Weakly Supervised Dual Learning for Facial Action Unit Recognition

***Abstract -*** Current research on facial action unit (AU) recognition typically requires fully AU-annotated facial images. Compared to facial expression labeling, AU annotation is a time-consuming, expensive, and error-prone process. Inspired by dual learning, we propose a novel weakly supervised dual learning mechanism to train facial action unit classifiers from expression-annotated images. Specifically, we consider AU recognition from facial images as the main task, and face synthesis given AUs as the auxiliary task. For AU recognition, we enforce satisfaction between the expression-dependent and expression-independent AU dependencies, i.e., the domain knowledge about expressions and AUs. For face synthesis given AUs, we minimize the difference between the synthetic face and the ground truth face, which has identical recognized and given AUs. By optimizing the dual tasks simultaneously, we successfully leverage their intrinsic connections as well as domain knowledge about expressions and AUs to facilitate the learning of AU classifiers from expression-annotated image. Furthermore, we extend the proposed weakly supervised dual learning mechanism to semi-supervised dual learning scenarios with partially AU-annotated images. Experimental results on three benchmark databases demonstrate the effectiveness of the proposed approach for both tasks.

1. INTRODUCTION

Facial behavior is one of the most important channels for emotional communication between humans. Both facial expression categories and facial action units (AUs) are adopted by researchers to describe facial behavior. Recent research focuses on automatic facial expression recognition [1]–[5], facial action unit recognition [6]–[10], simultaneous facial expression and action unit recognition [11], or facial expression and action unit intensity estimation [12], [13]. Facial expression categories have not been clearly defined by researchers, but the number and definitions of AUs are clearly described in the facial action coding system (FACS) developed by Ekman and Friesen [14]. Furthermore, any facial expression can be decomposed into combinations of AUs. For example, a smile can be identified as the upward movement of the lip corners, which corresponds to AU12 [15]. Therefore, we focus on AU recognition in this paper.

Current AU recognition methods typically include supervised learning, and thus require fully AU-annotated facial images. Since AUs describe subtle local facial changes, AU labels typically must be provided by experts. Manual annotation of AUs is time consuming and prone to error. Facial expressions, on the other hand, describe global facial behaviors. They can be annotated quickly and accurately, even by non-experts. There are strong dependencies between facial expressions and AUs due to underlying facial structures and muscle movement patterns. The emotional facial action coding system (EMFACS) [17] lists emotion-related AU combinations. For example, people typically show happiness by raising their cheeks and stretching their mouths [16]. Prkachin et al. [18] found that pain intensity can be inferred from the combination of several AUs (i.e., AU4, AU6, AU7, AU9, AU10, and AU43). There are also dependencies among AUs. For example, when AU1 appears, AU2 is also highly likely to be present, and vice versa. AU12 and AU15 are rarely seen concurrently [19]. Weakly supervised AU classifier learning can leverage this domain knowledge using facial images with expression annotations, but not AU labels.

Two tasks, i.e., AU recognition from facial images and face synthesis given AUs, emerge in dual form. They are intrinsically connected, forming a closed loop [20]. Despite this connection, current research studies the two tasks separately. To the best of our knowledge, few works leverage their duality to jointly train the models of the two tasks. Inspired by He et al.’s work [20] on dual learning from natural language translation, we propose a novel weakly supervised dual learning mechanism to simultaneously train facial action unit classifiers from expression-annotated images and a face synthesis model given AUs. Specifically, when an iteration commences, one feature vector is inputted to the AU classifier and converted to the middle AU output. The AU evaluation model gives the first objective term according to the degree of consistency between the predicted AUs and the domain knowledge. The face generator generates the face from the predicted AU labels. This step produces the second objective term: the log likelihood of the original face given the synthetic face. We train the two tasks simultaneously by maximizing the total objective. The training process is shown as Figure 1. Before the dual learning process, the AU evaluation model is trained from the domain knowledge, considering three kinds of AU conditional probability (introduced in Section III-A1). We sample the pseudo-AU labels from the summarized domain knowledge. After that, a restricted Boltzmann machine (RBM) model is used to capture the label distribution from pseudo-AU data, and the likelihood of AU labels is used as the first objective term.

The rest of the paper is organized as follows. In Section 2, we review related works on dual learning and AU recognition. In Section 3, we propose our dual AU learning method and the model for learning AU evaluation from domain knowledge. In Section 4, sufficient experiments are conducted on three databases annotated with AUs and expressions simultaneously. The last section concludes our paper.

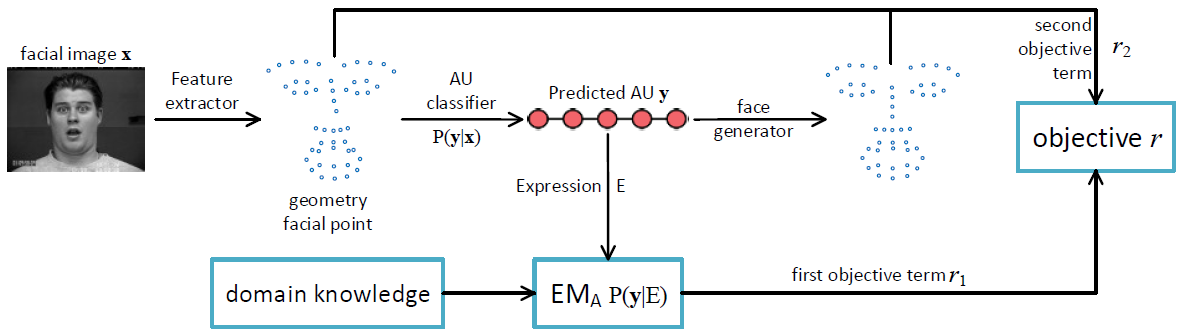


Fig. 1. The framework of weakly supervised dual AU learning. Facial feature points extracted from one facial image are inputted to the AU classifier, yielding the middle AU output. The AU evaluation model is trained from domain knowledge given the first objective term. The face generator uses the AU output to generate the face, and this step produces the second objective term.

1. RELATED WORK
2. Dual learning

Dual learning was first proposed in He et al.'s work [20] on neural machine translation. Specifically, they proposed a dual learning mechanism to reduce the requirement of labeled data in two translation tasks, i.e., English-to-French translation (primal) and French-to-English translation (dual). The learning process involves two agents. Each agent only understands one language. First, the primary translator converts the sentences from the first agent and sends them to the second agent. The second agent uses a pre-trained language model to evaluate confidence that the received message is a natural sentence. The dual translator converts the received message and sends it back to the first agent, and the first agent evaluates the consistency between the reconstructed and original sentences. In this way, two translators can simultaneously optimize learning from unlabeled data.

Considering the similarity between image translation and text translation, Yi et al. [21] extended this dual learning method to image-to-image translation using generative adversarial networks (GANs) [22]. They used two generators as the image translators and two discriminators as the image evaluation models. Unlike dual learning for machine translation [20] in which evaluation modes are pre-trained from language domain knowledge, Yi et al.’s unsupervised dual learning method for image-to-image translation trains the evaluation models and image translators through an adversarial framework.

Dual learning can also be used in supervised learning tasks. Xia et al*.* [23] proposed a dual supervised learning method that exploits the probabilistic correlation between two tasks to jointly regularize the training of dual tasks. Considering that many tasks have structural duality at the model level as well as the data level, Xia et al. [24] proposed model-level dual learning (MDL), which ties similar model parameters in the two tasks.

Although AU recognition from facial images and face synthesis given AUs can be regarded as dual tasks, to the best of our knowledge there has been no work considering dual learning to optimize these two tasks jointly and utilize their connections effectively. Therefore, in this paper we propose a novel weakly supervised dual learning method for AU recognition from facial images with expression annotations only, and face synthesis given AUs. The domain knowledge about AUs and expressions is leveraged to pre-train the AU evaluation model. Synthetic faces are evaluated by their difference from the ground truth faces. By optimizing these dual tasks jointly, we can successfully explore the domain knowledge inherent in facial structure and the connections between dual tasks to train AU classifiers using data labeled only with expressions. This significantly relieves the burden of AU annotations.

Current dual learning work is either unsupervised, requiring unlabeled data, or fully supervised, requiring fully labeled data. We propose a novel weakly supervised learning approach, and further extend it to a semi-supervised learning approach. This approach only needs weakly labeled data from the main task, i.e., facial images with expression annotations.

1. Action Unit Recognition

A comprehensive survey on facial action unit recognition can be found in [25]. In this section, we focus on AU recognition works that do not require fully AU-annotated images.

