Unsupervised Domain Adversarial Training in Angular Space for Facial Expression Recognition

Akihiko Takashima, Naoki Makishima, Mana Ihori, Tomohiro Tanaka, Shota Orihashi, Ryo Masumura
NTT Media Intelligence Laboratories, NTT Corporation,Japan
E-mail: akihiko.takashima.dg@hco.ntt.co.jp

***Abstract*—This paper presents unsupervised domain adversarial training in angular space (UDAT-AS), a novel unsupervised domain adversarial training method for facial expression recognition (FER). UDAT is effective as it can adapt existing neural network based classification models to the target domain by utilizing only unlabeled data sets. It is realized by forming a domain adversarial network consisting of a domain classifier and a gradient reversal layer. UDAT reduces the domain dependency of neural network based classification models by making them insensitive to domain labels. However, conventional unsupervised domain adversarial training is not suitable for FER because facial expressions strongly depend on the domain of the training data sets, e.g., race, gender, shooting environment, and pose. In order to learn domain invariance more clearly, our key advance in UDAT-AS is to perform unsupervised domain adversarial training with angular softmax loss. UDAT-AS is an extension of the domain adversarial network; its domain classifier uses angular softmax loss, a commonly utilized metric learning technique. This enables us to efficiently reduce domain bias in FER models and allow the effective use of unlabeled target domain data sets. We evaluate our approach using two different collection methods, and demonstrate that our method outperforms conventional alternatives.**

I. Introduction

Facial expressions are an important part of understanding the emotional state of people [1]. Automatic facial expression recognition (FER) is expected to be used in various applications such as human-computer-interaction, robotics, and healthcare. Recently, fully neural network based methods using a large number of face images have advanced recognition performance [2], [3]. One remaining issue is that most facial expression data sets suffer from the domain bias problem. This is because human faces have complex and wide domains due to differences in race, age, face pose, photograph condition, and facial expressions. Thus, it is difficult to train universal FER models that can be applied to any face data.

The domain bias problem can be often tackled by domain adaptation, which learns to shift an existing source domain to the target domain [4]. In particular, unsupervised domain adaptation can optimize the domain shift by using only unlabeled data sets. One of the most representative unsupervised domain adaptation methods compatible with fully neural network based classification models is unsupervised domain adversarial training (UDAT) [5]. The main strategy of UDAT is to add a domain classifier and a gradient reversal layer to the domain adversarial network. This helps to reduce the domain dependency of the neural network based classification models by making training insensitive to domain bias, e.g., learning domain invariant features. It is reported that UDAT offers good adaptation performance in speaker recognition [6].

However, conventional UDAT is not suitable for FER because facial expressions strongly depend on the training data domain, e.g., race, gender, shooting environment, and pose. In other words, it is difficult for FER models to be insensitive to domain bias. Therefore, a more sophisticated UDAT is needed if we are to acquire domain invariant features. To this end, our key idea is to utilize angular softmax loss in UDAT. This idea is motivated by its success in face verification; the angular softmax loss is leveraged for metric learning [7]–[9]. It is known that the angular softmax loss is superior to cross-entropy loss for clearly separating input features. Thus, we can expect to attain domain invariant features by introducing the angular softmax loss to the domain classification network in the domain adversarial network.

In this paper, we propose UDAT in angular space (UDAT-AS). UDAT-AS computes the angular softmax loss by utilizing the norm of features and weights in the last layer in the domain classifier. The angular softmax loss is leveraged for reducing domain bias in the FER models thorough the gradient reversal layer. In addition, UDAT-AS considers the angular margin [9] between a source domain and the target domain to efficiently learn domain invariant features. To the best of our knowledge, this paper is the first study to utilize angular space for UDAT. In our experiments, we use two FER data sets from different domains, and compare the proposal with conventional UDAT. We show that our angular softmax loss term and angular margin effectively improve UDAT.

