Pattern Recognition 114 (2021) 107837

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Multi-task face analyses through adversarial learning

Shangfei Wang*, Shi Yin, Longfei Hao, Guang Liang

University of Science and Technology of China, 443 HuangShan Rd, Hefei City, Anhui Province, 230027, China

ARTICLE INFO

Article history:
Received 29 November 2019
Revised 7 December 2020
Accepted 16 January 2021
Available online 26 January 2021

Keywords:
Multi-task learning
Adversarial learning
Face analyses

ABSTRACT

The inherent relations among multiple face analysis tasks, such as landmark detection, head pose estimation, gender recognition and face attribute estimation are crucial to boost the performance of each task, but have not been thoroughly explored since typically these multiple face analysis tasks are handled as separate tasks. In this paper, we propose a novel deep multi-task adversarial learning method to localize facial landmark, estimate head pose and recognize gender jointly or estimate multiple face attributes simultaneously through exploring their dependencies from both image representation-level and label-level. Specifically, the proposed method consists of a deep recognition network $\mathcal{R}$ and a discriminator $\mathcal{D}$. The deep recognition network is used to learn the shared middle-level image representation and conducts multiple face analysis tasks simultaneously. Through multi-task learning mechanism, the recognition network explores the dependencies among multiple face analysis tasks from image representation-level. The discriminator is introduced to enforce the distribution of the multiple face analysis tasks to converge to that inherent in the ground-truth labels. During training, the recognizer tries to confuse the discriminator, while the discriminator competes with the recognizer through distinguishing the predicted label combination from the ground-truth one. Though adversarial learning, we explore the dependencies among multiple face analysis tasks from label-level. Experimental results on benchmark databases demonstrate the effectiveness of the proposed method for multi-task face analyses.

1. Introduction

Face analyses have attracted increasing attention in recent years due to their wide applications in human computer interaction. Face analyses include several tasks, such as facial landmark detection, head pose estimation, face identification, facial expression classification, gender recognition and multiple face attribute estimation. These tasks are related to each other. For example, as shown in Fig. 1, a person who wears necklace and earrings is more likely to be a female, and is less likely to be a male; and a person with sideburns and goatee is more likely to be a male, and is less likely to be a female; the locations of landmark are affected by head poses; facial expression variations obviously influence the location of landmarks. Such inherent connections among facial landmarks, head poses and expressions or multiple face attributes can be leveraged for multiple face analysis tasks, but have not been thoroughly explored yet, since typically face analysis tasks are handled separately.

Only very recently, a few works have turned to solve several face analysis tasks jointly. Zhang et al. [1] and Ranjan et al. [2] modeled dependencies among several face analysis tasks from the learned representation-level. Zhang et al. [3] proposed a multi-task convolutional neural network (CNN) consisting of shared features for heterogeneous face attributes. But they failed to consider task dependencies inherent in label-level. Zhu and Ramanan [4] considered the task dependencies from the label-level, but ignored their dependencies in facial appearance. Instead of jointly learning multiple face analysis tasks in a parallel way, like the above work, Wu et al. [5] and Honari et al. [6] tried to leverage task dependencies in a serial manner, and mainly captured multi-task dependencies in representation-level.

To the best of our knowledge, although both representation-level dependencies and label-level dependencies are critical for multiple face analysis tasks, little work addresses them simultaneously till now. Therefore, in this paper, we propose a deep multi-task adversarial learning method for multiple face analysis tasks through exploring their dependencies from both representation-level and label-level. Specifically, we construct a deep network as a multi-task recognizer to explore connections among multiple facial analysis tasks through representation-level. These tasks include facial landmark related multi-task face analyses, which predict facial landmark lovisibilitiescations, landmark visibilities, face poses and genders jointly; and facial attribute estimation, which make predictions on multiple facial attributes, e.g., whether the subject

* Corresponding author.
E-mail address: sfwang@ustc.edu.cn (S. Wang).

https://doi.org/10.1016/j.patcog.2021.107837
0031-3203/

(a) (b) (c) (d) (e)

Fig. 1. The dependencies among multiple face analysis tasks.

is wearing hat, earrings or lipstick. Then, the recognizer competes with a discriminator under the adversarial learning framework, and the joint distribution of the labels predicted by the multi-task recognizer are converged to that inherent in the ground-truth labels. Thus, multi-task dependencies from the label-level are also captured. Experimental results on benchmark databases demonstrate that the proposed method successfully leverages task dependencies inherent in both representation and target label and thus achieves state of the art performance on multiple face analysis tasks.

The rest of this paper is organized in the following manner. Section 2 gives an overview of the related work on multi-task face analyses. Section 3 briefly gives the problem statement for our method. Section 4 elaborates on the proposed method for multi-task face analyses. Section 5 presents the experimental results and analyses on benchmark databases, and makes the comparison to related works. Section 6 concludes our work.

2. Related work

In this section, we summarize and analyze recent face analysis works. We divide these works into three categories: facial landmark detection, facial landmark related multi-task face analyses and multiple face attribute estimation. Furthermore, we discuss recent work on adversarial multi-task learning.

2.1. Facial landmark detection

Facial landmark detection is crucial for many face analysis tasks, such as head pose estimation, face recognition, facial expression recognition and gender recognition. Facial landmark detection methods can be classified into two types, template based and regression based method. Template based methods model facial shape by a parametric model, such as Active Appearance Model (AAM) [7], ASM [8] and Constrained Local Model (CLM) [9]. These methods assume an explicit form of facial parameters, which may have difficulty in handling “in the wild” facial appearances. Regression based methods train a regressor to precit landmark positions. These methods can be further divided into two types, i.e., coordinate regression methods [10,11] and heatmap regression methods[12–16]. The former directly maps facial appearances to landmark coordinates, while the latter outputs a spatial distribution for each landmark. Despite many works on landmark detection, most of them take locating landmarks as a single task, and the inherent dependencies among multiple facial analysis tasks are not explored thoroughly. To address this, the proposed method captures joint distributions among multiple tasks by an adversarial learning framework to assist each single task.

2.2. Facial landmark related multi-task face analyses

In this section, we discuss several recent works of landmark related multi-task face analyses. Zhang et al. [1] proposed a task-constrained deep convolutional network (TCDCN) to jointly optimize facial landmark detection with a set of related tasks, such as pose estimation, gender recognition, glasses detection, and smiling classification. They further systematically demonstrated that the representations learned from related tasks facilitate the learning of facial landmark detector. Zhang et al. [17] combined landmarks localization, pose estimation and 3D reconstruction in a multi-task learning framework. Ranjan et al. [2] strategically designed the network architecture to exploit both low-level and high-level features of the network. They proposed HyperFace and HF-Resnet, deep multi-task learning methods for simultaneous face detection, landmarks localization, pose estimation and gender recognition.

Unlike Zhang et al. [1]’s, Zhang et al. [17]’s and Ranjan et al. [2]’s works, which explored the inherent dependencies among multiple face analysis tasks from the learned representation-level, Zhu and Ramanan [4] considered the dependencies from the label-level, i.e. the topological changes due to related factors. They proposed a method for face detection, pose estimation, and landmark localization (FPLL) simultaneously. Specifically, they proposed a mixtures of trees with a shared pool of parts. Every facial landmark is modeled as a part, and the topological changes due to viewpoint are captured by the global mixtures.

Instead of jointly learning multiple face analysis tasks in a parallel way, like the above works, Wu et al. [5] proposed an iterative cascade method to simultaneously perform facial landmark detection, pose and deformation estimation. Their method iteratively updated the facial landmark locations, facial occlusion, head pose and facial deformation until convergence. Although the iterative cascade procedure can capture connections among multiple face analysis tasks at representation-level, the errors caused in the previous iteration may be propagated to the next iteration. Therefore, we prefer to jointly learning multiple face analysis tasks in a parallel way.

Other than exploring task dependencies in supervised learning scenarios, Honari et al. [6] leveraged task dependencies to improve landmark localization in semi-supervised learning scenarios. They proposed a framework of sequential multi-task learning for landmark localization and related face analysis tasks, such as expression recognition. Specifically, their proposed method first detected landmarks, and then the detected landmarks are the input of the related face analysis tasks, which are acted as an auxiliary signal to guide the landmark localization on unlabeled data. Although their proposed sequential multi-task learning framework successfully explores related face analysis tasks to boost facial landmark detection under partially labeled data, the dependencies among tasks are mainly exploited in the learned representation-level, not in the label-level. Furthermore, the errors caused by the first stage could be propagated to the next stage, and vice versa.

To the best of our knowledge, few works leverage inherent dependencies among landmark-related multiple face analysis tasks from both representation-level and label-level. Therefore, we propose a deep multi-task adversarial learning method for facial landmark detection enhanced by multiple face analysis tasks through

exploring their dependencies from both representation-level and label-level. Specifically, we first construct a deep network as a multi-task recognizer $\mathcal{R}$ to jointly detect facial landmarks, estimate landmark visibility, recognize head pose and classify gender. Through multi-task learning, the designed deep network can explore connections among multiple tasks through representation-level. Then, we introduce a discriminator $\mathcal{D}$ to distinguish the ground-truth label combination from the output of the recognizer $\mathcal{R}$. During training, $\mathcal{R}$ maximums the probability of mistake made by $\mathcal{D}$, while $\mathcal{D}$ does the opposite. Through such adversarial learning, the proposed method enforces the joint distribution of the labels predicted by $\mathcal{R}$ converge to that inherent in the ground-truth labels, and thus leverages multi-task dependencies from the label-level.