A few recent works have focused on AU recognition from partially AU-annotated samples. Song et al. [8] proposed a novel Bayesian graphical model (BGCS) that encodes sparsity and co-occurrence structure of facial action units via compressed sensing and group-wise sparsity-inducing priors. Their proposed method can handle partially observed labels by marginalizing over the unobserved values as a part of the inference procedure. Wu et al. [26] proposed a multi-label learning method (MLML) that explicitly handles missing labels by enforcing consistency between the predicted and provided labels, as well as local smoothness among the label assignments. Instead of using the same features for all AU classes as in Wu et al.’s work, Li et al. [27] extended the MLML method to discriminate each AU based on the most related features. All of these works require at least partial AU labels to learn AU classifiers.

To the best of our knowledge, there have been only six works on expression-assisted AU recognition. Wanget al. [28] proposed an expression-assisted AU recognition method under incomplete AU labeling. They constructed a Bayesian network (BN) to capture the dependencies among AUs and the dependencies between AUs and expressions. After training, the AUs of testing images are inferred by combining the measurements and the AU relations in the BN model. Their proposed method successfully uses expression labels as hidden knowledge to complement the missing AU labels. However, both expression labels and AU labels are required to learn AU classifiers, although AU labels may be partially missing. Wang et al. [7] proposed an AU recognition method using expression as privileged information required during training only. They used a 3-way RBM to capture the global dependencies between expressions and AUs. Their method requires complete AU and expression labels during training.

The other four works learn AU classifiers without AU labels. Ruiz et al. [6] proposed hidden-task learning (HTL) to learn the AU classifier from the image and the expression classifier from the AU when AU annotations are unavailable. They exploited the domain knowledge of expression-AU relations using extra large-scale facial images labeled with expressions. They also extended HTL to semi-hidden task learning (SHTL) when partially AU-annotated samples are provided. Ruizet al.'s method successfully leverages the domain knowledge of expression-AU relations to train the AU classifier from the training data when AU labels are limited or unavailable. However, their method only considers the conditional probability of a single AU under a single expression, and requires a large-scale database of expression-labeled images. Furthermore, any error caused by the expression classifier will propagate to the AU classifier.

Wang et al. [29] proposed an RBM prior (RBM-P) model that learns the prior joint AU distribution from the pseudo AU labels generated from the summarized domain knowledge. It then learns the AU classifier by maximizing its log likelihood. They considered not only the conditional probability of one AU under one expression, but also its inverse, as well as the conditional probability of one AU under another AU.

Zhang et al. [30] proposed a multiple AU classifier learning method (LP-SM) that incorporated AU domain knowledge into the objective as constraints. They comprehensively summarized the domain knowledge, and represented it as the inequality relations among the AU probabilities.

Peng et al [10] proposed an AU recognition adversarial network (RAN) based on GAN. They used the same domain knowledge and pseudo AU label sampling method as [29], and utilized an adversarial framework to achieve similarity between the distribution of the predicted labels and the distribution of the pseudo AU labels. The main difference between RAN and our method is that RAN focuses only on the AU recognition task. Our method leverages the dual task of the AU recognition task, i.e., the face synthesis task, to improve the performance of both AU recognition and face synthesis.

Although these four weakly supervised AU recognition methods leverage the domain knowledge about expressions and AUs to facilitate the learning process of AU classifiers from expression-annotated facial images, they ignore the duality between the tasks of AU recognition and face synthesis. In this paper, we consider the primary AU recognition task and dual face synthesis task, and train them simultaneously by pre-training an AU evaluation model learned from the domain knowledge about AUs and expressions.

1. PROPOSED METHOD

In this section, we propose a weakly supervised dual learning method that considers the AU classification task and the face synthesis task simultaneously. One facial feature vector is converted to an AU vector using an AU classifier, and we give the first objective term according to an AU evaluation model learned from the summarized domain knowledge. The face is generated from the middle AU output; this step has an objective term indicating the difference between the original and generated faces. In Section III-A, we introduce our weakly supervised dual AU learning method and extend it to semi-supervised learning. In Section III-B, we introduce the learning of AU evaluation model EMA.

We obey the procedure in [29] to learn AU evaluation model EMA. Specifically, we first thoroughly summarize the domain knowledge about AUs and expressions. Then, we sample pseudo AU labels for each expression from the summarized domain knowledge. Lastly, we train an RBM prior model to capture the AU distribution, and use the likelihood of AU labels as the first objective term. Training of the AU evaluation model is shown in Figure 2.

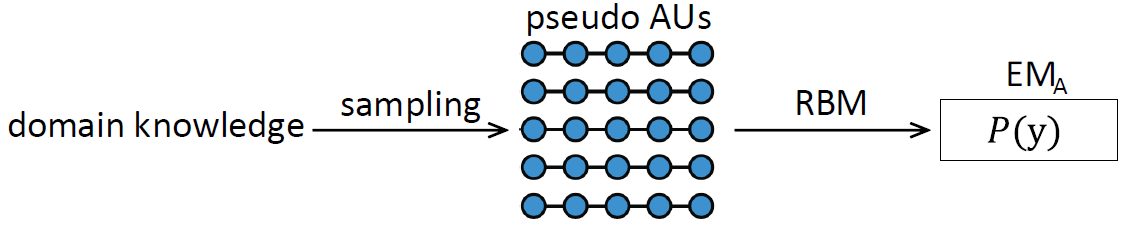
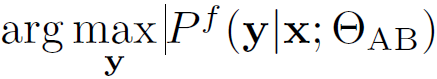
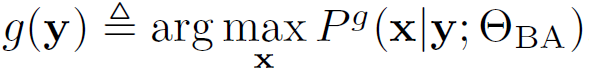
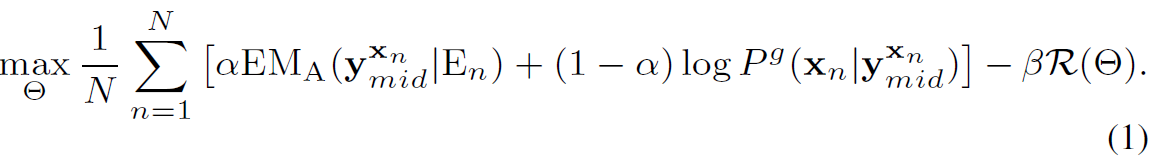


Fig. 2. As in Wang et al.’s work [29], we learn an RBM model with the pseudo AU data sampled from domain knowledge as the AU evaluation model.

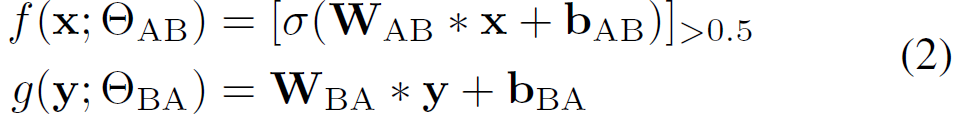
1. Dual AU Classifier Learning

*1) Weakly Supervised Learning:* Let  denote the training set, where  represents the *d*-dimensional facial feature vector and  is the expression label. *P* is the number of expression classes and *N* is the number of training samples. We consider two tasks. The main task is to learn an AU classifier  from expression-labeled facial image samples. The auxiliary task is to learn a face generator , where *L* represents the dimension of AU vector. We propose a weakly supervised dual learning machine to train the two tasks simultaneously. In our work, we learn two conditional distributions  and  for the main and auxiliary tasks respectively, where y is the AU vector. *f* and *g* are obtained by maximum posteriori inference  and .

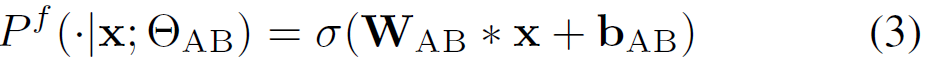
For one feature vector x randomly sampled from the training set, we convert it to an AU vector with the AU classifier. Let  denote the output of the AU classifier, which is actually a probability vector. The evaluation model EMA trained in Section III-B will evaluate the consistency between the AU output and the domain knowledge. We set the first objective term , where Ex is the expression label of x. Next, the face is generated from the output of the AU classifier and the second objective term for face synthesis is set as . The goal for feature vector x is to maximize the total objective  is a hyper-parameter. The objective for the whole training set is shown as follows:

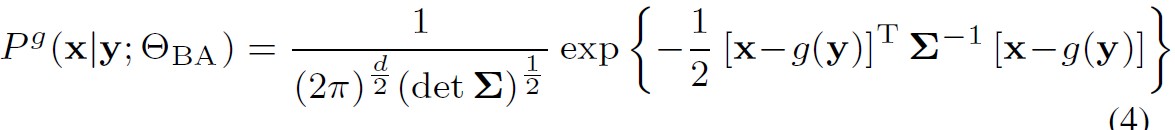


Where  is the regularized term and *β* is the weight of . Specifically, we use standard L2-regularization . For simplicity, we generate feature points to represent the synthetic faces, so the following linear function is used for AU classifier *f* and face generator *g*:

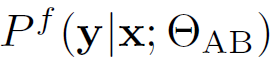


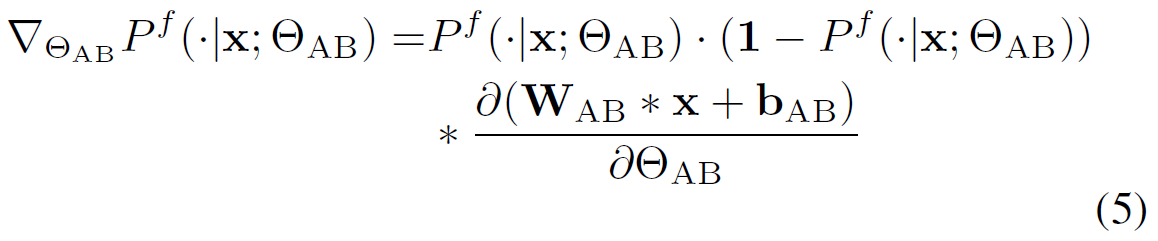
Where *σ* is the sigmoid function, [t]>0:5 = sign (t–0.5), and . The conditional distribution for the main task and the conditional distribution for the auxiliary task are shown as Equations 3 and 4, respectively.

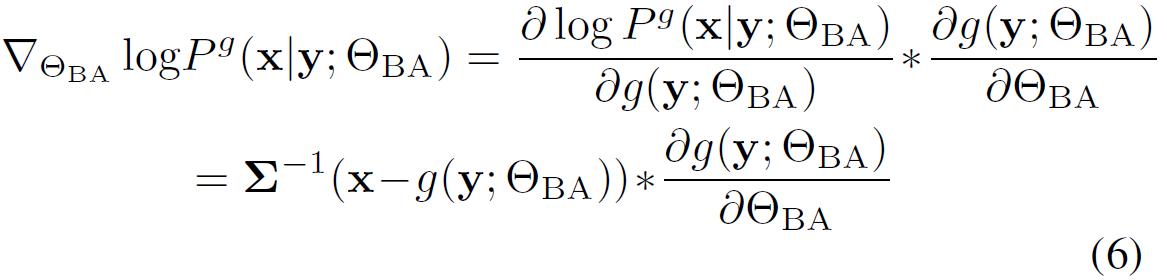


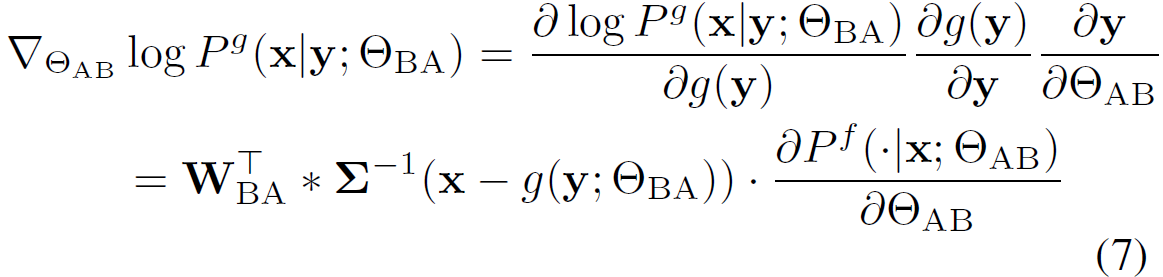


We assume a multivariate Gaussian distribution for the learned condition distribution of the face synthesis task. To ensure the conditional probability of x is only affected by *g*(y), we let the Σ be a constant matrix, like an identity matrix.

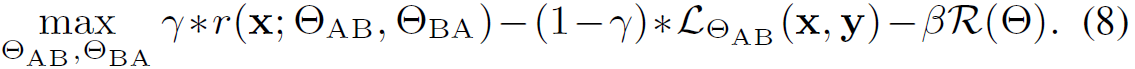
The derivation of  over  is shown as Equation 5; the derivations of  over  and  are shown as Equations 6 and 7, respectively. The parameters  and  can be updated using the stochastic gradient ascent method.



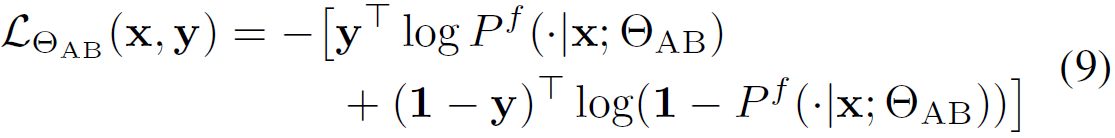


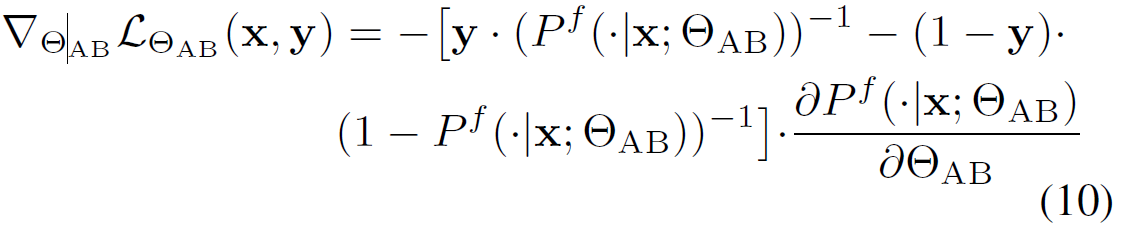


*2) Extension to Semi-Supervised Learning:* The proposed method can be extended to learn in a semi-supervised manner when partial AU annotations are available. Let  denote a subset of *D* that are annotated with AU labels, where  and *L* is the number of AUs. For samples with no AU annotations, the objective does not change. However, for samples annotated with AU labels, the objective is updated with an additional term minimizing the error between the predicted and ground truth AUs, as follows:



Where  is the cross-entropy loss as shown in Equation 9, and  is a tradeoff between two terms. When, the maximization problem is equivalent to learning in an unsupervised manner. Conversely, when , the maximization problem is equivalent to traditional supervised learning; the dependencies among AU labels and the assistance of the auxiliary task are not considered.

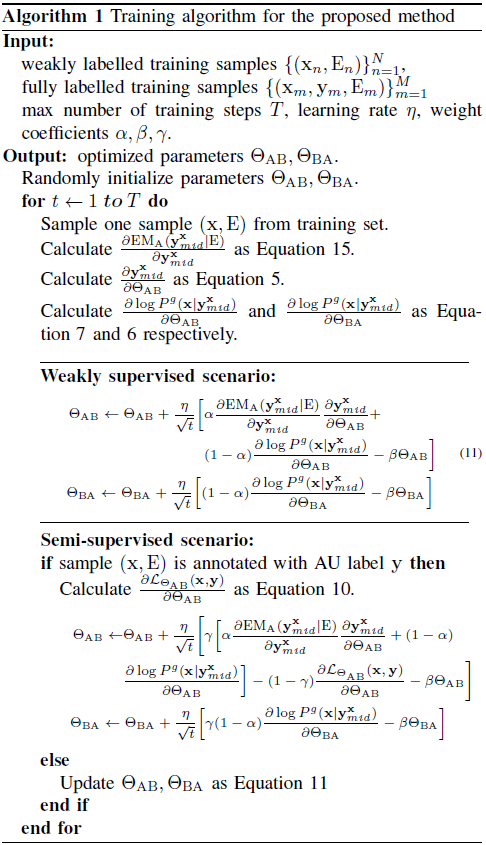




The derivation of  over  is shown as Equation 10. Parameters are updated using stochastic gradient ascent. The detailed training procedure is shown as Algorithm 1.

1. Learning the AU Evaluation Model

*1) Summary of Domain Knowledge:* Wang et al*.* [29] introduced detailed domain knowledge about AUs and expressions. Following their work, we give a brief summary of domain knowledge in this section. Domain knowledge about AUs and expressions appears as conditional probabilities of one AU. There are three kinds of conditional probability: the conditional probability of one AU under one expression, the conditional probability of one AU under another AU, and the conditional probability of one AU under one expression and another AU.



The first seven rows of Table I show the conditional probability of one AU under one expression. The conditional probabilities of the AUs under six basic expressions (the first six rows) are from [16]. For example, P(AU1=1|sadness)=0.6 and P(AU4=1|sadness)≥0.7. The blanks indicate that the conditional probabilities of AUs are less than 0.2, like P (AU1|happiness)<0.2. The seventh row of Table I shows the conditional probabilities of six AUs under the pain expression according to Prkachin and Solomon pain intensity (PSPI) [31], like P(AU4=1|pain)≥0.5. There is no information available for the other AUs.