II. Related work

A. Neural Facial Expression Recognition

Many FER tasks set the goal of recognizing seven facial expressions, including the seven basic facial expressions (angry, disgust, fear, happy, sad, surprise, and neutral) . In recent years, methods using convolutional neural networks (CNNs) such as VGG [10], ResNet [11] has been proposed. It is reported that CNN is robust to facial position and orientation in FER [12]. Other studies consider face distortion components using face alignment based on landmark information [13] and covariance pooling [14]. Furthermore, FER performance can be improved by fine tuning of pre-trained CNN trained with a large amount of face images; examples include FaceNet [15], VGGface [16], and VGGface2 [17]. In this work, we examine UDAT with the utilization of pre-trained CNN networks.

B. Unsupervised Domain Adaptation for Neural Networks

Unsupervised domain adaptation is the approach of transferring machine learning models into the target domain by utilizing only unlabeled data sets. Recently, several studies have described unsupervised domain adaptation specific to fully neural network based methods. Representative methods attempt to match the distribution of the source representations with that of the target without considering sample category [5], [18]–[21]. Of particular interest, domain classifier-based adaptation algorithms have been applied to many tasks [6]. This paper also adopts the domain classifier-based approach for unsupervised domain adaptation. In addition, some methods specific to FER have been proposed. While existing methods use limited 3D facial expression data sets and enable domain adaptation independent of subject and pose [22], [23], UDAT-AS, our proposed unsupervised domain adaptation method, can be applied to any classification modeling by using 2D facial expression data sets.

C. Angular Softmax Loss

Angular softmax loss is now being utilized for face verification tasks [7]–[9]. The face images are mapped into angular space and the discriminative features in the angular space are acquired by using the angular softmax loss. It is known that angular softmax loss can make inter-class variances larger and intra-class variances smaller compared with standard softmax loss. The angular softmax loss has some variants. CosFace and SphereFace use the extended angular softmax loss in which a non-linear angular margin is used in learning hidden representations on a hypersphere [7], [8]. Similarly, ArcFace learns hidden representations on a hypersphere and adds a constant linear angular margin between classes [9]. In this work, we utilize the angular softmax loss with the linear angular margin, which is used in the ArcFace, for UDAT-AS.

III. Neural Facial Expression Recognition

This section describes fully neural network based FER. The problem is to use neural networks for estimating the probability distribution of facial expressions $\boldsymbol{y} = [y_1, \cdots, y_{|Y|}]^\top$ and so categorize input face image $\boldsymbol{x}$. The face images are represented in the fundamental RGB color space and Y represents a set of classes. In this case, the neural FER models are composed of a feature extraction network and a label prediction network. The neural networks first produce pre-output representation $\boldsymbol{y}'$ by

$$\boldsymbol{y}' = \mathcal{G}_y(\mathcal{G}_f(\boldsymbol{x}; \boldsymbol{\theta}_f); \boldsymbol{\theta}_y), \tag{1}$$

where $\mathcal{G}_f()$ denotes the feature extraction network that converts an input feature into a hidden feature representation and $\mathcal{G}_y()$ denotes a part of the label prediction network that converts the hidden feature representation into the pre-output representation. $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_y$ are model parameters for each network. For both functions, arbitrary network structures such as VGG, ResNet, and fully-connected network can be leveraged. In the output layer in the label prediction network, the predicted probability for the i-th class is computed as

$$y_i = \frac{\exp(\boldsymbol{w}_i^\top \boldsymbol{y}')}{\sum_{k=1}^{|Y|} \exp(\boldsymbol{w}_k^\top \boldsymbol{y}')}, \tag{2}$$

where $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{|Y|}\} \in \boldsymbol{\theta}_y$ are the parameters in the output layer with the use of softmax activation.

Model parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_f, \boldsymbol{\theta}_y\}$ can be optimized by preparing training data set $\mathcal{D} = \{(\boldsymbol{x}^1, \bar{\boldsymbol{y}}^1), \cdots, (\boldsymbol{x}^N, \bar{\boldsymbol{y}}^N)\}$ where $\bar{\boldsymbol{y}}^n = [\bar{y}_1^n, \cdots, \bar{y}_{|Y|}^n]^\top$ is represented as a one-hot vector. In this case, cross-entropy loss, i.e., softmax loss, is computed as

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{|Y|} \bar{y}_i^n \log y_i^n, \tag{3}$$

where $\boldsymbol{y}^n = [y_1^n, \cdots, y_{|Y|}^n]^\top$ is the estimated probability distribution for the n-th face image $\boldsymbol{x}^n$. The optimization is conducted by mini-batch stochastic gradient descent (SGD). The model parameters are updated by

$$\boldsymbol{\Theta} \longleftarrow \boldsymbol{\Theta} - \mu \frac{\partial \mathcal{L}^l}{\partial \boldsymbol{\Theta}}, \tag{4}$$

where μ is the learning rate and $\mathcal{L}^l$ is the softmax loss for the l-th mini-batch.

