F-Score, which is promising to increase its generalizability.

Note that the number of parameters in the forward pass of E-Net is about 130M. In the work [15], the authors also reported F-scores (55.9% on BP4D and 48.5% on DISFA on average) with the use of cropping layers (noted as EAC). As compared to the E-Net, EAC has an extra set of $\sim 78M$ parameters, which is a drastic increase in terms of model capacity. However, without increasing the extra set of parameters, our proposed ATF can obtain comparable AU recognition rate while still keeping the lower model complexity.

Table 2: F-score on DISFA in terms of 8 AUs.

AU	APL [34]	DRML [34]	FVGG [15]	E-Net [15]	**E-Net+ATF**
1	11.4	17.3	32.5	37.2	**45.2**
2	12.0	17.7	24.3	6.1	**39.7**
4	30.1	37.4	**61.0**	47.4	47.1
6	12.4	29.0	34.2	**52.5**	48.6
9	10.1	10.7	1.67	13.4	**32.0**
12	65.9	37.7	**72.1**	71.1	55.0
25	21.4	38.5	**87.3**	84.2	86.4
26	26.9	20.1	7.1	**43.5**	39.2
Avg.	23.8	26.7	40.2	44.4	**49.2**



(a) Peer Models

(b) Impact of ID layer

Figure 3: Learning curves on BP4D dataset for comparison among different models (better view in color). Upper row: y-axis stands for L_{ID} along with x-axis (iterations) during the training. Lower row: y-axis stands for corresponding F-score on the validation set.

4.3.2 Model Comparison and Learning Visualization

In Figure 3(a), we compare different models (*i.e.*, $Baseline$, JM, and ATF under different α) and visualize the learning performance by plotting the training and validation curves on one split (best $\alpha = 0.01$ here) of the cross validation. The rest two splits follow the same trend. The training curves show the trend of ID loss from JM and ATF. Since there is no ID layer embedded in the baseline model, the ID Loss does not contribute to the parameter update θ_{Net}, neither to the θ_{ID}, hence we do not plot it for baseline model.

As the ID Loss is combined in the JM model, the total loss is thus increased. It also increases the gradients of all trainable parameters θ. We can clearly see that the curve (in blue) decreases quickly during the training stage. As an illustration, we only show one curve from ATF to compare with Baseline and JM.

It is worth noting that because the two processes (i.e., ID identification and AU recognition) compete each other in an adversarial way, the ID loss is able to maintain in a certain level, thus facilitating the entire feature learning procedure. More specifically, if without ID loss as the regularizer, the baseline model can only attend to those frequently occurred AUs and leave the rest unoptimized, resulting in low F scores (as illustrated by the green curves before $20k$ iterations).

Impact of ID layer: In the adversarial training model, the weight α indicates the impact of ID layer. By grid search, we get α in the range from 0.008 to 0.08, and compare their learning performance as shown in Figure 3(b). As we can see, different choices of scalar α could influence the learning procedure dramatically. A larger value of α means a more intensive competition from the ID layer while keeping the ID loss in a higher level. Extremely when α is 0, the adversarial training degrades to the baseline model. In general, a good choice of α (in this case, $\alpha = 0.01$) can make the convergence of AU learning perform better. We also notice that the ID loss decreases at the beginning of training process. We attribute it to the constant setting of α given that different relative magnitudes of loss between AU and ID during the training.

5 Conclusion

In this paper we have presented a novel adversarial training framework to tackle the issue of facial action unit detection for facial expression analysis. The core of this framework is to enable the underlying feature layers to be trained by minimizing the AU loss while maximizing the ID loss. This adversarial training essentially makes the learned features effective for AU detection, while invariant to subject. The results are promising as compared to the Baseline model and the JM model when tested on benchmark datasets.

Our future work will improve the proposed framework with extending the application to the temporal domain, thus enabling us to further conduct comparison with more state of the art approaches. We also hope that our work would inspire further explorations in deep multi-task learning methodologies for a broader range of audience. Notice that the same adversarial training framework is also applicable to facial expression recognition by minimizing the expression label loss while maximizing the ID loss for improving the person-independent expression classification, thus giving rise to our new development in the future.

6 Acknowledgement

The material is based upon the work supported by the National Science Foundation under grants CNS-1629898 and CNS-1205664.

References

[1] Timur Almaev, Brais Martinez, and Michel Valstar. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *International Conference on Computer Vision (ICCV)*, 2015.

[2] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. 1998.

[3] Jixu Chen, Xiaoming Liu, Peter Tu, and Amy Aragones. Learning person-specific models for facial expression and action unit recognition. *Pattern Recognition Letters*, 34(15):1964–1970, 2013.

[4] Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[5] Wen-Sheng Chu, Fernando De la Torre Frade, and Jeffrey Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *Automatic Face and Gesture Recognition (FG)*, 2017.

[6] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(8):1548–1568, 2016.

[7] Paul Ekman and Wallace V Friesen. Facial action coding system. 1977.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[10] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[11] Amogh Gudi, H Emrah Tasli, Tim M den Uyl, and Andreas Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *Automatic Face and Gesture Recognition Workshops*, 2015.

[12] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8, 2016.

[13] Pooya Khorrami, Thomas Paine, and Thomas Huang. Do deep neural networks learn facial action units when doing expression recognition? In *International Conference on Computer Vision Workshops (ICCVW)*, pages 19–27, 2015.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*. 2012.

[15] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. In *Automatic Face and Gesture Recognition (FG)*, 2017.

[16] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2018.

[17] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, 2015.

[18] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *the IEEE conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010.

[19] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[20] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, 2010.

[21] Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10), 2012.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[24] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *International Conference on Computer Vision (ICCV)*, 2015.

[25] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] M Valstar, J Girard, T Almaev, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and J Cohn. FERA'15 2nd facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG)*, 2015.

[27] Yue Wu and Qiang Ji. Constrained joint cascade regression framework for simultaneous facial action unit recognition and facial landmark detection. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Xing Zhang, Lijun Yin, Jeffrey Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. BP4D-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[31] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision (ECCV)*, pages 94–108, 2014.

[32] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] Kaili Zhao, Wen-Sheng Chu, Fernando De la Torre, Jeffrey F Cohn, and Honggang Zhang. Joint patch and multi-label learning for facial action unit detection. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[34] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.