The last thirteen rows of Table I show the conditional probability of one AU under another AU, summarized from [14] and [19]. The relations between two AUs include both co-existent and mutually exclusive relations. If two AUs are co-existent, the conditional probabilities should be greater than 0.5; for example, P (AU1=1 |AU2=1)>0.5 and vice versa. If two AUs are mutually exclusive, the conditional probabilities should be less than 0.2; for example, P (AU12=1|AU15=1) <0.2 and vice versa.

Domain knowledge about the conditional probability of one AU under one expression and another AU is adopted from the emotion facial action coding system (EMFACS) [17], [32] as shown in Table II. Table II lists some AU combinations that frequently appear during the same expression. For example, AU4 and AU5 usually appear simultaneously during anger, so P (AU4=1|AU5=1, anger)>0.5 and P (AU5=1|AU4=1, anger)>0.5. The relations in Table II are all co-existent relations among AUs.

TABLE I

THE CONDITIONAL PROBABILITY OF ONE AU UNDER ONE EXPRESSION OR ANOTHER AU [29]. HORIZONTAL AXIS REPRESENTS AU, AND VERTICAL AXIS INCLUDES EXPRESSIONS AND AUS.

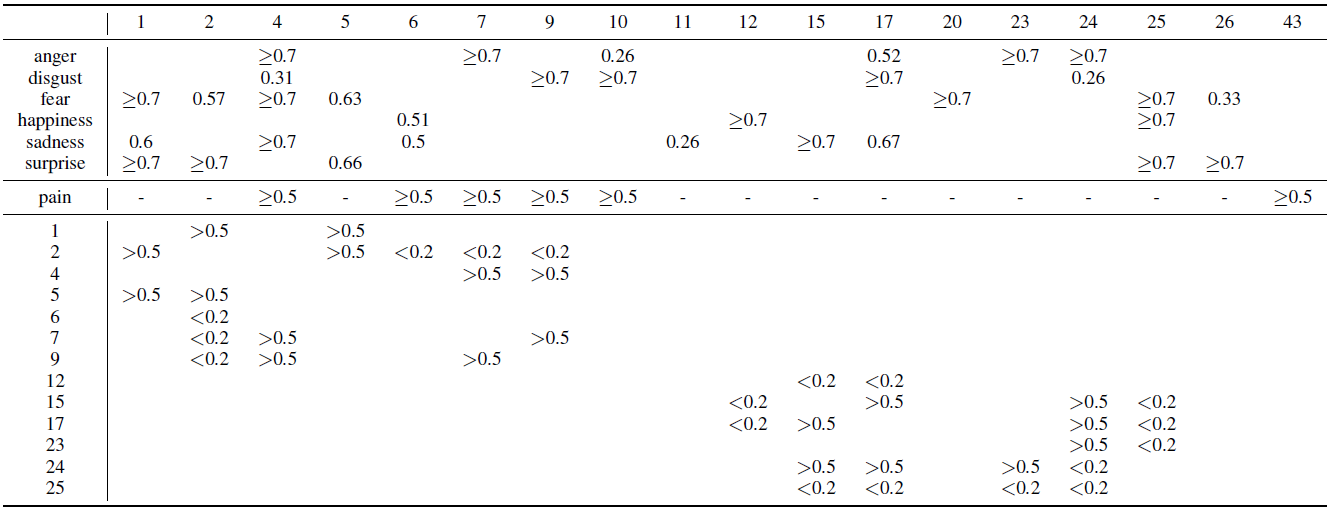
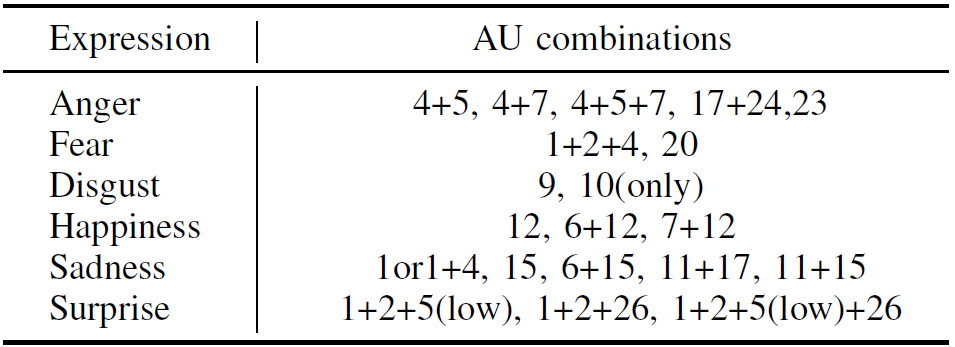


TABLE II

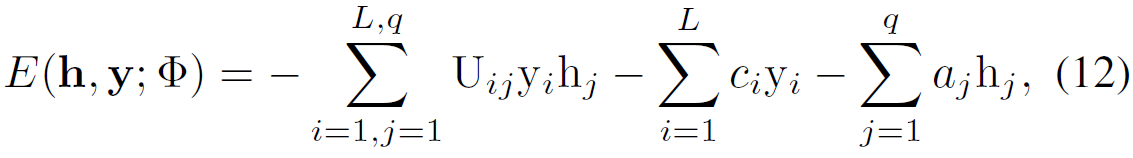
EXPRESSION-RELATED AU COMBINATIONS FROM EMFACS [17], [32].



*2) Pseudo AU Label Sampling:* In order to learn the AU evaluation model, pseudo AU labels are sampled according to the three kinds of AU conditional probability. Pseudo AU labels are generated for each expression. Probability parameters are generated before sampling, except for some concrete conditional probabilities of AUs under expressions in Table I, like P (AU1=1|sadness) =0.6. Specifically, for conditional probabilities of one AU under one expression that are greater than 0.7, like P (AU4=1|anger), probability parameters are drawn in the uniform distribution U(0.7, 1). For probabilities less than 0.2, like P(AU1=1|anger), parameters are drawn in the uniform distribution U(0, 0.2). Similarly, for the conditional probability of one AU under the pain expression, probability parameters are drawn from U(0.5, 1). For the conditional probability of one AU under another AU without loss of generality, which might as well be set as P (AUj = 1|AUi = 1). If AUi and AUj are co-existent, P(AUj = 1|AUi = 1) is drawn from U(0.5, 1). If they are mutually exclusive, P(AUj = 1|AUi = 1) is drawn from U(0, 0.2). After generating the probability parameters, the same sampling algorithm used in [29] is adopted to generate pseudo AU labels for each expression.

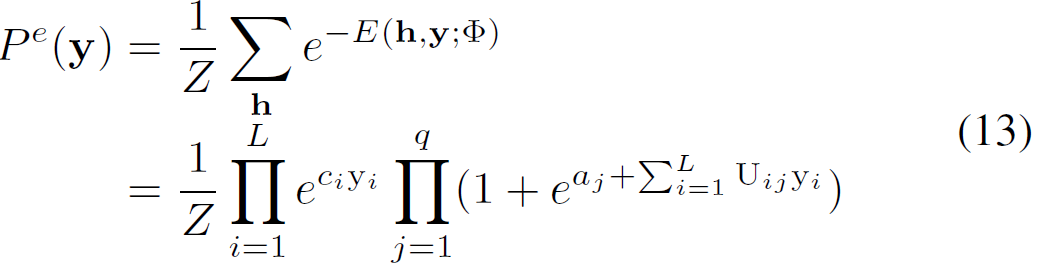
*3) Learning RBM model:* The joint AU probability is modeled by a restricted Boltzmann machine [33]. This RBM consists of two types of binary nodes: visible units y and hidden units h. The visible nodes represent pseudo AU labels. In our work, an evaluation model is pre-trained under each expression prior to learning the AU classifier and face generator.

The energy function of the RBM is defined as:



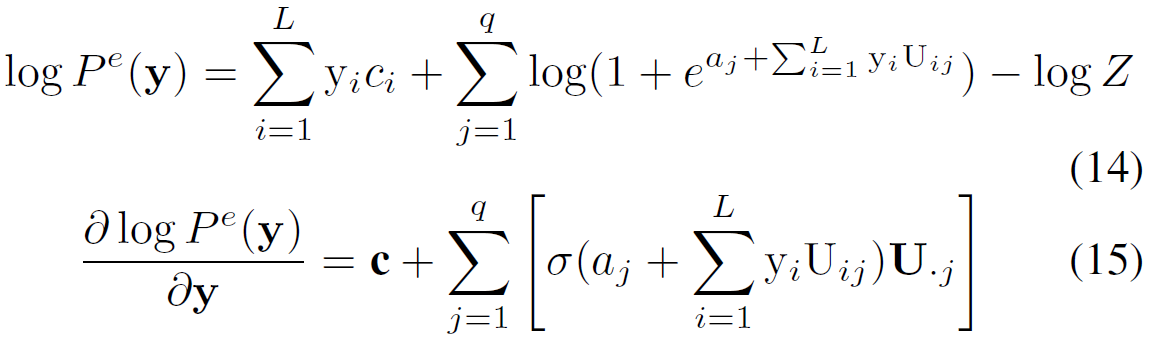
Where yi and hj are the components of y and h, respectively.  are parameters of RBM, c are biases of AU units, a are biases of hidden units, and U are weights between the AU units and the hidden units. L and q are the number of AUs and hidden units, respectively.

