Neural Processing Letters (2020) 51:2105–2121
https://doi.org/10.1007/s11063-019-09987-7

Inferring Personality Traits from Attentive Regions of User Liked Images Via Weakly Supervised Dual Convolutional Network

Hancheng Zhu[1] · Leida Li[1] · Hongyan Jiang[2] · Allen Tan[3]

Published online: 22 January 2019

Abstract

In social media, users usually unconsciously their preferences on images, which can be considered as the personal cues for inferring their personality traits. Existing methods map the holistic image features into personality traits. However, users' attention on their liked images is typically localized, which should be taken into account in modeling personality traits. In this paper, we propose an end-to-end weakly supervised dual convolutional network (WSDCN) for personality prediction, which consists of a classification network and a regression network. The classification network captures personality class-specific attentive image regions while only requiring the image-level personality class labels. The regression network is used for predicting personality traits. Firstly, the users' Big-Five (BF) traits are converted into ten personality class labels for their liked images. Secondly, the Multi-Personality Class Activation Map (MPCAM) is generated based on the classification network and utilized as the localized activation to produce local deep features, which are then combined with the holistic deep features for the regression task. Finally, the user liked images and the associated personality traits are used to train the end-to-end WSDCN model. The proposed method is able to predict the BF personality traits simultaneously by training the WSDCN network only once. Experimental results on the annotated PsychoFlickr database show that the proposed method is superior to the state-of-the-art approaches.

Keywords Attentive image regions · Multi-personality class activation map · Personality prediction · Weakly supervised dual convolutional network

1 Introduction

Personality trait is a psychological state capturing stable individual characteristics, which can be explained and predicted by observable behavioral differences [1]. The automatic

✉ Leida Li
lileida@cumt.edu.cn

[1] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

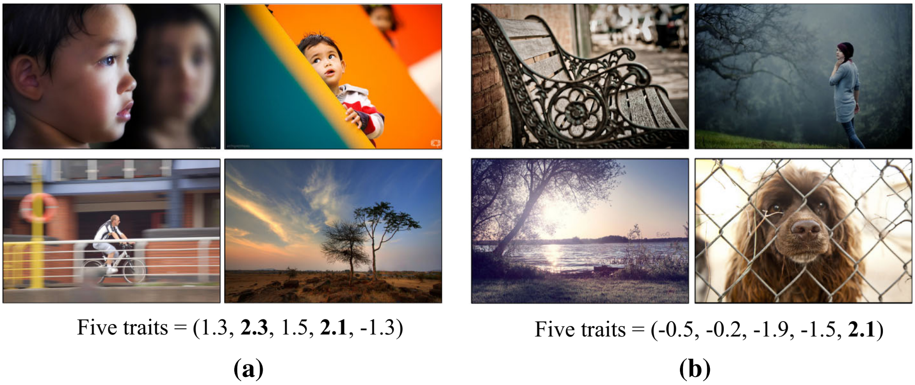
[2] School of Management, China University of Mining and Technology, Xuzhou 221116, China

[3] Tencent Media Lab, Tencent, Shenzhen 518000, China

computing of personality traits has many applications, such as job performance and sales ability evaluation [2], personalized recommendation [3], mental health evaluation [4], etc. With the prevalence of social media, multimedia has become increasingly popular on online photo-sharing platforms [5,6]. Multimedia data can be easily uploaded by users to the social networks such as Facebook, Flickr, and WeChat, and these data have various modalities, which contain rich semantic information [7–9]. Among these modalities, image is the most common and can be perceived and understood at a high level through the affective semantics [10]. There are two basic challenging tasks for visual sentiment analysis: image-centric prediction and user-centric prediction [11–13]. While image-centric prediction aims at image affective classification according to users' responses [14–17], user-centric prediction is to infer users' personalized emotions from some specific images [18–20]. With the demand for personalized customization and assessment, the analysis of users' personality traits is becoming increasingly important.

Studies on personality psychology show that stable individual characteristics result in stable behavior habits that people tend to express. That is, users externalize their personality traits through any behavior that can be deemed to personality-related cues [21]. Therefore, users' personality traits can be predicted through these personality-related cues [22]. Though there are various theories for personality analysis, the most widely used model is known as Big-Five (BF) or Five-Factor Model (FFM) [23]. The five personality traits are Openness (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), and Neuroticism (N). Assessing the personality of a user means to calculate the five scores corresponding to the traits above. Even if there are several kinds of questionnaires designed for such a task, the BFI-10 [24] is one of the most popular questionnaires. This is because that the BFI-10 can be filled in less than one minute while still providing reliable results. However, the questionnaires cannot meet real-time requirement. In addition, users often avoid the truth when the questionnaires have negative consequences for themselves. Hence, building computational models for personality prediction is highly desired. There are two fundamental problems in personality computing, namely Automatic Personality Recognition (APR) and Automatic Personality Perception (APP) [21]. The APR and APP indicate the personality prediction of self-assessed and attributed traits, respectively. The self-assessment traits, which can be obtained from asking users to answer the questionnaires for themselves, may not be objective. By contrast, the attributed traits are determined from the questionnaires that are filled by others and the results are usually more objective. For instance, when users are asked the question "I tend to be lazy" contained in the BFI-10 questionnaire [24], they often choose "disagree" to show that they are diligent person. It has been demonstrated that others' impressions are as important as the actual personality in the social identity of a person [25]. In this paper, we focus on modeling user's attributed personality traits.

Most of the works for modeling attributed personality traits take the approach of extracting two main personality-relates cues: nonverbal behavioral cues and social media cues [21]. The former approach means that users' attributed personality traits can be inferred from automatically detected nonverbal behavioral cues, such as interpersonal distances [26] and body movements [27]. The latter approach indicates that we can infer users' personality traits through the uploaded and tagged images or videos in social media [18–20,28,29]. Recently, social media has become one of the most popular channels for people to communicate with each other. There is a solid basis for research on personality computing in social media. Personality traits have been considered as the social signals sent with the social media cues, and personal inference is one of the meaningful challenges for social interaction among people. Taking advantage of the liking mechanisms, users can express their preferences on the online images, which can be used as their personality-relate cues. Thus, users convey


Five traits = (1.3, **2.3**, 1.5, **2.1**, -1.3)

(a)

Five traits = (-0.5, -0.2, -1.9, -1.5, **2.1**)

(b)

Fig. 1 Example images liked by two users from the PsychoFlickr database [18]: **a** a user with high conscientiousness and high agreeableness; **b** a user with high neuroticism. The order of five traits is (O, C, E, A, N) and the personality scores range from −4 to 4

their innate preferences through the personality-relate cues, which makes it possible to predict users' personality traits from the liked images in social media [30].

