Dual Semi-Supervised Learning for Facial Action Unit Recognition

Guozhu Peng, Shangfei Wang*
Key Lab of Computing and Communication Software of Anhui Province
School of Computer Science and Technology, University of Science and Technology of China
Hefei, Anhui, P.R.China, 230027
{gzpeng@mail., sfwang@}ustc.edu.cn

Abstract

Current works on facial action unit (AU) recognition typically require fully AU-labeled training samples. To reduce the reliance on time-consuming manual AU annotations, we propose a novel semi-supervised AU recognition method leveraging two kinds of readily available auxiliary information. The method leverages the dependencies between AUs and expressions as well as the dependencies among AUs, which are caused by facial anatomy and therefore embedded in all facial images, independent on their AU annotation status. The other auxiliary information is facial image synthesis given AUs, the dual task of AU recognition from facial images, and therefore has intrinsic probabilistic connections with AU recognition, regardless of AU annotations. Specifically, we propose a dual semi-supervised generative adversarial network for AU recognition from partially AU-labeled and fully expression-labeled facial images. The proposed network consists of an AU classifier C, an image generator G, and a discriminator D. In addition to minimize the supervised losses of the AU classifier and the face generator for labeled training data, we explore the probabilistic duality between the tasks using adversary learning to force the convergence of the face-AU-expression tuples generated from the AU classifier and the face generator, and the ground-truth distribution in labeled data for all training data. This joint distribution also includes the inherent AU dependencies. Furthermore, we reconstruct the facial image using the output of the AU classifier as the input of the face generator, and create AU labels by feeding the output of the face generator to the AU classifier. We minimize reconstruction losses for all training data, thus exploiting the informative feedback provided by the dual tasks. Within-database and cross-database experiments on three benchmark databases demonstrate the superiority of our method in both AU recognition and face synthesis compared to state-of-the-art works.

Introduction

Both facial expression and facial action units (AUs) are used to describe facial behavior. Ekman's six basic expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise) are commonly used to describe facial expressions. There are also many compound expressions, although their number and definition is not universally agreed upon. As the number and the definitions of AUs are absolutely clear in Ekman's facial action coding system (FACS) (Friesen and Ekman 1978), we focus on AU recognition in this paper.

Compared to expressions, which describe global facial behavior, AUs represent subtle local facial changes and are thus should be annotated by experts. To reduce reliance on AU labels, we propose a semi-supervised AU recognition method, which trains AU classifiers from partially AU-annotated images.

Each AU is controlled by one or more facial muscles. AUs are closely related due to underlying facial anatomy. For example, inner brow raiser (AU1) and outer brow raiser (AU2) almost always appear together, since they are both related to the muscle group *frontalis*. Such AU dependencies exist regardless of whether or not a facial image has been annotated, and thus can be leveraged for semi-supervised AU recognition. AUs are also strongly related to expressions. For example, Du *et al.* (Du, Tao, and Martinez 2014) found that people usually lower their jaws (AU26) when they show surprise, while the lip corner puller (AU12) rarely appears in sad faces. Prkachin *et al.*(Prkachin 1992) found that the pain expression is primarily conveyed by six AUs (AU4, AU6, AU7, AU9, AU10, and AU43). The emotion facial action coding system (EMFACS) (Friesen and Ekman 1983) lists expression-dependent AU combinations. Both expression-dependent and expression-independent AU relations are crucial to learn AU classifier. Expression labels are easier and less time-consuming to annotate than AU labels. Therefore, we prefer to learn AU classifiers from partially AU-annotated facial images with full expression labels.

AU recognition from facial images and face synthesis from the AU labels are dual tasks with intrinsic probabilistic connections. Such connections are independent of annotation status. Dual tasks help each other when they are trained together (Xia et al. 2017). Therefore, we consider the AU recognition task and the face synthesis task simultaneously, leveraging their links to improve the results of both tasks.

In this paper, we propose a dual semi-supervised generative adversarial network (DSGAN) to jointly learn an AU classifier and a facial image generator. The joint distribution of the input and the output from the AU classifier should be the same as that from the generator. This is referred to as probabilistic duality (Xia et al. 2017). Through adversarial learning, we force the joint distributions of the inputs and

*This is the corresponding author.

outputs from both the classifier and the generator to converge to the ground-truth distribution, which is embedded in the AU-expression-labeled training data. Such distribution includes AU dependencies. Furthermore, we reconstruct facial images by feeding the output of the AU classifier to the face generator, and minimize the reconstruction loss for all training data. Similarly, the face generator followed by the AU classifier forms a loop to reconstruct AUs. We also minimize the reconstruction loss for all training data. Thus, we explore the informative feedback provided by two dual tasks. Supervised losses are further minimized for AU-labeled training data.

Related Work

Dual Learning

Many learning tasks take dual forms (Xia et al. 2017). For example, an English-to-French translation task is the dual task of French-to-English translation task. AU recognition from facial images and face synthesis from AU labels are dual tasks. The primal task and the dual task form a closed loop, generating informative feedback signals that benefit both tasks.

Dual learning was first proposed by He *et al.* (He et al. 2016). They proposed dual learning to handle a neural machine translation problem from unpaired two monolingual corpora. Specifically, they simultaneously trained English-to-French and French-to-English translators from monolingual English and French corpora. Sentences of one language are translated by one translator and reconstructed by the other. Their proposed dual learning approach evaluates the similarity between sentences translated from the source language to natural sentences in the target language, and the extent to which reconstructed sentences are consistent with the original sentences. Dual learning allows two translators from monolingual English corpora and monolingual French corpora to be optimized simultaneously.