After learning the parameters of the RBM (the specific learning method refers to [34]), the marginal distribution of the visible units *P*e(y) can be calculated by marginalizing the hidden units as:



Where is a normalization constant. We set the evaluation model EMA(y) = log *P*e(y).

Equation 13 yields log *P*e(y) as Equation 14, and the derivation of log *P*e(y) over y is shown as Equation 15.



1. EXPERIMENTS
2. Experimental Conditions

The proposed method is evaluated on three publicly available databases: the Extended Cohn-Kanade database (CK+) [35], the MMI database [36], and the UNBC-McMaster Shoulder Pain Expression Archive database [37].

*1) Databases:* The CK+ database contains 593 video recordings from 123 subjects performing posed facial expressions. Each video begins with the expression onset frame and ends with the apex frame. The MMI database contains 2900 video recordings from 75 subjects displaying posed facial expressions. The UNBC database is a spontaneous expression database containing 200 video recordings from 25 patients suffering from shoulder pain. Each frame in the UNBC database is coded with PSPI. In this paper, the frames with a PSPI of five or higher are regard as “pain” frames, and those with a PSPI of zero are regard as “no pain” frames.

*2) Frames:* On the CK+ database, 309 video sequences from 106 subjects are annotated with six basic expressions. We use the apex frame of each video sequence (309 frames in total). On the MMI database, 171 video sequences from 27 subjects are annotated with six basic expressions and AUs, so 171 apex frames are used in our work. Since only 30 video sequences from 17 subjects have pain frames on the UNBC database, all pain and no pain frames are collected from these 30 video sequences (7319 frames in total).

*3) AUs:* On the CK+ database and the MMI database, we consider AUs in which the occurrence frequency of all samples is greater than 10%. Therefore, on the CK+ database, 12 AUs (1, 2, 4, 5, 6, 7, 9, 12, 17, 23, 24, and 25) are considered, and 13 AUs (1, 2, 4, 5, 6, 7, 9, 10, 12, 17, 23, 25, and 26) are considered on the MMI database. On the UNBC database, since we only have domain knowledge about six AUs (4, 6, 7, 9, 10, and 43), these six AUs are all considered.

*4) Features:* For a fair comparison with competing methods, we also use facial landmarks as features in our work. Some related works used texture-based features. However, facial landmarks are better able to visualize the face. There are 49 and 66 facial feature points provided by the database constructer on the CK+ and the UNBC databases, respectively. On the MMI database, 49 feature points were extracted with IntraFace [38]. Feature points were normalized using an affine transformation, so that the eye centers fall on the given position for all images. After that, we use a Gaussian normalization for each dimension of features. Common feature points are used for cross-database experiments.

*5) Settings:* We conduct weakly supervised and semi-supervised AU recognition experiments on three databases. For both weakly supervised and semi-supervised scenarios, we conduct within-database experiments via five-fold subject-independent cross-validation, and cross-database experiments. For semi-supervised experiments, we randomly miss AU labels according to rates varying from 0.1 to 0.9. Each experiment is conducted five times to reduce the influence of randomness, and the average F1 score (↑, the higher the better) is used as the evaluation metric. We also evaluate the face generator by root mean square error (RMSE) (↓, the lower the better). The hyper-parameters in Algorithm 1 are  , which are tuned on a validation set, and a grid search strategy is used. Specifically, for the maximum number of training steps on the CK+ and MMI database,, and on the UNBC database, . For learning rate . For weight coefficients,  .

*6) Comparisons:* For weakly supervised scenarios, we compare the proposed method with HTL, RBM-P, RAN, and LP-SM. The results of RBM-P [29] and RAN [10] are copied from the original paper, since the experimental setting of these two works are identical to our own. However, Ruiz et al. did not conduct experiments on the MMI database and adopted a different experimental strategy, so we use the results from [29], which re-conducted the experiments of HTL on the same databases and under the same experimental settings. Zhang et al. conducted experiments using two kinds of features, i.e., LBP and feature points, and did not conduct cross-database experiments. Therefore, the proposed method is compared to LP-SM using feature points, but not in cross-database experiments. The comparisons to LS-PM are for reference only, since the experimental settings of LS-PM are different from ours. Specifically, LS-PM considered 8 AUs on the CK+ and MMI databases and used 600 frames, and considered 3 AUs on the UNBC database. Our method is also compared to SVM, which trains with fully AU-labeled data, in cross-database experiments.

For semi-supervised scenarios, we compare our work with six state-of-the-art methods, i.e., RAN, RBM-P, SHTL, BGCS, MLML, and BN, in within-database experiments. For cross-database experiments, we compare the proposed method to SHTL [29], RBM-P [29], and RAN [10]. Since Zhanget al. conducted semi-supervised experiments with a missing rate of 0.5 only on the CK+ database, we do not compare to LP-SM.

As in the weakly supervised scenario, the results of RBM-P and RAN are copied from the original paper. In order to keep the same experimental conditions, the results of SHTL, BGCS, MLML, and BN are copied from [29], which re-conducted their experiments.

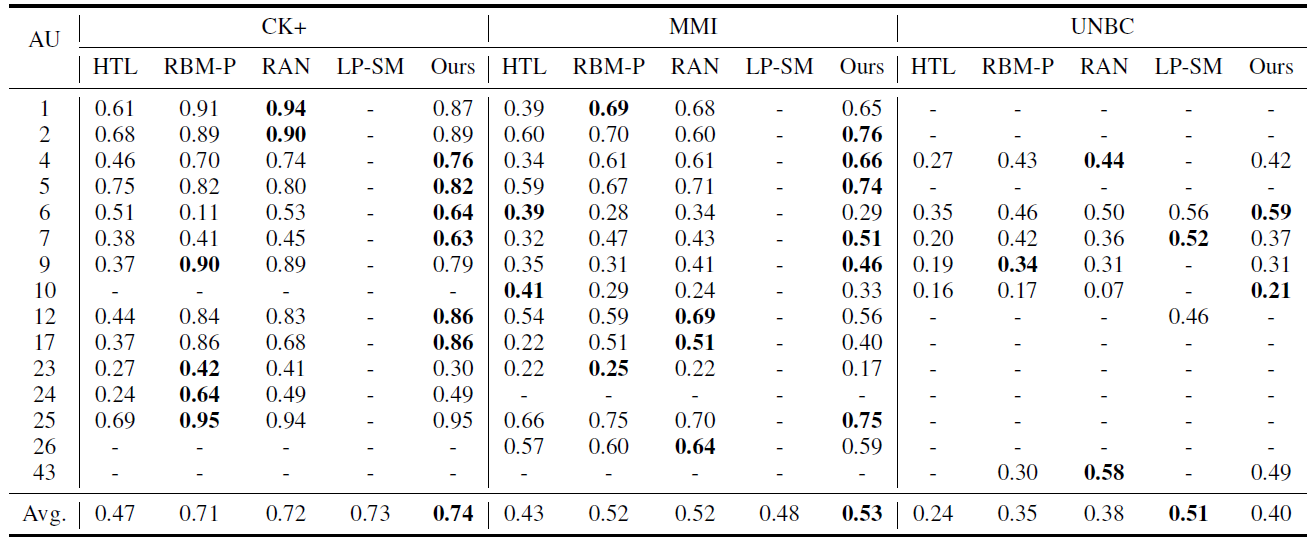
For face synthesis, we first validate the assumption of Gaussian distribution. Then we compare the proposed method with discriminative RBM (DRBM) [39], whose visible layer contains a feature vector and an AU label vector. We infer the facial features from the input AU labels through a Gibbs sampling method.

1. Weakly-Supervised AU Recognition

The results of weakly supervised within-database experiments are listed in Table III. From the results on the three databases, we find that the proposed method outperforms other methods with the highest average F1 score in most cases, demonstrating its superiority for weakly supervised AU recognition.

TABLE III

WITHIN-DATABASE EXPERIMENTAL RESULTS OF WEAKLY SUPERVISED AU RECOGNITION (F1 SCORES).



