Unpaired Multimodal Facial Expression Recognition

Bin Xia[1][0000−0001−6529−5434] and Shangfei Wang[1,2][0000−0003−1164−9895]

[1]Key Lab of Computing and Communication Software of Anhui Province,
School of Computer Science and Technology,
University of Science and Technology of China
[2]Anhui Robot Technology Standard Innovation Base
xiabin@mail.ustc.edu.cn, sfwang@ustc.edu.cn

Abstract. Current works on multimodal facial expression recognition typically require paired visible and thermal facial images. Although visible cameras are readily available in our daily life, thermal cameras are expensive and less prevalent. It is costly to collect a large quantity of synchronous visible and thermal facial images. To tackle this paired training data bottleneck, we propose an unpaired multimodal facial expression recognition method, which makes full use of the massive number of unpaired visible and thermal images by utilizing thermal images to construct better image representations and classifiers for visible images during training. Specifically, two deep neural networks are trained from visible and thermal images to learn image representations and expression classifiers for two modalities. Then, an adversarial strategy is adopted to force statistical similarity between the learned visible and thermal representations, and to minimize the distribution mismatch between the predictions of the visible and thermal images. Through adversarial learning, the proposed method leverages thermal images to construct better image representations and classifiers for visible images during training, without the requirement of paired data. A decoder network is built upon the visible hidden features in order to preserve some inherent features of the visible view. We also take the variability of the different images transferability into account via adaptive classification loss. During testing, only visible images are required and the visible network is used. Thus, the proposed method is appropriate for real-world scenarios, since thermal imaging is rare in these instances. Experiments on two benchmark multimodal expression databases and three visible facial expression databases demonstrate the superiority of the proposed method compared to state-of-the-art methods.

1 Introduction

Facial expression is one of the most important emotion communication channels for human-computer interaction. Great progress has recently been made on facial expression recognition due to its wide application in many user-centered fields. Due to their ubiquity, visible images are widely used to build facial expression

recognition systems. However, the light-sensitive property of the visible images prevents researchers from constructing better facial expression classifiers. To tackle this problem, researchers turn to thermal images, which record the temperature distribution of the face and are not light sensitive. Therefore, combining visible and thermal images could improve the facial expression recognition task.

The simplest method of multimodal facial expression recognition is feature-level or decision-level fusion [1–4]. However, the unbalanced quantity of visible and thermal images prevents us from applying this method in real-world scenarios. It would be more practical to utilize thermal images as privileged information [5] to assist the learning process of the visible classifier during training. During testing, thermal images are unavailable and only visible images are used to make predictions. Several works apply this learning framework and succeed in the task of visible facial expression recognition [6–8]. One assumption of these works is that the visible and thermal facial images must be paired during training. Since collecting paired visible and thermal facial images is often difficult, requiring paired data during training prevents the usage of the many available unpaired visible and thermal images, and thus may degenerate the learning effect of the visible facial expression classifier.

To address this, we propose an unpaired adversarial facial expression recognition method. We tackle the unbalanced quantity of visible and thermal images by utilizing thermal images as privileged information. We introduce adversarial learning on the feature-level and label-level spaces to cope with unpaired training data. Finally, we add a decoder network to preserve the inherent visible features.

2 Related Work

2.1 Learning with Privileged Information

Privileged information refers to extra information available during training, but not testing. Exploring privileged information can improve the learning process of the original classifier. Many different research fields have made progress by leveraging privileged information. For example, Vapnik *et al.* [5] first introduced privileged information to the support vector machine (SVM) algorithm, and proposed the SVM+ algorithm. Wang *et al.* [9] proposed to utilize privileged information as secondary features or secondary targets to improve classifier performance. Sharmanska *et al.* [10] proposed to close the mismatch between class error distributions in privileged and original data sets to address cross-data-set learning. Niu *et al.* [11] proposed a framework called multi-instance learning with privileged information (MIL-PI) for action and event recognition, which incorporated privileged information to learn from loosely labelled web data. Luo *et al.* [12] proposed a graph distillation framework for action detection, which not only transfers knowledge from the extra modalities to the target modality, but also transfers knowledge from the source domain to the target domain. Garcia *et al.* [13] proposed to learn a hallucination network within a multimodal-stream network architecture by utilizing depth images as privileged information. They

introduced an adversarial learning strategy that exploited multiple data modalities at training.

All of the above works demonstrate the benefits of leveraging privileged information. However, all works except for Sharmanska *et al.*'s method [10] require paired privileged and original information during training. This requirement prevents the adoption of large-scale unpaired data to learn better features and classifiers for the original information. Although Sharmanska *et al.*'s method leveraged the unpaired data from the target label by forcing a similarity between the classification errors of the privileged and original information, it didn't explore the feature-level dependencies between privileged and original information. Therefore, we propose to learn visible facial expressions classifier and leverage adversarial learning to force statistical similarity between the visible and thermal views.

