Heterogeneous Face Attribute Estimation:
A Deep Multi-Task Learning Approach

Hu Han, *Member, IEEE,* Anil K. Jain, *Fellow, IEEE,* Fang Wang,
Shiguang Shan, *Senior Member, IEEE* and Xilin Chen, *Fellow, IEEE*

Abstract—Face attribute estimation has many potential applications in video surveillance, face retrieval, and social media. While a number of methods have been proposed for face attribute estimation, most of them did not explicitly consider the attribute correlation and heterogeneity (*e.g.,* ordinal vs. nominal and holistic vs. local) during feature representation learning. In this paper, we present a *Deep Multi-Task Learning (DMTL)* approach to jointly estimate multiple heterogeneous attributes from a single face image. In DMTL, we tackle attribute correlation and heterogeneity with convolutional neural networks (CNNs) consisting of shared feature learning for all the attributes, and category-specific feature learning for heterogeneous attributes. We also introduce an unconstrained face database (LFW+), an extension of public-domain LFW, with heterogeneous demographic attributes (age, gender, and race) obtained via crowdsourcing. Experimental results on benchmarks with multiple face attributes (MORPH II, LFW+, CelebA, LFWA, and FotW) show that the proposed approach has superior performance compared to state of the art. Finally, evaluations on a public-domain face database (LAP) with a single attribute show that the proposed approach has excellent generalization ability.

Index Terms—Face recognition, heterogeneous attribute estimation, attribute correlation, attribute heterogeneity, multi-task learning

✦

1 INTRODUCTION

HUMAN face portrays important cues for social interaction, providing a wide variety of salient information, including the person's identity, demographic (age, gender, and race), hair style, clothing, etc. Over the past 50 years, significant advances have been made in extracting discriminative features in a face image to determine the subject's identity [3]. In more recent years, several applications have emerged that make use of face attributes, from demographic attributes (*e.g.,* age, gender, and race) to descriptive visual attributes (*e.g.,* clothing and hair style). These applications include (i) *video surveillance* [4] [5], *e.g.,* automatic detection of persons with sunglasses or mask observed at unusual hours or in unusual places; (ii) *face retrieval* [6] [7] [8], *e.g.,* automatic filtering of a face database to find person(s) of interest with given attributes; and (ii) *social media* [9] [10], *e.g.,* automatic recommendation of hair styles or makeups.

Despite recent progresses in face attribute prediction [7], [12]–[18], most prior work is limited to es-

- *Hu Han, Fang Wang, Shiguang Shan, and Xilin Chen are with the Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China, and the University of Chinese Academy of Sciences, Beijing 100049, China. Shiguang Shan is also with the CAS Center for Excellence in Brain Science and Intelligence Technology.*
  Anil K. Jain is with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA.

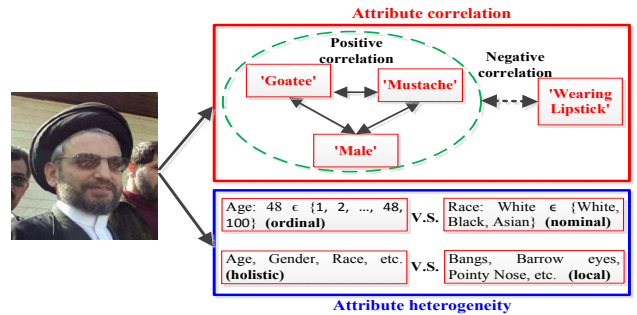 *E-mail: {hanhu, sgshan, xlchen}@ict.ac.cn; jain@cse.msu.edu; fang.wang14@vipl.ict.cn*

Fig. 1. Individual face attributes have both correlation and heterogeneity. While attribute correlation can be utilized to improve the robustness of attribute estimation, attribute heterogeneity should also be tackled by designing appropriate prediction models.