Compared to HTL, the proposed method achieves 57.66% improvement on the CK+ database, 23.62% improvement on the MMI database, and 69.01% improvement on the UNBC database. The proposed method considers more complete domain knowledge than HTL and introduces a face synthesis task to assist the learning of the AU classifier. Although HTL considers the AU-expression classification task (visible task) in addition to the feature-AU recognition task (hidden task), they are not dual tasks and are trained independently. The error caused by AU-expression classifiers may propagate to the feature-AU classifiers. Unlike HTL, we train the main AU recognition task and auxiliary face synthesis task simultaneously, so each task assists the other.

Compared to RBM-P, the proposed method achieves 5.06% improvement on the CK+ database, 3.24% improvement on the MMI database, and 13.25% improvement on the UNBC database. Both RBM-P and the proposed method use domain knowledge to capture the joint AU distribution through the RBM model. The proposed method also introduces the dual task of the AU recognition task (the face synthesis task) and utilizes the intrinsic connections between the two tasks to learn the AU classifier. The superior results of the proposed method demonstrate the effectiveness of leveraging the dual tasks to improve AU classifier learning.

RBM-P is a modified version of our method that sets the weight of the reconstruction term 1-*α* as zero. To analyze the impact of our reconstruction term, we run the same set of experiments on three databases, varying 1-*α* from 0.1 to 0.9. The results are shown in Figure 3, in which the first column shows the results of RBM-P. Optimal performance is reached when the weights for the reconstruction term are 0.3, 0.2, and 0.2 on the three databases, respectively. This indicates that the first objective term exploring domain knowledge has a significant impact, and the reconstruction term can improve AU recognition.

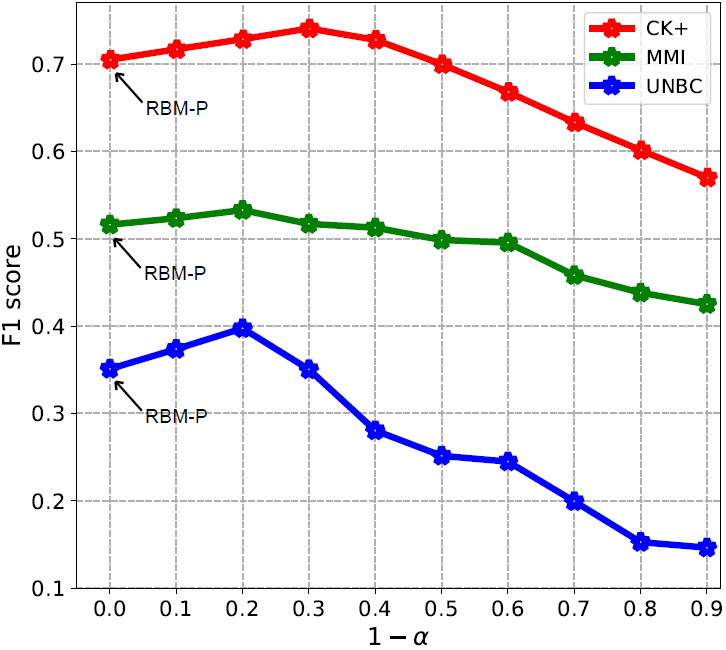


Fig. 3. AU recognition performance as the weight for the reconstruction objective term increases.

Compared to RAN, the proposed method achieves 3.51% improvement on the CK+ database, 2.34% improvement on the MMI database, and 5.63% improvement on the UNBC database. Compared to LP-SM, the proposed method also achieves better performances on the CK+ and MMI databases, demonstrating the superiority of the proposed method. Both RAN and LP-SM successfully leverage domain knowledge and expression labels. However, the proposed method is also assisted by the auxiliary face synthesis task, and thus achieves better performance. On the UNBC database, the average F1 of the proposed method is worse than that achieved by LP-SM. However, LP-SM only selected 300 apex frames with pain and 300 frames without pain on the UNBC database. This is significantly less data than the proposed method considered. Furthermore, the recognized AUs of LP-SM are AU6, AU7, and AU12, which are different from ours. Therefore, the comparison to LP-SM is only for reference.

TABLE IV

CROSS-DATABASE EXPERIMENTAL RESULTS (F1) OF WEAKLY

SUPERVISED AU RECOGNITION ON THE CK+ AND MMI DATABASES.

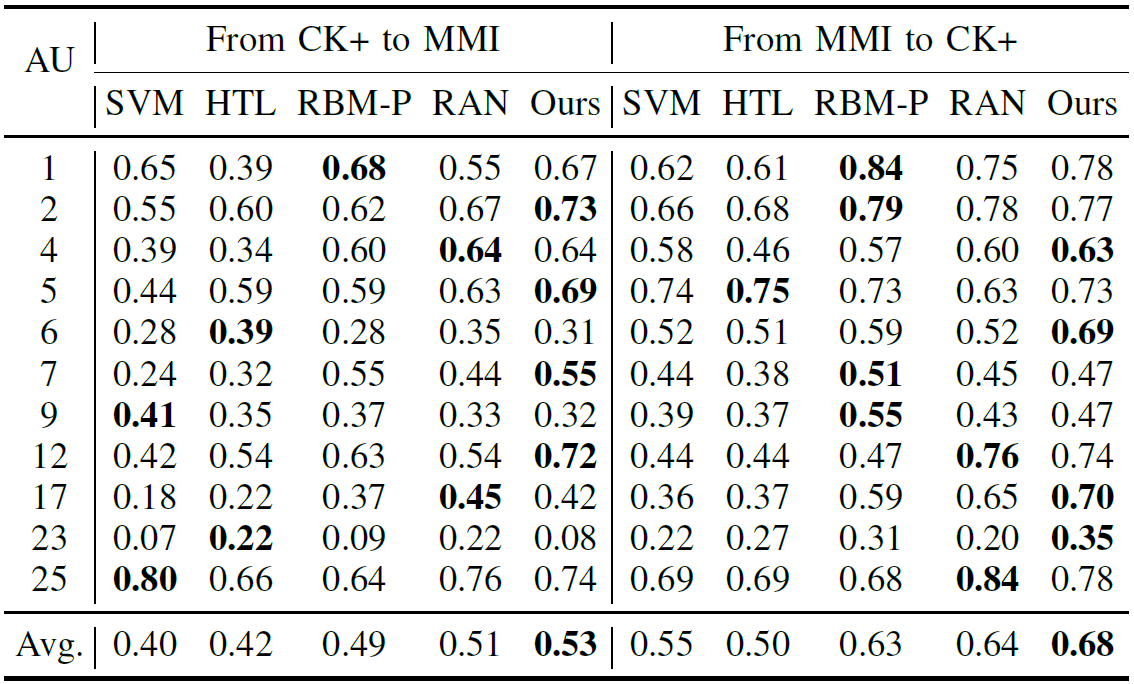


TABLE V

CROSS-DATABASE EXPERIMENTAL RESULTS (F1) OF WEAKLY

SUPERVISED AU RECOGNITION ON THE CK+ AND UNBC DATABASES.

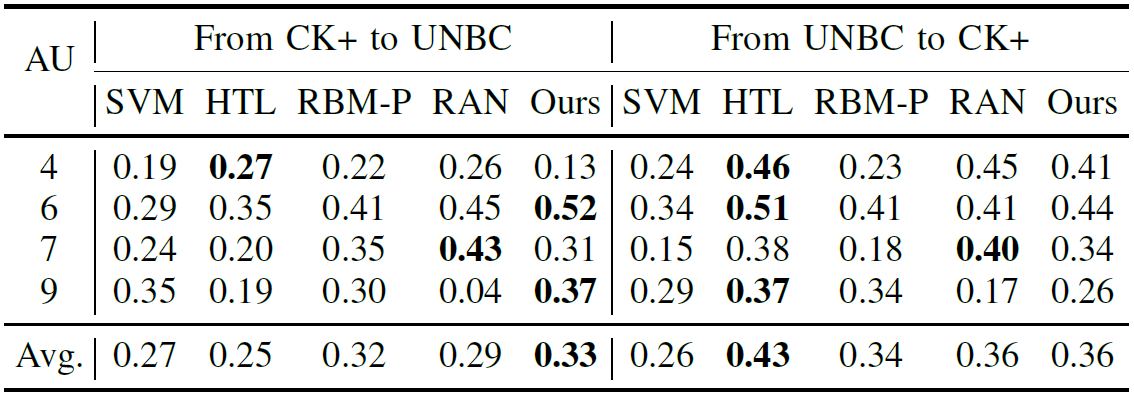
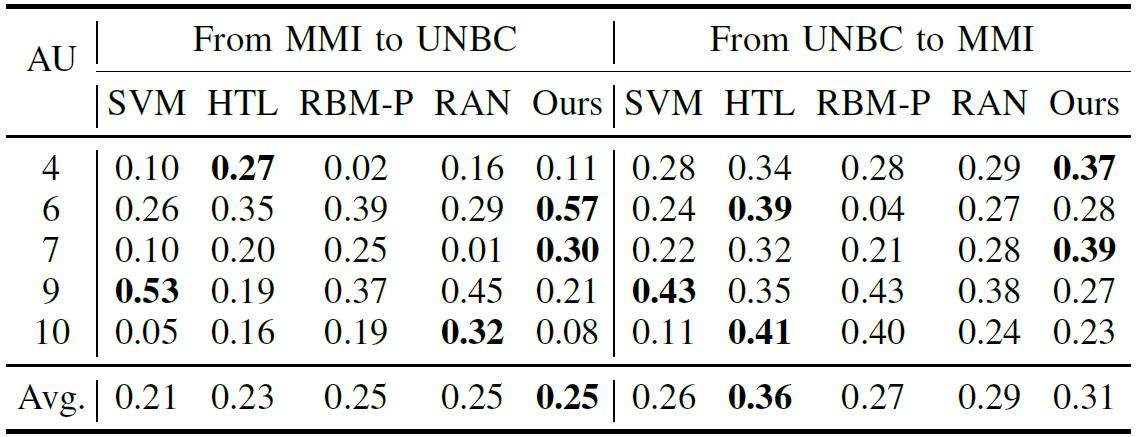