2.2 Facial Expression Recognition

Facial expression recognition (FER) has remained as an active research topic during the past decades. The main goal of FER is to learn expression-related features that is discriminative and invariant to variations such as pose, illumination, and identity-related information. Traditionally, previous works have used handcrafted features to study facial expression recognition, including Histograms of Oriented Gradients (HOG) [14], Scale Invariant Feature Transform (SIFT) [15], histograms of Local Binary Patterns (LBP) [16] and histograms of Local Phase Quantization (LPQ) [17].

Recently, deep CNN based methods have been employed to increase the robustness of FER. Identity-Aware CNN (IACNN) [18] was proposed to enhance FER performance by reducing the effect of identity related information with the help of an expression-sensitive contrastive loss and an identity-sensitive contrastive loss. Cai *et al.* [19] transferred facial expressions to a fixed identity to mitigate the effect of identity-related information. De-expression Residue Learning (DeRL) [20] utilized the cGAN to synthesize a neutral facial image of the same identity from any input expressive image, while the person-independent expression information can be extracted from the intermediate layers of the generative model. Although these works mitigate the influence of inter-subject variations, the light-sensitive property of the visible images prevents these works from constructing robust facial expression classifiers under different illumination.

2.3 Multimodal Facial Expression Recognition

Early methods of multimodal facial expression recognition are based on strategies including feature-level and decision-level fusions. For example, Sharma *et al.* [2] concatenated the visible and thermal features and fed them into an SVM to detect the pressure of people. Yoshitomi *et al.* [1] trained three classifiers with voice features, visible images, and thermal images, and adopted decision-level fusion. Wesley *et al.* [3] trained visible and thermal networks and fused the outputs to make predictions for facial expression recognition. Wang *et al.*

[4] adopted both feature-level and decision-level fusions to recognize facial expressions. All of the above works ignore the unbalanced quantity of visible and thermal images. In fact, visible cameras are widely used in our daily life, while thermal cameras are only available in laboratory environment. It prevents us from applying the fusion method into real practice.

To address this problem, some researchers view thermal images as privileged information, which is only required during training to help visible images construct a better expression classifier. Unlike fusion methods, which depend on visible and thermal representations containing complementary information, using thermal images as privileged information allows for more robust visible representations. Such visible representations contain unified and view-irrelevant information, rather than complementary information, so only visible data is required during testing. Shi *et al.* [6] proposed a method of expression recognition from visible images with the help of thermal images as privileged information. They combined canonical correlation analysis (CCA) and SVM. Through CCA, a thermal-augmented visible feature space is obtained. The SVM is used as the classifier on the learned subspace. The shortcoming of this method is that the learned subspace has no direct relation to the target label, since the subspace and the classifier are trained separately. To address this, Wang *et al.* [7] proposed to train two deep neural networks to extract feature representations from visible and thermal images. Then, two SVMs were trained for classification. Training of the deep networks and SVMs is integrated into a single optimization problem through the use of a similarity constraint on the label space. The main drawback of Wang *et al.*'s method is that the visible and thermal networks merely interact with each other by means of the similarity constraint, which works on the label space of the two views. There is still great freedom for the feature space below, weakening the correlation between visible and thermal views. Pan *et al.* [8] improved Wang *et al.*'s framework by introducing a discriminator to the hidden features of the two-view networks in order to learn view-irrelevant feature representations and enhance the correlation of the visible and thermal networks in the feature representations. Sankaran *et al.* [21] proposed cross-modality supervised representation learning for facial action unit recognition. They used a latent representation decoder to reconstruct thermal images from visible images. The generated thermal images were applied to construct action unit classifier.

All of the above works require paired visible and thermal images during training. However, it is impractical to collect a great number of paired images in real-life scenarios. Fortunately, recent advances in adversarial learning allow us to deal with multimodal data in terms of distributions rather than pair-wise samples.

Therefore, in this paper we propose a novel unpaired multimodal facial expression recognition method enhanced by thermal images through adversarial learning. Specifically, we first learn two deep neural networks to map the unpaired visible and thermal images to their ground truth labels. Then we introduce two modality discriminators and impose adversarial learning on the feature and label levels. This forces statistical similarity between the learned visible and