timating a single face attribute (*e.g.,* age), or learning a separate model for each face attribute. To address these limitations, attempts have been made to develop new approaches that explore attribute correlation for *joint* estimation of multiple face attributes [19]–[23]. Even these methods have some serious limitations. For example, approaches in [19], [20], [22] used the same features for estimating all the attributes without considering the attribute heterogeneity. The sum-product network (SPN) adopted in [21] for modeling attribute correlations may not be feasible because of the exponentially growing number of attribute group
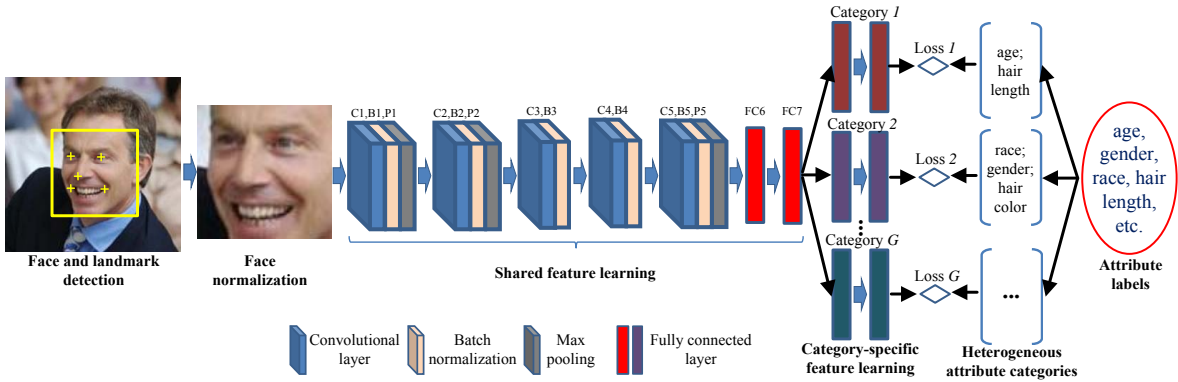
Fig. 2. Overview of the proposed deep multi-task learning (DMTL) network consisting of an early-stage shared feature learning for all the attributes, followed by category-specific feature learning for heterogeneous attribute categories. We use a modified AlexNet [11] with a batch normalization (BN) layer inserted after each Conv. layer for shared feature learning. The subnetworks are used to fine-tune the shared features towards the optimal estimation of individual heterogeneous attributes, *e.g.*, nominal vs. ordinal and holistic vs. local.

combinations. The cascade network in [23] also required learning a separate Support Vector Machine (SVM) classifier for each face attribute, and is not an end-to-end learning approach.

Figure 1 shows that a face image portrays a wide variety of attributes, which are both *correlated* and *heterogeneous*. Attribute correlation can be either positive or negative. For example, a person with goatee and mustache is more likely to be a male, and is less likely to wear lipstick. Meanwhile, individual attributes can be heterogeneous in terms of *data type and scale* [24], and *semantic meaning* [25]. While attributes like age and hair length are ordinal, attributes like gender and race are nominal; these two categories of attributes are heterogeneous in terms of data type and scale. Similarly, while attributes such as age, gender, and race describe the characteristics of the whole face, attributes such as pointy nose and big lips, mainly describe the characteristics of local facial components; these two categories of attributes are heterogeneous in terms of semantic meaning. Such attribute correlation and heterogeneity should be considered in designing face attribute estimation models.

Though a number of commercial systems (*e.g.*, Affectiva, Emotient, Face++, and Microsoft)[1] provide estimates of attributes like age, gender and expression, the underlying algorithms used in commercial systems are proprietary; in addition, the databases used by these commercial engines are not (or no longer) available to the research community. Robust estimation of a large number of heterogeneous attributes from a face image remains a challenging problem, particularly under unconstrained sensing and uncooperative subject scenarios.

1.1 Proposed Approach

We present a Deep Multi-Task Learning (DMTL) approach to jointly estimate multiple heterogeneous attributes from a single face image. The proposed approach is motivated by recent advances in face attribute prediction, but takes into account both attribute correlation and attribute heterogeneity in a single convolutional neural network (CNN). The proposed DMTL consists of an early-stage shared feature learning for all the attributes, followed by category-specific feature learning for heterogeneous attribute categories (see Fig. 2). The shared feature learning naturally exploits the relationship between tasks to achieve robust and discriminative feature representation. The category-specific feature learning aims at fine-tuning the shared features towards the optimal estimation of each heterogeneous attribute category. Given the effective shared feature learning and category-specific feature learning, the proposed DMTL achieves promising attribute estimation accuracy while retaining low computational cost, making it of value in many face recognition applications.

The main contributions of this paper include: (i) an efficient multi-task learning (MTL) method for joint estimation of a large number of face attributes; (ii) modeling both attribute correlation and attribute heterogeneity in a single network; (iii) studying the generalization ability of the proposed approach under cross-database testing scenarios; and (iii) compiling the LFW+ database[2] with face images in the wild (LFW), and heterogeneous demographic attributes (age, gender, and race) via crowdsourcing.

Some of the preliminary work is described in [1], [2]. Essential improvements in this work include: (i) extensions in category-specific feature learning for handling attribute heterogeneities in terms of data type and scale, and semantic meaning; (ii) additional technical and implementation details; and (iii) extensive evaluations using 6 different attribute databases, and comparisons with additional state of the art.

