Dynamic Multi-channel Metric Network for Joint Pose-aware and Identity-invariant Facial Expression Recognition

Yuanyuan Liu[a,b], Wei Dai[a], Fang Fang[a], Yongquan Chen[b,c,*], Rui Huang[b,c], Run Wang[a], Bo Wan[a]

[a] *Faculty of Information Engineering, China University of Geosciences, Wuhan, China*
[b] *The Chinese University of Hong Kong, Shenzhen, China*
[c] *Shenzhen Institute of Artificial Intelligence and Robotics for Society, China*

Abstract

Facial expression recognition (FER) is challenging because the appearance of an expression varies significantly depending on head pose and inter-subject characteristics. With existing techniques, it is often difficult to learn both pose-aware and identity-invariant representations of facial expressions effectively due to the complex distribution of intra-class variation and similarity caused by these two factors. In this study, we propose a dynamic multi-channel metric learning network for pose-aware and identity-invariant FER, called DML-Net, which can reduce the effects of pose and identity for robust FER performance. Specifically, DML-Net uses three parallel multi-channel convolutional networks to learn fused global and local features from different facial regions. Then it uses joint embedded feature learning to explore identity-invariant and pose-aware expression representations from fused region-based features in an embedding space. DML-Net is end-to-end trainable by minimizing deep multiple metric losses, FER loss, and pose estimation loss with dynamically learned loss weights, thereby suppressing overfitting and significantly improving recognition. We evaluate DML-Net on three widely-used multi-view facial expression datasets, namely, KDEF, BU-3DFE, and Multi-PIE, as well as a wild dataset SFEW2.0. Extensive

*Corresponding author at: The Chinese University of Hong Kong, Shenzhen, China
Email address: yqchen@cuhk.edu.cn (Yongquan Chen)

experiments demonstrate that our approach outperforms several other popular methods with accuracies of 88.2% on KDEF, 83.5% on BU-3DFE, 93.5% on Multi-PIE, and 54.36% on SFEW.

Keywords: `multi-view facial expression recognition; pose-aware; identity-invariant; multi-channel metric learning; dynamic weight; multi-task learning`

1. Introduction

Facial expressions are an important nonverbal way for human beings to convey emotions and intentions. Automated facial expression recognition (FER) is crucial to applications involving human-computer interaction, such as emotion robots, automated customer service, interactive games, and driver fatigue monitor systems. Despite tremendous progress in the past decade [1], most existing FER methods focus on near-frontal face evaluation in a constrained environment. Robust multi-view FER remains challenging because of pose variation and inter-subject variation (i.e., identity-specific attributes). These factors cause two difficulties. First, learning representations good for distinguishing different expressions rather than different poses and identities is difficult; second, a great deal of expression-related information is lost because of self-occlusion and inter-subject variation due to pose and identity variation. Since facial expressions often involve only subtle facial muscle movements, expression-unrelated features, and expression-related features couple nonlinearly, degrading FER performance.

Existing methods address the above challenges to improve expression-related features' learning regarding pose variation and individual identity variation. For pose variation, existing methods for multi-view FER are generally divided into three categories: pose-robust feature extraction, pose-specific classification, and pose normalization. Pose-robust features depend on well-designed hand-crafted features or local feature points [2, 3], which have a limited effect on nonlinear facial texture distortion. Pose-specific classification requires a large amount

of data to train a classifier for discrete poses that works synchronously with the expression classifier [4, 5]. Pose normalization typically synthesizes a two-dimensional (2D) or three-dimensional (3D) frontal facial image from a postural facial image through a generative adversarial network (GAN) before expression classification.

Existing methods can be divided into two broad categories for individual identity variation: GAN-based and metric-learning-based methods. GAN-based methods usually generate new facial expression images through adversarial learning to reduce the effect of identity features [6, 7]. Metric-learning-based methods incorporate metric learning schemes within a convolutional neural network (CNN) framework for clustering embedded representations of facial expressions [8, 9]. Although these methods have achieved promising results for FER, most of them address the two interference factors separately. However, pose and identity variations have a joint effect on FER performance and are difficult to separate.

To deal with these limitations and produce facial expression representations that have greater discriminating power regarding both pose and identity variations, we propose a novel dynamically multi-channel metric network for pose-aware and identity-invariant FER, termed DML-Net. Fig. 1 illustrates the motivation for our work. In Fig. 1(a) and 1(b), x_1 and x_2 represent samples with different poses and the same facial expression (Happiness), whereas x_3 and x_1 represent samples with different facial expressions (Afraid and Happiness) and the same pose. Similarly, in Fig. 1(c) and 1(d), x_4 and x_5 represent samples with different identities and the same expression (Happiness), whereas x_6 and x_4 are samples with different facial expressions (Sadness and Happiness) and the same identity. $f(x_i)$ represents the facial expression representation extracted from the i^{th} sample. D_1, D_2, D_3 and D_4 denote the Euclidean distances between the samples' expression representations. In our work, DML-Net aims to learn more discriminative expression-related features, even those in which facial movement is subtle; this means that different facial expressions should be farther apart in the feature space. However, features resulting from pose and

identity differences typically have higher discriminating power than those resulting from differences in facial expression; in Fig. 1(a) and 1(c), $D_1 > D_2$ and $D_3 > D_4$ in the feature space. Hence, to eliminate the effect of pose and identity variation, DML-Net clusters instances of the same expression with different poses and identities while enlarging the distance between different expressions in the embedding feature space. That is, features with the same expressions are closer to each other, even if they have different poses or identities. Features with different expressions are far away from each other, even if they have the same pose or identity: in Fig. 1(b) and 1(d), $D_1 < D_2$ and $D_3 < D_4$ in the embedding feature space.

Fig. 2 provides an overview of DML-Net for pose-aware and identity-invariant FER. It consists of two stages: five-tuple set construction and dynamically multi-channel metric learning. In the first stage, inputs are constructed as five-tuple input. That is, each input contains five samples: a shared anchor sample, a pose-based positive sample (i.e., an image with the same expression as the anchor, but a different pose), a pose-based negative sample (i.e., an image with the same pose as the anchor, but a different expression), an identity-based positive sample (i.e., an image of a different subject with the same expression as the anchor), and an identity-based negative sample (i.e., an image of the same subject as the anchor, but with a different expression).

In the second stage, DML-Net learns pose-aware and identity-invariant expression representations in the embedding space. This stage has three components: multi-channel feature extraction (MFE), jointly embedding feature learning (JEFL), and dynamically weighted multi-task learning (DWML). Specifically, MFE uses three parallel Resnet50 [10] models to extract multi-channel convolutional features from three different face regions: the mouth region, the eye region, and the entire face. Then, the extracted region-based multi-channel fusion features are fed into the JEFL module for further learning and clustering of identity-invariant and pose-aware facial expression representations in the embedding space. Here, the embedded features with the same expression category are closer to each other, whereas those with different expression categories are

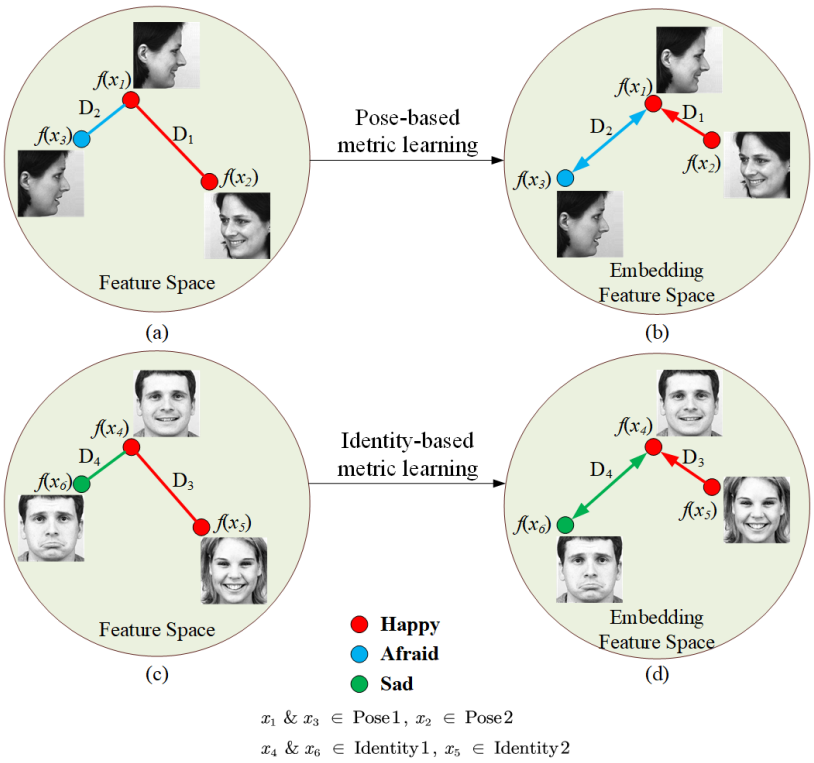

Figure 1: Illustrations of representations in feature space learned by (a) existing methods in pose variation, (b) DML-Net in pose variation, (c) existing methods in inter-identity variation, and (d) DML-Net in inter-identity variation. $f(x_i)$ is the facial expression representation extracted from the i^{th} sample. D_1, D_2, D_3 and D_4 denote the Euclidean distances between the expression representations of the samples.

farther apart. Finally, the DWML module simultaneously focuses pose estimation and FER tasks by minimizing the deep multiple metric losses, the FER loss, and pose estimation loss, with dynamically learned loss weights, which suppresses the overfitting, and vanishing gradient problems and significantly improves recognition.