In the literature, a few works have been done to address the personality prediction based on the liked images. As far as we known, the first attempt to predict users' personality traits from their liked images was presented in [18], where PsychoFlickr dataset was proposed to investigate the relationship between the low-level image features (color, composition, texture, etc.) and personality traits. The dataset contained 60,000 liked images of 300 Flickr users (200 images per user). Furthermore, 12 independent assessors were hired to fill the BFI-10 [24] questionnaire for collecting the attributed personality traits of the 300 users. The questionnaire contained 10 items, every two of which were related to a trait, and each item had five options from −2 ("Strongly disagree") to 2 ("Strongly agree"). Therefore, 12 assessments available for each user were averaged to obtain the attributed traits, which ranged from −4 to 4. For example, Fig. 1 shows example images liked by two users as well as the corresponding personality traits from the PsychoFlickr database. Images liked by the user with high conscientiousness (2.3) and high agreeableness (2.1) are shown in Fig. 1a. Figure 1b shows the liked images of the user with high neuroticism (2.1). Based on [18], Segalin et al. [19] further investigated different methodologies based on LASSO Regression [31] and presented a more extensive correlational analysis of features adopted. In [20], Guntuku found that users' personality traits were profoundly influenced by high-level image features (objects, scenes, people, etc.). The experiments on PsychoFlickr dataset had demonstrated that modeling users' personality traits from high-level image features outperformed methods employing low-level image features. Recently, Convolutional Neural Networks (CNNs) are becoming more and more popular in personality analysis [29]. CNN is a more efficient way for affective feature extraction. In [29], Segalin et al. adopted CNNs pre-trained for image classification and fine-tuned them to divide users liked images into two classes of each personality trait. Consequently, five binary classification models were trained for the liked images classification instead of users' personality prediction.

The aforementioned personality prediction methods have achieved notable success in inferring personality traits based on the liked images. However, several issues remain in addressing such a challenging task. First, most literatures on personality prediction map hand-crafted features (e.g. color, texture, composition, objects, and scenes) [18–20] into users' personality traits. Nevertheless, hand-crafted features may not be sufficient for representing

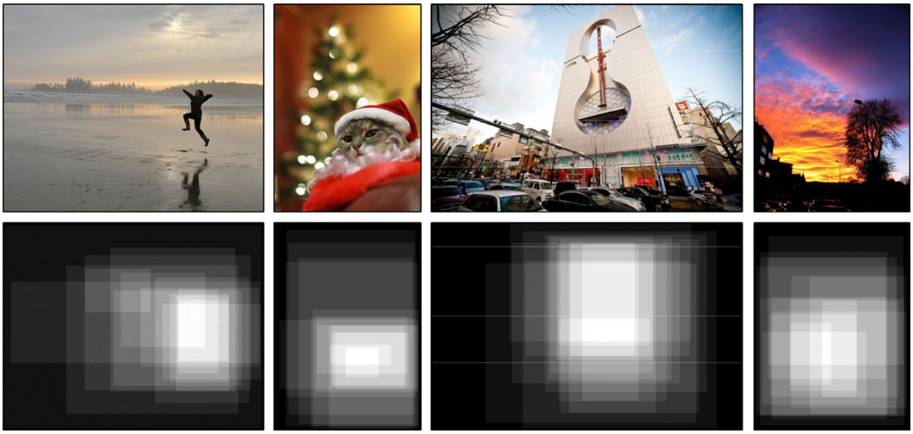


Fig. 2 Example images from the EmotionROI database [32]. The top row are four example images, and the bottom row are corresponding attentive regions annotated by 15 users

the affective aspects of images, which are typically more abstract. Therefore, it is necessary to employ a deep learning framework, which can extract more high-level semantic features for personality prediction. Second, the existing methods leverage the holistic image features for personality prediction, which has not dug deeply on users' affective attention on images. Third, the previous works can only predict a single personality trait by a trained model. This means that at least five networks need to be trained separately for predicting the five personality traits. Hence, it would be better a single trained network can predict the five personality traits simultaneously in an end-to-end manner.

Local regional information has been shown useful for capturing users' emotional attention in image affective analysis [32–34]. In [32], Peng et al. have shown that different image regions contribute differently to the users' evoked sentiment. The EmotionROI database, which collected users' attentive regions on images, was proposed in this work. In [33], You et al. found that people tended to pay attention to the local visual regions instead of the entire image because of their preferences. Yang et al. [34] proposed a method for image affective classification using localized sentiment information, which was proven to be effective in detecting emotion-related regions. Figure 2 shows some images and the corresponding attentive regions annotated by 15 users from the EmotionROI database [32]. As can be seen, users' affective attention on images is typically determined by local regions. Therefore, the attentive image regions, which are effective in image-centric affective prediction, can be taken into account for user-centric personality prediction. With the success of deep learning [35] and transfer learning [36] on image classification, a weakly supervised CNN was developed to discriminate the localization of objects [37]. Based on [37], several weakly supervised CNNs were further proposed to learn multiple local regions specific to different class modalities [38,39]. The weakly supervised CNNs for object detection aims to find salient map of objects in images supervised by image-level labels instead of localization-level labels. Therefore, the image-level personality class labels can be used to extract the attentive image regions of different personality classes by the weakly supervised CNNs. Owing to the diversity of user's personality traits, we believe that the personality distribution labels may be more suitable for generating attentive image regions of multiple personality classes.

Based on the finding that users' affective attention on images is typically localized, we propose an end-to-end personality prediction method based on weakly supervised dual con-

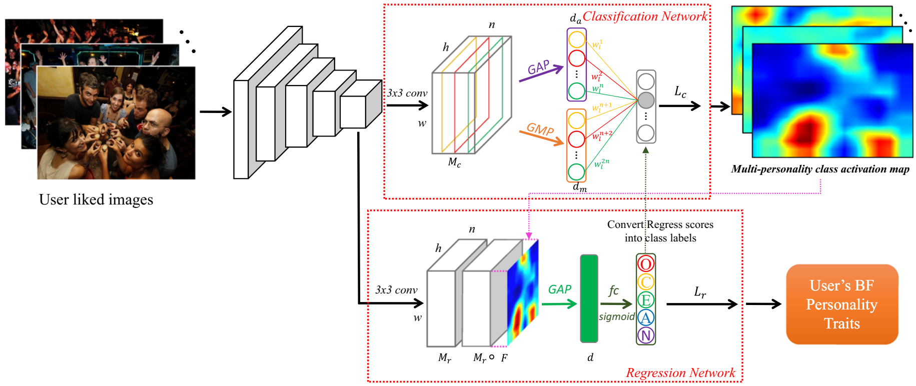


Fig. 3 The framework of the proposed personality prediction method based on WSDCN

volutional network (WSDCN), which can leverage both the local and holistic representations to predict five personality traits simultaneously by training the network only once. The experiments on the PsychoFlickr database show that our approach outperforms the state-of-the-art methods. The WSDCN consists of a classification sub-network and a regression sub-network. The classification sub-network is designed to generate the attentive image regions related to different personality classes. To address the lack of class labels, the user's BF personality traits [23] are first converted to image-level personality class labels, based on which the Class Activation Map (CAM) [37] of each personality class is obtained by training a fully convolutional network. Then, the Multi-Personality Class Activation Map (MPCAM) is generated and utilized to highlight the attentive regions of different personality classes. The regression sub-network uses the MPCAM as the localized activation of deep feature maps to obtain local representation, which is then combined with the holistic representation to achieve the BF personality scores.