IV. Conventional Unsupervised Domain Adversarial Training

This section briefly describes conventional UDAT [5]. Its training uses both source domain data sets with annotated class labels and target domain data sets without class labels to optimize a fully neural network FER model for the target domain. To this end, a domain adversarial network is formed from not only the feature extraction network and the label prediction network, but also a domain classification network and a gradient reversal layer. The gradient reversal layer outputs the input vectors without any conversion during forward propagation, and sign inversion of the gradients during back propagation [5].

The feature extraction network and the label prediction network are defined as Eqs. (1) and (2), respectively. The domain classification network estimates the probability distribution of domain labels $\boldsymbol{d} = [d_\mathrm{s}, d_\mathrm{t}]^\top$ from the hidden feature representation of the input face image. A pre-output representation in the domain classification network is given by

$$\boldsymbol{d}' = \mathcal{G}_d(\mathcal{G}_f(\boldsymbol{x}; \boldsymbol{\theta}_f); \boldsymbol{\theta}_d), \tag{5}$$

where $\mathcal{G}_d()$ denotes that part of the domain classification network that converts the hidden feature representation into the pre-output representation, and $\boldsymbol{\theta}_d$ represents the model parameters for the domain classification network. In the output layer of the domain classification network, the predicted probability of the target domain label is computed by

$$d_\mathrm{t} = \frac{\exp(\boldsymbol{z}_\mathrm{t}^\top \boldsymbol{d}')}{\exp(\boldsymbol{z}_\mathrm{s}^\top \boldsymbol{d}') + \exp(\boldsymbol{z}_\mathrm{t}^\top \boldsymbol{d}')}, \tag{6}$$

where $\{\boldsymbol{z}_\mathrm{s}, \boldsymbol{z}_\mathrm{t}\} \in \boldsymbol{\theta}_d$ are the parameters in the output layer assuming the use of softmax activation.

In UDAT, model parameters are optimized by preparing both source domain training data set $\mathcal{D}_\mathrm{s} = \{(\boldsymbol{x}^1, \bar{\boldsymbol{y}}^1), \cdots, (\boldsymbol{x}^N, \bar{\boldsymbol{y}}^N)\}$ and target domain training data set $\mathcal{D}_\mathrm{t} = \{\boldsymbol{x}^{N+1}, \cdots, \boldsymbol{x}^{N+M}\}$. UDAT attempts to make the distributions of hidden feature representations for the source domain data sets similar to those for the target domain data sets. In order to achieve domain invariant feature extraction in UDAT, we define the objective probability distributions of domain labels for the source domain training data set as $\bar{\boldsymbol{d}}^n = [1.0, 0.0]^\top$. On the other hand, we define those for the target domain training data set as $\bar{\boldsymbol{d}}^n = [0.0, 1.0]^\top$. In this case, loss functions for label prediction and domain classification can be defined as

$$\mathcal{L}_y = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{|Y|}\bar{y}_i^n \log y_i^n, \tag{7}$$

$$\mathcal{L}_d = -\frac{1}{N+M}\sum_{n=1}^{N+M}(\bar{d}_\mathrm{s}^n \log d_\mathrm{s}^n + \bar{d}_\mathrm{t}^n \log d_\mathrm{t}^n), \tag{8}$$

where $\boldsymbol{y}^n = [y_1^n, \cdots, y_{|Y|}^n]^\top$ and $\boldsymbol{d}^n = [d_\mathrm{s}^n, d_\mathrm{t}^n]^\top$ are the estimated probability distributions for the n-th face image. The optimization is conducted by mini-batch SGD. Due to use of the gradient reversal layer, the model parameters are updated as follows:

$$\boldsymbol{\theta}_y \longleftarrow \boldsymbol{\theta}_y - \mu\frac{\partial \mathcal{L}_y^l}{\partial \boldsymbol{\theta}_y}, \tag{9}$$

$$\boldsymbol{\theta}_d \longleftarrow \boldsymbol{\theta}_d - \mu\frac{\partial \mathcal{L}_d^l}{\partial \boldsymbol{\theta}_d}, \tag{10}$$

$$\boldsymbol{\theta}_f \longleftarrow \boldsymbol{\theta}_f - \mu(\frac{\partial \mathcal{L}_y^l}{\partial \boldsymbol{\theta}_f} - \lambda\frac{\partial \mathcal{L}_d^l}{\partial \boldsymbol{\theta}_f}), \tag{11}$$

where μ is the learning rate, hyper parameter λ has the role of adjusting the trade-off between the two predictions, and $\mathcal{L}_y^l$ and $\mathcal{L}_d^l$ are the softmax losses for the label prediction and the domain classification, respectively, for the l-th mini-batch. Note that UDAT is suppressed by setting λ to 0.0.

V. Proposed Method

This section details our proposal, UDAT-AS. Its main difference from UDAT is its use of angular softmax loss in the domain classification network. UDAT-AS computes the angular softmax loss by utilizing the norm of features and weights in the last layer of the domain classification network. The angular softmax loss is leveraged for reducing domain bias in the FER models through the gradient reversal layer.

Figure 1 shows the network structure of UDAT-AS. Its domain classification network estimates the probability distribution of domain labels $\boldsymbol{r} = [r_\mathrm{s}, r_\mathrm{t}]^\top$ from the hidden feature representation of the input face image. As in conventional UDAT, the domain classification network produces the pre-output representation by

$$\boldsymbol{r}' = \mathcal{G}_d(\mathcal{G}_f(\boldsymbol{x}; \boldsymbol{\theta}_f); \boldsymbol{\theta}_r), \tag{12}$$

where $\mathcal{G}_d$ is the same function as in Eq. (5) and $\boldsymbol{\theta}_r$ are the model parameters for the domain classification network with angular softmax activation. The output layer of the domain classification network computes the predicted probability of the target domain label by

$$r_\mathrm{t} = \frac{\exp(s\cos\rho_\mathrm{t})}{\exp(s\cos\rho_\mathrm{s}) + \exp(s\cos\rho_\mathrm{t})}, \tag{13}$$

where s is a scaling factor for the angular softmax function. $\cos\rho_\mathrm{t}$ is computed by

$$\cos\rho_\mathrm{t} = \frac{\boldsymbol{z}_\mathrm{t}^\top \boldsymbol{r}_\mathrm{t}'}{||\boldsymbol{z}_\mathrm{t}'|| \cdot ||\boldsymbol{r}_\mathrm{t}'||}. \tag{14}$$

In addition, we introduce an angular margin term to increase the between-class distances in the angular space. Figure 2 shows how the angular margin is used. The angular margin corresponds to the geodesic distance margin penalty in the normalized hypersphere; it takes account of the ground-truth domain labels. When the ground-truth is the target domain, the predicted probability for the target domain is computed by

$$r_\mathrm{t} = \frac{\exp(s\cos(\rho_\mathrm{t} + m))}{\exp(s\cos\rho_\mathrm{s}) + \exp(s\cos(\rho_\mathrm{t} + m))}, \tag{15}$$

where m denotes the angular margin. On the other hand, when the ground-truth is the source domain, the predicted probability for the target domain is computed by

$$r_\mathrm{t} = \frac{\exp(s\cos\rho_\mathrm{t})}{\exp(s\cos(\rho_\mathrm{s} + m)) + \exp(s\cos\rho_\mathrm{t})}. \tag{16}$$