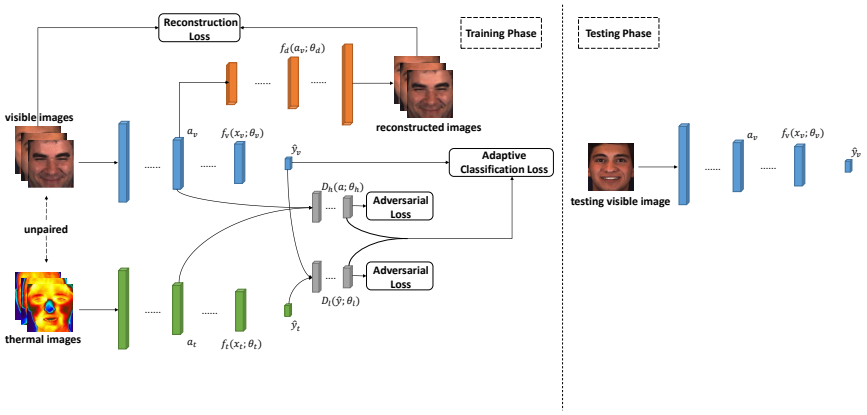


Fig. 1. The framework of the proposed unpaired facial expression recognition method.

thermal representations and minimizes the distribution mismatch between predictions of the visible and thermal images. Because there may be a few unpaired images that are significantly dissimilar across views, forcefully aligning them may have deleterious effects on the visible expression classifier. To remedy this, the variability of the different images transferability is taken into account via adaptive classification loss. Finally, a decoder network is built upon the visible hidden features in order to preserve some inherent features of the visible view. During training, the two-view neural networks and the two discriminators are optimized alternately, and the visible network is expected to be enhanced by the thermal network with unpaired visible and thermal images. During testing, the prediction of a visible testing image is given by the learned visible neural network.

Compared to related work, our contribution can be summarized as follows: (1) we are the first to tackle the task of facial expression recognition with unpaired visible and thermal images. (2) We propose to close the distributions of the unpaired visible and thermal data through adversarial learning at the feature and label levels. Experimental results demonstrate the effectiveness of our proposed method.

3 Problem Statement

Suppose we have two unpaired data sets $\mathcal{D}_v = \left\{x_v^{(i)}, y_v^{(i)}\right\}_{i=1}^{N_1}$ and $\mathcal{D}_t = \left\{x_t^{(i)}, y_t^{(i)}\right\}_{i=1}^{N_2}$. The first is the visible data set containing N_1 training instances and the second is the thermal data set containing N_2 training instances. $x_v \in \mathbb{R}^{d_v}$ and $x_t \in \mathbb{R}^{d_t}$ represent the visible and thermal images respectively, where d_v and d_t represent

their dimensions. $y \in \{0, 1, \cdots, K-1\}$ are the ground truth expression labels of images. Our task is to learn a visible expression classifier by utilizing the two unpaired data sets $\mathcal{D}_v$ and $\mathcal{D}_t$ during training phase. When testing a new visible image, the prediction is given by the learned visible expression classifier. No thermal data is involved during the testing phase.

4 Proposed Method

The framework of the proposed unpaired multimodal facial expression recognition method is summarized in Figure 1. As shown in Figure 1, there are five networks in the proposed method: the visible network $f_v : \mathbb{R}^{d_v} \rightarrow [0,1]^K$ with parameter θ_v, the thermal network $f_t : \mathbb{R}^{d_t} \rightarrow [0,1]^K$ with parameter θ_t, the discriminator in the image representation layer $D_h : \mathbb{R}^{d_h} \rightarrow [0,1]$ with parameter θ_h, the discriminator in the label layer $D_l : [0,1]^K \rightarrow [0,1]$ with parameter θ_l and the decoder network $f_d : \mathbb{R}^{d_h} \rightarrow \mathbb{R}^{d_v}$ with parameter θ_d.

The visible and thermal networks capture the mapping functions from the visible and thermal facial images to the expression labels. The two discriminators compete with the two-view networks, regularizing them in order to learn the modal-irrelevant image representations and output similar predictions. Since adversarial learning focuses on the statistical similarity of the learned representations and classification errors from two modalities, synchronous visible and thermal imaging is not required. The decoder network ensures the preservation of some inherent features of the visible view.

4.1 Basic Classification Loss for Two Views

We build duplicate neural networks for visible and thermal datasets respectively. The output layers of the visible and thermal networks are K-way softmax layers. The outputs of the visible and thermal networks are denoted as $\hat{y}_v = f_v(x_v)$ and $\hat{y}_t = f_t(x_t)$. Therefore, the supervised classification losses for two views are $L_v(y_v, \hat{y}_v)$ and $L_t(y_t, \hat{y}_t)$ respectively, we use common cross-entropy loss. The visible and thermal networks are first trained with their corresponding datasets. Then we fix the parameters of thermal network, and fine-tuning the visible network by exploiting thermal images as privileged information.