The rest of the paper is organized as follows: Sect. 2 proposes a personality prediction method from attentive regions of user liked images via weakly supervised dual convolutional network. The details of our experiments, including the dataset and a complete set of experimental results are presented in Sect. 3. Conclusions are drawn in Sect. 4.

2 Proposed Approach

In this paper, we propose a personality prediction method based on weakly supervised dual convolutional network (WSDCN). The framework of the proposed approach is illustrated in Fig. 3. The WSDCN is designed to learn an end-to-end model for predicting five personality scores simultaneously, which consists of a classification network and a regression network. As a personality-related cue, the liked images can be labeled with user's personality traits for training the WSDCN model. The classification network is designed to detect the attentive image regions (i.e. MPCAM) of a set of user liked images. The regression network aims at using the local and holistic deep features of all liked images for predicting user's BF personality traits.

The proposed method mainly consists of four parts. First, the image-level personality class labels are obtained based on user's BF personality scores (Sect. 2.1). Second, a set of MPCAMs are generated from the user liked images, which are fed into classification network supervised by the personality class labels (Sect. 2.2). Third, we construct a regression network to predict the user's personality scores based on the fusion of holistic and localized feature representations of the liked images (Sect. 2.3). Fourth, the end-to-end dual networks are trained jointly using a set of user liked images associated with the personality class labels and five personality scores (Sect. 2.4). The remainder of this section discusses these four parts in detail.

2.1 Personality Class Labels

In the classification network, the image-level personality class labels are required. Hence, the user's BF personality scores for regression task need to be converted into the personality class labels of their liked images for classification task. Each personality trait can be divided into two classes according to the personality scores [29]. Hence, the BF personality traits can be converted into ten classes, i.e., High Openness (HO), High Conscientiousness (HC), High Extroversion (HE), High Agreeableness (HA), High Neuroticism (HN), Low Openness (LO), Low Conscientiousness (LC), Low Extroversion (LE), Low Agreeableness (LA), and Low Neuroticism (LN).

Let $\{\boldsymbol{X}_i, \boldsymbol{Y}_i^r\}_{i=1}^N$ denote the N training users, where $\boldsymbol{X}_i = \{x_{i,j}\}_{j=1}^M$ indicate a set of M images liked by ith user, and $\boldsymbol{Y}_i^r = \{\boldsymbol{y}_{i,j}^r\}_{j=1}^M$ are the corresponding five normalized personality scores. Suppose $\boldsymbol{Y}_i^c = \{\boldsymbol{y}_{i,j}^c\}_{j=1}^M$ denote ten converted labels of ith user, then the ten personality classes $\boldsymbol{y}_{i,j}^c = \{y_{i,j,l}\}_{l=1}^{10}$ of all the ith user liked images can be obtained according to the original BF traits:

$$\boldsymbol{y}_{i,j}^c = \left[max\left(0, \left(\boldsymbol{y}_{i,j}^r - 0.5\right)\right); max\left(0, \left(0.5 - \boldsymbol{y}_{i,j}^r\right)\right)\right], \tag{1}$$

where $max(0, x)$ is the $Relu$ operation, and the threshold of low and high normalized personality scores is 0.5. For the jth image liked by the ith training user, the ten personality classes $\boldsymbol{y}_{i,j}^c \in \mathbb{R}^{10}$ are obtained from the BF personality scores $\boldsymbol{y}_{i,j}^r \in \mathbb{R}^5$, where c and r denote classification and regression, respectively.

There are two strategies for label learning : dominant label learning and label distribution learning. The dominant label learning strategy is widely used in early image affective classification tasks [14,15]. For an image liked by the ith user, the dominant personality class label can be calculated by

$$c_{i,j} = \arg\max \boldsymbol{y}_{i,j}^c. \tag{2}$$

where $\arg\max(f(x))$ means to find the index x that leads to the maximum value of $f(x)$, and $c_{i,j} \in \{1, 2, 3, \ldots, 10\}$ is the dominant personality class label of the jth image liked by the ith user.

In [40–44], the label distribution learning strategy has been shown effective in reflecting the affective content of images. Hence, the personality distributions $\{y_{i,j}^{(l)}\}_{l=1}^{10}$ can be obtained by

$$\{y_{i,j}^{(l)}\}_{l=1}^{10} = \frac{y_{i,j,l}}{\sum_{l'=1}^{10} y_{i,j,l'}}, \tag{3}$$

which is a normalization to make sure $\sum_{l=1}^{10} y_{i,j}^{(l)} = 1$.

Therefore, we can use both $c_{i,j} \in \mathbb{R}^1$ and $\{y_{i,j}^{(l)}\}_{l=1}^{10} \in \mathbb{R}^{10}$ as the supervision labels of the jth image liked by the ith user to learn the personality-specific activation maps, which can capture the attentive image regions related to different personality classes.

2.2 Classification Network

In this work, our model is built on the basic CNN models, which are pre-trained on the ImageNet dataset [45]. We remove the fully-connected layers after the last convolutional layer and add n kernels of size 3 × 3, stride 1, pad 1 to generate a new convolutional layer $\boldsymbol{M}_c \in \mathbb{R}^{w \times h \times n}$ for classification, where w and h are the width and height of the convolutional layer, respectively. Following $\boldsymbol{M}_c$, the Global Average Pooling (GAP) operation and Global Maximum Pooling (GMP) operation are employed to identify the localized and holistic part for each feature map in the same personality class. We use the coupled GAP vector $\boldsymbol{d}_a \in \mathbb{R}^{n \times 1}$ and GMP vector $\boldsymbol{d}_m \in \mathbb{R}^{n \times 1}$ as the fully-connected layer with a softmax operator to predict the ten-class personality probabilities $\{p_{i,j}^{(l)}\}_{l=1}^{10}$, which are defined as

$$\{p_{i,j}^{(l)}\}_{l=1}^{10} = \frac{e^{\boldsymbol{w}_l^{\mathrm{T}} \boldsymbol{d}_{a,m}}}{\sum_{c'=1}^{10} e^{\boldsymbol{w}_{c'}^{\mathrm{T}} \boldsymbol{d}_{a,m}}}, \tag{4}$$

where $\{\boldsymbol{w}_l\}_{l=1}^{10} \in \mathbb{R}^{2n \times 10}$ are the weights of ten personality classes from the coupled feature vector $\boldsymbol{d}_{a,m} = [\boldsymbol{d}_a; \boldsymbol{d}_m]$.