2.3. Multiple face attribute estimation

Face attribute estimation has attracted increasing attentions, since face attributes are middle-level abstraction between the low-level facial features and the high-level labels. FaceTracker [18] used a combination of support vector machines and Adaboost to select the optimal features for each attribute, and train each attribute classifier separately. It ignores the relations among multiple face attributes, which can be leveraged to boost the performance of multiple face attribute estimation. Zhang et al. [3] proposed Pose Alignment Networks for Deep Attribute modeling (PANDA) to obtain a pose-normalized deep representation for multiple face attribute estimation. Liu et al. [19] believed that face localization can improve the performance of multiple face attribute estimation, and thus cascaded face localization networks (LNets) and the attribute network (ANet). Mao et al. [20] proposed DMM-CNN, which learns facial landmark detection and facial attributes classification jointly with shared representations. DMM-CNN also splits facial attributes as two group of attributes, i.e., objective attributes and subjective attributes, and adopts different networks as well as a dynamic weighting strategy to learn attribute-specific representations for them. Zhong et al. [21] combined several off-the-shelf convolutional neural networks (i.e., CTS-CNN), which are trained for face recognition to estimate multiple face attributes simultaneously. These above works explore representation-level connections for multiple face attributes, but ignore the dependencies among multiple face attributes from the label-level.

Han et al. [22] tried to model both representation-level and label-level dependencies. They proposed a CNN to capture representation-level dependencies through the shared low-level features for all attributes and task specific high-level features for heterogeneous attributes. They further proposed constraints according to prior knowledge to capture the fixed label-level dependencies. Instead of using constrains to model fixed dependencies, Hand et al. [23] proposed a multi-task CNN (MCNN-AUX) to learn the label-level dependencies through an auxiliary network stacked on the top. Cao et al. [24] considered the identity information and attribute relationships jointly. They proposed a partially Shared Multi-task CNN (PS-MCNN) to learn the task specific and shared features, and then utilized the identity information to improve the performance of face attribute estimation (PS-MCNN-LC). Although the above three works can explore dependencies from both representation-level and label-level for multiple face attribute estimation, the captured label-level dependencies are either fixed or represented by fixed form through the structure and parameters of a network.

To address the above issues, the proposed work employs an adversarial strategy to capture label distributions directly without the assumption of the distribution form. Specifically, we first construct a deep multi-task network $\mathcal{R}$ to estimate multiple face attributes simultaneously. Then, we introduce a discriminator $\mathcal{D}$ to distinguish the ground-truth label combination from the output of the recognizer $\mathcal{R}$. Through adversarial learning, the proposed method leverages multi-task dependencies from both label-level and representation-level to facilitate multiple face attribute estimation.

2.4. Adversarial multi-task learning

Recent years have seen a few works incorporating adversarial learning with multi-task learning. For example, Bai et al. [25] introduced a generator to up-sample small blurred images into fine-scale ones for more accurate detection, and a discriminator describes each super-resolution image patch with multiple scores. Liu et al. [26] proposed to alleviate the shared and private latent feature spaces from interfering with each other by using adversarial training and orthogonality constraints. The adversarial training is used to construct common and task-invariant shared latent spaces, while the orthogonality constraint is used to eliminate redundant features from the private and shared spaces. Liu et al. [27] proposed an encoder to extract a disentangled feature representation for the factors of interest, and the discriminators to classify each of the factors as individual tasks. The encoder and the discriminators are trained cooperatively on factors of interest, but in an adversarial way on factors of distraction. All above works leverage adversarial learning for better input data or representations for multi-task learning, but ignore the dependencies among target labels. We are the first to explore dependencies among multiple tasks from both representation and label-level through adversarial mechanism. In our method, both the recognizer and the discriminator are deep networks with the capability to model complex distributions. Through adversarial training, the discriminator and the recognizer are improved together on capturing label patterns by their competition with each other. Hence, when the training is converged, the discriminator and the recognizer can fully capture the joint distributions among these labels.

3. Problem statement

Let $\boldsymbol{T} = \{\boldsymbol{x}, \boldsymbol{y}\}^N$ denotes N training samples, where $\boldsymbol{x}$ represents the facial image, $\boldsymbol{y} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}$ represents the ground-truth labels, such as facial landmark locations, visibility of each landmark, head pose angle and gender information or multiple face attributes. The purpose of the paper is to learn a multi-task recognizer $\mathcal{R} : \boldsymbol{x} \rightarrow \boldsymbol{y}$ through optimizing the following formula:

$$\min_{\Theta_{\mathcal{R}}} \alpha_1 * \mathcal{L}_s(\mathcal{R}(\boldsymbol{x}; \Theta), \boldsymbol{y}) + \alpha_2 * \mathcal{L}_d(P_{\boldsymbol{y}}, P_{\boldsymbol{y}'}) , \tag{1}$$

where $\mathcal{L}s$ is the supervised loss of multiple tasks, $\Theta_{\mathcal{R}}$ are parameters of multi-task recognizer $\mathcal{R}$, $\boldsymbol{y}' = \mathcal{R}(\boldsymbol{x})$, $P_{\boldsymbol{y}}$ and $P_{\boldsymbol{y}'}$ are the distribution of the ground-truth label and the distribution of the predicted labels from $\mathcal{R}$, respectively, $\mathcal{L}_d$ is the distance between two distributions. The first term minimizes the recognition errors of multi-tasks that sharing common representations, and the second term closes the joint distribution of the predicted label combination to the ground-truth label combination. α_1 and α_2 balance these two terms. Therefore, the proposed method can successfully explore connections among multiple face analysis tasks through both representation-level and label-level.

4. Proposed method

The framework of the proposed deep multi-task adversarial learning method is shown in Fig. 2. It consists of a deep multi-task recognizer $\mathcal{R}$ and a discriminator $\mathcal{D}$. The goal of $\mathcal{R}$ is to learn shared image representation and predict landmark location, landmark visibility, pose and gender simultaneously or multiple face

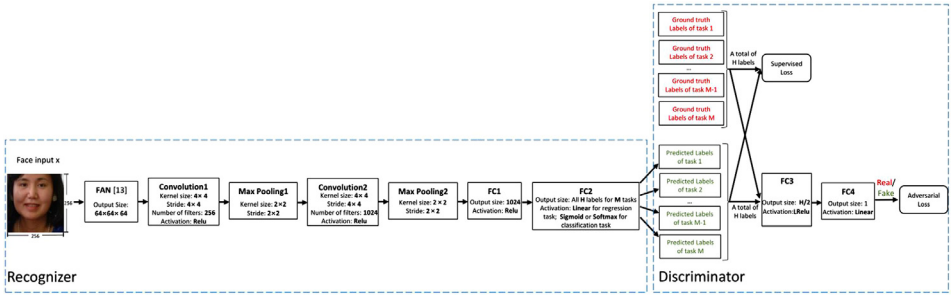

Fig. 2. Framework for the proposed method. For the multi-task recognizer $\mathcal{R}$, we use FAN network [13] to encode facial representations and use CNN and FC networks to convert the representations as the predicted labels. For the detailed structure of FAN, please refer to the source codes in the following website: https://www.adrianbulat.com/face-alignment. The discriminator $\mathcal{D}$ distinguishes the ground-truth label combination from the output of $\mathcal{R}$ by FC networks.

attributes simultaneously. $\mathcal{D}$ is to distinguish the ground-truth label combination from the label combination predicted by $\mathcal{R}$. With the supervisory information of the ground-truth label combinations, the recognizer $\mathcal{R}$ can successfully capture the connections among multiple face analysis tasks by sharing feature representations. Through the competition between $\mathcal{R}$ and $\mathcal{D}$, the distribution of the predicted label combination could converge to the label distribution of the ground-truth. Thus, the proposed method can model the dependencies among landmark, visibility, pose, gender and the dependencies among multiple face attributes.

WGAN [28] is adopted by our adversarial learning method for its better convergence. Through adversarial learning, we can minimize the distance of two distributions, i.e., the second term of Eq. (1), but do not need to model $P_{\boldsymbol{y}}$ and $P_{\boldsymbol{y}'}$ directly, which are complex and error prone processes. We replace $\mathcal{L}_d(P_{\boldsymbol{y}}, P_{\boldsymbol{y}'})$ as the following adversarial loss:

$$\min_{\mathcal{R}} \max_{\mathcal{D}} \mathcal{L}_{adv} = \mathbb{E}_{\boldsymbol{y}}[\mathcal{D}(\boldsymbol{y})] + \mathbb{E}_{\hat{\boldsymbol{y}}}[-\mathcal{D}(\hat{\boldsymbol{y}})]\,, \tag{2}$$

where $\hat{\boldsymbol{y}} = \mathcal{R}(\boldsymbol{x})$ is the predicted label combination of facial image $\boldsymbol{x}$ that is regard as "fake", $\boldsymbol{y}$ is the ground-truth label combination regarded as "real". It's hard to optimize the above problem directly. We seek individual objective for $\mathcal{R}$ and $\mathcal{D}$ and utilize an alternate training procedure as described in the following sections.