This study is an extension of a paper presented at the conference FG2018 [11]. The contributions of this study that differ from those of the conference study are as follows:

1) In the conference study, we proposed an end-to-end trainable MPCNN

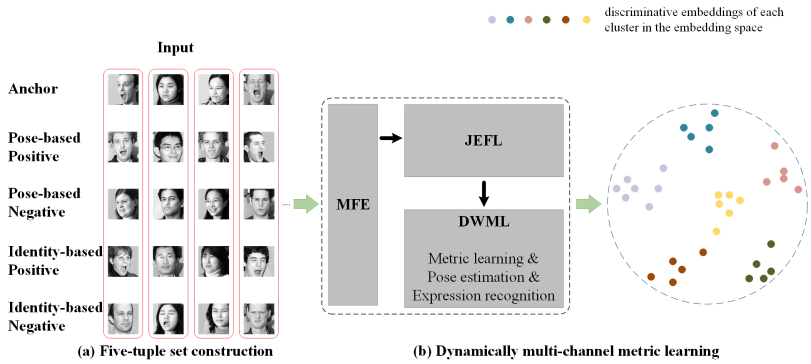

Figure 2: Overview of DML-Net for jointly pose-aware and identity-invariant facial expression recognition. (a) A constructed five-tuple set with five samples in each group. (b) Dynamically multi-channel metric learning. In the multi-channel metric learning stage, MFE extracts global, and local features; then, JEFL aims to map the original features into an embedding feature space so that features with the same expression tend to form clusters, whereas those with different expressions are far apart. Finally, DWML performs three tasks, namely, metric learning, pose estimation, and expression recognition by optimizing dynamic weighting multi-losses. The network aims to acquire more discriminative pose-aware and identity-invariant expression representations by clustering features in the embedding space.

with three components: MFE, multi-scale high-layer feature fusion, and pose-aware recognizer. This study proposes an end-to-end trainable network called DML-Net with three components, MFE, JEFL, and DWML, which jointly recognize facial expressions and estimate poses via multi-channel metric learning. In contrast to the multi-scale feature fusion in MPCNN, DML-Net incorporates a new JEFL module into a multi-task learning framework to acquire the most discriminative expression-relevant representations by clustering fusion features in the embedding space. Our proposed DML-Net is, to the best of our knowledge, the first FER method using deep metric learning (DML) for handling both identity and head pose variations.

2) This work proposes a new JEFL module for further learning identity-invariant and pose-aware facial expression representations via clustering in an embedding space.

3) This work introduces a joint multi-loss function with dynamic loss weights to optimize the entire network and prevent overfitting in the training procedure.

4) Evaluation on three typical and challenging multi-view facial expression datasets shows the advantages of DML-Net over existing state-of-the-art methods.

The rest of this paper is organized as follows: Section 2 introduces related work. Section 3 presents our DML-Net approach for FER. Section 4 discusses our experimental results using publicly available datasets. Section 5 concludes this paper.

2. Related Work

In this section, we mainly discuss methods related to FER, DML, and multi-task network architectures.

Facial expression recognition: As elaborated in the surveys [1], FER technology has advanced considerably over the past decade. However, performing pose-adaptive and identity-invariant FER with limited training data in spontaneous environments remains a challenge. Benefiting from advancements in deep learning (DL) in recent years, more effective methods are emerging to deal with pose and identity issues; these have achieved promising results in some settings. Jung *et al.* [3] trained two CNNs jointly with facial landmarks and color images to reduce the effects of poses. The works [2, 4] employed a deep neural network (DNN) with SIFT. For the multi-view facial expression dataset BU-3DFE, GAN-based approaches have also yielded good results. Zhang *et al.* [12] proposed an end-to-end model for pose-aware FER based on synthesizing multi-view facial images simultaneously using a GAN and achieved an average accuracy of 81.20%. Zhang *et al.* [13] combined facial landmarks with a GAN for FER and achieved an average accuracy of 81.95%. FERAtt [14] further improved the accuracy using an attention mechanism on the BU-3DFE dataset. Identity-adaptive generation (IA-gen) [6] generated six facial expressions with

the same subject from any input image using six cGANs and employed a regular CNN branch for FER. De-expression residue learning (DeRL) [7] learned residual person-independent expression information by generating a standardized neutral face. Exchange-GAN [15] separated identity and expression features using groups of GANs. IPFR [16] proposed a GAN-based structure and achieved good results on the SFEW in-the-wild dataset for a more challenging FER task in-the-wild. Hu *et al.* [17] realized facial de-expression and expression component extraction on the unpaired in-the-wild datasets based on DeRL [7]. Shao *et al.* [18] proved that a shallow CNN could also achieve good scores for FER in-the-wild. However, GAN-based models are costly and depond on the amount of training data. In general, most existing methods focus only on one of the two factors. Their FER performance depends heavily on the results of their pose or identity estimation or the quality of the generated samples. Therefore, it is necessary to build an end-to-end network for multi-view FER with full use of limited data and synthetically consider both pose and identity variation.

DML: Unlike traditional metric learning, DML uses deep learning techniques to learn nonlinear embedding data features in the embedding space. Many researchers have been interested in combining softmax loss and DML for FER in the last few years. IL-CNN [19] proposed the island loss to learn more discriminative deep features by compacting clusters and simultaneously pushing clusters away from each other. These methods are based on global features and do not consider specific factors in FER. Liu *et al.* [9] and [20] introduced identity-invariant FER in metric learning. The method proposed by [9], based on $(N + M)$ tuples, is an improvement of the triplet loss [21]; it optimizes the loss function by pulling M positive examples close to the anchor and pushing N negative examples away from the anchor with a dynamic margin. [20] proposed a hard negative generation network, combining GAN with identity-invariant $(N + M)$ tuples [9]. These DML methods only consider identity-aware feature learning in an embedding space but do not consider pose variation. Inspired by this, we propose a new JEFL module to simultaneously diminish the impacts of pose and identity variation.

Multi-task networks: Since information is shared among related tasks, multi-task learning has been introduced to solve problems caused by various factors in FER, such as poses, identities, and illumination. The identity-invariant CNN (IACNN) [8], based on a distance metric, used a pair of images as input data and learned features for identity-invariant FER by developing two identical sub-CNNs for expression-sensitive constructive loss and identity-sensitive constructive loss, respectively. The multi-signal CNN (MSCNN) [22], which is similar to [6], not only used two images as input data but also employed cross-entropy loss and identity-sensitive constructive loss by the same full-connection layer in a unified network. These works usually use fixed loss weighting parameters or train all tasks equally, which can easily result in overfitting and are time-consuming. Generally, adaptive dynamic weights (ADW) (i.e., Zheng *et al.* [23]) will be more effective. To address this problem, inspired by [24], we introduce a DWML into the framework, adopting an adaptive weighting method more effective training.

3. Methodology

This section presents a novel dynamically multi-channel metric network for pose-aware and identity-invariant FER (DML-Net), which aims to diminish the effect of both pose and inter-subject variation for better FER performance. We first introduce five-tuple set construction for learning the embedding space, then describe the architecture, and implementation details of DML-Net in the training and inference procedures.

3.1. Five-tuple set construction

Instead of the traditional triplet construction used in metric learning, we propose a five-tuple set construction strategy to simultaneously learn information about poses and identities, i.e., the embedding distances of pose variation and identity variation. As shown in Figs. 2 and 3, a five-tuple contains a pose-based triplet and an identity-based triplet with a shared anchor sample.

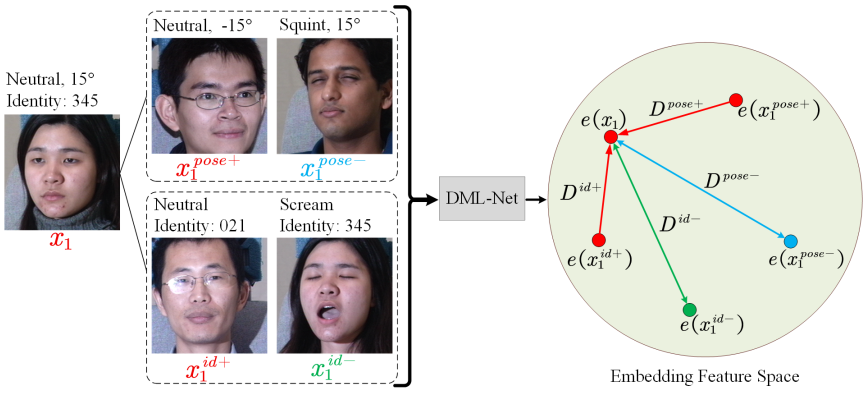

Figure 3: Example of a five-tuple. It contains a pose-based triplet and an identity-based triplet, including a negative sample, a positive sample, and a shared anchor sample. The constructed five-tuples will be fed into DML-Net, where positive images will be pulled closer to the anchor in the embedding space, whereas the negative images are pushed away from the anchor.

When using random sampling for each five-tuple, due to an imbalance of positive and negative facial expression samples in the training data, the number of positive samples is significantly less than the number of negative samples in a training mini-batch. This will make the network ignore large absolute positive distances during training. The five-tuple construction strategy is applied before each training epoch to address this imbalance problem and equalize the number of positive and negative samples.

Algorithm 1 shows the details of the five-tuple set construction strategy. Formally, given a shared anchor sample x_i, we first define and find another four samples in the training set: x_i^{pose+} (the pose-based positive sample of x_i), x_i^{pose-} (the pose-based negative sample of x_i), x_i^{id+} (the identity-based positive sample of x_i), and x_i^{id-} (the identity-based negative sample of x_i). These will form a five-tuple set. We then repeat the process, traversing all samples until no new five-tuple set can be found. Finally, we construct the five-tuple set as follows.