Unlike the aforementioned unsupervised dual learning, Xia *et al.* (Xia et al. 2017) proposed dual supervised leaning (DSL) from paired data. They minimized the empirical risk of dual tasks under a necessary condition, i.e., *probabilistic duality*, which means that the joint distribution of the input and the output of one task should be equal to that of its dual task. They minimized the distance of the two distributions. In order to represent the joint distribution of the input and the output, the marginal distribution of the input must be estimated. This may lead to errors in the learning process.

Unlike the above two works, which explore dualities at the data level, Xia *et al.* (Xia et al. 2018) also proposed model-level dual learning to explore dualities by sharing partial parameters of dual models.

Until now, there has not been any research on dual learning in semi-supervised scenarios. In this paper, we propose a dual semi-supervised learning approach for AU recognition from partially AU-labeled facial images. Specifically, we formulate AU recognition and face synthesis as dual tasks and train the two tasks simultaneously, utilizing their connections. We leverage adversarial loss to force convergence between the distributions of the input and the output from the AU classifier and the facial image generator, and the ground-truth-labeled training data. Furthermore, we introduce two reconstruction losses for all training data to utilize the constraints of one task for its dual task, and two supervised losses for labeled training data.

Unlike DSL, which requires fully labeled training data, the proposed dual semi-supervised learning scenario only needs partially labeled training data. Instead of minimizing the distance of two joint distributions directly, which requires the estimation of the marginal distribution of the input, the proposed approach uses an adversarial strategy to exploit the probabilistic duality, thus avoiding the estimation of marginal distribution.

AU Recognition

A comprehensive survey of work on AU recognition can be found in (Martinez et al. 2017). In this section, we focus on works studying AU recognition training from partially AU-labeled images. Current semi-supervised AU recognition work can be categorized into two approaches according to the availability of expression labels: semi-supervised AU recognition with the help of expressions, and semi-supervised AU recognition without expressions.

For semi-supervised AU recognition scenarios without expressions, label smoothness or AU dependencies are exploited to handle missing AUs. For example, Song *et al.* (Song et al. 2015) developed a Bayesian group-sparse compressed sensing (BGCS) model to encode sparsity and co-occurrence structure of AUs for AU recognition. This method can be naturally extended to semi-supervised scenarios by marginalizing over the unobserved values as part of the inference procedure. Wu *et al.* (Wu et al. 2015) proposed a multi-label learning method with missing labels (MLML). They handled the missing labels by enforcing consistency between the predicted and provided labels, as well as the local smoothness among the label assignments. Wu *et al.* (Wu et al. 2017) proposed a deep AU recognition network from partially AU-annotated data. They utilized a restricted Boltzmann machine (RBM) model to capture AU label distribution from given AU labels. The objective is to simultaneously maximize the log likelihood of the AU classifier with regard to the learned label distribution while minimizing the error between predicted AUs and ground-truth AUs for AU-labeled samples.

Expression-dependent AU dependencies are exploited to complement missing AUs for semi-supervised AU recognition methods enhanced by expressions. Wang *et al.* (Wang, Gan, and Ji 2017) constructed a Bayesian network (BN) to capture the relations among facial expressions and AUs as well as relations among AUs, and then use expression labels as hidden knowledge to complement the missing AU labels. In the testing phase, AU labels are inferred by combining the AU-expression relationships encoded in the BN and the AU measurements obtained from a basic classifier (SVM). Ruiz *et al.* (Ruiz, Van de Weijer, and Binefa 2015) proposed the semi-hidden task learning (SHTL) method for AU recognition from partially AU-labeled images and an extra-large set of facial images labeled only with expressions. Their approach uses an AU classifier from facial images and an ex-

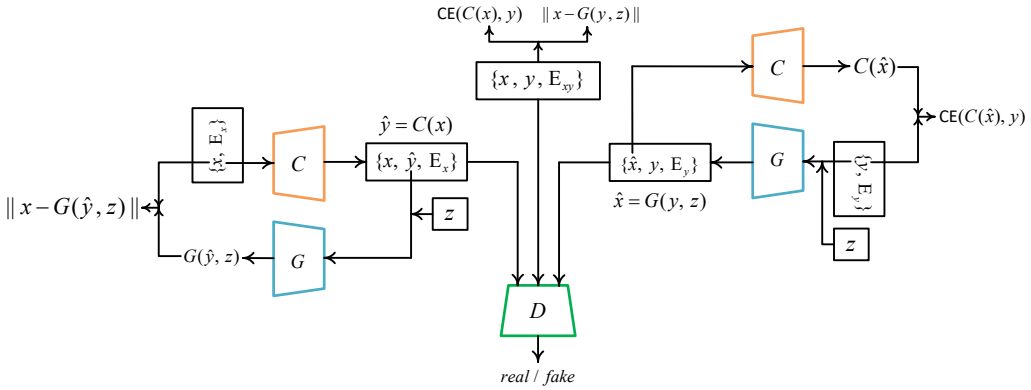

Figure 1: The framework of the proposed dual semi-supervised GAN, consisting of three modules: a discriminator D, a classifier C, and a generator G.