4.1. Recognizer

One objective of recognizer $\mathcal{R}$ is to minimize $\mathcal{L}_{adv}$ in Eq. (2). It means recognizer $\mathcal{R}$ tries to 'fool' discriminator $\mathcal{D}$ and let it classify the predicted label combination $\hat{\boldsymbol{y}}$ as "real". Therefore, the adversarial objective for recognizer $\mathcal{R}$ is as follows:

$$\mathcal{L}^{\mathcal{R}}_{adv} = -\mathcal{D}(\hat{\boldsymbol{y}})\,. \tag{3}$$

We construct a multi-task recognizer $\mathcal{R}$ to learn the sharing representation and explore dependencies among these multiple tasks, i.e., landmark detection, visibility recognition, head pose estimation and gender recognition, or multiple face attribute estimation. The supervised loss $\mathcal{L}_s$ for multi-task recognizer $\mathcal{R}$ contains the following losses:

Landmark Detection: The supervised loss for landmark detector is described as Eq. (4).

$$\mathcal{L}^L_s = \frac{1}{2m}\sum_{i=1}^{m} v_i((\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2)\,, \tag{4}$$

where (x_i, y_i) is the location of ith landmark, $(\hat{x}_i, \hat{y}_i)$ is corresponding estimation, m is the total number of landmark points in one image. The visibility factor v_i is 1 if the ith landmark is visible, otherwise is 0, which implies that the ground-truth location for ith landmark is not provided.

Visibility Recognition: We learn the visibility recognizer to predict the visibility of all landmarks. $\boldsymbol{v}$ is a multiple binary-value label vector. Hence, the supervised loss for visibility recognizer is shown as in Eq. (5):

$$\mathcal{L}^V_s = -\frac{1}{m}\sum_{i=1}^{m}(v_i \log \hat{v}_i + (1 - v_i)\log(1 - \hat{v}_i))\,, \tag{5}$$

where $\hat{v}_i$ and v_i are the predicted visibility and the ground-truth visibility of the ith landmark, respectively.

Pose Estimation: Since the pose information provided by database constructors are either continuous or discrete, the form of loss for pose estimator varies by databases. For continuous pose information (i.e., roll, pitch and yaw), the L2 loss function is used:

$$\mathcal{L}^P_s = \frac{1}{3}\left[(\hat{p}_1 - p_1)^2 + (\hat{p}_2 - p_2)^2 + (\hat{p}_3 - p_3)^2\right], \tag{6}$$

where (p_1, p_2, p_3) are the ground-truth roll, pitch and yaw respectively, and $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$ are the estimated pose angles. For discrete pose information, we view the pose estimation as a multi-class classification problem and the cross-entropy loss is used.

$$\mathcal{L}^P_s = -\sum_{i=1}^{K} p_i \log(\hat{p}_i)\,, \tag{7}$$

where $(p_1, p_2, \ldots, p_K)$ is the one-hot code of the ground-truth pose angle and $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$ is the one-hot code of corresponding estimated angle. K is the number of angles.

Gender Classification: Gender classification is a binary classification problem. Hence, the supervised loss for gender classifier is as shown in Eq. (8).

$$\mathcal{L}^G_s = -\left[g\log(\hat{g}) + (1 - g)\log(1 - \hat{g})\right], \tag{8}$$

where $\hat{g}$ and g are the predicted gender and the ground-truth gender, respectively.

Face Attribute Estimation: The face attributes are all binary. Therefore, the supervised loss for multiple face attribute estimator is shown as Eq. (9):

$$\mathcal{L}^A_s = -\frac{1}{n}\sum_{i=1}^{n} a_i \log \hat{a}_i + (1 - a_i)\log(1 - \hat{a}_i)\,, \tag{9}$$

where $\hat{a}_i$ and a_i are the ith predicted face attribute and the ground-truth attribute, respectively. n is the number of attributes.

Finally, the full supervised loss $\mathcal{L}_s$ can be written as follows:

$$\mathcal{L}_s = \alpha_L * \mathcal{L}^L_s + \alpha_V * \mathcal{L}^V_s + \alpha_P * \mathcal{L}^P_s + \alpha_G * \mathcal{L}^G_s\,, \tag{10}$$

or

$$\mathcal{L}_s = \alpha_{attr} * \mathcal{L}^A_s\,. \tag{11}$$

We combine the supervised loss $\mathcal{L}_s$ and the adversarial loss $\mathcal{L}^{\mathcal{R}}_{adv}$ as the full objective of multi-task recognizer $\mathcal{R}$, shown as Eq. (12):

$$\mathcal{L}^R = \mathcal{L}_s + \alpha_A * \mathcal{L}^{\mathcal{R}}_{adv}\,, \tag{12}$$

where α_L, α_V, α_P, α_G, α_{attr} and α_A are weight coefficients of supervised losses of the corresponding subtasks and adversarial loss, respectively.

4.2. Discriminator

As shown in the right part of Fig. 2, we construct a discriminator $\mathcal{D}$. The purpose of the discriminator is to classify the ground-truth label combination as "real" and the predicted label combination as "fake". Therefore, the adversarial loss for $\mathcal{D}$ is shown as Eq. (13):

$$\mathcal{L}^{\mathcal{D}} = -[\mathcal{D}(\boldsymbol{y}) - \mathcal{D}(\hat{\boldsymbol{y}})] \tag{13}$$

The multi-task recognizer $\mathcal{R}$ and the discriminator $\mathcal{D}$ are updated with an alternate procedure: fix $\mathcal{R}$, update $\mathcal{D}$ according to Eq. (13), and then fix $\mathcal{D}$, update $\mathcal{R}$ according to Eq. (12). This process repeats until convergence. The detailed training procedure is described in Algorithm 1.

Algorithm 1 Training algorithm of the proposed multi-task adversarial learning.

Input The training set $\boldsymbol{T}$, the batch size s, the number of training step K and the hyper parameter k.
Output The multi-task recognizer $\mathcal{R}$.

1: Initialize the parameters $\Theta_{\mathcal{R}}$ and $\Theta_{\mathcal{D}}$ of $\mathcal{R}$ and $\mathcal{D}$, respectively.
2: **for** $i = 1 \rightarrow K$ **do**
3: **for** $j = 1 \rightarrow k$ **do**
4: Randomly sample mini-batch of s facial images $\{\boldsymbol{x}\}_{i=1}^{s}$ from feature space and sample mini-batch of s labels $\{\boldsymbol{y}\}_{i=1}^{s}$ from label space.
5: Update the parameters of discriminator $\mathcal{D}$ according to:

$$\nabla_{\Theta_{\mathcal{D}}}\left(-\frac{1}{s}\sum_{i=1}^{s}[\mathcal{D}(\boldsymbol{y}) - \mathcal{D}(\mathcal{R}(\boldsymbol{x}))]\right)$$

6: Clip the absolute value of$\mathcal{D}$'s weights to not more than δ.
7: **end for**
8: Randomly sample a mini-batch of s samples $\{\boldsymbol{x}, \boldsymbol{y}\}_{i=1}^{s}$ from training set $\boldsymbol{T}$
9: Update multi-task recognizer $\mathcal{R}$ according to Equation (12).
10: **end for**

The adopted WGAN framework is an improved version of GAN [29] with three significant modifications. First, WGAN uses linear activation instead of the Sigmoid activation in the last layer of $\mathcal{D}$, as shown in the right part of Fig. 2. Second, WGAN directly takes the difference between $\mathcal{D}$'s outputs on the "real" and "fake" inputs as the training loss without using a log function, as shown in Eq. (13). Third, each time $\mathcal{D}$ is updated, the absolute values of its parameters are clipped to not more than a threshold (δ), as shown in the sixth line of Algorithm 1. According to Arjovsky et al. [28], such modifications improve the convergence of adversarial learning.

5. Experiments

5.1. Experimental conditions

The Annotated Facial Landmark in the Wild (AFLW) database [30] and the CMU Multi-PIE Face (Multi-PIE) database [31] contain facial landmarks, corresponding visibility, head poses and gender information simultaneously. The IBUG database [32] contains facial landmark labels. We evaluate the proposed adversarial multi-task learning approach for facial landmark related multi-task face analyses on the AFLW and the Multi-PLE databases, and for landmark detection on the IBUG database. Furthermore, CelebA database and the LFWA database [19] are used to evaluate the proposed adversarial multi-task learning approach for multiple face attribute estimation.

The AFLW [30] database contains 25, 993 faces in 21, 997 real-world images with full pose, expression, ethnicity, age and gender variations. It provides annotations for 21 landmark points per face, along with the face bounding-box, face pose (i.e., roll, pitch and yaw) and gender. We follow the same strategy as Ranjan et al.'s [2] to divide training and testing sets, i.e., 1000 images for testing and the other for training. The testing set is divided into three subsets by their absolute yaw angles.