4.2 Adversarial Learning in Feature and Label Levels

After training the visible and thermal networks, the two-view networks are combined and fine-tuned simultaneously. Since the visible and thermal training images are unpaired, we cannot adopt a pair-wise similarity constraint as in Wang *et al.*'s work [7]. Motivated by He *et al.*'s work [22], we propose to build discriminators in the feature and label levels in order to make full use of the thermal data as privileged information.

Let a_v and a_t represent the activations of the visible and thermal data in a certain hidden layer. The learning objective of the feature-level discriminator

$D_h(a;\theta_h)$ is to classify the source of the activations as accurately as possible. The visible network serves as a generator which tries to fool the discriminator. We treat the visible view as a positive class and the thermal view as a negative class. The minimax objective can be formulated as follows:

$$\min_{\theta_v}\max_{\theta_h}\mathbb{E}_{a_v\sim P_{a_v}}\log D_h\left(a_v\right)+\mathbb{E}_{a_t\sim P_{a_t}}\log\left(1-D_h\left(a_t\right)\right) \tag{1}$$

Adversarial learning is introduced in the feature level to reduce the statistical gap between the visible and thermal views and learn view-irrelevant features. At the label level, we also introduce adversarial learning because the tasks of facial expression recognition from visible and thermal views are highly correlated. Therefore, the decision-making behaviour of the two-view networks should be similar. The learning objective is formulated as Eq. 2.

$$\min_{\theta_v}\max_{\theta_l}\mathbb{E}_{\hat{y}_v\sim P_{\hat{y}_v}}\log D_l\left(\hat{y}_v\right)+\mathbb{E}_{\hat{y}_t\sim P_{\hat{y}_t}}\log\left(1-D_l\left(\hat{y}_t\right)\right) \tag{2}$$

In addition to minimizing the classification errors, the learning objective of the visible network is to fool the two discriminators into making mistakes, so its loss function in the feature and label space can be formulated as Eq. 3 and 4.

$$L_h(\theta_v)=-\log D_h\left(a_t\right) \tag{3}$$

$$L_l(\theta_v)=-\log D_l\left(\hat{y}_t\right) \tag{4}$$

4.3 Visible Reconstruction Loss

Although we introduce a feature-level discriminator in order to learn view-irrelevant feature representations, some inherent features of the original two views may be missing during adversarial learning. We want to preserve some inherent features of the visible view to learn a highly performing visible facial expression classifier. To this end, we set a decoder network f_d upon the visible hidden features a_v and force the visible network f_v to learn some feature representations which can be decoded into the original visible images. The decoder network outputs a reconstructed image $\hat{x}_v=f_d(a_v)$, which has the same size as the original visible image x_v. Then we evaluate the difference between $\hat{x}_v$ and x_v using mean squared error as shown in Eq. 5.

$$^{1}L_r(\theta_v)=\left\|x_v-f_d\left(a_v\right)\right\|^2 \tag{5}$$

4.4 Adaptive Classification Loss Adjustment

Since there may be a few unpaired visible and thermal images that are significantly dissimilar with each other, and forcefully aligning these images may introduce irrelevant knowledge to the visible network. Therefore, we utilize the feature

[1] This term should be written as $L_r(\theta_v,\theta_d)$. In fact, the decoder network can be viewed as a branch of the visible network. We omit the θ_d for convenience.

discriminator's output $p_h = D_h(a_v)$ and label discriminator's output $p_l = D_l(\hat{y}_v)$ to generate attention values to alleviate these effects. In information theory, the entropy function is an uncertainty measure defined as $H(p) = -\sum_j p_j \log p_j$, which we can use to quantify the transferability. We thus utilize the entropy criterion to generate the attention value for each image as:

$$w = 1 + (H(p_h) + H(p_l))/2 \tag{6}$$

Embedding the attention value into the cross entropy loss of visible network $L_v\left(y_v, \hat{y}_v\right)$, the adaptive classification loss can be formulated as:

$$L_a\left(y_v, \hat{y}_v\right) = wL_v\left(y_v, \hat{y}_v\right) \tag{7}$$

4.5 Overall Loss Function

The loss function of the visible networks is defined as Eq. 8.

$$L(\theta_v) = L_a + \lambda_1 L_h + \lambda_2 L_l + \lambda_3 L_r \tag{8}$$

where λ_1, λ_2 and λ_3 are hyper-parameters, controlling the weights of feature-level adversarial loss, label-level adversarial loss and reconstruction loss respectively.

4.6 Optimization

The visible and thermal networks play the role of "generator" in the proposed framework. The optimization procedures of the feature-level and label-level discriminators are mutually independent. Therefore, we can apply alternate optimization steps, as in the original GAN framework [23].

5 Experiment

5.1 Experimental Conditions

We perform our experiments on multimodal databases containing visible and thermal facial images. Currently, available databases include the NVIE database [24], the MAHNOB Laughter database [25], and the MMSE database [26]. The NVIE database is unsuitable for deep learning due to a limited number of training instances. We also perform experiments on facial expression databases containing visible facial images, i.e., CK+ [27], Oulu-CASIA [28] and MMI [29] databases.