An example of a constructed five-tuple set is shown in Fig. 3. x_1 represents the shared anchor sample, with pose-based triplet containing x_1^{pose+} and

x_1^{pose-} and identity-based triplet containing x_1^{id+} and x_1^{id-}. Points $e(\cdot)$ are the embedding representations of the samples learned by DML-Net. D^{pose+} and D^{id+} are the distances between the images of x_1 and the two positive samples in the embedding space, whereas D^{pose-} and D^{id-} are the distances between the images of x_1 and the two negative samples. DML-Net reduces D^{pose+} and D^{id+} while increasing D^{pose-} and D^{id-} in the embedding space.

Algorithm 1: Five-tuple set construction strategy

Input:

the original dataset S: $\{Sample(x_i, y_i^e, y_i^p, y_i^s)\}_{i=1}^{M}$;

M: the number of images in a dataset; i, j: the iterator index;

x_i: the i^{th} sample; y_i^e: expression label of x_i;

y_i^p: pose label of x_i; y_i^s: identity label of x_i;

initialized an empty set T of five-tuples.

1. shuffle S
2. **For each** x_i **in** S **do**
3. **if** $\exists x_j \in S$, $y_i^e = y_j^e$ **and** $y_i^p \neq y_j^p$ **then:** $x_i^{pose+} \leftarrow x_j$
4. **if** $\exists x_j \in S$, $y_i^e \neq y_j^e$ **and** $y_i^p = y_j^p$ **then:** $x_i^{pose-} \leftarrow x_j$
5. **if** $\exists x_j \in S$, $y_i^e = y_j^e$ **and** $y_i^s \neq y_j^s$ **then:** $x_i^{id+} \leftarrow x_j$
6. **if** $\exists x_j \in S$, $y_i^e \neq y_j^e$ **and** $y_i^s = y_j^s$ **then:** $x_i^{id-} \leftarrow x_j$
7. **if** all of x_i^{pose+}, x_i^{pose-}, x_i^{id+} and x_i^{id-} exsit **then:**
8. put them into a five-tuple with and move the five samples to T
9. update S
10. **end**
11. **end**

Output

the set T of N five-tuples: $\{Tuple(x_i, x_i^{pose+}, x_i^{pose-}, x_i^{id+}, x_i^{id-})\}_{i=1}^{N}$

3.2. Network architecture

In this section, we first introduce the overall architecture of DML-Net with its three main components, then describe the learning process of each component

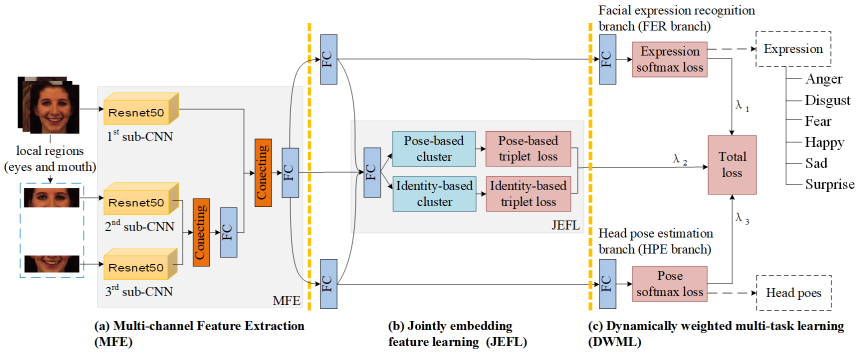

Figure 4: The architecture of DML-Net, with three components: (a) MFE, (b) JEFL, and (c) DWML. First, given a facial sample from the five-tuple set as the input, three parallel sub-CNNs are used to extract the combined global and local facial features. Second, JEFL maps fused features to the embedding space to learn pose-aware and identity-invariant embedding features. Finally, DWML performs joint FER, pose estimation, and embedding distance calculation by minimizing the FER loss, pose estimation loss, and deep multiple metric losses (pose-based triplet loss and identity-based triplet loss) with dynamic weights.

and the optimization of the entire network in detail.

3.2.1. DML-Net framework

The overall architecture of the proposed network is illustrated in Fig. 4. As described above, DML-Net consists of three parts: MFE, JEFL, and DWML. Specifically, MFE uses three parallel convolutional networks to learn fused global and local features from different facial regions; then, JEFL is used to learn further identity-invariant and pose-aware expression representations from the fused region-based features in an embedding space. Finally, DWML jointly calculates feature distances, recognizes facial expressions, and estimates head poses by minimizing the deep multiple metric losses, the FER loss and pose-estimation loss with dynamically learned loss weights.

3.2.2. MFE

Unlike the traditional pre-trained CNN backbone for feature extraction, MFE uses three parallel sub-CNNs to extract multi-channel convolutional features from different facial regions. We want a small-scale feature extractor that

can extract local features from smaller facial regions. However, we also want a large-scale feature extractor that can exploit detailed global features from the entire face to increase accuracy. For this goal, we train separate multi-channel convolution feature extractors, using Resnet50 [10] as the backbone for different facial regions. The input data are cropped and sent into each channel sub-CNN, making feature maps more efficient and robust to limited training data. We let $(M_1; M_2; M_1)$ represent the three channel sub-CNNs; they are defined as follows.

The first channel, M_1, for the entire face: M_1 is used to learn global features of the whole face. It first standardizes the input data and resizes the data to $\times 224$. Then, the standardized data are sent into first sub-CNN to extract global features.

The second channel, M_2, for the eye region: M_2 is used to learn local features from the eye region. Based on the resized data from the first channel, M_2 first crops the eye region, approximately the uppermost third of the face, and then sends the cropped data to the second sub-CNN to extract features.

The third channel, M_3, for the mouth region: M_3 is used to learn local features from the mouth region. Based on the resized data from the first channel, M_3 first crops the mouth region, approximately the lowest third of the face, and then sends the cropped data to the third sub-CNN to extract features.

To enhance the representation ability of the features obtained from limited training data, we join global and local features using two feature fusion layers. The fusion procedure is detailed in Algorithm 2. We first extract multi-channel features using three sub-CNNs, as described above. Then, in the first fusion layer, p^1 and f^1 are computed for local feature representation from v_2 and v_3 by connection and activation, respectively. Finally, the second fusion layer improves the feature representation from the global feature vector v_1. The first fusion feature f^1. f^2 is the region-based fusion feature for output, which will subsequently be fed into the JEFL module.

Algorithm 2: Joint multi-channel feature fusions

Input:

Images of the whole face, the eye region, and the mouth region;

W: weight matrix; b: bias vector;

1. extract feature vectors in each sub-CNN: v_i, with $i = \{1, 2, 3\}$
2. fuse local features (eye and mouth regions) and activate in the first fusion layer:
   $p^1 = Connect(v_2, v_3)$
   $f^1 = Relu(p^1 W^1 + b^1)$
3. fuse global features and local features and activate in the second fusion layer:
   $p^2 = Connect(v_1, f_1)$
   $f^2 = Relu(p^2 W^2 + b^2)$

Output

the fused features f^2

3.2.3. JEFL

To learn the identity-invariant and pose-aware facial expression representations in the embedding space, the JEFL component maps region-based features fused by the MFE component to the embedding features and cluster the embedding features according to joint pose-based and identity-based triplets simultaneously. As shown in Fig. 4(b), the joint clustering operations in the JEFL module can effectively reduce the effect of pose and identity variation based on the similarity between embedding features.

To measure the similarity between embedding features, we introduce the squared Euclidean distance. Given two facial expression images x_1 and x_2, $e(x_1)$ and $e(x_2)$, respectively, denote the embedding representations of x_1 and x_2 learned by JEFL. The squared distance between x_1 and x_2 in the embedding space is defined as

$$D(e(x_1), e(x_2)) = \|e(x_1) - e(x_2)\|_2^2. \quad (1)$$

To reduce the distances between the anchor and its positive samples, JEFL designs the pose-based triplet loss L^{pose} and the identity-based triplet loss L^{id} to cluster the shared anchor with its pose-based positive sample and identity-based positive sample in the embedding space. Due to the complex distribution of intra-class variation and inter-class similarity in FER, using the conventional triplet loss may not yield satisfactory performance. Therefore, instead of the conventional triplet loss, we introduce an online multiple triplet loss within a mini-batch and select the hardest positive/negative samples to compute the online distance loss. Taking pose-based triplets as an example, given an anchor x_i with P pose-based positive samples and N pose-based negative samples, we calculate the hardest pose-based positive distance D_i^{HP+} between x_i and its farthest pose-based positive sample, and the hardest pose-based negative distance D_i^{HP-} between x_i and its nearest pose-based negative sample. That is, the quantities D_i^{HP+} and D_i^{HP-} are defined as follows:

$$D_i^{HP+} = \max D(e(x_i), e(x_p^{pose+})), p = 1, 2, ..., P, \tag{2}$$

$$D_i^{HP-} = \min D(e(x_i), e(x_n^{pose-})), n = 1, 2, ..., N, \tag{3}$$

where P and N represent the number of pose-based positive and negative samples in the mini-batch, respectively. With M anchors in a mini-batch, the pose-based triplet loss L^{pose} is calculated as follows:

$$L^{pose} = \sum_{i=1}^{M} (m^{pose} + D_i^{HP+} - D_i^{HP-}), \tag{4}$$

where m^{pose} indicates the minimum margin values for pose-based triplets, which force the negative samples to be a certain distance away from the positive samples.