For the dominant personality class label $c_{i,j}$, the classification network is trained using the following loss function:

$$L_c = -\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{10} \boldsymbol{I}[l = c_{i,j}] \log p_{i,j}^{(l)}, \tag{5}$$

where $\boldsymbol{I}(x)$ is indicator function, and $\boldsymbol{I}(x) = 1$ if the condition x is true, and 0 otherwise. For the personality distributions $\{y_{i,j}^{(l)}\}_{l=1}^{10}$, the classification network is trained using the following loss function:

$$L_c = -\frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{10} y_{i,j}^{(l)} \log p_{i,j}^{(l)}. \tag{6}$$

The Eqs. (5) and (6) are all cross entropy loss functions, which are calculated using the dominant personality class label and the personality distributions, respectively. We choose one of the loss functions to optimize the classification network and validate the performance for personality prediction.

Motivated by the ability of deep features for discriminative localization [37], the MPCAM can be obtained with supervision from different personality class labels. Therefore, we can capture the attentive image regions preferred by users with different personality classes. Different from [37], both GAP operation and GMP operation are adopted to capture the localized and holistic weights of each feature map for a personality class. Figure 4 shows an example of how to generate the class activation map. As illustrated in Fig. 4, a weighted sum of GAP vector $\boldsymbol{d}_a$ and GMP vector $\boldsymbol{d}_m$ is used to generate the predicted probability of each personality class $p_{i,j}^{(l)}$. Similarly, the CAM $\boldsymbol{f}_l \in \mathbb{R}^{w \times h}$ of each personality class can be obtained by computing the weighted sum of the feature maps of the last convolutional layer $\boldsymbol{M}_c$, which is formulated as

$$\boldsymbol{f}_l = \frac{1}{2}\sum_{t=1}^{n}(w_l^t + w_l^{n+t}) f_t(x, y), \quad l \in \{1, 2, \ldots, 10\}, \tag{7}$$

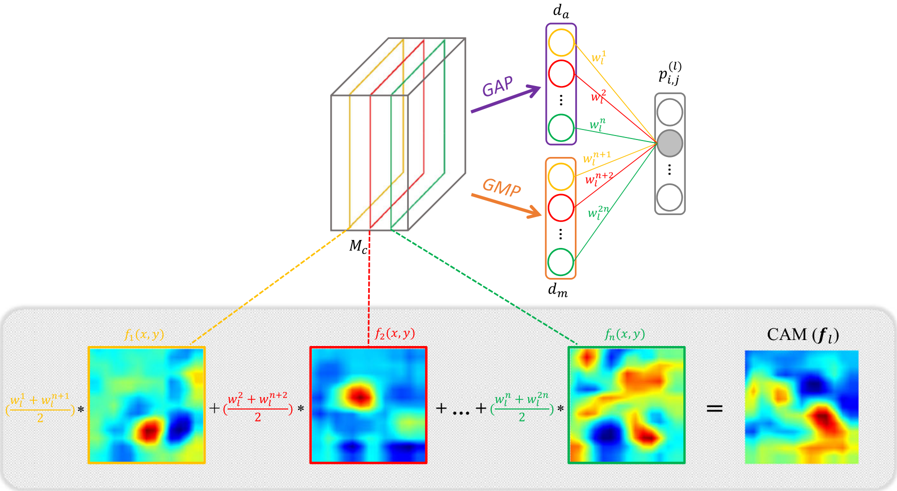


Fig. 4 Generation of the class activation map. The CAM can highlight the attentive regions of different personality classes

where $f_t(x, y)$ represents the tth feature map from $\boldsymbol{M}_c$, and $\{w_l^t\}_{t=1}^{2n}$ are the weights of each personality class from the coupled feature vector $\boldsymbol{d}_{a,m}$. Then, the MPCAM can be calculated by

$$\boldsymbol{F} = \sum_{l=1}^{10} p_{i,j}^{(l)} \boldsymbol{f}_l, \tag{8}$$

where $\{p_{i,j}^{(l)}\}_{l=1}^{10}$ denote the prediction outputs of ten personality classes. The MPCAM $\boldsymbol{F} \in \mathbb{R}^{w \times h}$ is a weighted linear sum of different attentive regions related to each personality class.

2.3 Regression Network

The goal of the regression network is to learn the BF personality scores by integrating the output MPCAM of the classification network. We also add n kernels of size 3×3, stride 1, pad 1 to generate a new convolutional layer $\boldsymbol{M}_r \in \mathbb{R}^{w \times h \times n}$, which can be regarded as holistic representation. The $\boldsymbol{F}$ is utilized to produce the local representation by combining with the convolutional features $\boldsymbol{M}_r$. Thus, the local feature map $\boldsymbol{S} \in \mathbb{R}^{w \times h \times n}$ can be obtained by taking the element-wise (Hadamard) product of the MPCAM and holistic feature map, which is defined as

$$\boldsymbol{S} = \boldsymbol{M}_r \cdot \boldsymbol{F}, \tag{9}$$

where $\cdot$ represents the element-wise multiplication. The local representation $\boldsymbol{S}$ and the holistic representation $\boldsymbol{M}_r$, which have been demonstrated to be more effective in image sentiment classification [34], are jointly used for modeling the BF personality scores. Then, the joint feature vector $\boldsymbol{d} = P_{gav}(\boldsymbol{M_r} \uplus \boldsymbol{S})$ can be calculated by the GAP operation of both holistic feature maps and local feature maps, where the P_{gav} and $\uplus$ denote the GAP operation and the concatenation of different feature maps, respectively.

For the image liked by the ith user, the joint feature vector $\boldsymbol{d} \in \mathbb{R}^{2n\times 1}$ is considered as the full-connected layer with a sigmoid operator to produce the predicted five personality scores $\boldsymbol{s}^r_{i,j} = \{s_{i,j,k}\}^5_{k=1}$, which is calculated by

$$\boldsymbol{s}^r_{i,j} = \frac{1}{1 + e^{-\boldsymbol{W}_r^{\mathrm{T}}\boldsymbol{d}}}, \tag{10}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{2n\times 5}$ denotes the weight of predicted BF personality scores from the joint feature vector $\boldsymbol{d}$. The learning objective is formulated as a regression problem, and we employ the Euclidean distance as the multi-personality regression loss function:

$$L_r = \frac{1}{N}\frac{1}{M}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{5}\|s_{i,j,k} - y_{i,j,k}\|_2^2, \tag{11}$$

where $\boldsymbol{y}^r_{i,j} = \{y_{i,j,k}\}^5_{k=1}$ denote the actual five personality scores. In this regression network, five personality scores can be predicted simultaneously by optimizing the multi-personality regression loss function.