The Multi-PIE database contains 337 subjects, captured under 13 yaw angles and 19 illuminations in four recording sessions for a total of more than 750,000 images. Among them, 6152 images are labeled with landmarks, whose number varies from 39 to 68, depending on their visibility. Following the same sample selecting strategy as Wu et al.'s [5] work, we use the facial images from the first 150 subjects as training data and use the subjects with IDs between 151 and 200 as testing data.

The IBUG [32] database contains 135 images, each of them annotated with 68 facial landmarks. The landmark definition of the IBUG database is the same as the Multi-PIE database. All of the images in the IBUG database are used as testing samples.

The CelebA database is a large scale unconstrained face attribute database and contains more than 10, 000 identities, each of which has twenty images. There are more than 200, 000 images total. The LFWA database has 13, 233 images of 5749 identities. Each image in the CelebA database and LFWA database is annotated with forty face attributes. Both databases are challenging for attribute estimation, with large variations in expressions, poses, races, illumination, background, etc. Following Liu et al. [19], for the CelebA database, we use the images of first 8,000 identities as the training set and the images of the last 1,000 identities as the testing set; for the LFWA database, we randomly split images into half and half as the training and testing set.

The facial images are cropped from their bounding boxes and resized to $256 \times 256 \times 3$. In order to obtain enough data and improve their generalization performance, for the AFLW, the IBUG and the Multi-PIE databases, we augment the training data through random shifting bounding box, resizing bounding box and jittering the color of facial images. For the CelebA and LFWA databases, the images are processed through resizing and color jittering. For each image in the training set, we generate one augmented sample by a combination of the data augmentation operations. Therefore, the number of images obtained by data augmentations is as the same as the size of the training set.

On the AFLW and the IBUG databases, we adopt the normalized mean error (NME) as the accuracy metric. For the AFLW database, following Jourabloo et al. [33], face size is used to normalize the prediction error. For the IBUG database, the inter-ocular distance is used to normalize the prediction error. On the Multi-PIE database, the absolute pixel distance (APD) [34] is adopted for landmark detection. For pose estimation, the average degree error (ADE) is adopted on the AFLW database and the accuracy is adopted on the Multi-PIE database. For visibility and gender, the accuracy is adopted. For the multiple face attribute estimation, the accuracy is adopted.

To validate the effectiveness of the proposed adversarial multi-task network for landmark-related multi-task face analyses, six methods are compared: the method considering task dependencies from representation-level only (Ours_{no}), which employs the first term of Eq. (1); the method considering landmark dependencies (Ours_{l}), where only the landmark is fed into discriminator; the method considering joint distribution of landmark and visibility (Ours_{lv}), where the landmark and visibility are fed into discrimi-

nator; the method considering joint distribution of landmark, visibility, and gender (Ours_{lvg}), where the landmark, visibility and gender are fed into discriminator; the method considering joint distribution of landmark, visibility, and pose (Ours_{lvp}), where the landmark, visibility, and pose are fed into discriminator; and the proposed method considering joint distribution of landmark, visibility, gender and pose (Ours_{all}), where the landmark, visibility, gender and pose are all fed into discriminator. For multiple face attribute estimation on the CelebA and LFWA databases, we compare Ours_{no} and Ours_{gan}. Ours_{gan} considers the joint distribution of all facial attributes by feeding all of them into the discriminator.

We conduct within-database experiments for the AFLW, the Multi-PIE, the CelebA, and the LFWA databases. Since the IBUG database only contains 135 images, we do not conduct within-database experiment for it. To further validate the generalization performance of our method, we also conduct cross-database experiments. For landmark related face analyses, since the AFLW and the Multi-PIE databases have different landmark definitions, we could not conduct cross-database experiments between them. We just train our method on the Multi-PIE database and evaluate it on the IBUG database. For facial attribute estimation, we conduct cross-database experiments between the CelebA and the LFWA database by training our method on one database and evaluate it on the other database. For the experiment on each database, the training set is firstly splitted as 10 folds to conduct cross validation and select optimal values for all hyper-parameters. Then, these hyper-parameters are assigned with their optimal values and the method is re-trained on the whole training set. Since randomly sampling images and labels in the training process may cause randomness on the experimental results, we repeat each experiment ten times independently and reported their average results in our paper to offset the randomness.

5.2. Implementation details

For the multi-task recognizer $\mathcal{R}$, we follow Yin et al. [35] to use FAN network as the feature encoder, as shown in the left part of Fig. 2. FAN [13] is a kind of stacked hourglass network which captures both global and local patterns from a facial image by fusing features at different down-sampling and up-sampling stages with a residual mechanism. FAN further improves the original version of stacked hourglass network [12] by using hierarchical parallel and multi-scale convolutional bottleneck blocks. The learned representation from FAN is compressed by CNN networks as a feature vector, then converted to the predicted labels by fully connected (FC) networks. All convolution layers are followed by a Batch normalization layer [36] and ReLU activation unit. FC networks are used to convert the facial representations to the output labels for each task. For the regression tasks, i.e., landmark detection and pose regression, we adopt linear activation unit as the activation function for the last FC layer. For the gender classification and facial attribute estimation tasks, we adopt Sigmoid activation unit; For pose classification, we adopt Softmax activation.

As for the structure of discriminator $\mathcal{D}$, we adopt a FC network. For facial landmark related multi-task face analyses, the input dimensionality of $\mathcal{D}$ is the total size of labels from each individual task. On the AFLW database, the input dimensionality is 21+21+3+1=46. On the Multi-PIE database, the input dimensionality is 68+68+13+1=150. For multiple face attribute estimation, the input dimensionality of the discriminator is the number (40) of all facial attributes. The size of the hidden states for the FC network of D is half of its input size.

The recognizer $\mathcal{R}$ only outputs continuous values, whereas some of their ground truth labels are discrete. To prevent $\mathcal{D}$ from learning something trivial rather than the real label distributions, we add a fractional Gaussian perturbation onto these discrete labels to convert them as continuous values, then fed them into the discriminator. For the continuous labels, i.e., values of landmark coordinates and head poses, we normalize them into $[-1, 1]$, then fed them into the discriminator.

5.3. Experimental results and analyses of facial landmark related multi-task face analyses

Experimental results of landmark detection, visibility recognition and gender recognition on the AFLW, the Multi-PIE, and the IBUG databases are shown in Tables 1–3, respectively. On the AFLW and the Multi-PIE databases, we display the results of multi-task face analyses grouped by the absolute value of head's yaw angle. On the IBUG database, since it only provides landmark annotations, we just display the results of landmark detection.

From Tables 1–3, we observe the following:

First, the experimental results for near frontal faces are better than those for other poses for all methods. Specifically, the experimental results for [0, 30] yaw angle on the AFLW database and the experimental results for 0 angle on the Multi-PIE database are the best, with the lowest error and the highest accuracy in the most cases. It is reasonable since face analyses from facial images with extreme head pose are more changeling than those from near frontal views.

Second, the proposed method considering both shared representation and label-level connection significantly outperforms the proposed method only exploiting representation-level connection. Specifically, on the AFLW database, compared to the proposed method only exploiting representation-level constraint, Ours_{l} decreases the average NME(%) of the landmark detection by 0.18, and increases the accuracy of visibility and gender recognition by 2% and 2%, respectively. On the Multi-PIE database, Ours_{l} decreases the average APD of the landmark detection by 0.16, and increases the accuracy of visibility and gender recognition by 1% and 5%. Ours_{l} also decreases the standard deviation of performances among different intervals of yaw angle on both the AFLW and the Multi-PIE databases. On the IBUG database, Ours_{l} decreases the average NME of the landmark detection by 0.36. It is reasonable, since the method considering both shared representation and label-level constraint models the inherent dependencies among multiple face analysis tasks more faithfully and completely than the method only exploiting shared representation.

Third, more label-level constraints may achieve better performance. Specifically, Ours_{all} performs best among the six methods, with the lowest NME or APD for landmark detection, and the highest accuracy for visibility, gender recognition and pose estimation on both databases. It indicates that capturing more task relations from label-level can boost the performance of multiple tasks better.

Fourth, different label combinations lead to different effects. For instance, compared to Ours_{lvg}, Ours_{lvp} achieves lower error for landmark detection on both databases. The possible reason is that there exist more close relations between landmark and head poses than landmark and genders.

Fifth, the proposed multi-task adversarial learning method trained on the Multi-PIE database can improve the performance for landmark detection on the IBUG database. This result demonstrates the generalization performance of our method.

Visualization techniques are used to demonstrate the effectiveness of the proposed method (Ours_{all}). First, we visualize the features of CNN filters trained on the AFLW database in Fig. 3. Each sample in Fig. 3 is the channel-wise average of feature maps from the fifth convolutional layer in the FAN encoder. The value of each pixel in a feature map is normalized between 0 and 256 for visualization. From Fig. 3, we find that our method can effectively capture the spatial patterns of facial boundaries and encode them as

Table 1
Results for within-database experiments on the AFLW database grouped by the absolute value of yaw angle.