The identity-based triplet loss L^{id} is calculated in the same way as L^{pose}. Given the shared anchor x_i, D_i^{HI+} denotes its hardest identity-based positive distance and D_i^{HI-} denotes its hardest identity-based negative distance. With M anchors in a mini-batch, L^{id} is defined as

$$L^{id} = \sum_{i=1}^{M} (m^{id} + D_i^{HI+} - D_i^{HI-}), \tag{5}$$

where m^{id} represents the minimum margin values for identity-based triplets.

The deep multiple metric loss $L^{triplet}$ is defined as a combination of the pose-based triplet loss and identity-based triplet loss:

$$L^{triplet} = L^{pose} + L^{id}. \tag{6}$$

Since pose and identity variation have the same importance, we set the weight of each loss type as 1. When $L^{triplet}$ is reduced, embedding features with the same facial expression move closer to the cluster center, whereas embedding features with different expressions move away from each other. Thus, JEFL can learn pose-aware and identity-invariant embedded expression representations.

3.2.4. DWML

To improve FER performance and suppress overfitting during training, we introduce DWML, which jointly recognizes facial expressions and estimates head poses based on multi-channel metric learning. As shown in Fig. 4(c), FER branch and head pose estimation (HPE) branch have similar structures: each uses a fully connected layer to extract high-dimensional features, then classifies the features using the softmax function. DWML simultaneously minimizes the deep multiple metric losses, the FER loss, and pose-estimation loss.

In traditional multi-task learning, how to set the loss weight for each task is an open problem. To solve this problem, DWML employs a dynamic weighting strategy to adaptively assign weights to each task's loss to balance the importance of the tasks. It suppresses overfitting during training; [24] inspired this idea. The multi-task objective function in DWML includes a weighting expression cross-entropy loss L^{ecls}, a weighting pose cross-entropy loss L^{pcls}, and a weighting deep multiple metric loss $L^{triplet}$; it is defined as

$$L = \lambda_1 L^{ecls} + \lambda_2 L^{pcls} + \lambda_3 L^{triplet}, \tag{7}$$

where λ_1, λ_2 and λ_3 are the loss weights of the tasks, which are adaptively calculated in each mini-batch. Based on the changes in the importance and interactions of the task loss types during training, the multiple joint losses are

balanced and re-weighted via dynamic learning of the loss weights, which suppresses the problems of overfitting and gradient vanishing. In this paper, for $k = 1, 2$, and 3, λ_k is calculated as

$$\lambda_k(t) = \frac{K \exp(\omega_k(t-1)/T)}{\sum_i \exp(\omega_i(t-1)/T)}, \omega_k(t-1) = \frac{L_k(t-1)}{L_k(t-2)}, \tag{8}$$

where $\omega_k(\cdot)$ denotes the relative loss descent rate, t represents the iterator index, and T and K represent the adjusters. λ_k approaches 1 as T increases; this controls the smoothness of weight distribution. The factor K ensures that $\sum \lambda_k(t) = K$. In our experiments, ω_k is calculated from the average loss of five mini-batches.

In summary, DWML can jointly perform FER and pose estimation based on significant discriminative expression representations in the embedding space. Furthermore, it can effectively suppress overfitting problems using an ADW strategy based on online loss values in the training procedure.

4. Experiments

4.1. Datasets and settings

We perform evaluations on four multi-view public datasets: KDEF [25], BU-3DFE [26], Multi-PIE [27] and SFEW2.0 [28]. Some samples from these datasets are shown in Fig. 5.

KDEF is a multi-view facial expression dataset consisting of two groups of images depicting 70 individuals (35 females and 35 males) displaying seven facial expressions (Anger (AN), Disgust (DI), Afraid (AF), Happiness (HA), Sadness (SA), Surprise (SU), and Neutral (NE)) under five pan angles ($-90°$, $-45°$, $0°$, $45°$, and $90°$).

BU-3DFE is a multi-view facial expression dataset containing images of 100 individuals (56 females and 44 males) displaying six facial expressions (AN, DI, Fear (FE), HA, SA, and SU) under nine pan angles ($-90°$, $-60°$,$-45°$, $-30°$, $0°$, $30°$, $45°$, $60°$, and $90°$).

Multi-PIE is a multi-view facial expression dataset containing images of 337 individuals with varying illuminations in a controlled setting. We use partial

Figure 5: Examples of facial expression images from the datasets used in our experiment: (a) KDEF, (b) BU-3DFE, (c) Multi-PIE, and (d) SFEW

images displaying six facial expressions (NE, Smile (SM), SU, Squint (SQ), DI, and Scream (SC)) captured under five pan angles ($-30°$, $-15°$, $0°$, $15°$, and $30°$).

SFEW2.0 is an in-the-wild facial expression dataset, which was the benchmark data for the SReco sub-challenge in EmotiW 2015 [28]. It is constructed using static frames chosen from movies and divided into three sets: a training set (958 images), a validation set (436 images), and a testing set (372 images). The dataset displays seven facial expressions (AN, DI, FE, HA, NE, SA, and SU) without pose labels and identity labels. Unlike the previous three lab datasets, the expressions in the SFEW dataset are more realistic and diverse under unconstrained conditions. In addition to identity and poses, it contains other challenging factors of FER, such as illuminations and occlusions.

Our experiments' training and validation datasets include 3,914 images of 56 subjects from KDEF, 14,112 images of 70 subjects from BU-3DFE, 8,100 images of 293 subjects from Multi-PIE, and 958 images of the training set from SFEW. For testing, we use another 979 images from KDEF, 6,264 images from

BU-3DFE, 660 images from Multi-PIE, and 436 images of the validation set from SFEW. We guarantee that the subjects in the training and test sets are independent. Due to the lack of identity and pose labels in the SFEW dataset, each image's identity was first labeled with face clustering [29]. Then, we used a method [30] to label the yaw angle of head poses in each image. Finally, we manually checked and corrected labeling errors.

We implemented DML-Net using the TensorFlow DL framework. The key training parameters involved in the work are presented in Table 1. The table shows that the ADAM optimizer [31] was used for training with an initial learning rate of 0.01 and a learning rate decay of 0.8. The exponential moving average decay was 0.9999. The weights were initialized from a zero-centered normal distribution with a standard deviation (STD.) of $\frac{1}{512}$ for softmax layers and 0.04 for other FC layers. For dynamic weight calculation, we set K=2 and T=2 empirically; these values achieved the optimum results. The mini-batch size was 15, which contained three different five-tuples. Since the PC branch typically converges faster than the others, we first froze the PC branch and trained the FER branch and the JEFL component interactively. The default value of the epoch was 500, and training stopped when the total loss no longer decreased. The experiments were conducted on a personal computer with Intel(R) Core(TM) i7-8750H CPU at 2.20GHz and 32GB memory, and NVIDIA GeForce GTX 1070Ti.

Table 1: The key training parameters involved in the work.

Parameters		Settings
Optimization	Optimizer	ADAM
	Init learning rate	0.01
	Learning rate decay	0.8
Moving average decay		0.9999
STD. of initial weights	Softmax layers	$\frac{1}{512}$
	Other FC layers	0.04
Dynamic weight	K	2
	T	2
Mini-batch size		15
Epoch		500

4.2. Experiments with KDEF dataset

Fig. 6 shows the confusion matrices for FER and pose estimation using our method on the KDEF dataset. The overall classification accuracy of FER is 88.2%, and pose estimation accuracy is 99.9%. HA is the easiest to recognize among the seven expressions, with the highest accuracy (98.6%). AN and AF are difficult to distinguish because they involve less facial movement, and thus the recognition rates are significantly lower.

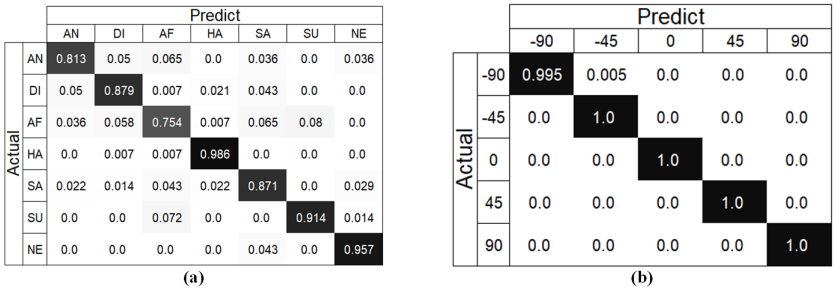

Figure 6: Performance on the KDEF dataset: (a) the FER confusion matrix; (b) the pose estimation confusion matrix.

We compared the performance of our method on the KDEF dataset with that of several state-of-the-art methods: DCNN [32], MPCNN [11], DenseNet

[33], TLCNN [34], SURF boosting [35], and SVM [36]. Comparison results are shown in Table 2. Compared to MPCNN [11], which has the best performance among the other methods, our method improves FER accuracy by more than 1.3% and achieves the highest pose estimation accuracy.

More details are provided in Table 3. The rightmost column shows the average recognition rate for each expression across all five poses, and the bottom row provides the average recognition rate of all expressions under each angle. DML-Net performed well in recognizing HA and NE under all yaw angles. The most difficult expression to recognize was AF from an angle of $-90°$; this may be because the facial movement for this expression is extremely weak and is easily affected by a large change in the head pose.

Since some state-of-the-art methods only use frontal images from the KDFE dataset for FER, Table 4 shows the frontal accuracy on the KDEF dataset using our method, CNN [37], GAN [38], AlexNet [39], and RCFN [40]. Compared to the state-of-the-art methods, our method also achieved an increase of up to 3.5%.