The remainder of this paper is structured as follows. In Section 2, we briefly review related literature. In

---

1. Affectiva: www.affectiva.com; Emotient: www.emotient.com; Face++: www.faceplusplus.com; Microsoft: www.how-old.net

2. We plan to place the LFW+ dataset in the public domain.

TABLE 1
A summary of published methods on multi-attribute estimation from a face image.

Publication	Approach (feature and prediction model)	Face database #images (training; testing)	Accuracy
Cottrell and Metcalfe [26]	Autoencoder; One backpropagation network per attribute	Private dataset (160, 40)	**Private dataset** Emotion: $< 50\%$ (Avg. of eight classes) Gender: 100% (on training set)
Kumar et al. [27]	Grayscale and color pixel values, edge magnitude, and gradient direction; One SVM classifier per attribute	CelebA (public) ($180K$, $20K$) LFW[1] (public) (n/a; 13,143)	**CelebA**: 81% (Avg. of 40 attrs.) [23] **LFW**[1] Gender: 92.7%; Race: 90.3%
Guo and Mu [19]	Biologically-inspired features (BIFs); multi-label regression with CCA and PLS	MORPH II (public) (10,530, 44,602)	**MORPH II** Age: 70.0% CS(5)[2], 3.92 yrs. MAE Gender: 98.5%, Race: 99.0% (Black vs. White)
Yi et al. [20]	Concatenated features by multi-scale CNN (3-layer network); multi-label loss	MORPH II (public) (10,530, 44,602)	**MORPH II** Age: 3.63 yrs. MAE Gender: 98.0%, Race: 99.1% (Black vs. White)
Eidinger et al. [15]	LBP and four-patch LBP; One SVM classifier per attribute	*Images of Groups* (public) (3, 500; 1, 050) Adience (public) (13,000; 3,300)	**Images of Groups** Age group: 66.6%, Gender: 88.6% **Adience** Age group: 45.1%, Gender: 76.1%
Han et al. [16]	BIFs with feature selection; One SVM classifier per attribute	MORPH II (public) (20,569; 78,207) PCSO (private) (81,533; 100,012) LFW Frontal (public) (4211, 4211)	**MORPH II** Age: 77.4% CS(5)[2], 3.6 yrs. MAE Gender: 97.6%, Race: 99.1% (Black vs. White) **PCSO** Age: 72.6% CS(5)[2], 4.1 yrs. MAE Gender: 97.1%, Race: 98.7% (Black vs. White) **LFW Frontal** Age: 42.5% CS(5)[2], 7.8 yrs. MAE Gender: 94%, Race: 90% (White vs. Other)
Levi and Hassner [28]	CNN with 3 Conv. layers and 2 FC layers; One CNN classifier per attribute	Adience (public) (15,590; 3,897)[3]	**Adience** Age group: 50.7%, Gender: 86.8%
Liu et al. [23]	Multi-patch features by a cascade of LNets (5 Conv. layers) and ANet (4 Conv. layers); One SVM classifier per attribute	CelebA (public) ($180K$, $20K$) LFWA (public) (6, 263; 6, 970)	**CelebA**: 87% (Avg. of 40 attributes) **LFWA**: 84% (Avg. of 40 attributes)
Huang et al. [29]	CNN features by DeepID2 with large margin local embedding; kNN classifier	CelebA (public) ($180K$, $20K$)	**CelebA**: 84%[4] (Avg. of 40 attributes)
Uřičář et al. [30]	CNN features by VGG-16 [31]; One SVM classifier per attribute	ChaLearn LAP 2016 (public) (4,113; 1500 (validation set))	**ChaLearn LAP2016 (validation set)** Age: 0.24 ϵ-error, Gender: 89.2%, Smile: 79.03%
Ehrlich et al. [32]	Multi-task Restricted Boltzmann Machines with PCA and keypoint features; Multi-task classifier	CelebA (public) ($180K$, $20K$) ChaLearn FotW (6,171; 3,087)	**CelebA**: 87% (Avg. of 40 attributes) **FotW**: Smile and gender: 76.3% (Avg.)
Hand and Chellappa [33]	Multi-task CNN features (3 Conv. layers and 2 FC layers); Joint regression of multiple binary attributes	CelebA (public) ($180K$, $20K$) LFWA (public) (6, 263; 6, 970)	**CelebA** 91% (Avg. of 40 attributes) **LFWA** 86% (Avg. of 40 attributes)
Zhong et al. [34]	Off-the-shelf CNN features by FaceNet and VGG-16 [31] One SVM classifier per attribute	CelebA (public) ($180K$, $20K$) LFWA (public) (6, 263; 6, 970)	**CelebA** 86.6% (Avg. of 40 attributes) **LFWA** 84.7% (Avg. of 40 attributes)
Proposed method	Deep multi-task feature learning (DMTL) with shared feature learning (modified AlexNet) and category-specific feature learning (2 FC layers) Joint estimation of multiple heterogeneous attributes	MORPH II (public) (62, 566; 15, 641)[3] LFW+ (created by authors) (12, 559; 3, 140)[3] CelebA (public) ($180K$, $20K$) LFWA (public) (6,263; 6,970) LAPAge2015 (public) (2,476; 1,136) ChaLearn FotW (public) (6,171; 3,087)	**MORPH II** (w/o pre-training on IMDB-WIKI) Age: 85.3% CS(5)[2], 3.0 yrs. MAE; Gender: 98.0%, Race: 96.6% (Black, White, Other) **LFW+** Age: 75.0% CS(5)[2], 4.5 yrs MAE; Gender: 96.7%; Race: 94.9% **CelebA** 92.1% (Avg. of 40 attributes); **LFWA** 86% (Avg. of 40 attributes) **CLAP2015** (w/o pre-training on IMDB-WIKI) Age: 5.2 yr. MAE, ϵ-error: 0.449 **FotW** Accessory: 94.0% (Avg. of 7 attributes); Smile and gender: 86.1% (Avg.)