Methods	Tasks	Metrics	[0, 30]	[30, 60]	[60, 90]	mean	std
$Ours_{no}$	landmark	NME(↓)	2.63	2.74	3.02	2.80	0.201
	visibility	Acc(↑)	0.95	0.92	0.91	0.93	0.021
	gender	Acc(↑)	0.93	0.91	0.86	0.90	0.036
	R/P/Y	ADE(↓)	3.38/3.62/5.14	3.64/4.19/5.32	4.26/4.28/5.79	3.76/4.03/5.42	0.452/0.358/0.336
$Ours_{l}$	landmark	NME(↓)	2.48	2.61	2.76	2.62	0.140
	visibility	Acc(↑)	0.97	0.94	0.94	0.95	0.017
	gender	Acc(↑)	0.95	0.91	0.89	0.92	0.031
	R/P/Y	ADE(↓)	3.16/3.42/4.76	3.40/3.84/4.93	3.92/3.97/5.35	3.49/3.74/5.01	0.389/0.287/0.304
$Ours_{lv}$	landmark	NME(↓)	2.42	2.47	2.68	2.52	0.138
	visibility	Acc(↑)	0.97	0.95	0.94	0.95	0.015
	gender	Acc(↑)	0.97	0.96	0.94	0.96	0.015
	R/P/Y	ADE(↓)	3.09/3.28/4.72	3.31/3.43/4.88	3.68/3.77/5.27	3.36/3.49/4.96	0.298/0.251/0.283
$Ours_{lvg}$	landmark	NME(↓)	2.36	2.40	2.61	2.46	0.134
	visibility	Acc(↑)	0.98	0.98	0.96	0.97	0.012
	gender	Acc(↑)	0.97	0.96	0.95	0.96	0.010
	R/P/Y	ADE(↓)	3.01/3.04/4.64	3.22/3.17/4.86	3.50/3.52/5.18	3.24/3.24/4.89	0.246/0.248/0.271
$Ours_{lvp}$	landmark	NME(↓)	2.33	2.35	2.55	2.41	0.122
	visibility	Acc(↑)	0.98	0.98	0.97	0.98	0.006
	gender	Acc(↑)	0.97	0.95	0.95	0.96	0.012
	R/P/Y	ADE(↓)	2.97/2.92/4.58	3.16/3.00/4.78	3.42/3.29/5.11	3.18/3.07/4.82	0.226/0.195/0.268
$Ours_{all}$	landmark	NME(↓)	2.28	2.30	2.49	2.36	0.116
	visibility	Acc(↑)	0.98	0.98	0.98	0.98	0.000
	gender	Acc(↑)	0.99	0.99	0.98	0.99	0.006
	R/P/Y	ADE(↓)	2.92/2.83/4.49	3.04/2.95/4.61	3.25/3.06/4.94	3.07/2.95/4.68	0.167/0.115/0.233

Note: ↑ represents that the higher value indicates better performance and ↓ represents that the smaller value indicates better performance. R, P, Y are the abbreviations of roll, pitch and yaw, respectively.

Table 2
Results for within-database experiments on the Multi-PIE database grouped by the absolute value of yaw angle.

Methods	Tasks	Metrics	0	15	30	45	60	75	90	mean	std
$Ours_{no}$	landmark	APD(↓)	2.42	2.51	2.53	2.71	2.66	2.86	3.08	2.68	0.228
	visibility	Acc(↑)	0.99	0.98	0.97	0.94	0.92	0.92	0.83	0.94	0.054
	pose	Acc(↑)	0.97	0.98	0.98	0.96	0.93	0.95	0.93	0.96	0.021
	gender	Acc(↑)	0.95	0.96	0.92	0.83	0.82	0.79	0.73	0.86	0.088
$Ours_{l}$	landmark	APD(↓)	2.33	2.42	2.40	2.58	2.52	2.60	2.82	2.52	0.163
	visibility	Acc(↑)	0.99	1.00	0.97	0.96	0.97	0.91	0.88	0.95	0.044
	pose	Acc(↑)	0.98	0.98	0.97	0.97	0.94	0.95	0.93	0.96	0.020
	gender	Acc(↑)	0.97	0.96	0.94	0.91	0.88	0.86	0.83	0.91	0.053
$Ours_{lv}$	landmark	APD(↓)	2.23	2.26	2.34	2.38	2.46	2.56	2.64	2.41	0.152
	visibility	Acc(↑)	1.00	0.99	0.96	0.98	0.95	0.93	0.92	0.96	0.030
	pose	Acc(↑)	0.99	1.00	0.97	0.99	0.97	0.97	0.95	0.98	0.017
	gender	Acc(↑)	0.97	0.96	0.95	0.91	0.89	0.91	0.85	0.92	0.043
$Ours_{lvg}$	landmark	APD(↓)	2.18	2.24	2.29	2.31	2.36	2.50	2.57	2.35	0.140
	visibility	Acc(↑)	1.00	0.99	0.98	0.98	0.96	0.95	0.94	0.97	0.022
	pose	Acc(↑)	1.00	1.00	0.98	0.99	0.98	0.97	0.96	0.98	0.015
	gender	Acc(↑)	0.97	0.97	0.96	0.93	0.92	0.92	0.91	0.94	0.026
$Ours_{lvp}$	landmark	APD(↓)	2.13	2.19	2.26	2.30	2.33	2.48	2.46	2.31	0.130
	visibility	Acc(↑)	1.00	0.99	1.00	0.98	0.98	0.97	0.96	0.98	0.015
	pose	Acc(↑)	1.00	1.00	0.99	0.99	0.98	0.97	0.97	0.99	0.013
	gender	Acc(↑)	0.96	0.96	0.97	0.94	0.94	0.93	0.91	0.94	0.021
$Ours_{all}$	landmark	APD(↓)	2.09	2.10	2.14	2.24	2.28	2.34	2.41	2.23	0.124
	visibility	Acc(↑)	1.00	0.99	1.00	0.99	0.98	0.97	0.97	0.99	0.013
	pose	Acc(↑)	1.00	1.00	1.00	0.99	0.99	0.99	0.98	0.99	0.008
	gender	Acc(↑)	0.97	0.97	0.97	0.96	0.95	0.94	0.94	0.96	0.014

Note: ↑ represents that the higher value indicates better performance and ↓ represents that the smaller value indicates better performance.

Fig. 3. Visualization of the representations learned by the CNN filters on the AFLW database.

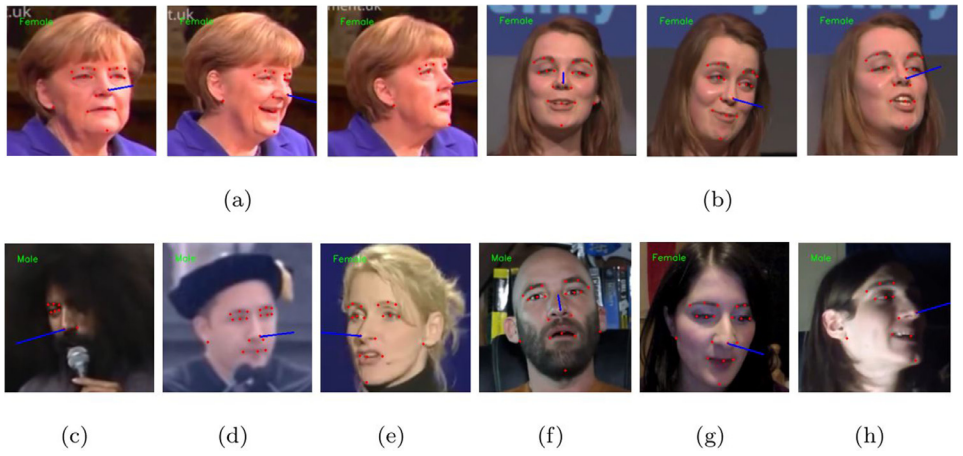

Fig. 4. Visualization for recognition results of landmark coordinates, landmark visibilities, poses, and genders on some new facial videos and images. For landmarks predicted as visible, their positions are visualized as red dots. For landmarks predicted as invisible, their positions are not depicted. The predicted poses are visualized as the blue normal lines of faces. The predicted genders are shown in the top left corners of these figures. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Experimental results for landmark detection on the IBUG database. The method is trained on the Multi-PIE database and evaluate on the IBUG database.

Methods	Ours_{no}	Ours_{l}	Ours_{lv}	Ours_{lvg}	Ours_{lvp}	Ours_{all}
NME	7.24	6.88	6.72	6.68	6.65	6.59

informative features. Second, to validate the generalization of our method, we adopt the proposed method trained on AFLW to predict landmark coordinates, landmark visibilities, poses, and genders on some new facial videos and images, and visualize the recognition results in Fig. 4. From Fig. 4, we could observe that the proposed method can well generalize to new face samples under different imaging conditions.

5.4. Experimental results and analyses of multiple face attribute estimation

The experimental results of face attribute estimation on the CelebA database and LFWA database are listed in Table 8. From Table 8, we have the following observations.