TABLE VI

CROSS-DATABASE EXPERIMENTAL RESULTS (F1) OF WEAKLY

SUPERVISED AU RECOGNITION ON THE MMI AND UNBC DATABASES.



The results of weakly supervised cross-database experiments are listed in Tables IV, V, and VI. The proposed method outperforms RBM-P and RAN, with the highest average F1 scores in all cross-database experiments. This demonstrates the superior generalization ability of the proposed method, which successfully utilizes the auxiliary face synthesis task to improve AU recognition.

The experimental results in Table IV are generally higher than the results in Tables V and VI. The CK+ and MMI databases incorporate six basic emotions, while the UNBC database only contains pain or no pain expressions. The biases between the CK+ and MMI databases are much less than biases between the UNBC database and the CK+ or MMI databases. The cross-database experiments that train or test on the UNBC database are difficult scenarios for AU recognition due to these biases. The proposed method performs best in the experiments that train on the CK+ and MMI databases and test on the UNBC database, which further demonstrates that the proposed method can be more easily generalized to other databases, even when they are very different from the training set. The proposed method performs worse than HTL in the experiments that train on the UNBC database. This is because HTL trains on another large-scale facial expression database of six basic emotions. The proposed method still performs better than RBM-P and RAN, which also only train on the UNBC database.

We also compare the proposed method with SVM, which trains with fully AU-labeled data. The proposed method achieves better performances in all scenarios. The completely data-driven learning manner of SVM limits its generalization ability, while the proposed method learns the AU classifier with domain knowledge but not ground truth AU labels.

1. Semi-Supervised AU Recognition

The results of semi-supervised within-database experiments on three databases are shown as Figure 4. From Figure 4, we can obtain the following observations:

First, the proposed method performs best in all scenarios except for the experiment on the CK+ database with a missing rate of 0.6, in which RBM-P achieves the best performance. This demonstrates that the proposed method successfully leverages the auxiliary face synthesis task to improve the performance of AU recognition.

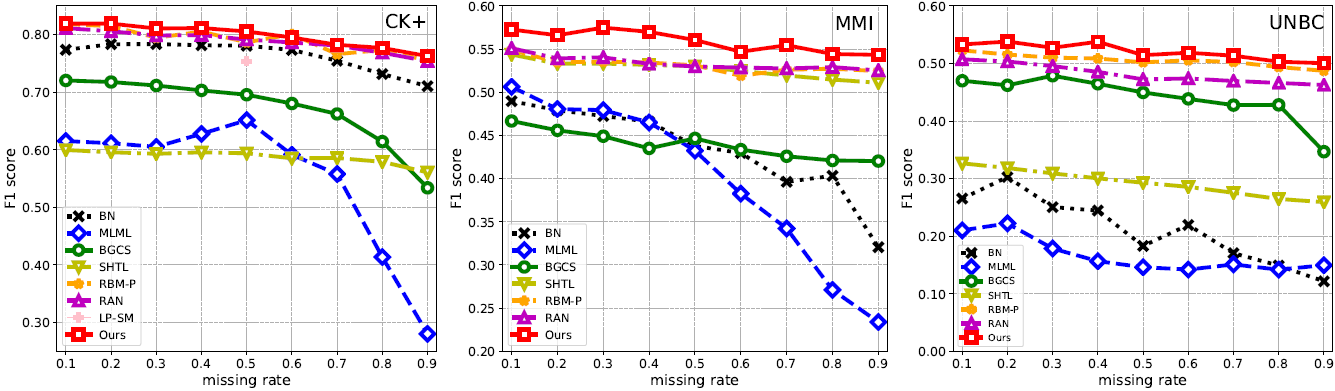


Fig. 4. Within-database experiments of semi-supervised AU recognition. Left: results on the CK+ database; Middle: results on the MMI database; Right: results on the UNBC database.

Secondly, as the missing rate increases, the trend of the experimental results varies for different methods. The performances of HTL, RBM-P, RAN, and the proposed method fall very slowly on the three databases. However, the performances of BN, MLML, and BGCS decline faster as the missing rate increases. This is particularly noticeable in the performances of BN on the MMI and UNBC databases, the performances of MLML on the CK+ and MMI databases, and the performances of BGCS on the CK+ and UNBC databases. This may be because all of these methods learn AU label correlation from the ground truth labels. Specifically, BN learns the relations between AUs and expressions from ground truth labels, MLML captures label smoothness and label consistency from ground truth labels, and BGCS exploits both sparsity and co-occurrence structure of ground truth labels. When there are few ground truth labels, these methods cannot optimally learn label correlations, while the methods learning label correlation from domain knowledge do not suffer from this problem.

Thirdly, RBM-P performs better than BN, MLML, BGCS, and SHTL in all scenarios. MLML and BGCS do not consider the assistance of expression labels. Although BN uses expression labels to improve AU classifier learning, it can only capture pairwise label correlations. SHTL also considers the assistance of expression labels, but the domain knowledge it uses is incomplete. However, RBM-P thoroughly summarizes domain knowledge and uses an RBM model to capture global AU relations, thus achieving better performance that these methods.

Lastly, the performances of RBM-P and RAN are similar, since both methods leverage comprehensive domain knowledge and enforce similar distribution of the predicted labels and the pseudo AU labels. The proposed method outperforms RBM-P in most cases and RAN in all cases. Although RBM-P, RAN, and the proposed method capture global AU relations from complete domain knowledge, the proposed method also considers the intrinsic connections of the dual task of the AU recognition task. It learns two tasks simultaneously to further enhance the performance of AU recognition.

The results on semi-supervised cross-database experiments are shown in Figure 5. In most cases, the proposed method performs best in the experiments that test on the UNBC database, demonstrating its generalization ability. However, the proposed method performs poorly on experiments that train on the UNBC database. There are two possible reasons for this. The first is that our method only trains on the UNBC database and does not use the other facial expression databases, with six basic emotions. The second reason is that only six AUs are considered on the UNBC database, which may be too few AUs to generate a face.