pression classifier from AUs. Weak supervisory information from expression labels can be propagated to the AU classifier via the expression classifier for samples lacking AU annotations. In their approach, the expression classifier is trained from the expression labels and the pseudo-AU labels sampled according to relations between expressions and AUs. Wang *et al.* (Wang, Peng, and Ji 2018) proposed an RBM prior (RBM-P) model to learn the joint distribution of all considered AUs conditioned on each expression, with the pseudo-AU labels sampled from the summarized domain knowledge. In this approach, the AU classifiers are trained by maximizing the log likelihood of the AU classifier with regard to the learned expression-conditioned AU label prior while minimizing the supervised loss from partially labeled data. Peng *et al.* (Peng and Wang 2018) summarized the domain knowledge and sampled pseudo AU labels as Wang *et al.* (Wang, Peng, and Ji 2018) did. Then, they minimized the distance between the distribution of the predicted AU labels and the pseudo-AU labels through a recognition adversarial network (RAN). For AU-labeled samples, an extra supervised loss is incorporated into the full objective. Zhang *et al.* (Zhang et al. 2018) proposed a multiple AU classifier learning method (LP-SM) that incorporates the domain knowledge (represented as the inequality relations among the AU probabilities) into the objective as the constraints. They simultaneously learned multiple AU classifiers and AU labels of training samples through an iterative optimization algorithm. The AU labels in the loss of AU classifier are ground truth AU labels for AU-labeled samples and estimated AU labels for samples without AU annotations.

The above works demonstrate the potential for AU dependencies to improve semi-supervised AU recognition. However, none of them considers the connections between AU recognition and face synthesis. The primary task and its dual task can provide effective information to each other, since they have intrinsic probabilistic connections. Therefore, we propose a semi-supervised AU recognition method that leverages both the inherent AU dependencies and imbedded connections of dual tasks.

Compared to related works, our contributions are as follows: (1) We are the first to leverage face synthesis to improve AU recognition. (2) We propose a dual semi-supervised GAN for AU recognition.

Problem Statement

Let $\Omega = T \cup U$ denote the training set, where $T = \{x^i, y^i, \mathrm{E}^i_{xy}\}_{i=1}^N$ contains N training samples with feature vectors $x \in \mathbb{R}^d$, expression label E_{xy} and AU labels $y \in \{1,0\}^l$, d is the dimension of x and l is the number of AUs. $U = \{x^j, \mathrm{E}^j_x\}_{j=1}^M$ contains M training samples annotated with expression labels only. Let $X = \{x^i, \mathrm{E}^i_x\}_{i=1}^N$ store all feature vectors in T and their corresponding expression labels, and $B = \{y^i, \mathrm{E}^i_y\}_{i=1}^N$ store all AU labels in T and their corresponding expression labels. $A = X \cup U$ stores all training feature vectors and their corresponding expression labels. Given T and U, our goal is to jointly train an AU classifier $C\colon \mathbb{R}^d \to \{1,0\}^l$ and a facial image generator $G\colon \{1,0\}^l \to \mathbb{R}^d$. Thus we can explore the connections between the two tasks to complement the missing AU labels and boost the performance of both tasks. We use facial feature points as feature vectors x , which is the input of AU classifier and the output of face generator, to represent facial image.

Proposed Approach

Figure 1 illustrates the framework of our proposed approach. Specifically, sample $\{x, \mathrm{E}_x\}$ is sampled from A. Through the AU classifier C, we get the predicted AU labels $\hat{y} = C(x)$. Similarly, sample $\{y, \mathrm{E}_y\}$ is sampled from B. Through the facial image generator G, we get the generated feature vector $\hat{x} = G(y, z)$, where $z \sim p_z(z)$ is random noise. The predicted AU label $\hat{y}$ is inputted into the generator G, and the output $G(\hat{y}, z)$ is the reconstruction of x. Similarly, the generated feature $\hat{x}$ is inputted into classifier C, and the output $C(\hat{x})$ is the reconstruction of y. According to the above procedure, three kinds of losses are considered:

Adversarial loss. As shown in Figure 1, the pseudo feature-AU-expression tuples $\{x, \hat{y}, \mathrm{E}_x\}$ and $\{\hat{x}, y, \mathrm{E}_y\}$ generated by C and G respectively are regard as "fake" samples and are sent to discriminator D for judgement. D also samples true data $\{x, y, \mathrm{E}_{xy}\}$ from T as "real" samples. To

Algorithm 1 The training of dual semi-supervised GAN in AU recognition and face synthesis

Require: Training set T, A, and B, max number of training step K, batch size s, hyper parameters α, λ_c, λ_{cl}, λ_g, and λ_{sup}.
Ensure: Classifier C and generator G

1: Randomly initialize parameters θ_d, θ_c, and θ_g of discriminator D, classifier C, and generator G, respectively.
2: **for** $k = 1, 2, ..., K$ **do**
3: Sample mini-batch of s samples $\{(x_i^d, y_i^d, \mathrm{E}_{xy_i^d})\}_{i=1}^s$ from T, sample mini-batch of s samples $\{(x_i^c, \mathrm{E}_{x_i^c})\}_{i=1}^s$ from A, sample mini-batch of s samples $\{(y_i^g, \mathrm{E}_{y_i^g})\}_{i=1}^s$ from B, sample mini-batch of s noise samples $\{z_i\}_{i=1}^s$ from $p_z(z)$.
4: Update discriminator D by descending its gradient:

$$\nabla_{\theta_d}\left[-\frac{1}{s}\sum_{i=1}^{s}\Big(\log D(x_i^d, y_i^d, \mathrm{E}_{xy_i^d}) + \alpha\log(1 - D(x_i^c, C(x_i^c), \mathrm{E}_{x_i^c})) + (1-\alpha)\log(1 - D(G(y_i^g, z_i), y_i^g, \mathrm{E}_{y_i^g}))\Big)\right]$$