First, we find that compared to Ours_{no}, Ours_{gan} brings a consistent accuracy improvement for both within-database and cross-database experiments on the two databases. For the 40 face attributes, there exist complex correlations. For instance, a person who wears lipstick and necklace is less likely to be a male, while a person with mustache or goatee is more likely to be a male. On the other hand, some face attributes are mutually exclusive. For instance, at most one of black hair, blond hair, brown hair and gray hair appears. These correlations are crucial for improving the performance on multiple attribute estimation simultaneously. The proposed method Ours_{gan} considers the distribution among the predicted face attributes and the ground-truth face attributes. Through this way, the positive correlation and negative correlation among attributes can be exploited. The improvement demonstrates that the proposed method can successfully capture the label-level dependencies and results in better performance.

Second, we find that the results of cross-database experiments are lower than the results of within-database experiments on the two databases. This observation is also consistent with the experimental results of Han et al. [22]. According to Han et al. [22], the reason of the performance drop in cross-database experiments compared to within-database experiments may be that the attribute distributions and image styles are different between the LFWA and the CelebA databases, bringing some difficulties to adapting the learned model in one database to the other. Despite these difficulties, Ours_{gan}(cross) still outperforms Ours_{no}(cross) on prediction accuracy, demonstrating a better generalization performance of the proposed adversarial learning method.

Table 4
NME(%) performance of the proposed method and the related works for landmark detection on the AFLW database.

Methods	[0, 30]	[30, 60]	[60, 90]	mean	std
CDM [37]	8.15	13.02	16.17	12.44	4.04
RCPR [38]	5.43	6.58	11.53	7.85	3.24
ESR [11]	5.66	7.12	11.94	8.24	3.29
SDM [10]	4.75	5.55	9.34	6.55	2.45
3DDFA [39]	5.00	5.06	6.74	5.60	0.99
3DDFA+SDM [39]	4.75	4.83	6.38	5.32	0.92
Zhang et al. [17]	3.90	4.10	4.70	4.24	0.42
FAN [13]	3.45	3.06	4.24	3.58	0.60
SAN [14]	2.80	2.92	3.32	3.01	0.27
FHR [15]	2.75	2.96	3.18	2.96	0.22
$\text{HyperFace}_{\text{no}}$ [2]	3.93	4.14	4.71	4.26	0.40
$\text{HyperFace}_{\text{all}}$	3.19	3.28	3.49	3.32	0.15
$\text{HF} - \text{Resnet}_{\text{no}}$ [2]	2.71	2.88	3.19	2.93	0.24
$\text{HF} - \text{Resnet}_{\text{all}}$	2.34	2.45	2.61	2.47	0.14
Ours_{no}	2.63	2.74	3.02	2.80	0.20
Ours_{all}	**2.28**	**2.30**	**2.49**	**2.36**	**0.12**

To validate the effectiveness of the proposed method in capturing relationships among multiple face attributes, we graphically illustrate the captured dependencies in Fig. 5. The values are the output of the last layer of the proposed multi-task recognizer $\mathcal{R}$. A larger output value indicates a high confidence of the occurrence for the attribute, and a smaller output value indicates a high confidence of the absence for the attribute. The first figure shows that the sample, which encodes a pattern for a person who is likely to be with sideburns, 5 o'clock shadow and without wearing earrings. This combination is more likely to represent the attribute relationships for a male. The second figure shows that the sample, which encodes a pattern for a person who is likely to wear lipstick and earrings and with no beard. This combination is more likely to represent the attribute relationships for a heavy makeup female. The two figures show that the proposed method is able to effectively capture the relationships among multiple face attributes.

5.5. Significance test for the proposes method

Significance test is conducted to demonstrate the statistical significance of performance boosts brought by the proposed method.

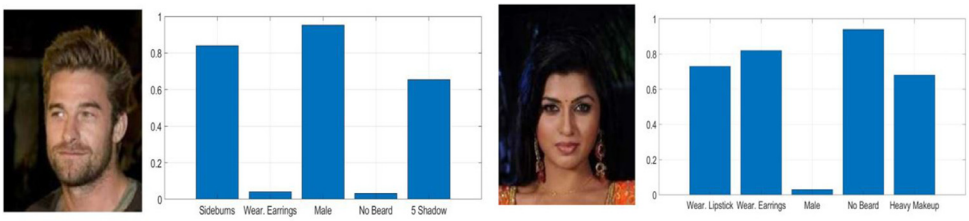

Fig. 5. Example showing: some face attribute combinations are frequently observed. Each bar shows the output value from the Sigmoid unit of the recognizer.

Table 5
APD performance of our method and the related works for landmark detection on the Multi-PIE database.

Methods	Wu et al. [5]	FAN	SAN	FHR	$\text{HyperFace}_{\text{no}}$	$\text{HyperFace}_{\text{all}}$	$\text{HF}-\text{Resnet}_{\text{no}}$	$\text{HF}-\text{Resnet}_{\text{all}}$	Ours_{no}	Ours_{all}
APD	3.62	2.93	2.84	2.59	3.12	2.85	2.81	2.44	2.68	**2.23**

Table 6
ADE performance of the proposed method and the related works for pose estimation on the AFLW database. R, P, Y are the abbreviations of roll, pitch and yaw, respectively.

Methods	$\text{HyperFace}_{\text{no}}$	$\text{HyperFace}_{\text{all}}$	$\text{HF}-\text{Resnet}_{\text{no}}$	$\text{HF}-\text{Resnet}_{\text{all}}$	Ours_{no}	Ours_{all}
R	3.92	3.54	3.29	3.18	3.76	**3.07**
P	6.13	3.08	5.33	3.02	4.03	**2.95**
Y	7.62	4.81	6.24	4.76	5.42	**4.68**

Table 7
Pose estimation accuracy of the proposed method and the related works on the Multi-PIE database.

Methods	PCR [40]	Linear PLS [40]	KPLS [40]	Wu et al. [5]	FPLL [4]	Ours_{all}
Accuracy	0.48	0.57	0.79	0.77	0.91	**0.99**

Specifically, on the AFLW, the Multi-PIE, and the IBUG databases, we conduct t-test for the performance difference between Ours_{all} and Ours_{no}. On the CelebA and LFWA databases, t-test for the performance difference between Ours_{gan}(within) and Ours_{no}(within) is conducted, while on the AFLW database, P-values for landmark localization, visibility detection, gender classification, roll regression, pitch regression and yaw regression are 2.30e-4, 2.62e-6, 1.40e-12, 1.56e-8, 3.43e-6, 2.90e-3, respectively. On the Multi-PIE database, P-values for landmark localization, visibility detection, gender classification and pose classification are 1.60e-4, 2.15e-7, 3.54e-5 and 2.42e-6, respectively. On the IBUG database, P-value for landmark localization is 1.83e-5. On the CelebA and LFWA databases, P-values for facial attribute estimations are 1.62e-7 and 3.43e-5, respectively. All of these p-values are much less than 0.05, which demonstrates the significance of the proposed adversarial learning method.

5.6. Comparison with related works on accuracy of facial landmark related multi-task face analyses

As mentioned in the introduction, several works handle landmark-related multi-task face analysis tasks jointly. Among them, Ranjan et al.'s work [2] (HyperFace and HF-Resnet) and Zhang et al.'s work [17] provided landmark detection experimental results on the AFLW database. Thus, we compare our work on landmark detection with these methods. We also train HyperFace and HF-Resnet with the proposed adversarial learning framework and compare our method with them. To be consistent with the denotation of our method, the original HyperFace and HF-Resnet are denoted as $\text{HyperFace}_{\text{no}}$ and $\text{HF}-\text{Resnet}_{\text{no}}$, respectively, since they were not trained in an adversarial way; while the networks trained by our adversarial learning framework are denoted as $\text{HyperFace}_{\text{all}}$ and $\text{HF}-\text{Resnet}_{\text{all}}$, respectively. We compare Ours_{no} with $\text{HyperFace}_{\text{no}}$ and $\text{HF}-\text{Resnet}_{\text{no}}$; and compare Ours_{all} with $\text{HyperFace}_{\text{all}}$ and $\text{HF}-\text{Resnet}_{\text{all}}$. Furthermore, Ranjan et al. [2] provided landmark detection results of some facial landmark detection methods, i.e., CDM [37], RCPR [38], ESR [11], SDM [10], 3DDFA [39] and 3DDFA+SDM [39]. We also compare our method with these methods, although they are not trained under a multi-task learning framework. As state-of-the-art landmark detection methods, FAN [13], SAN [14], and FHR [15] were conducted under different experimental conditions with ours, we just re-implement them and re-conduct experiments with our training set. Their open source codes[1] are used to facilitate the re-implementation.

On the Multi-PIE database, we compare our method with Wu et al.'s work [5]. For landmark detection, since Wu et al. [5] only provided the detection performance on the inner landmarks instead of all landmarks, we just re-implement their method and compare with them on all 68 landmarks. We also implement $\text{HyperFace}_{\text{no}}$, $\text{HF}-\text{Resnet}_{\text{no}}$, $\text{HyperFace}_{\text{all}}$, $\text{HF}-\text{Resnet}_{\text{all}}$, FAN, SAN, and FHR on the Multi-PIE database and compare their results with ours.