UDAT-AS operates in the same manner as conventional UDAT. Thus, we prepare source domain training data set $\mathcal{D}_\mathrm{s} = \{(\boldsymbol{x}^1, \bar{\boldsymbol{y}}^1), \cdots, (\boldsymbol{x}^N, \bar{\boldsymbol{y}}^N)\}$ and target domain training data set $\mathcal{D}_\mathrm{t} = \{\boldsymbol{x}^{N+1}, \cdots, \boldsymbol{x}^{N+M}\}$. In addition, we define the objective probability distributions of domain labels for the source domain training data set as $\bar{\boldsymbol{r}}^n = [1.0, 0.0]^\top$ and those for the target domain training data set as $\bar{\boldsymbol{r}}^n = [0.0, 1.0]^\top$. In this case, loss functions for the label prediction and the domain classification are defined as

$$\mathcal{L}_y = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{|Y|}\bar{y}_i^n \log y_i^n, \tag{17}$$

$$\mathcal{L}_r = -\frac{1}{N+M}\sum_{n=1}^{N+M}(\bar{r}_\mathrm{s}^n \log r_\mathrm{s}^n + \bar{r}_\mathrm{t}^n \log r_\mathrm{t}^n), \tag{18}$$

where $\boldsymbol{y}^n = [y_1^n, \cdots, y_{|Y|}^n]^\top$ and $\boldsymbol{r}^n = [r_\mathrm{s}^n, r_\mathrm{t}^n]^\top$ are the estimated probability distributions for the n-th face image. Optimization is performed by mini-batch SGD. Due to use of the gradient reversal layer, the model parameters are updated as follows

$$\boldsymbol{\theta}_y \longleftarrow \boldsymbol{\theta}_y - \mu\frac{\partial \mathcal{L}_y^l}{\partial \boldsymbol{\theta}_y}, \tag{19}$$

$$\boldsymbol{\theta}_r \longleftarrow \boldsymbol{\theta}_r - \mu\frac{\partial \mathcal{L}_r^l}{\partial \boldsymbol{\theta}_r}, \tag{20}$$

$$\boldsymbol{\theta}_f \longleftarrow \boldsymbol{\theta}_f - \mu(\frac{\partial \mathcal{L}_y^l}{\partial \boldsymbol{\theta}_f} - \lambda\frac{\partial \mathcal{L}_r^l}{\partial \boldsymbol{\theta}_f}), \tag{21}$$

where $\mathcal{L}_r^l$ is the angular softmax loss for the l-th mini-batch.

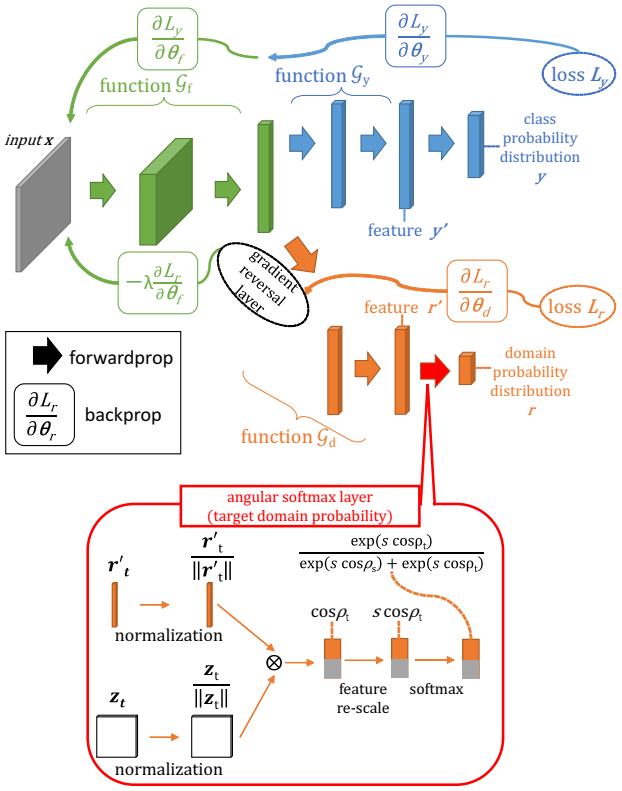

Fig. 1. The pipeline of unsupervised domain adversarial training in angular space

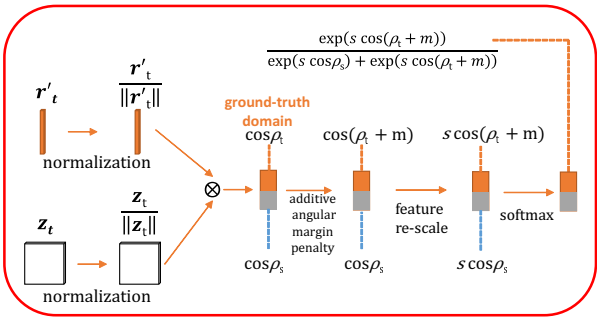

Fig. 2. Unsupervised domain adversarial training in angular space with margin added to ground-truth domain