5: Sample mini-batch of s samples $\{(x_i^c, \mathrm{E}_{x_i^c})\}_{i=1}^s$ from A, $s_1 (s_1 \leq s)$ of which are annotated with AU labels, $\{(x_j^c, y_j^c, \mathrm{E}_{x_j^c})\}_{j=1}^{s_1}$, sample mini-batch of s noise samples $\{z_i\}_{i=1}^s$ from $p_z(z)$.
6: Update classifier C by descending its gradient:

$$\nabla_{\theta_c}\left[-\frac{1}{s}\sum_{i=1}^{s}\log D(x_i^c, C(x_i^c), \mathrm{E}_{x_i^c}) + \frac{\lambda_c}{s}\sum_{i=1}^{s}||x_i^c - G(C(x_i^c), z_i)||_1 + \frac{\lambda_{cl}}{s_1}\sum_{j=1}^{s_1}\mathrm{CE}(C(x_j^c), y_j^c)\right]$$

7: Sample mini-batch of s samples $\{(x_i^g, y_i^g, \mathrm{E}_{y_i^g})\}_{i=1}^s$ from T, sample mini-batch of s noise samples $\{z_i\}_{i=1}^s$ from $p_z(z)$.
8: Update generator G by descending its gradient:

$$\nabla_{\theta_g}\left[-\frac{1}{s}\sum_{i=1}^{s}\log D(G(y_i^g, z_i), y_i^g, \mathrm{E}_{y_i^g}) + \frac{\lambda_g}{s}\sum_{i=1}^{s}\mathrm{CE}(C(G(y_i^g, z_i)), y_i^g) + \frac{\lambda_{reg}}{s}\sum_{i=1}^{s}||x_i^g - G(y_i^g, z_i)||_1\right]$$

9: **end for**

explore the duality between the two tasks and the dependencies among facial features, AUs, and expressions, we reduce the distance between the joint distribution of the generated pseudo feature-AU-expression tuples and the distribution of ground-truth tuples through the following adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_{(x,y,\mathrm{E}_{xy})\sim T}[\log D(x, y, \mathrm{E}_{xy})] + \\ & \alpha\mathbb{E}_{(x,\mathrm{E}_x)\sim A}[\log(1 - D(x, C(x), \mathrm{E}_x))] + \\ & (1-\alpha)\mathbb{E}_{(y,\mathrm{E}_y)\sim B, z\sim p_z(z)}[\log(1 - D(G(y, z), y, \mathrm{E}_y))]\end{aligned} \tag{1}$$

Where $\alpha \in (0, 1)$ weighs the importance of the distribution of pseudo-data generated by C in the mixed distribution. We set $\alpha = 0.5$ in our experiments to balance the distributions of pseudo-tuples generated from C and G. The discriminator D tries to maximize this loss, while classifier C and generator G try to minimize it. Since the objectives for D, C, and G are different, we define $\mathcal{L}_{adv}^d = \mathcal{L}_{adv}$, and the $\mathcal{L}_{adv}^c$ for C and $\mathcal{L}_{adv}^g$ for G as follows:

$$\mathcal{L}_{adv}^c = -\mathbb{E}_{(x,\mathrm{E}_x)\sim A}[\log D(x, C(x), \mathrm{E}_x)] \tag{2}$$

$$\mathcal{L}_{adv}^g = -\mathbb{E}_{(y,\mathrm{E}_y)\sim B, z\sim p_z(z)}[\log D(G(y, z), y, \mathrm{E}_y)] \tag{3}$$

Reconstruction loss. The AU recognition task and the face synthesis task emerge as dual forms. Considering the constraints of the dual task on the primary task, we apply a reconstruction loss (Yi et al. 2017; Zhu et al. 2017) to both AU classifier C and facial image generator G. As shown in Figure 1, the reconstruction losses for C ($\mathcal{L}_{rec}^c$) and G ($\mathcal{L}_{rec}^g$) are as follows:

$$\mathcal{L}_{rec}^c = \mathbb{E}_{(x,\mathrm{E}_x)\sim A, z\sim p_z(z)}[||x - G(C(x), z)||_1] \tag{4}$$

$$\mathcal{L}_{rec}^g = \mathbb{E}_{(y,\mathrm{E}_y)\sim B, z\sim p_z(z)}[\mathrm{CE}(C(G(y, z), y)] \tag{5}$$

We adopt L_1 distance for $\mathcal{L}_{rec}^c$, and CE in $\mathcal{L}_{rec}^g$ represents cross-entropy loss, since the AU labels are binary vectors.