The MAHNOB Laughter database consists of audio, visible videos, and thermal videos of spontaneous laughter from 22 subjects captured while the subjects watched funny video clips. Subjects were also asked to produce posed laughter and to speak in their native languages. We cannot conduct expression recognition on this database because it only provides visible and thermal images for laughter. In our experiment, two sub-data sets were used: the laughter versus speech data set (L vs S) which contains 8252 laughter images and 12914 speech

images, and the spontaneous laughter versus posed laughter data set (L vs PL) which contains 2124 spontaneous laughter images and 1437 posed laughter images. Following the same experimental conditions as Wang *et al.*'s work [7], a leave-one-subject-out cross validation methodology is adopted. Accuracy and F1-score are used for performance evaluation.

The MMSE database consists of 3D dynamic imaging, 2D visible videos, thermal videos, and physiological records from 140 subjects induced by 10 emotion tasks. We use the same images as Yang *et al.*'s work [20], they semi-automatically select 2468 frames from 72 subjects (45 female and 27 male) on four tasks based on the FACS codes, which contains 676 happiness images, 715 surprise images, 593 pain images and 484 neutral images. A 10-fold cross validation is performed, and the split is subject independent. Accuracy and Macro-F1 are used for performance evaluation.

The CK+ database contains 593 video sequences collected from 123 subjects. Among them, 327 video sequences with 118 subjects are labeled as one of seven expressions, i.e., anger, contempt, disgust, fear, happiness, sadness and surprise. We use the last three frames of each sequence with the provided label, which results in 981 images. A 10-fold cross-validation is performed, and the split is subject independent. Accuracy is used for performance evaluation.

The Oulu-CASIA database contains data captured under three different illumination conditions. During the experiment, only the data captured under strong illumination condition with the VIS camera is used. The Oulu-CASIA VIS has 480 video sequences taken from 80 subjects, and each video sequence is labeled as one of the six basic expressions. The last three frames of each sequence are selected, a 10-fold subject-independent cross validation is performed. Accuracy is used for performance evaluation.

The MMI database consists of 236 image sequences from 31 subjects. Each sequence is labeled as one of the six basic facial expressions. We selected 208 sequences captured in frontal view. We selected three frames in the middle of each sequence as peak frames and associated them with the provided labels. This results a dataset with 624 images. A 10-fold subject-independent cross validation is performed. Accuracy is used for performance evaluation.

On all databases, we crop the facial regions from the visible and thermal images with the help of landmark points and resize the facial regions to 224×224. We use ResNet-34 [30] as the basic architecture for the visible and thermal networks. The last layer of 1000 units in the original ResNet-34 is replaced by fully connected layers with K units. The last layer of the "conv3_x"[2], with a dimension of $28 \times 28 \times 128$, is selected to add feature-level adversarial learning and build the decoder network. The learning rate of the discriminator is 10^{-4}, and the learning rates of the two-view networks start from at 2×10^{-3} and use cosine annealing strategy.

In order to evaluate the influence of each proposed loss function, we conduct a series of ablation experiments to verify our methods when images are unpaired. Firstly, a standard ResNet-34 is trained as baseline using only visible

[2] See Table 1 in the original ResNet paper [30].

Table 1. Experimental results on the MAHNOB Laughter and MMSE databases.

Scenario	**Methods**	L vs PL		L vs S		MMSE	
		Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**
single view	visible neural network [7]	93.98	92.08	83.97	87.01	88.74	89.01
	ResNet	93.93	93.44	85.77	88.96	90.62	90.21
paired images	SVM2K [7]	91.40	88.52	72.40	77.61	83.01	82.73
	DCCA+SVM [7]	86.42	81.17	65.07	72.29	72.36	71.97
	DCCAE+SVM [7]	86.44	82.21	68.29	74.91	81.30	80.94
	Wang *et al.*'s method [7]	94.14	92.30	85.54	88.38	92.18	91.96
	Pan *et al.*'s method [8]	95.77	94.51	90.23	91.77	93.15	92.88
unpaired images	L_v+L_h	95.80	94.57	88.60	90.98	92.99	92.78
	L_v+L_l	95.63	94.33	88.35	90.91	92.87	92.75
	L_v+L_h+L_l	96.43	95.38	90.08	92.16	93.82	93.78
	L_a+L_h+L_l	96.73	95.88	90.72	92.54	94.92	94.79
	Ours:L_a+L_h+L_l+L_r	**96.92**	**96.02**	**91.62**	**93.54**	**95.83**	**95.74**

Table 2. Experimental results on the CK+, Oulu-CASIA and MMI databases.