2.4 Training Process

We leverage the two loss functions jointly by end-to-end stochastic gradient descent (SGD) optimization. Given the N training users $\{\boldsymbol{X}_i, \boldsymbol{Y}^r_i\}^N_{i=1}$, we explicitly train the proposed dual networks to optimize the joint loss function:

$$L = L_r + \gamma L_c, \tag{12}$$

where γ is the hyper-parameter to balance two loss functions. In this case, the MPCAM $\boldsymbol{F}$ can be captured during the training process. $\boldsymbol{F}$ is just the one-way activation from the classification network to the regression network for producing local features. In regression network, both the holistic and local features are employed to model the BF personality scores.

In the WSDCN model, users' personality scores are deemed to be the supervision of their liked images during training. Conversely, the liked images are users' personal information for inferring their personality scores. Therefore, users' personality traits can be inferred from the fusion of the predicted results of their liked images. To predict the BF personality traits of a given user, all the liked images are fed into the trained WSDCN to calculate a set of predicted BF personality scores $\{\boldsymbol{s}^r_{i,j}\}^M_{j=1}$. Thus, the predicted BF personality scores $\boldsymbol{S}^r_i \in \mathbb{R}^5$ of ith user can be obtained by

$$\boldsymbol{S}^r_i = \frac{1}{M}\sum_{j=1}^{M}\boldsymbol{s}^r_{i,j}. \tag{13}$$

In this way, the user's BF personality scores can be predicted simultaneously via this end-to-end WSDCN model.

3 Experiments

3.1 Database and Baseline Methods

We evaluate our method on the public PsychoFlickr database [18], which consists of 300 Flickr users and their 60,000 liked images (200 images per user). The users' BF personality traits are Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N). The attributed personality scores of users are obtained by 12 independent observers answering the BFI-10 questionnaire [24]. The questionnaire contains 10 items, every two of which are related to a trait, and each item has five options. Therefore, the attributed personality scores of users are in the range [− 4, 4]. In this work, we normalize the attributed personality scores into the range [0, 1].

The purpose of this paper is to map user liked images into personality traits. As far as we know, only two works have been proposed on personality prediction based on the liked images, namely [19] and [20]. Therefore, we compare the proposed method with these two state-of-the-arts on the PsychoFlickr database. In [19], a bag of the low-level image features (color, composition, texture, etc.) are employed as the user's personality-related cues, which are used for personality prediction based on the LASSO regression [31]. Besides the low-level image features, the high-level features of image contents (objects, scenes, people, etc.) are more related to users' personality traits. Based on this finding, a F2A+A2P (Features to Answers + Answers to Personality) approach [20] is adopted to predict personality traits.

3.2 Implementation Details and Performance Criteria

We build the proposed method using three basic deep learning architectures: AlexNet [35], VGGNet [46] with 16 layers and ResNet101 [47], which have been pre-trained on ImageNet [45]. We replace the fully connected layer with classification and regression networks. Images have been resized to 227 × 227 (AlexNet) and 224 × 224 (VGGNet and ResNet101) to feed into the architecture of proposed network, and the number of convolutional kernels n is set to 512. The network parameters are set as follows: weight decay of 0.0001, momentum of 0.9, batch size of 50, initial learning rate of 0.005, drops to a factor of 0.98 every epoch, and total epoch of 50. The 300 users are split into two groups, i.e., 90% users for model training and the remaining 10% for test. Thus, ten-fold cross-validation is used to avoid bias, and the average of 10 test results is reported. The proposed method is implemented using Tensorflow [48].

Similar to [19,20], Spearman Rank Order Correlation Coefficient (SROCC) and Root Mean Square Error (RMSE) are employed to evaluate the prediction monotonicity and accuracy of the proposed method, respectively. In order to measure the dispersion degree of 10 test results, the Coefficient of Variation (CV), which indicates the percentage of standard deviation and average value, is also reported. Higher value represents better performance for SROCC, while lower value indicates better performance for RMSE and CV.

3.3 Evaluation on Personality Prediction

To evaluate the performance of the proposed personality prediction method, we compare our method against the state-of-the-art methods [19,20] on the PsychoFlickr database. In this experiment, $\{y_{i,j}^{(l)}\}_{l=1}^{10}$ and ResNet101 are chosen as the classification labels and the basic CNN, respectively. Table 1 summarizes the performance of three methods for personality prediction and the best results for each personality are marked in bold font. The CV_s and CV_r denote the dispersion degree of SROCC values and RMSE values, respectively. It can be observed that the proposed method can achieve highest SROCC values and lowest RMSE values, which indicates the prediction monotonicity and accuracy of our method are significantly better than the other two methods. In particular, even for the personality traits (e.g., openness, conscientiousness and agreeableness), which are relatively difficult to predict, the SROCC values of the proposed method are higher than the other two methods by at least 10%.

Table 1 Comparison of prediction performance on PsychoFlickr database

Personality	Metric	SROCC	CV_s	RMSE	CV_r
O	Segalin [19]	0.3543	20.98%	0.5549	19.11%
	Guntuku [20]	0.3984	21.76%	0.5386	18.67%
	Proposed	**0.5802**	**15.64%**	**0.3766**	**15.42%**
C	Segalin [19]	0.5348	17.21%	0.3901	21.45%
	Guntuku [20]	0.5518	15.27%	0.3745	20.75%
	Proposed	**0.6451**	**14.12%**	**0.3465**	**16.84%**
E	Segalin [19]	0.6248	19.42%	0.7642	11.24%
	Guntuku [20]	0.6785	14.58%	0.5864	14.89%
	Proposed	**0.7294**	**9.72%**	**0.4764**	**10.65%**
A	Segalin [19]	0.4763	18.84%	0.5264	20.21%
	Guntuku [20]	0.5247	17.54%	0.4678	19.42%
	Proposed	**0.6582**	**10.85%**	**0.3743**	**16.85%**
N	Segalin [19]	0.6125	13.89%	0.4996	17.22%
	Guntuku [20]	0.6357	11.37%	0.4876	16.96%
	Proposed	**0.7255**	**7.98%**	**0.3824**	**15.73%**

This is mainly because of the effectiveness of our approach for capturing the personality-related attentive image regions. In addition, the CV values of the proposed method are the smallest for each trait, which indicates that the proposed method based on deep features can achieve more stable performance than the other two methods based on hand-crafted features.