Standard supervised loss. Since all samples in T are annotated with AU labels, standard supervised loss should be included in the whole objective for AU-labeled data $\{x, y, \mathrm{E}_{xy}\}$. The supervised losses for C ($\mathcal{L}_{cl}$) and G ($\mathcal{L}_{reg}$) are defined as:

$$\mathcal{L}_{cl} = \mathbb{E}_{(x,y,\mathrm{E}_{xy})\sim T}[\mathrm{CE}(C(x), y)] \tag{6}$$

$$\mathcal{L}_{reg} = \mathbb{E}_{(x,y,\mathrm{E}_{xy})\sim T, z\sim p_z(z)}[||x - G(y)||_1] \tag{7}$$

Full objective. Finally, the objectives for D, C, and G are respectively written as:

$$\begin{aligned}\mathcal{L}_D &= -\mathcal{L}_{adv}^d \\ \mathcal{L}_C &= \mathcal{L}_{adv}^c + \lambda_c\mathcal{L}_{rec}^c + \lambda_{cl}\mathcal{L}_{cl} \\ \mathcal{L}_G &= \mathcal{L}_{adv}^g + \lambda_g\mathcal{L}_{rec}^g + \lambda_{reg}\mathcal{L}_{reg}\end{aligned} \tag{8}$$

Where λ_c and λ_g are weight coefficients of reconstruction loss for C and G, respectively, and λ_{cl} and λ_{reg} are weight coefficients of supervised loss for C and G, respectively. As in the training procedure of Vanilla GAN (Goodfellow et al. 2014), D, C, and G are updated alternately: C and G are fixed, D is updated, D and G are fixed, C is updated, D and C are fixed, G is updated. The process repeats until convergence. The training procedure is shown as Algorithm 1.

Discriminator D, classifier C, and generator G are parameterized through a four-layer feedforward network. We implement the proposed method using the TensorFlow framework. Any gradient-based learning rule could be used to

update parameters for the optimization method. We use the Adam (Kingma and Ba 2014) algorithm to optimize D, C, and G in our experiments. Other hyper parameters, such as weight coefficients λ_c, λ_{cl}, λ_g, and λ_{sup}, training step K, and batch size s, vary by databases, and are determined by a validation set.

Compared to other recent works of GAN variations, DualGAN (Yi et al. 2017) is the most similar to the proposed DSGAN, although there are some important differences. To handle dual generative tasks, DualGAN consists of two GANs. DualGAN is trained from unpaired data of two domains. Through adversarial learning, DualGAN makes the distribution of the generated data from one domain converge to the distribution of the true data of another domain. It uses the reconstruction loss to utilize the constraint of one task to its dual task. Unlike DualGAN, which handles unpaired data of two domains, the proposed DSGAN handles paired data for partially labeled samples. It considers the joint distribution of the two domains (combined with expression label), but not the marginal distribution of one domain. Therefore, there is only one discriminator in our framework but not two discriminators as DualGAN has. We also consider reconstruction loss as DualGAN does. In addition, DSGAN uses the annotations of labeled samples to provide supervisory information during learning.

Experiments

Experimental Conditions

Three benchmark databases are used in our experiments. The Extended Cohn-Kanade database (CK+) (Lucey et al. 2010) is a posed expression database from which we select 309 apex frames from 309 sequences of 106 subjects that are annotated with both AU and one of six basic expressions. Following the same AU selecting strategy as Peng *et al.*'s work (Peng and Wang 2018). We consider 12 AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU12, AU17, AU23, AU24, and AU25) with occurrence frequencies greater than 10%. The MMI database (Pantic et al. 2005) is another posed expression database. We use 171 apex frames from 171 sequences of 27 subjects. Like the CK+ database, frames are annotated with both AUs and one of six expressions. We consider 13 AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU17, AU23, AU25, and AU26) with occurrence frequencies greater than 10%. The UNBC-McMaster Shoulder Pain Expression Archive database (Pain) (Lucey et al. 2011) is a spontaneous database containing 200 video sequences from 25 subjects performing "pain" or "no pain" expressions. Prkachin and Solomon Pain Intensity (PSPI) (Prkachin and Solomon 2008) is used to evaluate the pain intensity. Similar to (Peng and Wang 2018), frames with PSPI>4 are regard as "pain", and frames with PSPI=0 are regard as "no pain". We select all pain and no pain frames (7319 frames in total) from 30 sequences of 17 subjects in which pain frames are present. We consider the six AUs related to pain expression, i.e., AU4, AU6, AU7, AU9, AU10, and AU43.

We use the feature points provided by the database constructers as features on the CK+ and Pain databases. Feature points are not provided on the MMI database, so we extracted them with IntraFace (De la Torre et al. 2015). We use 49, 49, and 66 feature points on the CK+, MMI, and Pain databases, respectively. We normalize feature points through an affine transformation to make the eye centers fall on the appropriate positions for all images and use Gaussian normalization for each feature dimension. We evaluate our results using average F1 score (↑, the higher the better) of all AUs for AU recognition and root mean square error (RMSE) (↓, the lower the better) for facial feature synthesis.