Scenario	**Methods**	CK+	Oulu-CASIA	MMI
single view	LBP-TOP [16]	88.99	68.13	59.51
	HOG 3D [14]	91.44	70.63	60.89
	ResNet	94.80	84.58	74.04
	IACNN [18]	95.37	-	71.55
	DeRL [20]	97.30	88.00	73.23
	IF-GAN [19]	95.90	-	74.52
unpaired images	L_v+L_h	96.64	86.81	77.56
	L_v+L_l	96.33	86.67	77.40
	L_v+L_h+L_l	97.15	87.85	78.04
	L_a+L_h+L_l	97.86	88.40	78.68
	Ours:L_a+L_h+L_l+L_r	**98.37**	**89.11**	**79.33**

images. Another ResNet-34 is also trained using thermal images and is fixed as guidance later. Secondly, the methods with feature-level adversarial learning L_h, label-level adversarial learning L_l and both of them are trained for comparison. Thirdly, the method with L_h, L_l and adaptive classification loss L_a is trained. Finally, our proposed method which combines L_a, L_h, L_l and L_r is trained.

Note that the visible and thermal training images on the MAHNOB and MMSE databases are all paired. Since we want to conduct unpaired experiment, the most intuitive way is doing cross-dataset experiment. However these two databases have different expression categories, we can't conduct cross-dataset experiment directly. Therefore, in order to simulate the unpaired scenario, we randomly sample visible and thermal training samples $\{x_v^{(i)}, y_v^{(i)}\}_{i=1}^m$ and $\{x_t^{(i)}, y_t^{(i)}\}_{i=1}^m$ from visible and thermal train set of the same database, and ensure they come from disjoint partitions. On the facial expression databases that only contain visible images, we use the thermal images from MMSE database as privileged information.

5.2 Experimental Results and Analysis

Experimental results are shown in Table 1 and 2. From the tables, we can find the following observations:

Firstly, adopting the introduced losses from the feature and label spaces both lead to a great improvement comparing with the the single-view baseline using ResNet. Specifically, the acc/f1 of L_v+L_h and L_v+L_l are 1.87%/1.13%, 1.70%/0.89% higher than ResNet on the L vs PL data set, 2.83%/2.02%, 2.58%/ 1.95% higher than ResNet on the L vs S data set, 2.37%/2.57%, 2.25%/2.54% higher than ResNet on the MMSE database. The experimental results on the visible facial expression databases, i.e., CK+, Oulu-CASIA and MMI database, show similar trend. Current methods cannot utilize unpaired data, so only one data view can be used. However, our method effectively explores both visible and thermal data to achieve superior results.

Secondly, our method can combine the strengths of adversarial loss from feature and label space to achieve better performance. For example, L_v+L_h+L_l outperforms L_v+L_h and L_v+L_l by 0.83%/1.00% and 0.95%/1.03% of acc/f1 on the MMSE database. Other databases have similar results, which indicate the different privileged informations will not cause the inter-view discrepancy. Guidance in both feature and label spaces can help visible classifier to learn more robust feature representations and make better predictions.

Thirdly, our approach can reduce the irrelevant knowledge impact of some dissimilar visible and thermal images. To be specific, the method of L_a+L_h+L_l is 0.71%, 0.55% and 0.64% better than L_v+L_h+L_l on the CK+, Oulu-CASIA and MMI database. The experimental results demonstrate our introduced adaptive classification loss can highlight transferable images and reduce negative transformation.

Fourthly, our proposed method achieves the best performance by using feature adversarial loss, label adversarial loss, adaptive classification loss and reconstructed loss together. Thermal images are used as privileged information to reduce the statistical gap between the visible and thermal views in both feature and label levels during training. However, adversarial learning focuses on learning view-irrelevant features and may discard the original information of the visible images. Our method adds a decoder network upon the visible feature space, forcing the visible network to preserve inherent features of the visible view. Specifically, the accuracy of our method is 3.57%, 4.53% and 5.29% higher than the baseline using ResNet on the CK+, Oulu-CASIA and MMI database. The experimental results demonstrate that both feature-level and label-level adversarial learning are effective for exploring the dependencies between visible and thermal images, and the decoder network is able to preserve the inherent features of the visible view during adversarial learning.

5.3 Comparison to Related Methods

Comparison to Multimodal FER Methods. As shown in Table 1, our method achieves better performance than three traditional multimodal learning

methods, i.e., SVM2K, DCCA+SVM, and DCCAE+SVM. On the L vs S data set, the accuracy of our method is 19.22%, 26.55%, and 23.33% higher than those of SVM2K, DCCA+SVM, and DCCAE+SVM, respectively. SVM2K can be viewed as a shallow version of the ResNet method that includes a similarity constraint. DCCA and DCCAE are learned with an unsupervised objective. Our method is based on the deep convolutional neural network and is learned in an end-to-end manner, resulting in superior accuracy.