The comparison of the proposed method to the related works on landmark detection and pose estimation accuracy is shown in Tables 4–7. From this table, we have three observations. First, we find $\text{HyperFace}_{\text{all}}$, $\text{HF}-\text{Resnet}_{\text{all}}$ and Ours_{all} outperform $\text{HyperFace}_{\text{no}}$, $\text{HF}-\text{Resnet}_{\text{no}}$ and Ours_{no} respectively, which means adversarial learning can achieve a better performance compared to supervised regression for all these network structures. This may be because adversarial learning well captures spatial patterns from target label-level, such that the recognizer can make prediction based on dependencies among all labels. Second, we

[1] https://github.com/1adrianb/face-alignment, https://github.com/D-X-Y/SAN, https://github.com/tyshiwo/FHR_alignment.

Table 8
Comparison of Attribution Estimation on the CelebA and LFWA databases. For our methods, $Ours_{no}$ (within) and $Ours_{gan}$ (within) are the results of within-database experiments, while $Ours_{no}$ (cross) and $Ours_{gan}$ (cross) are the results of cross-database experiments.

Database		5 Shadow	Arch. Eyebrows	Attractive	Bags Un. Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	H. Checkbones	Male
CelebA	PANDA [3]	88	78	81	79	96	92	67	75	85	93	86	77	86	86	88	98	93	94	90	86	97
	LNets+ANet [19]	91	79	81	79	98	95	68	78	88	95	84	80	90	91	92	99	95	97	90	87	98
	MCNN-AUX [23]	95	83	83	85	**99**	96	71	85	90	96	96	89	93	96	96	90	**100**	97	**98**	**99**	92
	PS-MCNN-LC [24]	**97**	86	84	87	**99**	98	73	86	**92**	**98**	98	91	**95**	**98**	**98**	**100**	98	**99**	93	90	**99**
	DMM-CNN [20]	95	85	83	86	**99**	96	73	85	91	96	96	89	93	96	96	**100**	98	98	92	88	**99**
	DMTL (cross) [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	$Ours_{no}$ (within)	96	85	85	87	**99**	96	89	89	93	97	98	92	**95**	96	97	**100**	98	98	93	91	98
	$Ours_{gan}$ (within)	**97**	**87**	**88**	**92**	**99**	**99**	**92**	**93**	**92**	97	**99**	**97**	**95**	**98**	**98**	**100**	**100**	**99**	97	98	**99**
	$Ours_{no}$ (cross)	73	69	69	66	74	77	66	72	74	73	79	67	66	62	65	72	68	78	74	72	76
	$Ours_{gan}$ (cross)	69	69	68	67	74	78	72	72	75	78	77	70	68	68	68	73	71	76	73	76	76
LFWA	PANDA [3]	84	79	81	80	84	84	73	79	87	94	74	74	79	69	75	89	75	81	93	86	92
	LNets+ANet [19]	84	82	83	83	88	88	75	81	90	97	74	77	82	73	78	**95**	78	84	95	88	94
	MCNN-AUX [23]	77	82	80	83	92	90	79	85	93	97	85	81	85	77	82	**95**	91	83	89	90	**96**
	PS-MCNN-LC [24]	78	84	82	**87**	93	91	83	86	93	**99**	87	82	86	78	**87**	93	84	91	97	89	95
	DMM-CNN [20]	79	83	81	83	92	91	80	84	92	97	88	82	85	78	81	93	83	89	96	88	94
	DMTL (cross) [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	$Ours_{no}$ (within)	89	88	85	84	97	**94**	85	92	94	**99**	**98**	78	82	78	81	93	87	92	**98**	90	92
	$Ours_{gan}$ (within)	**91**	**89**	**85**	84	**98**	**94**	**91**	**93**	**95**	**99**	**98**	**84**	**88**	**87**	86	92	**93**	**94**	97	**91**	93
	$Ours_{no}$ (cross)	75	70	66	64	77	74	69	69	70	76	77	70	76	73	73	75	80	77	75	76	78
	$Ours_{gan}$ (cross)	74	71	74	73	79	77	72	71	78	76	79	75	74	75	79	72	79	81	82	80	79
		Mouth S. O.	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Reced. Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wear. Earrings	Wear. Hat	Wear. Lipstick	Wear. Necklace	Wear. Neck-tie	Young		**Average**
CelebA	PANDA [3]	93	93	84	93	65	91	71	85	87	93	92	69	77	78	96	93	67	91	84		85
	LNets+ANet [19]	92	95	81	95	66	91	72	89	90	96	92	73	80	82	**99**	93	71	93	87		87
	MCNN-AUX [23]	88	94	**98**	94	97	87	**87**	**97**	96	76	**97**	77	**94**	**95**	98	93	84	84	88		91
	PS-MCNN-LC [24]	96	**99**	89	**98**	77	**99**	79	96	**97**	98	95	86	86	93	**99**	**96**	89	**99**	**91**		93
	DMM-CNN [20]	94	97	88	96	76	97	77	94	95	98	93	85	86	91	99	94	88	97	89		92
	DMTL (cross) [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		70
	$Ours_{no}$ (within)	96	97	97	95	94	**99**	82	96	96	98	94	86	88	91	**99**	94	94	98	87		94
	$Ours_{gan}$ (within)	**97**	**99**	**98**	95	**98**	**99**	85	96	**97**	**99**	**97**	**91**	91	92	**99**	**96**	**97**	98	**91**		**96**
	$Ours_{no}$ (cross)	69	80	73	69	65	70	65	65	74	64	75	72	71	72	78	76	76	71	70		71
	$Ours_{gan}$ (cross)	69	76	74	74	71	74	70	67	77	71	74	65	72	74	77	77	71	73	69		72
LFWA	PANDA [3]	78	87	73	75	72	84	76	84	73	76	89	73	75	92	82	93	86	79	82		81
	LNets+ANet [19]	82	92	81	79	74	84	80	85	78	77	91	76	76	94	88	95	88	79	86		84
	MCNN-AUX [23]	**88**	95	**94**	84	**93**	83	**90**	81	82	77	**93**	84	**86**	88	83	92	79	82	86		86
	PS-MCNN-LC [24]	85	94	84	82	78	**95**	88	88	89	84	**93**	80	83	**96**	91	96	91	82	87		88
	DMM-CNN [20]	84	94	84	82	77	92	85	86	86	83	92	79	80	94	91	95	89	81	89		87
	DMTL (cross) [22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		73
	$Ours_{no}$ (within)	86	96	88	84	85	84	87	88	96	85	90	**86**	82	87	95	97	**94**	84	93		89
	$Ours_{gan}$ (within)	87	**98**	92	**86**	89	84	88	**89**	**97**	**88**	**93**	84	**86**	88	**95**	**98**	**94**	**86**	**95**		**91**
	$Ours_{no}$ (cross)	79	76	80	76	77	83	68	76	73	82	73	74	73	73	78	79	74	77	71		75
	$Ours_{gan}$ (cross)	81	80	82	79	78	79	70	77	79	77	74	73	74	77	80	80	76	81	75		77

Table 9
Comparison on parameter numbers among $\text{HyperFace}_{\text{all}}$, $\text{HF} - \text{Resnet}_{\text{all}}$ and Ours_{all}.

Databases	$\text{HyperFace}_{\text{all}}$	$\text{HF} - \text{Resnet}_{\text{all}}$	Ours_{all}
AFLW, Multi-PIE, LFWA, CelebA	35M	97M	**30M**

find that Ours_{no} outperforms $\text{HyperFace}_{\text{no}}$ and $\text{HF} - \text{Resnet}_{\text{no}}$, and meanwhile, Ours_{all} outperforms $\text{HyperFace}_{\text{all}}$ and $\text{HF} - \text{Resnet}_{\text{all}}$. This result demonstrates the proposed recognizer network is more adept at the facial analysis tasks compared to the HyperFace and the HF – Resnet networks. The reason may be that, the adopted Hourglass network is better at spatial modeling by its mechanism to mix global and local representations. Third, the proposed method, i.e., Ours_{all}, performs the best on the AFLW database and the Multi-PIE database for landmark detection and pose estimation. The proposed method simultaneously exploits the shared representation and label-level constraint through multi-task network and adversarial mechanism, and thus achieves the best performance.