Table 1: Within database experimental results of semi-supervised AU recognition with five missing rates on the three databases. (Bold numbers indicate the best performance.)

	methods	0.1	0.2	0.3	0.4	0.5
CK+	MLML	.6152	.6115	.6052	.6278	.6515
	BGCS	.7205	.7178	.7117	.7032	.6957
	DSGAN$_{ne}$	.8141	.8018	.7923	.7844	.7824
	BN	.7738	.7835	.7837	.7817	.7808
	SHTL	.5997	.5958	.5931	.5957	.5940
	RBM-P	.8186	.8148	.7948	**.8053**	.7868
	RAN	.8114	.8059	.7986	.7993	.7916
	DSGAN	**.8287**	**.8184**	**.8057**	.8015	**.7917**
	DSGAN$_{nr}$	.8131	.8062	.8032	.7929	.7838
MMI	MLML	.5063	.4806	.4793	.4651	.4323
	BGCS	.4667	.4559	.4491	.4350	.4466
	DSGAN$_{ne}$	.5672	.5583	.5489	.5349	.5218
	BN	.4897	.4792	.4725	.4659	.4378
	SHTL	.5437	.5331	.5332	.5317	.5301
	RBM-P	.5489	.5348	.5355	.5344	.5312
	RAN	.5510	.5392	.5405	.5328	.5299
	DSGAN	**.5732**	**.5609**	**.5550**	**.5464**	**.5373**
	DSGAN$_{nr}$	.5634	.5519	.5467	.5368	.5264
Pain	MLML	.2101	.2222	.1786	.1566	.1461
	BGCS	.4700	.4621	.4787	.4647	.4497
	DSGAN$_{ne}$	.5145	.5002	.4970	.5026	.4939
	BN	.2654	.3027	.2505	.2445	.1831
	SHTL	.3266	.3184	.3091	.3005	.2929
	RBM-P	.5288	.5155	.5101	.5087	.5020
	RAN	.5072	.5034	.4955	.4854	.4724
	DSGAN	**.5368**	**.5279**	**.5187**	**.5161**	**.5195**
	DSGAN$_{nr}$	.4992	.4965	.4784	.4535	.4355

We conduct within-database experiments via five fold subject-independent cross-validation and cross-database experiments. To simulate semi-supervised scenarios, we randomly exclude AU labels with certain probabilities, i.e., 0.1, 0.2, 0.3, 0.4, and 0.5, and conduct each experiment five times. The proposed method employs expression labels as auxiliary information to enhance the learning of the dual tasks. We compare the proposed method to a method that does not use the help of expression labels, referred to as DSGAN$_{ne}$. This method only considers the joint distribution of features and AUs in Equations 1, 2, and 3.

In addition, we compare the proposed method with re-

lated works. For AU recognition, the proposed DSGAN and DSGAN$_{ne}$ are compared to BGCS, MLML, BN, SHTL, RBM-P, and RAN on within-database experiments, and to SHTL, RBM-P, and RAN on cross-database experiments. To retain experimental integrity, we copy the results of RBM-P, SHTL, BGCS, MLML, and BN from (Wang, Peng, and Ji 2018), as their experimental conditions are identical to ours. Since Zhang *et al.* (Zhang et al. 2018) conducted semi-supervised experiments on the CK+ database with a missing rate of 0.5 only and Wu *et al.* (Wu et al. 2017) did not conduct experiments on the CK+, MMI, and Pain databases, we do not compare our results to theirs.

For face synthesis, we compare the proposed method with the discriminative RBM (DRBM) (Larochelle and Bengio 2008), in which the visible layer contains feature and AU label vectors. We infer facial features from the input AU labels through the Gibbs sampling method.

Experimental Results and Analyses of Within-Database Experiments

The within-database experimental results of the semi-supervised AU recognition task on the three databases are shown in Table 1. Among the first eight methods, the first three (i.e., MLML, BGCS, and DSGAN$_{ne}$) do not take advantage of expressions; the latter five learn the AU classifier with the help of expressions. As for last method, we will analyze it in later section(Evaluation of Reconstruction Loss). From Table 1, we can obtain the following observations.

First, on the whole, methods considering expressions perform better than methods ignoring expression labels. For example, DSGAN performs better than DSGAN$_{ne}$ in all cases, indicating that expression is explicitly helpful for AU recognition, since expression and AUs are strongly related. When AUs are missing, expression labels can provide weak supervisory information.

Second, DSGAN$_{ne}$ performs better in all scenarios than the other two methods that don't take expression into account (MLML and BGCS). For example, when the missing rate is 0.1, the performance of DSGAN$_{ne}$ on the CK+ database is 12.99% and 32.33% higher than that of BGCS and MLML, respectively, which demonstrates the superiority of the proposed method. Although expression labels are not present, the joint distribution of features and AUs captured from partial samples with ground-truth AU labels can provide weak supervisory information for samples without AU annotations. BGCS and MLML do not utilize the distribution to constrain AU predictions of training samples when samples lack AU annotations. More importantly, BGCS and MLML only consider the AU recognition task, while DSGAN$_{ne}$ considers and trains the face synthesis task and AU recognition simultaneously, thus achieving better performance.

Third, compared to BN, SHTL, RBM-P, and RAN, the proposed DSGAN performs best in most cases. For example, when the missing rate is 0.1, the performances of DSGAN are 17.05%, 5.43%, 4.43%, and 4.03% higher than that of BN, SHTL, RBM-P, and RAN respectively, demonstrating the effectiveness of DSGAN. BN can only explore pairwise dependencies among AUs, while DSGAN considers the joint distribution of features, AUs, and expression, capturing global dependencies among all AUs. Since SHTL trains the AU recognition task and expression classification task separately, any expression classifier errors propagate to the AU classifier. We consider both tasks in our work and train them simultaneously. Both RBM-P and RAN only consider the joint distribution of AUs conditioned on expression. Our method realizes the relations between features and AU labels and considers the joint distribution of features, AUs, and expression. Furthermore, all four methods only handle the AU recognition problem. Our model recognizes that the intrinsic connections between AU recognition and face synthesis are very helpful for learning both tasks. We optimize the dual tasks to achieve better performance.