Table 2: Performance comparison on the KDEF dataset in terms of average accuracy for the seven expressions. The best results are in bold.

Methods	Features	Acc. on Poses (%)	Acc. on FER(%)
DML-Net	Multi-channel metric learning	**99.9**	**88.2**
DCNN [32]	Deep CNN features	-	86.44
MPCNN [11]	Fused multi-scale representations	99.8	86.9
DenseNet [33]	Facial image	99.23	85.10
TLCNN [34]	Action Unit Selective Feature	97.55	86.43
SURF boosting [35]	SURF	-	74.05
SVM [36]	LBP and LGBP	86.67	70.5

Table 3: Recognition rate (%) of all expression-angle pairs performed on the KDEF dataset.

Exp./pose	$-90°$	$-45°$	$0°$	$45°$	$90°$	Average
Anger (AN)	78.6	82.1	82.1	89.3	74.1	81.3
Disgust (DI)	96.4	82.1	92.9	75.0	92.9	87.9
Afraid (AF)	70.4	85.7	78.6	75.0	66.7	75.4
Happiness (HA)	100.0	100.0	100.0	92.9	100.0	98.6
Sadness (SA)	85.7	89.3	89.3	85.2	85.7	87.1
Surprise (SU)	92.9	89.3	96.4	88.9	89.3	91.4
Neutral (NE)	100.0	92.9	100.0	92.9	92.9	95.7
Average	89.2	88.8	91.3	85.6	86.1	88.2

Table 4: Frontal FER Performance comparison on the KDEF dataset in terms of average accuracy for the seven expressions. The best results are in bold.

Methods/Exp.	AN	DI	FE	HA	NE	SA	SU	Average
CNN [37]	-	-	-	-	-	-	-	89.4
GAN [38]	80.00	88.67	**97.00**	85.34	**96.00**	87.57	90.23	89.68
AlexNet [39]	78.6	85.7	83.3	**100.00**	92.9	83.3	90.5	87.8
RCFN [40]	-	-	-	-	-	-	-	91.01
DML-Net	**82.1**	**92.9**	78.6	**100.00**	89.3	**96.4**	**100.00**	**91.3**

4.3. Experiments with BU-3DFE dataset

Fig. 7 shows the confusion matrices for FER and pose estimation using our method on the BU-3DFE dataset. Our method achieves average accuracies of 83.5% on FER and 99.4% on pose estimation. Among the six facial expressions in this dataset, two (HA and SU) are identified with more than 90% accuracy. The BU-3DFE results show HA is the most accurate, 96.9%, whereas FE is the least accurate, 55%. Furthermore, most FE instances are misclassified as HA, and 70% are classified as FE in the misclassified HA instances. One possible reason is that the subtle facial movements of these expressions are mainly concentrated in the mouth area and thus are difficult to distinguish.

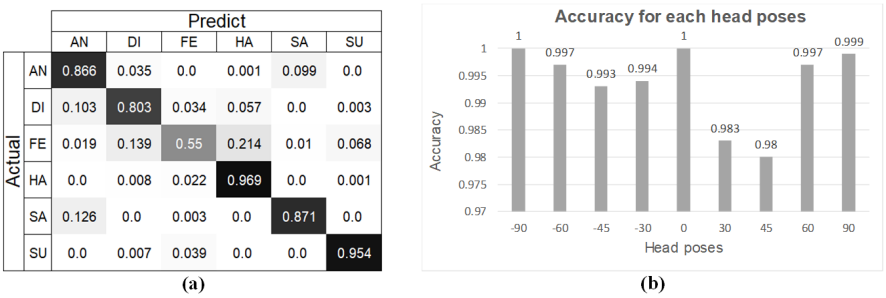

Figure 7: Performance on the BU-3DFE dataset: (a) the FER confusion matrix; (b) the pose estimation confusion matrix.

Table 5: Performance comparison on the BU-3DFE database in terms of average accuracy across seven expressions. The best results are in bold.

Methods	Features	Yaw	Acc. on Poses (%)	Acc. on FER(%)
DML-Net	Multi-channel metric learning	$(-90°, 90°)$	**99.4**	**83.5**
FERAtt [14]	Attention Net	$(-90°, 90°)$	-	82.11
IPFR [16]	Identity and Pose	$(-90°, 90°)$	-	80.9
DenseNet [33]	Dense features	$(-90°, 90°)$	94.45	80.39
PC-RF [5]	Heterogeneity	$(-90°, 90°)$	87.15	76.1
JFDNN [3]	Image and landmarks	$(-90°, 90°)$	-	72.5
GSRRR [2]	Sparse SIFT	$(-90°, 90°)$	87.36	78.9
DNN-D [4]	SIFT	$(-90°, 90°)$	92.26	80.1
SVM [36]	LBP and LGBP	$(-90°, 90°)$	-	71.1
DML-Net	Multi-channel metric learning	$(-45°, 45°)$	**99.2**	**84.8**
IPFR [16]	Identity and Pose	$(-45°, 45°)$	-	84.0
GAN [13]	Image and landmarks	$(-45°, 45°)$	-	81.95
GAN [12]	Convolutional features	$(-45°, 45°)$	95.38	81.2
LLRS [41]	Sparse features	$(-45°, 45°)$	-	78.64
SSE [42]	Supervised super-vector encoding	$(-45°, 45°)$	-	76.60
MMGL [43]	Soft Vector Quantization	$(-45°, 45°)$	-	76.34

Comparison results with existing state-of-the-art methods are given in Ta-

ble 5. We performed multi-view FER on a set of discrete poses, including nine yaw angles for comparison with FERAtt [14], IPFR [16], DenseNet [33], PC-RF [5], JFDNN [3], GSRRR [2], DNN-D [4] and SVM [36]; meanwhile, we used a set, including five yaw angles for comparison with IPFR [16], GAN [13], GAN [12], LLRS [41], SSE [42] and MMGL [43]. Note that the IPFR results [16] were achieved without a generator. The results show that our method performs competitively without additional sample generation, achieving the highest accuracies (99.4% and 99.2%, respectively) for pose estimation.

4.4. Experiments with Multi-PIE dataset

Fig. 8 shows the confusion matrices for FER and pose estimation using our method on the Multi-PIE dataset. The average FER accuracy across all poses is 93.5%, which shows that our method can reduce the effect of pose variation and inter-identity variation. In addition, the average accuracy of pose estimation is 99.7%.

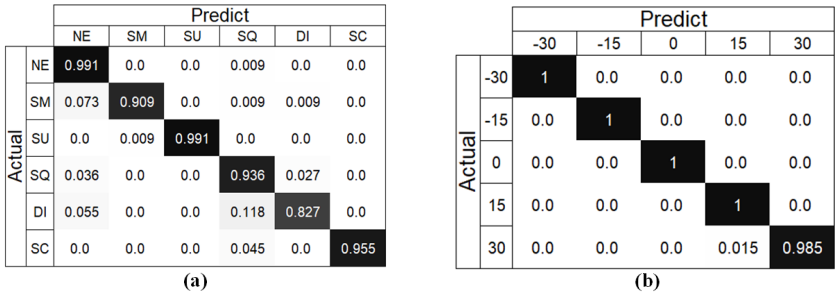

Figure 8: Performance on the Multi-PIE dataset: (a) the FER confusion matrix; (b) the pose estimation confusion matrix.

Table 6 shows the comparison results with the state-of-the-art methods. We thoroughly evaluate our method by comparing its performance with that of GAN [13], Exchange-GAN [15], IPFR [16], MPCNN [11], GAN [12], DenseNet [33], and other state-of-the-art methods reported in [44], namely KNN, LDA, LPP, D-GPLVM, GPLRF, GMLDA, GMLPP, MvDA, and DS-GPLVM. Note that the IPFR [16] results were achieved without a generator. In the results,

there is a clear gap between our method and the other methods at a head pose angle of $-15°$. Most of these methods achieved better results at $-15°$, which may be caused by the greater deformation of training samples generated by a GAN-based structure. However, our method outperforms the other methods by 17.35%–1.2% in terms of FER accuracy, attributable to the pose-aware, and identity-invariant representations learned by our model. The experimental results show that our method achieves a more robust performance despite poses and identity variation.

Table 6: Performance comparison on the Multi-PIE database in terms of accuracy (%) in each pose and average accuracy (%) across six expressions. The best results are in bold.