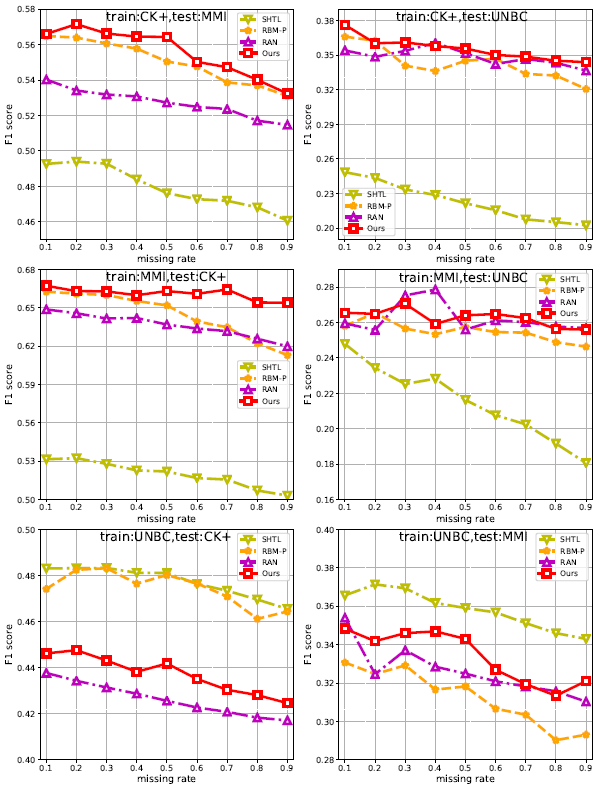


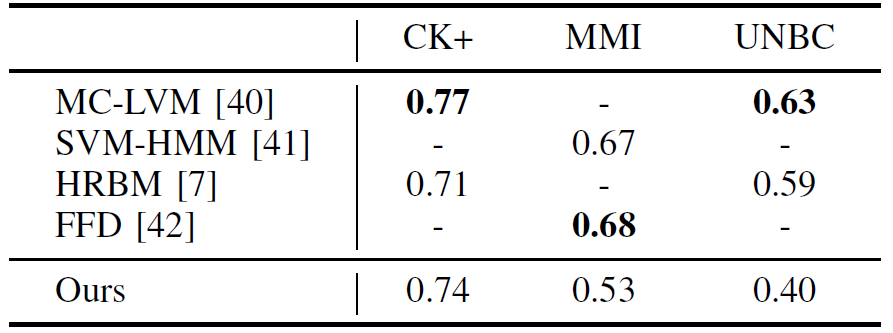
Fig. 5. Cross-database experimental results (F1) of semi-supervised AU recognition.

1. Comparison to Fully Supervised Methods

We compare the proposed weakly supervised method to four supervised methods that require fully AU-labeled data. The comparisons are shown in Table VII. The results of HRBM are from [40]. Since the experimental conditions of these methods are different from ours, these comparisons are only for reference. It's expected that the proposed method will perform more poorly than fully supervised methods, since it trains multiple label classifiers without any AU annotations, while these four methods use supervisory information. Surprisingly, the proposed method performs comparably. It even achieves superior performance over HRBM on the CK+ database. These results demonstrate the effectiveness of the proposed AU recognition method. Note that due to database biases, fully supervised information from ground truth labels may limit the generalization ability of algorithms to some extent; for example, SVM performs poorly in cross-database experiments. Our method is better able to generalize, as it uses domain knowledge that is independent from the databases.

TABLE VII

COMPARISON OF F1 SCORES TO STATE-OF-THE-ART FULLY SUPERVISED METHODS



1. Face Synthesis

In Section III-A1, we assume that the distribution of features given expressions satisfies Gaussian distribution. This section validates this assumption. We use the quantile-quantile plot (Q-Q plot) [43] with a 0.05 significance level, and randomly select one dimensional feature on the three databases. Additionally, we use Michael goodness of fit test [44] to give the boundary of acceptance interval. The results are shown in Figure 6. All points are very close to the line y = x, and most of them fall in the acceptance interval, so we can accept the assumption of normality.

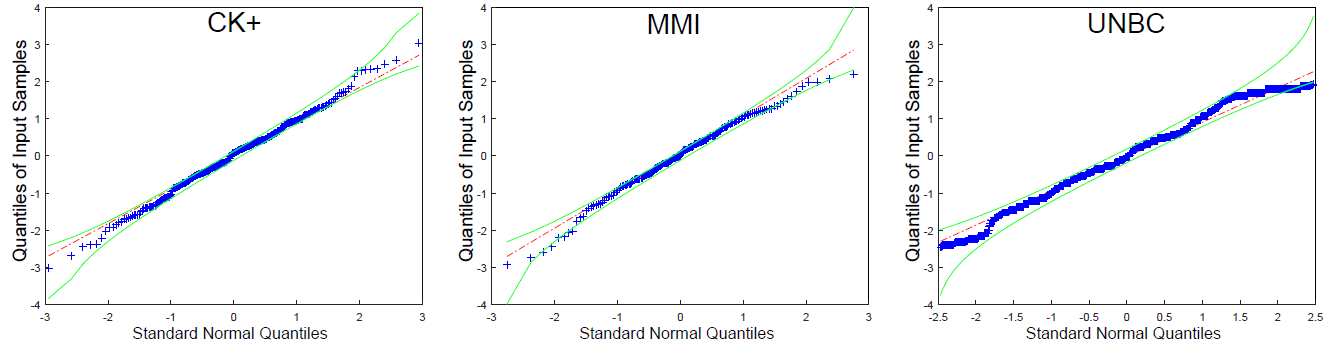
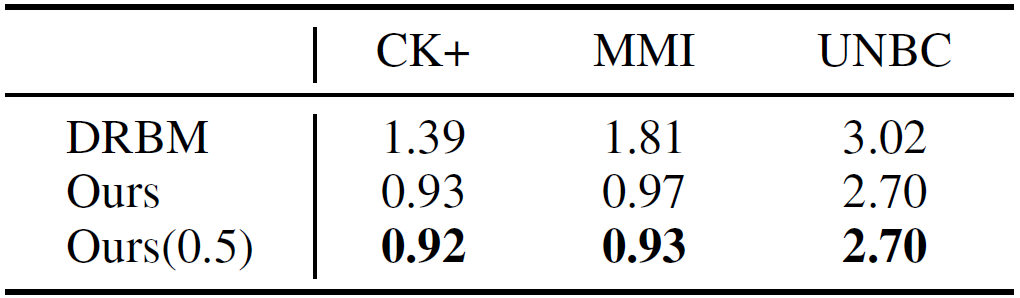


Fig. 6. The normality test of geometry features on the three databases (Q-Q plot). Blue points represent database samples. The broken red line is y = x. Two green lines are the boundaries of acceptance intervals. The closer the plot is to the line y = x, the more it accepts the assumption of normality.

TABLE VIII

RMSE OF THE DRBM AND THE PROPOSED WEAKLY SUPERVISED AND

SEMI-SUPERVISED METHODS.



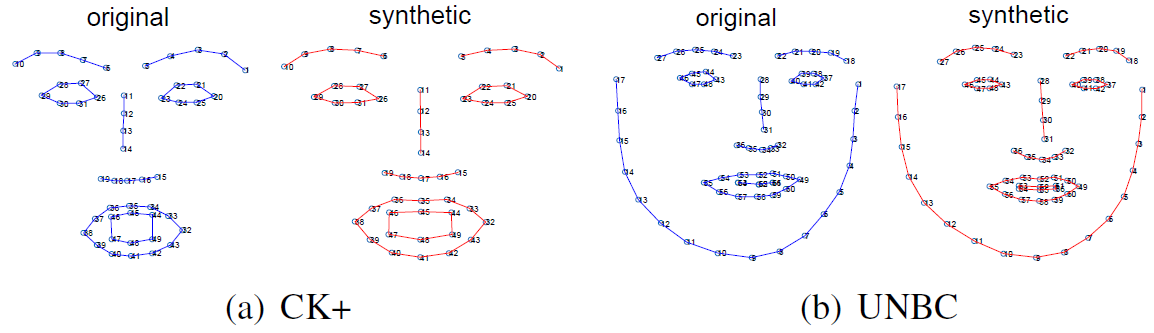


Fig. 7. Examples of original face and corresponding synthetic face.

The performance of the facial image generator is evaluated with RMSE. Table VIII lists the results of the compared method (DRBM), the proposed weakly supervised method, and the proposed semi-supervised method (with a missing rate of 0.5) on the three databases. Figure 7 shows examples of the original face and its corresponding synthetic facial points on the CK+ and UNBC databases. There is very little difference between the two.

From Table VIII, we can obtain the following three observations. First, the method performs best on the CK+ database and more poorly on the UNBC database, which is consistent with the performance of AU recognition. Only six AUs are considered on the UNBC database; this is not enough information to generate 66 facial points. Secondly, the proposed weakly supervised method performs better than DRBM, although DRBM uses AU-labeled data. This demonstrates the superiority of the proposed method for face synthesis as well as AU recognition. Thirdly, the proposed semi-supervised method outperforms the proposed weakly supervised method, which is expected. Additional AU-labeled data can improve the performance of AU recognition, and better AU recognition performance can improve the performance of face synthesis.

1. CONCLUSION

In this paper, we introduce a face synthesis task to assist AU classifier learning. The AU recognition task and the face synthesis task are mutual dual tasks. The AU classifier and face generator are learned simultaneously based on two objective terms. The first is the log likelihood of the predicted AUs from the evaluation model, which is trained from the summarized domain knowledge. The second is the log likelihood of inputted features according to the synthetic face. Compared to state-of-the-art methods, our method achieves superior results in weakly supervised and semi-supervised experiments on three databases. This demonstrates the effectiveness of the proposed method in both AU recognition and face synthesis.