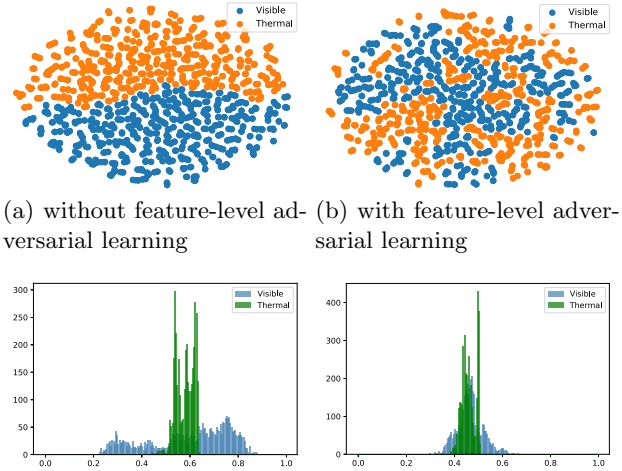
Compared to state-of-the-art multimodal FER works, i.e., Wang *et al.*'s method [7] and Pan *et al.*'s method [8], our method achieves better performance by exploiting unpaired images. For example, on the L vs PL data set, the accuracy of our method is 2.78% and 1.15% higher than those of Wang *et al.*'s method and Pan *et al.*'s method, respectively. Wang *et al.*'s method requires paired visible and thermal images during training, and the similarity constraint is imposed on the predictions of the two-view networks in order to make them similar. Our method achieves the same goal without this constraint by learning with unpaired images in an adversarial manner. Pan *et al.*'s method also requires paired images and used adversarial learning in the feature spaces of the visible and thermal views. However, some important information of the visible view may be missing. Our method incorporates a decoder network to ensure the preservation of the original information of the visible view, leading to more robust visible feature representations.

Comparison to Visible FER Methods. As shown in Table 2, our method get better results than the state-of-the-art FER methods which only use visible images. To be specific, our method is 3.00%, 1.07% and 2.47% higher than IACNN [18], DeEL [20] and IF-GAN [19] on the CK+ database, 7.78%, 6.10% and 4.81% higher than these methods on the MMI database. IACNN used expression-sensitive contrastive loss to reduce the effect of identity information, DeEL extracted the information of the expressive component through de-expression procedure, IF-GAN transferred facial expression to a fixed identity to mitigate the effect of identity-related information. Although these works concentrate on extracting the discriminative expressive feature, our method takes full advantage of the thermal images as privileged information to train more robust classifiers.

5.4 Evaluation of Adversarial Learning

To further evaluate the effectiveness of adversarial learning, we visualize the distributions of visible and thermal views. Figure 2 displays the visualization of data from the feature and label spaces with and without adversarial learning on the L vs PL data set. Specifically, we project the hidden feature representations onto a 2D space with t-SNE [31] and plot them on a two-dimensional plane. Predictions of visible and thermal views are plotted in a histogram. In Figure 2(a), feature points of visible and thermal views are separate, as feature-level adversarial learning is not used. Introducing adversarial learning on the feature

space leads to feature points that are mixed together, as shown in Figure 2(b). Similarly, when comparing Figures 2(c) and 2(d), the distributions of visible and thermal predictions become closer in the latter figure, demonstrating the effectiveness of reducing the statistical gap between visible and thermal views via adversarial learning.



(a) without feature-level adversarial learning

(b) with feature-level adversarial learning

(c) without label-level adversarial learning

(d) with label-level adversarial learning

Fig. 2. Visualization of the visible and thermal distributions on the laughter versus posed laughter data set of the MAHNOB Laughter database.

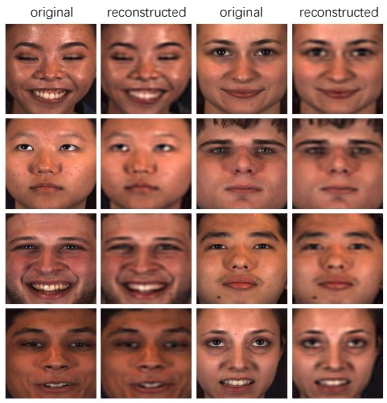


Fig. 3. Comparison between the original facial images and the reconstructed facial images on the MMSE database.

5.5 Visualization of the Decoder Network

As elaborated in Section 4.3, a decoder is built upon the visible feature space and the reconstruction loss is included in the overall loss. Thus, we can visualize the outputs of the decoder network, i.e., the reconstructed facial images, to see what the decoder learns. The visualization of the original and reconstructed facial images on the MMSE database is shown in Figure 3. We can see that the original and reconstructed facial images are nearly identical, indicating that the inherent features of the visible view are preserved during adversarial learning.