We now report the experimental results when two kinds of classification supervision labels and three basic CNNs are adopted in the proposed method. Table 2 summarizes the comparison results. For each personality, the best results are marked in bold font. From the results, it is known that our WSDCN model using the personality distributions $\{y_{i,j}^{(l)}\}_{l=1}^{10}$ can obtain better prediction performance than that using the dominant personality class label $c_{i,j}$, except for conscientiousness. This indicates that the label distribution learning strategy is more effective than the dominant label learning strategy. This confirms that the user's personality traits are so diverse that the personality distributions are more reasonable in representing these traits. We find that the proposed WSDCN model can achieve slightly better prediction performance with the deeper networks. This benefits from the better capacity of deeper networks for extracting high-level semantic features. To sum up, both the personality distributions and the deeper networks have contributed to the prediction performance. Therefore, we select the personality distributions and ResNet101 as the classification labels and basic CNN in our approach, respectively.

The effect of parameter γ in Eq. (12) is also evaluated, with results shown in Fig. 5. As can be seen, when γ increases from 0.05 to 1, the SROCC values tend to reach the maximum except for conscientiousness. Further increasing the γ leads the decreasing of the prediction performance. Besides, the proposed method can achieve the relatively stable performance when γ ranges from 0.05 to 20 for conscientiousness. This indicates that both the L_r and the L_c play an important role in the proposed WSDCN model learning. That is to say, the classification loss has a great effect on capturing the attentive regions of different personality classes, and the regression loss is crucial for predicting the five personality scores. Therefore, γ is set to 1 in our experiments for a trade-off between classification loss and regression loss.

Table 2 Prediction performances of the proposed method using two classification supervision labels (i.e. $c_{i,j}$ and $\{y_{i,j}^{(l)}\}_{l=1}^{10}$) and three basic CNNs (i.e. AlexNet, VGGNet, and ResNet101) on PsychoFlickr database

Personality	Label		Network			SROCC	RMSE
	$c_{i,j}$	$\{y_{i,j}^{(l)}\}_{l=1}^{10}$	AlexNet	VGGNet	ResNet101		
O	✓		✓			0.5445	0.4091
	✓			✓		0.5531	0.3966
	✓				✓	0.5576	0.3912
		✓	✓			0.5754	0.3814
		✓		✓		0.5799	0.3787
		✓			✓	**0.5802**	**0.3766**
C	✓		✓			0.6425	0.3489
	✓			✓		**0.6524**	**0.3387**
	✓				✓	0.6501	0.3401
		✓	✓			0.6374	0.3598
		✓		✓		0.6434	0.3511
		✓			✓	0.6451	0.3465
E	✓		✓			0.7067	0.5103
	✓			✓		0.7101	0.4998
	✓				✓	0.7121	0.4934
		✓	✓			0.7093	0.5044
		✓		✓		0.7265	0.4810
		✓			✓	**0.7294**	**0.4764**
A	✓		✓			0.6374	0.4022
	✓			✓		0.6410	0.3979
	✓				✓	0.6434	0.3934
		✓	✓			0.6498	0.3867
		✓		✓		0.6537	0.3801
		✓			✓	**0.6582**	**0.3743**
N	✓		✓			0.6984	0.4058
	✓			✓		0.7092	0.3987
	✓				✓	0.7123	0.3934
		✓	✓			0.7114	0.3965
		✓		✓		0.7221	0.3877
		✓			✓	**0.7255**	**0.3824**

In order to evaluate the relative contributions of local representation $\boldsymbol{S}$ and the holistic representation $\boldsymbol{M}_r$ in the whole model, comparative experiments are conducted with three settings (i.e. $\boldsymbol{S}$, $\boldsymbol{M}_r$, and $\boldsymbol{M_r} \uplus \boldsymbol{S}$). The experimental results are summarized in Table 3, where the best results for each personality are marked in bold font. We can find that jointly using the local and holistic representation in our framework delivers better performance than using each of them separately. The local representation has a better performance than the holistic representation except for extraversion. This is mainly because that users' preferences on images are mainly influenced by the localized regions instead of the entire images, which coincides with the results in [33,34].

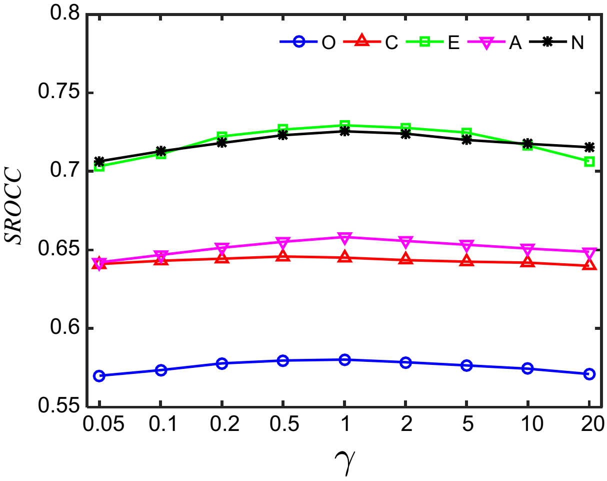


Fig. 5 The effect of parameter γ for personality prediction on PsychoFlickr database

Table 3 Prediction performance comparison of three representation (i.e. $\boldsymbol{S}$, $\boldsymbol{M}_r$, and $\boldsymbol{M}_r \uplus \boldsymbol{S}$) on PsychoFlickr database

Personality	$\boldsymbol{S}$	$\boldsymbol{M}_r$	SROCC	RMSE
O	✓		0.5631	0.3845
		✓	0.5447	0.4084
	✓	✓	**0.5802**	**0.3766**
C	✓		0.6336	0.3610
		✓	0.6237	0.3675
	✓	✓	**0.6451**	**0.3465**
E	✓		0.7061	0.5112
		✓	0.7089	0.5053
	✓	✓	**0.7294**	**0.4764**
A	✓		0.6348	0.4053
		✓	0.6264	0.4123
	✓	✓	**0.6582**	**0.3743**
N	✓		0.7054	0.4012
		✓	0.6899	0.4171
	✓	✓	**0.7255**	**0.3824**

To qualitatively analyze the prediction performance of the proposed method and the state-of-the-art methods, Fig. 6 shows example images liked by two users and the corresponding ground truth and predicted personality scores. The blue bar is the ground truth scores, the purple bar is the predicted scores of the proposed method, the green bar is the predicted scores of Segalin's [19] method, and the yellow bar is the predicted scores of Guntuku's [20] method. From the results, we can see that the proposed method more accurately predicts the personality scores than the other two methods, which demonstrates the effectiveness of the proposed method.