Methods	Poses					Average
	$-30°$	$-15°$	$0°$	$15°$	$30°$	
DML-Net	**95.5**	90.9	**93.2**	**96.2**	91.7	**93.5**
GAN [13]	92.58	93.74	89.65	89.97	**94.51**	92.09
Exchange-GAN [15]	-	-	-	-	-	91.08
IPFR [16]	-	-	-	-	-	91.3
MPCNN [11]	94.8	91.8	89.91	92.62	91.8	92.3
GAN [12]	90.97	94.72	89.11	93.09	91.3	91.8
DenseNet [33]	90	91.67	90.56	91.67	90.83	91.06
KNN [44]	80.88	81.74	68.36	75.03	74.78	76.15
LDA [44]	92.52	94.37	77.21	87.07	87.47	87.72
LPP [44]	92.42	94.56	77.33	87.06	87.68	87.81
D-GPLVM [44]	91.65	93.51	78.7	85.96	86.04	87.17
GPLRF [44]	91.65	93.77	77.59	85.66	86.01	86.93
GMLDA [44]	90.47	94.18	76.6	86.64	85.72	86.72
GMLPP [44]	91.86	94.13	78.16	87.22	87.36	87.74
MvDA [44]	92.49	94.22	77.51	87.1	87.89	87.84
DS-GPLVM [44]	93.55	**96.96**	82.42	89.97	90.11	90.6

4.5. Experiments with SFEW in-the-wild dataset

Fig. 9 shows the confusion matrices for FER and pose estimation using our method on the SFEW dataset. Owing to the lack of annotations of poses

and facial expressions in the SFEW dataset, we used the Multi-PIE dataset to pre-train the network for feature extraction. We then jointly trained the entire network using the SFEW dataset to model the changes of identities and poses effectively. The overall classification accuracy of FER is 54.39%, and the accuracy of pose estimation is 80.0%. All pose categories achieved an accuracy of more than 50%, especially for $-15°$ (95.7%) and $15°$ (94.9%). The comparison results with the state-of-the-art methods are shown in Table 7. We evaluated our method by comparing its performance with that of IACNN [8], IPFR [16], DLP-CNN [45], DDMTL [23], RAN (Resnet18) [46], DCNN [47], IL-CNN [19], and CNN [48]. Note that the results were achieved using single classifiers. As shown in Table 7, our approach outperformed most of the methods, achieved 83.5% accuracy for HA, 67.5% accuracy for AN, and attained the highest accuracy of 70.9% for NE and 49.1% for SU. The results show that DML-Net achieved a balanced recognition rate in each expression and improved the total FER accuracy as much as possible due to the jointly multi-task network architecture based DML.

Actual \ Predict	AN	DI	FE	HA	NE	SA	SU
AN	0.675	0.026	0.000	0.026	0.117	0.052	0.104
DI	0.130	0.087	0.000	0.174	0.391	0.087	0.130
FE	0.234	0.000	0.085	0.085	0.170	0.085	0.340
HA	0.000	0.000	0.000	0.835	0.096	0.055	0.014
NE	0.012	0.000	0.012	0.012	0.709	0.139	0.116
SA	0.110	0.000	0.069	0.055	0.164	0.397	0.205
SU	0.105	0.000	0.035	0.070	0.228	0.070	0.491

Actual \ Predict	-60	-40	-15	0	15	40	60
-60	0.583	0.375	0.042	0.0	0.0	0.0	0.0
-40	0.115	0.827	0.058	0.0	0.0	0.0	0.0
-15	0.0	0.0	0.957	0.035	0.009	0.0	0.0
0	0.0	0.0	0.010	0.796	0.194	0.0	0.0
15	0.0	0.0	0.0	0.051	0.949	0.0	0.0
40	0.0	0.0	0.0	0.0	0.282	0.590	0.128
60	0.0	0.0	0.0	0.0	0.143	0.347	0.51

Figure 9: Performance on the SFEW dataset: (a) the FER confusion matrix; (b) the pose estimation confusion matrix.

Table 7: Performance comparison on the SFEW dataset in terms of average accuracy for the seven expressions. The best results are in bold.

Methods/Exp.	AN	DI	FE	HA	NE	SA	SU	Average
IACNN [8]	70.7	0	8.9	70.4	60.3	58.8	28.9	50.98
IPFR [16]	73.7	**8.9**	8.9	**89.0**	69.9	**61.8**	47.1	**55.1**
DLP-CNN [45]	-	-	-	-	-	-	-	51.05
DDMTL [23]	**78.23**	8.05	**9.67**	76.3	61.9	35.8	47.44	51.21
RAN (Resnet18) [46]	-	-	-	-	-	-	-	54.19
DCNN [47]	-	-	-	-	-	-	-	52.5
IL-CNN [19]	-	-	-	-	-	-	-	52.52
CNN [48]	-	-	-	-	-	-	-	52.75
DML-Net	67.5	8.7	8.5	83.5	**70.9**	39.7	**49.1**	54.39

4.6. Ablation study and discussion

4.6.1. Effect of different components in DML-Net

In this section, we verify the impact of each component of DML-Net on its final performance on the three datasets (KDEF, BU-3DFE and Multi-PIE). The components we consider are MFE, HPE, JEFL, and ADW. The baseline used ResNet50 [10] as the backbone, with only one channel extracting global features for FER. Table 8 shows the results of an ablation study in which the above training components were added to the baseline framework one at a time.

Table 8: Ablation study of DML-Net. Impacts of integrating the three components (MFE, JEFL, and DWML) into the baseline on the three datasets. The best results are in bold.

Methods	KDEF			BU-3DFE			Multi-PIE		
	Acc. on Poses	Acc. on FER	FPS	Acc. on Poses	Acc. on FER	FPS	Acc. on Poses	Acc. on FER	FPS
ResNet50 [10]	-	83.7	64	-	78.4	136	-	90.5	98
MFE	-	84.0	54	-	81.8	100	-	90.5	77
MFE+HPE	100.0	85.5	55	99.2	82.3	100	99.7	91.2	74
MFE+HPE+JEFL	100.0	86.1	52	99.2	82.6	95	99.7	93.0	78
MFE+HPE+JEFL+ADW	**99.9**	**88.2**	53	**99.4**	**83.5**	98	**99.7**	**93.5**	77

First, integrating MFE improves FER accuracy to 84.0% on the KDEF dataset and 81.8% on the BU-3DFE dataset because MFE helps extract the fusion features from the entire face and expression-related local regions. How-

ever, the improvement is almost invisible on the Multi-PIE dataset. The main reason may be that the differences among the local facial regions between $-30°$ and $30°$ in the Multi-PIE dataset are extremely small to reflect the advantages of the enhanced fusion features. Since it has multiple channels with ResNet50 backbones, MFE reduces processing speed on the three datasets by 16%, 26%, and 21%, respectively.

Second, HPE improves FER accuracy by 1.5%, 0.5%, and 0.7%, respectively, on the three datasets. We believe that HPE aids in the learning of pose-aware representations for each region.

JEFL continuously improves FER accuracy by 0.6%, 0.3%, and 1.8%, respectively, on the three datasets. Finally, ADW brings further accuracy improvements of 2.1%, 0.9%, and 0.5%, respectively. The HPE results show that ADW slightly decreases HPE (only 0.1% reduction on the KDEF dataset). This is because of the tradeoff between the loss of HPE and the pose-aware metric loss. Thus, compared to the baseline, our model achieves overall FER accuracy improvements of 4.5% on the KDEF dataset, 5.1% on the BU-3DFE dataset and 3% on the Multi-PIE dataset. Furthermore, compared to MFE, integrating all components can achieve the best performance with a small additional computational cost (declined average 1 FPS). The proposed method can improve FER and pose estimation accuracy and efficiency.

4.6.2. Visualization of pose-aware and identity-invariant representations in JEFL

We visualized the expression features with different settings in 2D feature space using the Barnes-Hut t-SNE visualization scheme [49] on the BU-3DFE dataset to evaluate the impact of deep multiple metric learning in JEFL. The visualizations include the following four cases: without JEFL, with JEFL using only pose-based triplets, with JEFL using only identity-based triplets, and with JEFL using pose-based and identity-based triplets jointly.

Fig. 10 shows a 2D t-SNE [49] visualization of expression features without JEFL. The features in different expression categories overlap, whereas features with the same expression are divided into nine subgroups corresponding to the

nine head poses in the BU-3DFE dataset. This shows that without JEFL, pose and inter-subject variations cause significant intra-class variation and inter-class similarity between the extracted expression features.

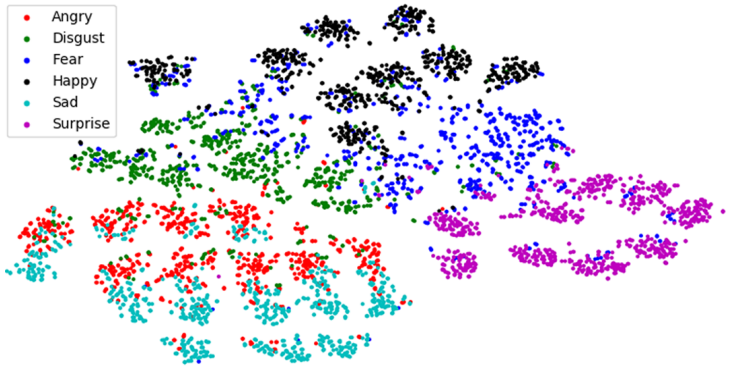

Figure 10: Barnes-Hut t-SNE [49] visualization of the expression features extracted by the FER branch in DWML without JEFL on the BU-3DFE dataset. Each color represents one of the six emotions.

Fig. 11 shows the visualization of expression features using JEFL with only pose-based triplets. Compared to Fig. 10, intra-class differences are significantly lower, and the effect of pose variation is effectively reduced. However, features with the same expression are not sufficiently cohesive, and inter-class overlap is still large due to the influence of identity.

Fig. 12 shows the visualization of expression features using JEFL with only identity-based triplets. Compared to Fig. 10, classes are more separable and the overlap between different expression categories is reduced, indicating that the influence of identity has been effectively reduced. However, there are still sub-classes within each expression due to the influence of pose.