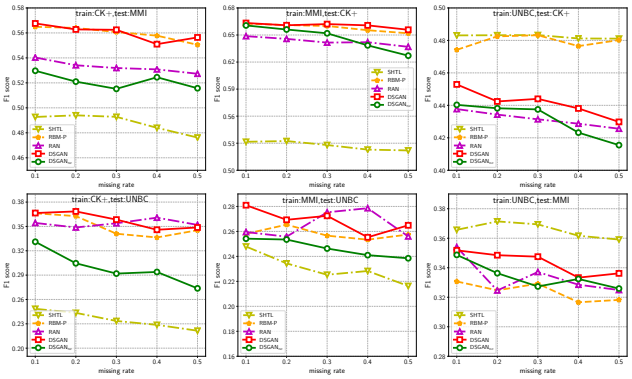

Figure 2: Cross-database experiment results of semi-supervised AU recognition.

Finally, when comparing DSGAN$_{ne}$ to the four methods considering expression (i.e., BN, SHTL, RBM-P, and RAN), we find that the proposed DSGAN$_{ne}$ performs better than some of them. Specifically, DSGAN$_{ne}$ performs better than BN and SHTL on the CK+ database; better than all four methods on the MMI database; and better than BN, SHTL, and RAN on the Pain database. This further demonstrates that the proposed method successfully exploits the duality between the two tasks to enhance the learning of the AU classifier.

Experimental Results and Analyses of Cross-Database Experiments

The results of cross-database experiments are shown in Figure 2. When the methods are trained on the CK+ and MMI databases (the first two columns of Figure 2), the proposed DSGAN performs best in most cases. Especially for the experiments that test on the Pain database, which are difficult scenarios for AU recognition since the emotion setting of the Pain database is different from that of the CK+ and MMI databases, the better performances of DSGAN demonstrate the better generalization ability of DSGAN. Although DSGAN$_{ne}$ is not assisted by expressions, it takes full advantage of the duality between face synthesis and AU recognition task, thus still performing better than SHTL in all cases.

When the methods are trained on the Pain database (the last column of Figure 2), the proposed method performs

Table 2: Comparison to state-of-the-art fully supervised methods.

	CK+	MMI	UNBC
MC-LVM	.7707	-	**.6345**
SVM-HMM	-	**.6712**	-
HRBM	.7147	-	.5942
FFD	-	.6652	-
DSGAN	**.7917**	.5373	.5195

poorly. DSGAN is inferior to SHTL in two scenarios and inferior to RBM-P when testing on the CK+ database. When training on the Pain database, SHTL achieves the best performance since it uses not only the partially available AU labels in the Pain database, but also expression labels in an extra-large facial image database with six basic emotion settings. Furthermore, we only consider six AUs on the Pain database. This may not be enough AUs for the face synthesis task, so its assistance is not significant.

Comparisons to Fully-Supervised Methods

We also compare the proposed semi-supervised method (with 0.5 missing rate) to fully-supervised methods. On the CK+ and Pain databases, we compare our method to MC-LVM (Eleftheriadis, Rudovic, and Pantic 2015) and HRBM (Wang et al. 2013). On the MMI database, we compare DSGAN to SVM-HMM (Valstar and Pantic 2012) and FFD (Koelstra, Pantic, and Patras 2010). The results of HRBM are from (Eleftheriadis, Rudovic, and Pantic 2015). The results are shown in Table 2. Since the experimental conditions of these methods are different from ours, these comparisons are only for reference.

Table 2 shows that DSGAN performs worse than other methods on the MMI and Pain databases. This is expected, since only half the samples in our training set are annotated with AU labels, while the other four methods use fully AU-labeled training samples. Supervisory information of AU labels is very helpful tool for learning the AU classifier. It's surprising that DSGAN performs better than MC-LVM and HRBM on the CK+ database. This demonstrates the effectiveness of the proposed method for leveraging the face synthesis task and expression labels.

Evaluation of Reconstruction Loss

In order to evaluate the contribution of reconstruction loss, we remove the reconstruction loss in the full objective (DSGAN$_{nr}$) by setting $\lambda_c = \lambda_g = 0$. We then conduct experiments on the three databases and compare with DSGAN. The results of DSGAN$_{nr}$ are shown in Table 1, which shows that DSGAN$_{nr}$ performs worse than DSGAN in all scenarios. The reconstruction loss reflects the constraint of the dual task to the primary task. When reconstruction loss is removed, AU recognition performance typically decreases. The deterioration tends to be gradual, particularly on the Pain database. This indicates that the role of reconstruction losses is more important as the number of missing AU labels increases.

Table 3: RMSE of DRBM and the proposed methods for face synthesis.