VI. Experiments

A. Datasets

We use two public facial expression datasets to evaluate UDAT-AS.

a) FER2013: The FER2013 dataset is a face image data set having seven classes of facial expressions: the six basic facial expressions and neutral expression [24]. All 35,887 images are 48x48 resolution grayscale images without background. The images were automatically collected from Google image search using 184 keywords related to emotions; the keywords were manually associated with the seven facial expression categories. We split all images into training data set (28,707 images), and test data set (3,589 images).

b) RAF-DB: The RAF-DB dataset is also a face image data set having the same seven categories [25], [26]. The 15,339 images are 100x100 resolution RGB color images without background. The images were collected from SNS sites using keywords related to emotions and facial expression categories were manually annotated. We split all images into training data set (12,271 images), and test data set (3,068 images).

B. Implementation Details

In our experiments, we used a unified CNN network structure for FER. We used the modified version of the VGG16 architecture, which introduced 13 convolutional layers and 4 fully-connected layers. Input shape of the network was set to $(224, 224, 3)$. Note that the last layer corresponds to a softmax layer with linear transformation. When we performed UDAT, the domain classification network was connected to the second fully-connected layer. The domain classification network was composed of 3 fully-connected layers where the last layer corresponds to the softmax layer or the angular softmax layer.

To train these networks, we used the SGD optimizer where the learning rate was set to 0.001 and batch size was set to 64. For UDAT, we altered parameter λ, which controls the influence of adversarial training, from 0.0-1.0 in steps of 0.1 and selected the value with the highest performance. For the angular softmax loss, we set the scaling parameter s to 64 and margin parameter m to 0.5. For training, we examined two setups. One is flat-start training in which all initial parameters were randomly initialized. The other is fine-tuning of pre-trained model that was trained by VGGface [16].

C. Results

Results of evaluating RAF-DB and FER2013 data sets with exchange of the source and target data are shown on Tables 1 and 2. Table 1 shows results without pre-training while Table 2 shows those with VGGface pre-training.

In each table, S:FER2013 means that FER2013 was set to the source domain. On the other hand, T:RAF-DB means that RAF-DB was set to the target domain. Line 1 shows the ideal results gained by utilizing the labeled target domain data sets. Line 2 shows the results gained by utilizing labeled source domain data sets. The results show that there is a performance gap between line 1 and line 2. This indicates that domain bias is clearly present in the different facial expression datasets. In each table, Lines 3 show results of UDAT using standard softmax loss that used both labeled source domain data sets and unlabeled target domain data sets. Lines 4-5 are for UDAT-AS. The results show that each UDAT method outperformed the use of only labeled source domain data sets. This confirms that UDAT methods are an effective way of improving performance in the target domain. In addition, lines 4 and 5 (UDAT-AS) yielded better performance than line 3 (UDAT). This verified that angular softmax loss was

TABLE I
RECOGNITION ACCURACY ACHIEVED WITHOUT PRE-TRAINED MODEL (%)

Method	S:FER2013 T:RAF-DB	S:RAF-DB T:FER2013
Train on target	74.13	60.15
Source only	58.55	39.04
UDAT (using softmax loss)	59.68	42.21
UDAT-AS	**60.63**	**43.63**
UDAT-AS with margin	**61.34**	**44.41**

TABLE II
RECOGNITION ACCURACY ACHIEVED WITH PRE-TRAINED MODEL (%)

Method	S:FER2013 T:RAF-DB	S:RAF-DB T:FER2013
Train on target	84.78	70.49
Source only	66.10	53.78
UDAT (using softmax loss)	67.57	54.81
UDAT-AS	**68.25**	**55.50**
UDAT-AS with margin	**68.61**	**56.34**

more effective than standard softmax loss. It is thought that the angular softmax loss helped to achieve domain invariant features in the FER model. The highest results were achieved by the angular softmax loss with angular margin. This indicates that increasing the between-class distances in the angular space can improve the performance of unsupervised domain adaptation. Furthermore, Table 1 and 2 show that UDAT-AS offered performance improvements whether the pre-trained model was used or not. Thus, we can conclude that pre-trained methods can be combined with unsupervised domain adversarial training.