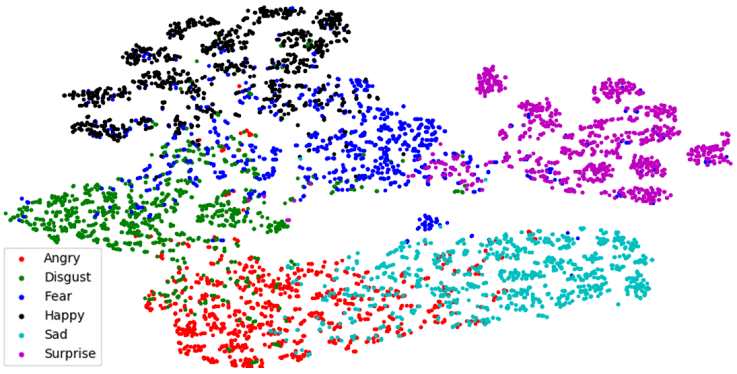

Figure 11: Barnes-Hut t-SNE [49] visualization of the expression features extracted by the FER branch in DWML using JEFL with only pose-based triplets on the BU-3DFE dataset. Each color represents one of the six emotions.

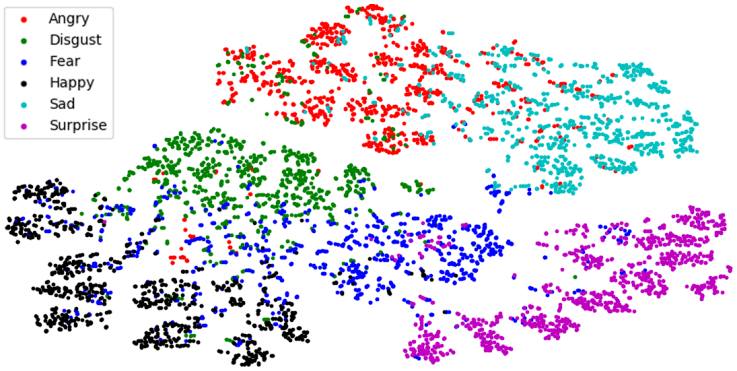

Figure 12: Barnes-Hut t-SNE [49] visualization of the expression features extracted by the FER branch in DWML using JEFL with only identity-based triplets on the BU-3DFE dataset. Each color represents one of the six emotions.

Fig. 13 shows the visualization of expression features using JEFL with pose-base and identity-based triplets simultaneously. Compared to Figs. 11 and 12, the results present better clustering, with a more compact intra-class distance and less overlap between classes. This means that deep multiple triplet metric learning successfully reduces the influence of both pose and identity variation, and thus, learned embedding expression features are more discriminative.

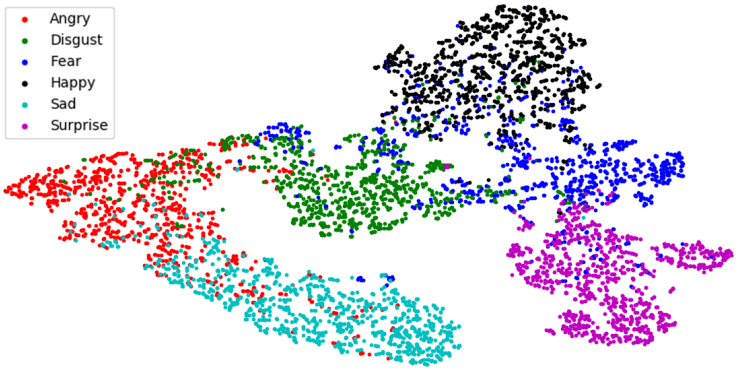

Figure 13: Barnes-Hut t-SNE [49] visualization of the expression features extracted by the FER branch in DWML using JEFL with pose-based and identity-based triplets jointly on the BU-3DFE dataset. Each color represents one of the six emotions.

In addition, to evaluate the robustness of our DML-Net to identity variance using the BU-3DFE dataset, from facial images classified as having HA, we selected frontal ones for identity clustering. Fig. 14 shows the the top-10 images of the clustering. The 31 subjects in the images are divided into four categories, and most facial images of the same subject are clustered into the same category, which indicates that our model is identity-robust.

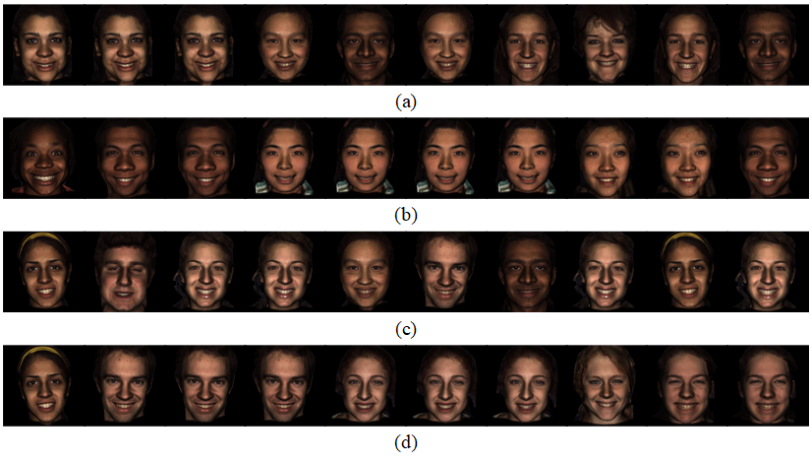

Figure 14: Top-10 images of identity clustering obtained from the class "happiness" on the BU-3DFE dataset.

4.6.3. Visualization of ADW in DWML

We further compared the loss changes for each task before and after adding ADW to DWML. Fig. 15 shows the training procedures of the deep multiple metric losses, the FER loss, and pose-estimation loss in the first 1000 mini-batches on the Multi-PIE and BU-3DFE datasets. Downtrends of the three losses are more similar and smoother when using ADW, which is in line with our expectations.

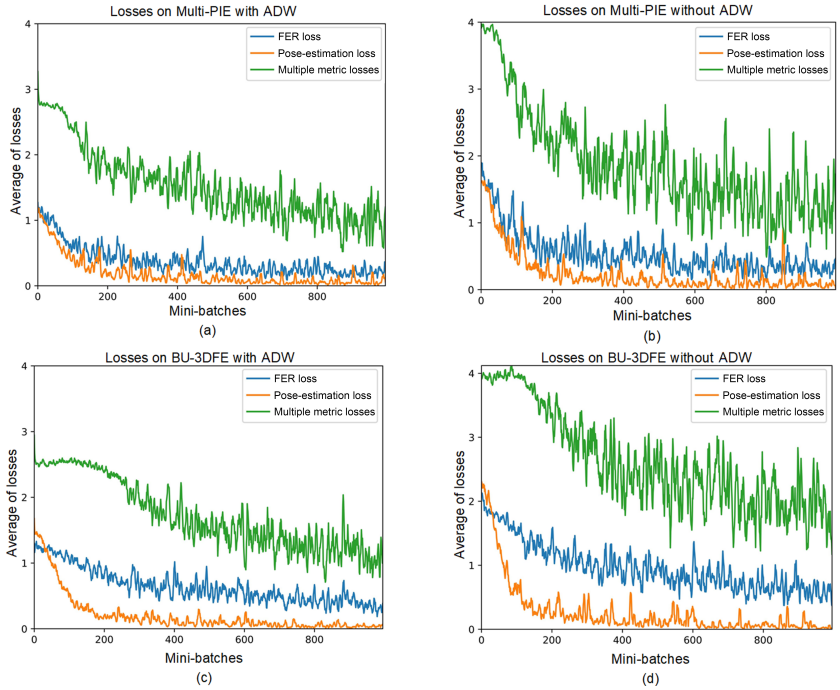

Figure 15: Training procedures of three losses in DWML on two datasets: (a) trained with ADW and (b) trained without ADW on the Multi-PIE dataset; (c) trained with ADW and (d) trained without ADW on the BU-3DFE dataset.

4.6.4. Cross-database experiment on BU-3DFE↦SFEW

In addition, to verify the generalizability of DML-Net, cross-database validation was conducted on the challenging in-the-wild SFEW dataset. First, images from the BU-3DFE dataset were used for training, whereas images from the SFEW validation set were used for testing without fine-tuning. Table 9

presents the comparison results of the proposed model and state-of-the-art methods, including MvDA, GMLDA, GMLPP, DS-GPLVM and GAN reported in [13]. Although the training and testing datasets have different settings (e.g., pose, lighting, ethnicity, glasses, age, etc.), the results of DML-Net demonstrate that it is reusable for expression recognition on the SFEW dataset. Our method achieved an average accuracy of 27.13% while significantly improving the recognition rate of HA (53.42%) and NE (37.21%).

Table 9: Cross-validation comparison with state-of-the-art methods. BU-3DFE → SFEW The best results are in bold.

Methods/Exp.	AN	DI	FE	HA	NE	SA	SU
MvDA [13]	23.21	17.65	27.27	40.35	27.00	10.10	13.19
GLMDA [13]	23.21	17.65	**29.29**	21.93	25.00	11.11	10.99
GLMPP [13]	19.07	21.18	27.27	39.47	20.00	19.19	16.48
DS-GPLVM [13]	25.89	28.24	17.17	42.98	14.00	**33.33**	10.99
GAN [13]	**29.09**	24.88	17.65	51.19	20.00	29.20	18.70
IPER [16]	27.30	**28.90**	24.30	38.70	19.70	26.20	**31.40**
DML-Net	23.38	21.74	17.02	**53.42**	**37.21**	17.81	19.30

5. Conclusion

This study proposes an effective end-to-end trainable network, DML-Net, for pose-aware and identity-invariant FER. DML-Net consists of two stages. A five-tuple set is constructed in the first stage. In the second stage, DML-Net first employs multi-channel sub-CNNs to extract region-based fused features, then maps the fused features to the embedding space for multiple triplet metric learning. Finally, DML-Net jointly recognizes facial expressions and estimates poses based on multi-channel metric learning by minimizing the deep multiple metric losses, the FER loss, and pose-estimation loss with dynamically learned loss weights. Our method outperforms existing state-of-the-art methods in terms of performance and robustness, with the highest accuracies of 93.5% in multi-view FER and 99.9% in pose estimation. In future work, we will introduce context attention mechanisms and apply the model in an unconstrained environment

(i.e., in-the-wild).