	CK+	MMI	Pain
DRBM	1.3866	1.8061	3.0192
DSGAN	**0.9687**	**0.9866**	**2.5238**
DSGAN$_{ne}$	1.0002	1.0056	2.5563
DSGAN$_{nr}$	1.0103	1.0177	2.5476

Experimental Results and Analyses of Facial Image Synthesis

In this section, we evaluate the performance of the face generator with RMSE. Table 3 lists the results of the proposed methods (DSGAN, DSGAN$_{ne}$, and DSGAN$_{nr}$) with missing rate set to 0.2, and the results of compared method (DRBM) on the three databases. From Table 3, we can obtain the following observations. First, the proposed methods all perform better than DRMB, even though DRBM uses fully AU-labeled data. DSGAN performs best on the three databases, demonstrating the superiority of the proposed method for face synthesis tasks. We explore the duality between two tasks and utilize the AU recognition task to enhance the face synthesis task. The better performance of DSGAN compared to DSGAN$_{ne}$, and DSGAN$_{nr}$ demonstrates the contributions of expression labels and reconstruction losses. Second, the methods perform best on the CK+ database. The performances on the MMI database are slightly worse than that on the CK+ database, but the performances on the Pain database are poorest by far. This may be because fewer AUs and more feature points are considered on the Pain database. Six AUs may be insufficient to generate 66 feature points.

Conclusion

In this paper, we propose a novel dual semi-supervised method (DSGAN) to handle AU recognition and face synthesis simultaneously. We consider the joint distribution of features, AUs, and expression, and propose an adversarial framework to explore the global dependencies among them and the probabilistic duality between two tasks. Expression labels can be used as auxiliary information to improve the performances of the learning system, and few framework changes are needed when expression labels are unavailable. Two reconstruction losses are leveraged to utilize the constraint of dual task to primary task. Standard supervised losses are added to the full objective for samples with AU annotations. Our method achieves better results than state-of-the-art methods on both within-database and cross-database experiments, demonstrating its superiority.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant No. 61473270, 917418129, 61727809), and the major project from Anhui Science and Technology Agency (1804a09020038).

References

De la Torre, F.; Chu, W.-S.; Xiong, X.; Vicente, F.; Ding, X.; and Cohn, J. F. 2015. Intraface. In *IEEE International Conference on Automatic Face and Gesture Recognition*.

Du, S.; Tao, Y.; and Martinez, A. M. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences* 111(15):E1454–E1462.

Eleftheriadis, S.; Rudovic, O.; and Pantic, M. 2015. Multi-conditional latent variable model for joint facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 3792–3800.

Friesen, E., and Ekman, P. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*.

Friesen, W. V., and Ekman, P. 1983. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco* 2(36):1.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.; and Ma, W.-Y. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, 820–828.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koelstra, S.; Pantic, M.; and Patras, I. 2010. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence* 32(11):1940–1954.

Larochelle, H., and Bengio, Y. 2008. Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, 536–543. ACM.

Lucey, P.; Cohn, J. F.; Kanade, T.; Saragih, J.; Ambadar, Z.; and Matthews, I. 2010. 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 (CVPRW)*, 94–101. IEEE.

Lucey, P.; Cohn, J. F.; Prkachin, K. M.; Solomon, P. E.; and Matthews, I. 2011. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011)*, 57–64. IEEE.

Martinez, B.; Valstar, M. F.; Jiang, B.; and Pantic, M. 2017. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*.

Pantic, M.; Valstar, M.; Rademaker, R.; and Maat, L. 2005. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo, (ICME 2005)*, 5. IEEE.

Peng, G., and Wang, S. 2018. Weakly supervised facial action unit recognition through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2188–2196.

Prkachin, K. M., and Solomon, P. E. 2008. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain* 139(2):267–274.

Prkachin, K. M. 1992. The consistency of facial expressions of pain: a comparison across modalities. *Pain* 51(3):297–306.

Ruiz, A.; Van de Weijer, J.; and Binefa, X. 2015. From emotions to action units with hidden and semi-hidden-task learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 3703–3711.

Song, Y.; McDuff, D.; Vasisht, D.; and Kapoor, A. 2015. Exploiting sparsity and co-occurrence structure for action unit recognition. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, 1–8. IEEE.

Valstar, M. F., and Pantic, M. 2012. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42(1):28–43.

Wang, Z.; Li, Y.; Wang, S.; and Ji, Q. 2013. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 3304–3311.

Wang, S.; Gan, Q.; and Ji, Q. 2017. Expression-assisted facial action unit recognition under incomplete au annotation. *Pattern Recognition* 61:78–91.

Wang, S.; Peng, G.; and Ji, Q. 2018. Exploring domain knowledge for facial expression-assisted action unit activation recognition. *IEEE Transactions on Affective Computing*.

Wu, B.; Lyu, S.; Hu, B.-G.; and Ji, Q. 2015. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition* 48(7):2279–2289.

Wu, S.; Wang, S.; Pan, B.; and Ji, Q. 2017. Deep facial action unit recognition from partially labeled data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 3971–3979. IEEE.

Xia, Y.; Qin, T.; Chen, W.; Bian, J.; Yu, N.; and Liu, T.-Y. 2017. Dual supervised learning. In *International Conference on Machine Learning*, 3789–3798.

Xia, Y.; Tan, X.; Tian, F.; Qin, T.; Yu, N.; and Liu, T.-Y. 2018. Model-level dual learning. In *International Conference on Machine Learning*, 5379–5388.

Yi, Z.; Zhang, H. R.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, 2868–2876.

Zhang, Y.; Dong, W.; Hu, B.-G.; and Ji, Q. 2018. Classifier learning with prior probabilities for facial action unit recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5108–5116.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251. IEEE.