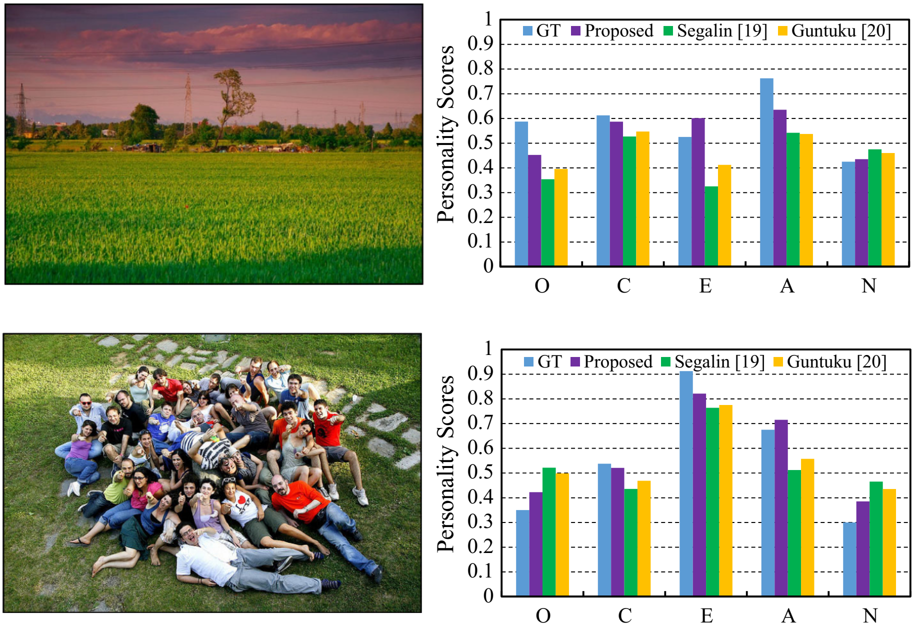


Fig. 6 Example images liked by two users and the corresponding ground truth ('GT') and predicted personality scores using three methods. Images are in the left column and the personality scores are in the right column, respectively

3.4 Visualizing Analysis

As mentioned in Sect. 3.3, the local deep feature maps are effective in representing attentive image regions of different personality classes. In order to explore the impact of two supervision labels (the dominant personality class label and the personality distributions) on the attentive regions, we visualize the detected MPCAM of an image liked by a test user in Fig. 7. The normalized BF personality scores are O: 0.65, C: 0.675, E: 0.75, A: 0.6625, and N: 0.4. The dominant personality class is high extroversion. As shown in Fig. 7b, we can observe that the region around a boat can be detected. The boat in the lake, meaning play in outdoor places, is associated with high extroversion. Besides the boat, the region of a tower is also detected in the MPCAM, which is shown in Fig. 7c. The reason is that the attentive regions relating to multiple personalities can be detected using the personality distributions, while the MPCAM using the dominant personality class only focuses on the region relating to the dominant personality. Thus, it is reasonable to adopt the personality distributions instead of the dominant personality class label.

In order to investigate how image local regions attract people with different personality traits, the corresponding CAM and predicted scores are shown in Fig. 8. The predicted scores less than 0.01 are set to 0.01. The ground truth of converted personality distributions are HO: 0.1791, HC: 0.2089, HE: 0.2985, HA: 0.1940, HN: 0, LO: 0, LC: 0, LE: 0, LA: 0, and LN: 0.1195. As shown in Fig. 8, the predicted scores are consistent with the ground truth. We can find that the predicted score of high extroversion is the highest, which indicates that the CAM of high extroversion has the most contribution to the MPCAM. In addition to high extroversion, the predicted scores of high openness, high conscientiousness, and high

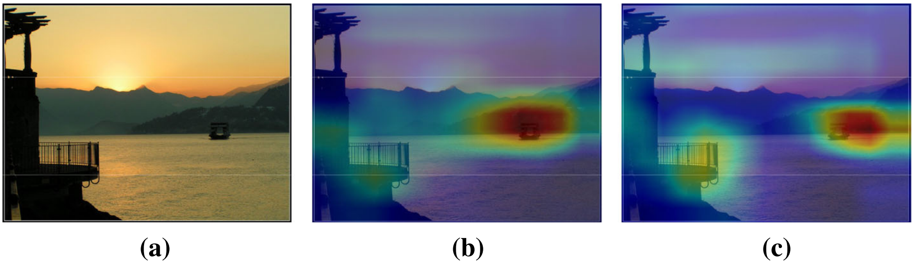
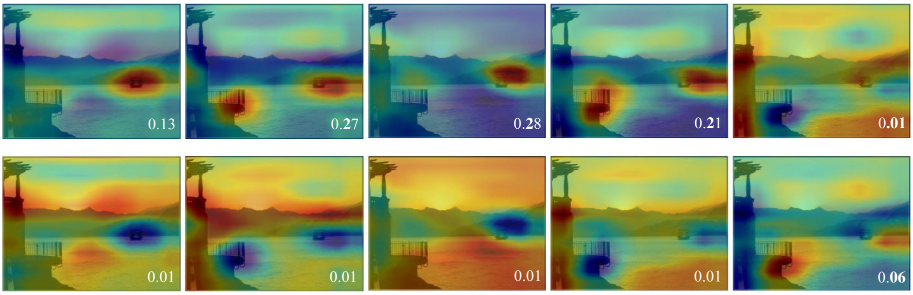


Fig. 7 An image liked by user with high extroversion and its MPCAM: **a** the liked image; **b** the MPCAM using dominant personality class label; **c** the MPCAM using personality distributions



(a) Openness **(b)** Conscientiousness **(c) Extroversion** **(d)** Agreeableness **(e)** Neuroticism

Fig. 8 The personality class activation map of each traits. High personality traits are in the first row, and low personality traits are in the second row: **a** Openness; **b** Conscientiousness; **c** Extraversion; **d** Agreeableness; **e** Neuroticism

agreeableness are also higher than other personality classes. The region around the sky also can be discriminated in the CAM of high openness, which is associated with users who have creative thinking. The tower can be detected in the CAM of high conscientiousness and high agreeableness. This indicates that the responsible and affable person tend to like the building, which is associated with a house to live in. It is worth noting that the opposite personality classes (e.g., high openness and low openness) also have the inverse CAM, which is consistent with the fact that the preferences of user with the high or low personality trait on image regions are opposite.

4 Conclusions

In this paper, we have proposed a method to predict the BF personality traits by using a weakly supervised dual convolutional network (WSDCN), which jointly uses the local and holistic representations from a set of user liked images. The CAM of each personality class has been shown effective in capturing the attentive image regions liked by the users with different personality classes. In addition, the proposed WSDCN model can predict the BF personality traits simultaneously in an end-to-end manner. Experimental results on public database have demonstrated that the performance of our approach outperforms the state-of-the-art approaches. While the proposed method has achieved the best performance, the prediction accuracy is still far from ideal (the highest SROCC value is 0.7294). We will try to construct more efficient deep learning models to predict personality traits in future work.

Acknowledgements This work was supported by Outstanding Innovation Scholarship for Doctoral Candidate of "Double First Rate" Construction Disciplines of CUMT.