6. Acknowledgments

This work was partially supported by Shenzhen Fundamental Research grant (JCYJ20180508162406177, JCYJ20190813170601651), National Natural Science Foundation of China grant (62076227), Wuhan Applied Fundamental Frontier Project Grant (2020010601012166), and the Shenzhen Institute of Artificial Intelligence and Robotics for Society (No. AC01202005024, No. AC01202108001-04, and No. AC01202101010).

References

References

[1] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Transactions on Affective Computing (2020) 1–1.

[2] W. Zheng, Multi-view facial expression recognition based on group sparse reduced-rank regression, IEEE Transactions on Affective Computing 5 (1) (2014-01) 71–85. doi:10.1109/TAFFC.2014.2304712.

[3] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, 2015-12, pp. 2983–2991. doi:10.1109/ICCV.2015.341.

[4] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, K. Yan, A deep neural network-driven feature learning method for multi-view facial expression recognition, IEEE Transactions on Multimedia 18 (12) (2016-12) 2528–2536. doi:10.1109/TMM.2016.2598092.

[5] A. Dapogny, K. Bailly, S. Dubuisson, Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests,

IEEE Transactions on Affective Computing 10 (2) (2019-04-01) 167–181. doi:10.1109/TAFFC.2017.2708106.

[6] H. Yang, Z. Zhang, L. Yin, Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018-05, pp. 294–301. doi:10.1109/FG.2018.00050.

[7] H. Yang, U. Ciftci, L. Yin, Facial expression recognition by de-expression residue learning, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018-06, pp. 2168–2177. doi:10.1109/CVPR.2018.00231.

[8] Z. Meng, P. Liu, J. Cai, S. Han, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, in: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE, 2017, pp. 558–565. doi:10.1109/FG.2017.140.

[9] X. Liu, B. V. K. V. Kumar, J. You, P. Jia, Adaptive deep metric learning for identity-aware facial expression recognition, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017-07, pp. 522–531. doi:10.1109/CVPRW.2017.79.

[10] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016-06, pp. 770–778. doi:10.1109/CVPR.2016.90.

[11] Y. Liu, J. Zeng, S. Shan, Z. Zheng, Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018-05, pp. 458–465. doi:10.1109/FG.2018.00074.

[12] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint pose and expression modeling for facial expression recognition, in: 2018 IEEE/CVF Conference on Computer

Vision and Pattern Recognition, IEEE, 2018-06, pp. 3359–3368. doi:10.1109/CVPR.2018.00354.

[13] F. Zhang, T. Zhang, Q. Mao, C. Xu, Geometry guided pose-invariant facial expression recognition, IEEE Transactions on Image Processing 29 (2020) 4445–4460. doi:10.1109/TIP.2020.2972114.

[14] P. D. Marrero Fernandez, F. A. Guerrero Pena, T. Ing Ren, A. Cunha, Feratt: Facial expression recognition with attention net, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019.

[15] L. Yang, Y. Tian, Y. Song, N. Yang, K. Ma, L. Xie, A novel feature separation model exchange-gan for facial expression recognition, Knowledge-Based Systems 204 (2020) 106217. doi:https://doi.org/10.1016/j.knosys.2020.106217.

[16] C. Wang, S. Wang, G. Liang, Identity- and pose-robust facial expression recognition through adversarial feature learning, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, 2019, p. 238–246. doi:10.1145/3343031.3350872.

[17] J. Hu, B. Yu, Y. Yang, B. Feng, Towards facial de-expression and expression recognition in the wild, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), 2019, pp. 157–163. doi:10.1109/ACII.2019.8925461.

[18] J. Shao, Y. Qian, Three convolutional neural network models for facial expression recognition in the wild, Neurocomputing 355 (2019) 82–92.

[19] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, Y. Tong, Island loss for learning discriminative features in facial expression recognition, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 302–309. doi:10.1109/FG.2018.00051.

[20] X. Liu, B. Vijaya Kumar, P. Jia, J. You, Hard negative generation for identity-disentangled facial expression recognition, Pattern Recognition 88 (2019) 1 – 12. doi:10.1016/j.patcog.2018.11.001.

[21] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognition 48 (10) (2015-10) 2993–3003. doi:10.1016/j.patcog.2015.04.005.

[22] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Transactions on Image Processing 26 (9) (2017) 4193–4203.

[23] H. Zheng, R. Wang, W. Ji, M. Zong, W. K. Wong, Z. Lai, H. Lv, Discriminative deep multi-task learning for facial expression recognition, Information Sciences 533 (2020) 60 – 71. doi:https://doi.org/10.1016/j.ins.2020.04.041.

[24] S. Liu, E. Johns, A. J. Davison, End-to-end multi-task learning with attention, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019-06, pp. 1871–1880. doi:10.1109/CVPR.2019.00197.

[25] D. Lundqvist, A. Flykt, A. Öhman, The karolinska directed emotional faces – KDEF, CD ROM from department of clinical neuroscience, psychology section, Karolinska Institutet (1998) 91–630.

[26] L. Yin, X. Wei, Y. Sun, J. Wang, M. Rosato, A 3d facial expression database for facial behavior research, in: 7th International Conference on Automatic Face and Gesture Recognition (FGR06), IEEE, 2006, pp. 211–216. doi:10.1109/FGR.2006.6.

[27] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, Image and Vision Computing 28 (5) (2010-05) 807–813. doi:10.1016/j.imavis.2009.08.002.

[28] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI '15, Association for Computing Machinery, 2015, p. 423–426. doi:10.1145/2818346.2829994.

[29] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823. doi:10.1109/CVPR.2015.7298682.

[30] M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods, Pattern Recognition 71 (2017) 132–143. doi:https://doi.org/10.1016/j.patcog.2017.06.009.

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2017). arXiv:1412.6980.

[32] M. I. Ul Haque, D. Valles, Facial expression recognition using dcnn and development of an ios app for children with asd to enhance communication abilities, in: 2019 IEEE 10th Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON), 2019, pp. 0476–0482. doi:10.1109/UEMCON47517.2019.8993051.

[33] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017-07, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[34] Y. Zhou, B. E. Shi, Action unit selective feature maps in deep networks for facial expression recognition, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, 2017-05, pp. 2031–2038. doi:10.1109/IJCNN.2017.7966100.

[35] Q. Rao, X. Qu, Q. Mao, Y. Zhan, Multi-pose facial expression recognition based on SURF boosting, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2015, pp. 630–635. doi:10.1109/ACII.2015.7344635.

[36] S. Moore, R. Bowden, Local binary patterns for multi-view facial expression recognition, Computer Vision and Image Understanding 115 (4) (2011-04) 541–558. doi:10.1016/j.cviu.2010.12.001.

[37] R. Melaugh, N. Siddique, S. Coleman, P. Yogarajah, Facial expression recognition on partial facial sections, in: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA), 2019, pp. 193–197. doi:10.1109/ISPA.2019.8868630.

[38] Q. Chu, M. Hu, X. Wang, Y. Gu, T. Chen, Facial expression recognition based on contextual generative adversarial network, in: 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS), 2019, pp. 120–125. doi:10.1109/CCIS48116.2019.9073699.

[39] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, H. Zhou, Deep convolution network based emotion analysis towards mental health care, Neurocomputing 388 (2020) 212–227. doi:https://doi.org/10.1016/j.neucom.2020.01.034.

[40] Y. Ye, X. Zhang, Y. Lin, H. Wang, Facial expression recognition via region-based convolutional fusion network, Journal of Visual Communication and Image Representation 62 (2019) 1–11. doi:https://doi.org/10.1016/j.jvcir.2019.04.009.

[41] M. Jampour, T. Mauthner, H. Bischof, Multi-view facial expressions recognition using local linear regression of sparse codes, in: Proceedings of the 20th Computer Vision Winter Workshop Paul Wohlhart, 2015.

[42] U. Tariq, J. Yang, T. S. Huang, Supervised super-vector encoding for facial

expression recognition, Pattern Recognition Letters 46 (2014-09) 89–95. doi:10.1016/j.patrec.2014.05.011.

[43] U. Tariq, J. Yang, T. S. Huang, Maximum margin GMM learning for facial expression recognition, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013-04, pp. 1–6. doi:10.1109/FG.2013.6553794.

[44] S. Eleftheriadis, O. Rudovic, M. Pantic, Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition, IEEE Transactions on Image Processing 24 (1) (2015-01) 189–204. doi:10.1109/TIP.2014.2375634.

[45] S. Li, W. Deng, Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition, IEEE Transactions on Image Processing 28 (1) (2019) 356–370. doi:10.1109/TIP.2018.2868382.

[46] K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, IEEE Transactions on Image Processing 29 (2020) 4057–4069. doi:10.1109/TIP.2019.2956143.

[47] B. K. Kim, J. Roh, S. Y. Dong, S. Y. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, Journal on Multimodal User Interfaces 10 (2) (2016) 1–17.

[48] Y. Gan, J. Chen, L. Xu, Facial expression recognition boosted by soft label with a diverse ensemble, Pattern Recognition Letters 125 (2019) 105–112. doi:https://doi.org/10.1016/j.patrec.2019.04.002.

[49] L. Van Der Maaten, Accelerating t-SNE using tree-based algorithms, The Journal of Machine Learning Research 15 (1) (2014) 3221–3245.