Figure 3 shows a visualization of the output feature vectors of the layer before the gradient reversal layer. The feature vectors were dimensionally reduced to 2 dimensional map by t-SNE [27]. In this experiment, we used FER2013 as the source data and RAF-DB as the target data, and each model was trained with pre-training. We used test data of source and target datasets as input data for the feature extraction. The mismatched distribution of target and source features indicates that domain invariant features are not acquired. In the results for "Source Only", features of target data do not overlap those of source data. This indicates that the trained model was not invariant to domain dependency. The results for "UDAT" showed that less mismatches were attained than "Source Only". The results for "UDAT-AS with margin" showed that features of target data match those of source data more clearly. This indicates that our proposed UDAT acquired domain invariant features over conventional UDAT.

VII. CONCLUSION

In this paper, we proposed an unsupervised domain adversarial training (UDAT) method in angular space (UDAT-AS) to improve the performance of unsupervised domain adaptation in facial expression recognition models. Main strength of the proposed method is that domain invariant feature extraction can be well trained by introducing just angular softmax loss to the domain classification network. Experiments showed that

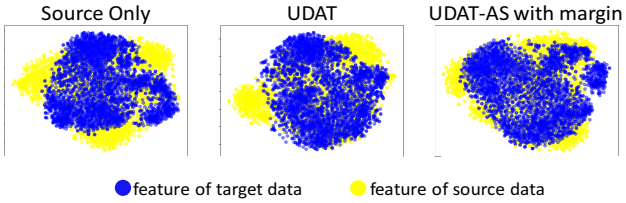

Fig. 3. Visualization of the feature map before the gradient reversal layer

our proposed method offers significant improvements over UDAT and yield better performance in the target domain. In future work, we will apply UDAT-AS to other classification tasks and generation tasks that involve unsupervised domain adaptation.

REFERENCES

[1] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. In *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124 – 129. 1971.
[2] S. Albanie and A. Vedaldi. Learning grimaces by watching tv. In *BMVC*, 2016.
[3] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. In *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18 – 31, Jan.-Mar. 2017.
[4] S. Jialin Pan and Q. Yang. A survey on transfer learning. In *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345 – 1359, Oct. 2010.
[5] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.
[6] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *ICASSP*, 2018.
[7] W. Liu, Y. Wen, Z. Yu, M. Li, BhikshaRaj, and L. Song. SphereFace: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.
[8] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large margin cosine loss for deep face recognition. In *CVPR*, 2018.
[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.
[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
[12] M. Matsugu, K. Mori, Y. Mitari, and Y. Kaneda. Subject independent facial expression recognition with robust face detection using a convolutional neural network. In *Neural Networks*, vol. 16, no. 5 – 6, pp. 555 – 559, Jun.-Jul. 2003.
[13] A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. In *WACV*, 2016.
[14] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool. Covariance pooling for facial expression recognition. In *CVPR Workshops*, 2018.
[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.
[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.
[17] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising face across pose and age. In *FG*, 2018.
[18] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan. Domain separation networks. In *NIPS*, 2016.
[19] J. Feng B. Sun and K. Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.
[20] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu. Variational recurrent adversarial deep domain adaptation. In *ICLR*, 2017.
[21] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018.

[22] X. Wei, H. Li, J. Sun, and L. Chen. Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+3d facial expression recognition. In *FG*, 2018.

[23] W.Zheng, Y. Zong, X. Zhou, and M. Xin. Cross-domain color facial expression recognition using transductive transfer subspace learning. In *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 21-37, Jan.-Mar. 2018.

[24] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, and et al. Challenges in representation learning: A report on three machine learning contests. In *NIPS*, 2013.

[25] L. Shan, D. Weihong, and D. JunPing. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

[26] L. Shan and D. Weihong. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. In *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356-370, Jan. 2019.

[27] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. In *Journal of Machine Learning Research*, vol. 9, no. 1, pp. 2579–2605, 2008.