References

1. Matthews G, Deary I, Whiteman M (2009) Personality traits. Cambridge University Press, Cambridge
2. Furnham A, Jackson CJ, Miller T (1999) Personality, learning style and work performance. Personal Individ Differ 27(6):1113–1122
3. Guntuku SC, Roy S, Lin W (2015) Personality modeling based image recommendation. In: Proceedings of the international conference on multimedia modeling, pp 171–182
4. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 18:43–49
5. House VN (2011) Personal photography, digital technologies and the uses of the visual. Vis Stud 26(2):125–134
6. Zhao S, Gao Y, Ding G, Chua TS (2018) Real-time multimedia social event detection in microblog. IEEE Trans Cybern 48(11):3218–3231
7. Deng C, Chen Z, Liu X, Gao X, Tao D (2018) Triplet-based deep hashing network for cross-modal retrieval. IEEE Trans Image Process 27(8):3893–3903
8. Li C, Deng C, Li N, Liu W, Gao X, Tao D (2018) Self-supervised adversarial hashing networks for cross-modal retrieval. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4242–4251
9. Yang E, Deng C, Li C, Liu W, Li J, Tao D (2018) Shared predictive cross-modal deep quantization. IEEE Trans Neural Netw 99:1–12
10. Joshi D, Datta R, Fedorovskaya E, Luong Q (2011) Aesthetics and emotions in images. IEEE Signal Proc Mag 28(5):94–115
11. Zhao S, Yao H, Gao Y, Ding G, Chua TS (2018) Predicting personalized image emotion perceptions in social networks. IEEE Trans Affect Comput 9(4):526–540
12. Zhu H, Li L, Zhao S, Jiang H (2018) Evaluating attributed personality traits from scene perception probability. Pattern Recognit Lett 116:121–126
13. Zhao S, Ding G, Han J, Gao Y (2018) Personality-aware personalized emotion recognition from physiological signals. In: Proceedings of the international joint conferences on artificial intelligence, pp 1660–1667
14. Machajdik J, Hanbury A (2010) Affective image classification using features inspired by psychology and art theory. In: Proceedings of the ACM international conference on multimedia, pp 83–92
15. Zhao S, Gao Y, Jiang X, Yao H, Chua TS , Sun X (2014) Exploring principles-of-art features for image emotion recognition. In: Proceedings of the ACM international conference on multimedia, pp 47–56
16. Peng KC, Chen T, Sadovnik A, Gallagher AC (2015) A mixed bag of emotions: model, predict, and transfer emotion distributions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 860–868
17. You Q, Luo J, Jin H, Yang J (2016) Building a large scale dataset for image emotion recognition: the fine print and the benchmark. In: Proceedings of the AAAI conference on artificial intelligence, pp 308–314
18. Cristani M, Vinciarelli A, Segalin C, Perina A (2013) Unveiling the multimedia unconscious: implicit cognitive processes and multimedia content analysis. In: Proceedings of the ACM international conference on multimedia, pp 213–222
19. Segalin C, Cristani M, Perina A, Vinciarelli A (2017) The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. IEEE Trans Affect Comput 8(2):268–285
20. Guntuku SC, Zhou JT, Roy S, Lin WS, Tsang IW (2018) Who likes what, and why? Insights into personality modeling based on image 'likes'. IEEE Trans Affect Comput 9(1):130–143
21. Vinciarelli A, Mohammadi G (2014) A survey of personality computing. IEEE Trans Affect Comput 5(3):273–291
22. Goldberg LR (1993) The structure of phenotypic personality traits. Am Psychol 48(1):26–34
23. Goldberg LR (1990) An alternative "description of personality": the big-five factor structure. J Pers Soc Psychol 59(6):1216
24. Rammstedt B, John O (2007) Measuring personality in one minute or less: a 10-item short version of the big five inventory in English and German. J Res Pers 41(1):203–212
25. Jenkins R (2014) Social identity. Routledge 6(1):1396

26. Zen G, Lepri B, Ricci E, Lanz O (2010) Space speaks: towards socially and personality aware visual surveillance. In: Proceedings of the ACM international workshop on multimodal pervasive video analysis, pp 37–42
27. Pianesi F, Mana N, Cappelletti A, Lepri B, Zancanaro M (2008) Multimodal recognition of personality traits in social interactions. In: Proceedings of the international conference on multimodal interfaces, pp 53–60
28. Wei X, Zhang C, Zhang H, Wu J (2018) Deep bimodal regression of apparent personality traits from short video sequences. IEEE Trans Affect Comput 9(3):303–315
29. Segalin C, Dong SC, Cristani M (2017) Social profiling through image understanding: personality inference using convolutional neural networks. Comput Vis and Image Und 156:34–50
30. Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci 110(15):5802–5805
31. Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. J Roy Stat Soc 73(3):273–282
32. Peng KC, Sadovnik A, Gallagher A, Chen T (2016) Where do emotions come from? Predicting the emotion stimuli map. In: Proceedings of the IEEE international conference on image processing, pp 614–618
33. You Q, Jin H, Luo J (2017) Visual sentiment analysis by attending on local image regions. In: Proceedings of the AAAI conference on artificial intelligence, pp 231–237
34. Yang J, She D, Lai YK, Rosin P, Yang MH (2018) Weakly supervised coupled networks for visual sentiment analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 231–237
35. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the international conference on neural information processing systems, pp 1097–1105
36. Deng C, Liu X, Li C, Tao D (2018) Active multi-kernel domain adaptation for hyperspectral image classification. Pattern Recognit 77:306–315
37. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
38. Diba A, Sharma V, Pazandeh A, Pirsiavash H, Gool LV (2017) Weakly supervised cascaded convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5131–5139
39. Durand T, Mordan T, Thome N, Cord M (2017) Wildcat: weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5957–5966
40. Zhao S, Ding G, Gao Y, Zhao X, Tang Y, Han J (2018) Discrete probability distribution prediction of image emotions with shared sparse learning. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2018.2818685
41. Yang J, She D, Sun M (2017) Joint image emotion classification and distribution learning via deep convolutional neural network. In: Proceedings of the international joint conference on artificial intelligence, pp 3266–3272
42. Zhao S, Yao H, Gao Y, Ji R, Ding G (2017) Continuous probability distribution prediction of image emotions via multi-task shared sparse regression. IEEE Trans Multimedia 19(3):632–645
43. Zhao S, Zhao X, Ding G, Keutzer, K (2018) EmotionGAN: unsupervised domain adaptation for learning discrete probability distributions of image emotions. In: Proceedings of ACM multimedia conference on multimedia conference, pp 1319–1327
44. Zhao S, Ding G, Gao Y, Han J (2017) Approximating discrete probability distribution of image emotions by multi-modal features fusion. Transfer 1000(1)
45. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115(3):1–42
46. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Comput Sci
47. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
48. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems, arXiv preprint. arXiv:1603.04467