6 Conclusions

In this paper, we propose an unpaired facial expression recognition method that utilizes thermal images as privileged information to enhance the visible classifier. Two deep neural networks are first trained with visible and thermal images. Two discriminators are introduced and compete with the two-view networks in the feature and label space to reduce the statistical gap between the learned visible and thermal feature representations and close the distributions between the predictions of the visible and thermal images. Furthermore, a decoder network is built upon the visible hidden features in order to preserve some inherent features of the visible view during adversarial learning. Experimental results on benchmark expression databases demonstrate that our method can achieve state-of-the-art performance on the task of facial expression recognition.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 917418129), and the major project from Anhui Science and Technology Agency (1804a09020038). Thank for Lijun Yin for providing the MMSE database.

References

1. Yoshitomi, Y., Kim, S.I., Kawano, T., Kilazoe, T.: Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In: Robot and Human Interactive Communication, 2000. RO-MAN 2000. Proceedings. 9th IEEE International Workshop on, IEEE (2000) 178–183
2. Sharma, N., Dhall, A., Gedeon, T., Goecke, R.: Modeling stress using thermal facial patterns: A spatio-temporal approach. In: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE (2013) 387–392
3. Wesley, A., Buddharaju, P., Pienta, R., Pavlidis, I.: A comparative analysis of thermal and visual modalities for automated facial expression recognition. In: International Symposium on Visual Computing, Springer (2012) 51–60
4. Wang, S., He, S., Wu, Y., He, M., Ji, Q.: Fusion of visible and thermal images for facial expression recognition. Frontiers of Computer Science **8** (2014) 232–242

5. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22** (2009) 544–557
6. Shi, X., Wang, S., Zhu, Y.: Expression recognition from visible images with the help of thermal images. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, ACM (2015) 563–566
7. Wang, S., Pan, B., Chen, H., Ji, Q.: Thermal augmented expression recognition. IEEE transactions on cybernetics (2018)
8. Pan, B., Wang, S.: Facial expression recognition enhanced by thermal images through adversarial learning. In: 2018 ACM Multimedia Conference on Multimedia Conference, ACM (2018) 1346–1353
9. Wang, Z., Ji, Q.: Classifier learning with hidden information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 4969–4977
10. Sharmanska, V., Quadrianto, N.: Learning from the mistakes of others: Matching errors in cross-dataset learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3967–3975
11. Niu, L., Li, W., Xu, D.: Exploiting privileged informtion from web dt for action nd event recognition. International Journal of Computer Vision **118** (2016) 130–150
12. Luo, Z., Hsieh, J.T., Jiang, L., Niebles, J.C., Fei-Fei, L.: Graph distillation for action detection with privileged modalities. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 166–183
13. Garcia, N.C., Morerio, P., Murino, V.: Learning with privileged information via adversarial discriminative modality distillation. IEEE transactions on pattern analysis and machine intelligence (2019)
14. Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association (2008) 275–1
15. Chu, W., La Torre, F.D., Cohn, J.F.: Selective transfer machine for personalized facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2017) 529–545
16. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 915–928
17. Jiang, B., Valstar, M.F., Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: Face and Gesture 2011, IEEE (2011) 314–321
18. Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y.: Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), IEEE (2017) 558–565
19. Cai, J., Meng, Z., Khan, A., Li, Z., Oreilly, J., Tong, Y.: Identity-free facial expression recognition using conditional generative adversarial network. arXiv: Computer Vision and Pattern Recognition (2019)
20. Yang, H., Ciftci, U., Yin, L.: Facial expression recognition by de-expression residue learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 2168–2177
21. Sankaran, N., Mohan, D.D., Setlur, S., Govindaraju, V., Fedorishin, D.: Representation learning through cross-modality supervision. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE (2019) 1–8

22. He, L., Xu, X., Lu, H., Yang, Y., Shen, F., Shen, H.T.: Unsupervised cross-modal retrieval through adversarial learning. In: Multimedia and Expo (ICME), 2017 IEEE International Conference on, IEEE (2017) 1153–1158
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
24. Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., Wang, X.: A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Transactions on Multimedia **12** (2010) 682–691
25. Petridis, S., Martinez, B., Pantic, M.: The mahnob laughter database. Image and Vision Computing **31** (2013) 186–202
26. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., et al.: Multimodal spontaneous emotion corpus for human behavior analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 3438–3446
27. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE (2010) 94–101
28. Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikainen, M.: Facial expression recognition from near-infrared videos. Image and Vision Computing **29** (2011) 607–619
29. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. (2005) 317–321
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
31. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9** (2008) 2579–2605