5.7. Comparison with related works on accuracy of multiple face attribute estimation

We compare the proposed method with other multiple face attribute estimation works in Table 8. For within-database experiments, we compare our method with PANDA [3], LNets+ANet [19], MCNN-AUX [23], PS-MCNN-LC [24] and DMM-CNN [20]. For cross-database experiments, we compare our method with DMTL (cross) [22]. Compared to PANDA, LNets+ANet and DMM-CNN, which also estimated multiple attributes through deep convolutional neural network, the proposed method performs better. For the three methods, multiple face attributes share common representation, and thus the dependencies among attributes can be exploited in a certain. However, the dependencies on label-level are not considered. Compared to these works, the proposed method considers the representation-level dependencies and the label-level dependencies jointly, and thus achieves better experimental results. The proposed method also outperforms MCNN-AUX, PS-MCNN-LC and DMTL (cross), which exploited both the representation-level and label-level dependencies. MCNN-AUX proposed an auxiliary network to obtain relationships among multiple face attributes. PS-MCNN-LC and DMTL (cross) grouped these multiple face attributes according to prior knowledge. Although these works exploited dependencies among multiple face attributes from both representation-level and label-level, the captured label-level relationships are either fixed groups or fixed form. Through multi-task adversarial network, the proposed method can capture the complex and global relationship among multiple face attributes. On both databases, the proposed method achieves the best performance. It further suggests that the proposed method has strong ability for multi-task analyses.

5.8. Comparison with related works on model size

Beside comparison on the accuracy of multi-task face analyses, we also compare our method (Ours_{all}) with $\text{HyperFace}_{\text{all}}$ and $\text{HF} - \text{Resnet}_{\text{all}}$ on model size. The numbers of parameters for the three methods are shown in Table 9. Sizes of the same model on different databases are almost the same, since the variance on the number of output labels has little effect on the total number of parameters. From this table, we find that our method is more lightweight than the two compared methods. This is because the size of fully connected networks in our method is smaller. With a small model size, our method can well fit to the application with limited machine memory.

6. Conclusion

In this paper, we propose a novel multiple facial analysis method through exploiting both representation-level and label-level dependencies. Specifically, we first utilize deep multi-task network as a recognizer $\mathcal{R}$ to capture representation-level dependencies. And then, we introduce a discriminator $\mathcal{D}$ to distinguish the label combinations from the ground-truth. Through optimizing the two networks in an adversarial manner, the proposed method manages to make predicted label combination closer to the distribution of the ground-truth. Experimental results demonstrate that the proposed method successfully captures both the shared representation-level and label-level constraint and thus outperforms related works. The current work models the spatial dependencies among multiple facial analysis tasks on a static image, while temporal dependencies existed in a facial video stream could also be exploited. We plan to incorporate sequential modeling techniques with the proposed adversarial learning framework to capture both spatial and temporal patterns from a facial image sequence.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by the National Science Foundation of China (Grant No. 91748129), and the major project from Anhui Science and Technology Agency (1804a09020038).

References

[1] Z. Zhang, P. Luo, C.C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: ECCV, 2014, pp. 94–108.
[2] R. Ranjan, V.M. Patel, R. Chellappa, HyperFace: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, TPAMI 41 (1) (2019) 121–135.
[3] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, L. Bourdev, PANDA: pose aligned networks for deep attribute modeling, in: CVPR, 2014, pp. 1637–1644.
[4] D. Ramanan, X. Zhu, Face detection, pose estimation, and landmark localization in the wild, in: CVPR, 2012, pp. 2879–2886.
[5] Y. Wu, C. Gou, Q. Ji, Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion, in: CVPR, 2017, pp. 3471–3480.
[6] S. Honari, P. Molchanov, S. Tyree, P. Vincent, C. Pal, J. Kautz, Improving landmark localization with semi-supervised learning, in: CVPR, 2018, pp. 1546–1555.
[7] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, TPAMI 23 (6) (2001) 681–685.
[8] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[9] X. Jin, X. Tan, Face alignment by robust discriminative hough voting, Pattern Recognit. 60 (2016) 318–333.
[10] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: CVPR, 2013, pp. 532–539.
[11] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, IJCV 107 (2) (2014) 177–190.
[12] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: ECCV, 2016, pp. 483–499.
[13] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks), in: ICCV, 2017, pp. 1021–1030.
[14] X. Dong, Y. Yan, W. Ouyang, Y. Yang, Style aggregated network for facial landmark detection, in: CVPR, 2018, pp. 379–388.
[15] Y. Tai, Y. Liang, X. Liu, L. Duan, J. Li, C. Wang, F. Huang, Y. Chen, Towards highly accurate and stable face alignment for high-resolution videos, in: AAAI, 2019, pp. 8893–8900.
[16] P. Chandran, D. Bradley, M.H. Gross, T. Beeler, Attention-driven cropping for very high resolution facial landmark detection, in: CVPR, 2020, pp. 5860–5869.
[17] G. Zhang, H. Han, S. Shan, X. Song, X. Chen, Face alignment across large pose via MT-CNN based 3D shape reconstruction, in: FG, 2018, pp. 210–217.
[18] N. Kumar, P. Belhumeur, S. Nayar, FaceTracer: a search engine for large collections of images with faces, in: ECCV, 2008, pp. 340–353.

[19] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: ICCV, 2015, pp. 3730–3738.
[20] L. Mao, Y. Yan, J. Xue, H. Wang, Deep multi-task multi-label CNN for effective facial attribute classification, IEEE Trans. Affect. Comput. (2018), doi:10.1109/TAFFC.2020.2969189.
[21] Y. Zhong, J. Sullivan, H. Li, Face attribute prediction using off-the-shelf CNN features, in: International Conference on Biometrics (ICB), 2016, pp. 1–7.
[22] H. Han, A.K. Jain, F. Wang, S. Shan, X. Chen, Heterogeneous face attribute estimation: a deep multi-task learning approach, TPAMI 40 (11) (2018) 2597–2609.
[23] E.M. Hand, R. Chellappa, Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification, in: AAAI, 2017, pp. 4068–4074.
[24] J. Cao, Y. Li, Z. Zhang, Partially shared multi-task convolutional neural network with local constraint for face attribute learning, in: CVPR, 2018, pp. 4290–4299.
[25] Y. Bai, Y. Zhang, M. Ding, B. Ghanem, SOD-MTGAN: small object detection via multi-task generative adversarial network, in: ECCV, 2018, pp. 206–221.
[26] P. Liu, X. Qiu, X. Huang, Adversarial multi-task learning for text classification, in: ACL, 2017, pp. 1–10.
[27] Y. Liu, Z. Wang, H. Jin, I. Wassell, Multi-task adversarial network for disentangled feature learning, in: CVPR, 2018, pp. 3743–3751.
[28] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks, in: ICML, 2017, pp. 214–223.
[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014, pp. 2672–2680.
[30] M. Koestinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: ICCV workshops, 2011, pp. 2144–2151.
[31] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (5) (2010) 807–813.
[32] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: the first facial landmark localization challenge, in: ICCV Workshop, 2013, pp. 397–403.
[33] A. Jourabloo, X. Liu, Pose-invariant 3D face alignment, in: ICCV, 2015, pp. 3694–3702.
[34] Z. Lei, Q. Bai, R. He, S.Z. Li, Face shape recovery from a single image using CCA mapping between tensor spaces, in: CVPR, 2008, pp. 1–7.
[35] S. Yin, S. Wang, G. Peng, X. Chen, B. Pan, Capturing spatial and temporal patterns for facial landmark tracking through adversarial learning, in: IJCAI, 2019, pp. 1010–1017.
[36] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: ICML, 2015, pp. 448–456.
[37] X. Yu, J. Huang, S. Zhang, W. Yan, D.N. Metaxas, Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model, in: ICCV, 2013, pp. 1944–1951.
[38] X.P. Burgos-Artizzu, P. Perona, P. Dollár, Robust face landmark estimation under occlusion, in: ICCV, 2013, pp. 1513–1520.
[39] X. Zhu, Z. Lei, X. Liu, H. Shi, S.Z. Li, Face alignment across large poses: a 3D solution, in: CVPR, 2016, pp. 146–155.
[40] M. Al Haj, J. Gonzalez, L.S. Davis, On partial least squares in head pose estimation: how to simultaneously deal with misalignment, in: CVPR, 2012, pp. 2602–2609.

Shangfei Wang (SM'15) received the B.S. degree in electronic engineering from Anhui University, Hefei, Anhui, China, in 1996, and the M.S. degree in circuits and systems and the Ph.D. degree in signal and information processing from the University of Science and Technology of China (USTC), Hefei, in 1999 and 2002, respectively. From 2004 to 2005, she was a Post-Doctoral Research Fellow with Kyushu University, Japan. From 2011 to 2012, she was a Visiting Scholar with the Rensselaer Polytechnic Institute, Troy, NY, USA. She is currently a Professor with the School of Computer Science and Technology and the School of Data Science, USTC. She has authored or co-authored over 90 publications. Her research interests cover affective computing and probabilistic graphical models. She is a member of the ACM.

Shi Yin received the B.S. degree in automation from Central South University in 2016. He is currently pursuing the Ph.D degree in computer science with the University of Science and Technology of China, Hefei, China. His research interest is in affective computing.

Longfei Hao received the B.S. degree in computer science from Anhui University in 2016. He is currently pursuing the M.S. degree in computer science with the University of Science and Technology of China, Hefei, China. His research interest is in affective computing.

Guang Liang received the B.S. degree in computer science from Shandong University, Jinan,China in 2015. He is now studying for a master's degree in cyberspace security at the University of Science and Technology of China, Hefei,China. His current research direction is emotional theory analysis.