[1]The ground-truth age, gender, and race information of the LFW face images was not provided in [27]; the accuracies reported for [27] are from [23]. [2]CS(5) denotes the age estimation accuracy @ 5-year absolute error. [3]The numbers of training and testing images reported here are the average in one-fold test. [4]A different metric is used: an average of true positive rate and true negative rate.

Section 3, we detail the proposed heterogeneous face attribute estimation approach. In Section 4, we introduce the LFW+ database which contains faces in the wild, and heterogeneous attributes of age, gender, and race obtained via crowdsourcing, and provide the experimental results and analysis. Finally, we conclude this work in Section 5.

2 ~~Related Work~~

2.1 Multi-attribute Estimation From Face

While there are a number of studies on face attribute estimation in the literature, many of them focus on estimating a single attribute, *e.g.,* age, expression, etc. The age estimation error with mean absolute error (MAE) metric has been reduced by a large margin from 8.8 years [35] to 2.68 years [17] on the MORPH II database [36]. Facial expression recognition accuracy has been substantially improved from less than 80% to over than 93% on the Cohn-Kanade database [37],

[38]. Due to limited space, we refer interested readers to reviews of the prior work on single facial attribute estimation in [12], [14], [16], [17], [38]–[41]. In the following, we briefly review the most recent literature on joint estimation of multiple face attributes, covering feature representation, prediction models, databases, and performance (see Table 1).

Attempts to design computational models based on psychological studies on multi-attribute estimation from a face image started in the 1990s [26]. Since then, a number of approaches have been reported in the literature, but the early work utilized hand-crafted features for attribute estimation. In [27], edge magnitude and gradient features were extracted from various face regions; the same features were used to learn a separate SVM classifier for each face attribute. Multi-label regressions using canonical correlation analysis (CCA) and partial least squares (PLS) based on BIF features were used in [19] for joint estimation of three face attributes (age, gender, and race); the joint estimation resulted in a better performance than separate models for age, gender, and race. In [15], per attribute dropout-SVM classifiers were trained using LBP features for estimating age and gender, respectively. BIF features with three separate SVM classifiers were used for age, gender, and race estimation in [16], but unlike [19], feature selection was applied to BIF features to find demographic informative features for each attribute.

Except for [26] which used autoencoder for feature learning, all the above approaches utilized hand-crafted features. Recently, the biologically inspired deep learning network has resulted in significant advances in many computer vision tasks [42], including face attribute prediction, due to their ability to learn compact and discriminative features [20], [23], [43]. In [20], CNN features extracted from multi-scale patches were concatenated together, and used for joint estimation of three face attributes (age, gender, and race). A CNN with three convolutional layers and two FC layers was proposed in [28], and per attribute CNNs were trained to handle age and gender estimation, respectively. In [23], a cascaded network of face localization (LNet) and attribute prediction (ANet) was used for face localization and feature extraction for individual SVM classifiers. Additionally, two face attribute databases (CelebA and LFWA) were presented in [23] along with face image labels. Per attribute SVM classifiers were also used in [30] for the estimation of age, gender, and smile, but the features were learned using the VGG-16 network [31]. In [34], a similar idea of per attribute SVM classifiers using FaceNet and VGG-16 features was applied for estimating 40 face attributes on the CelebA and LFWA databases [23]. In [29], a large margin local embedding kNN (LMLE-kNN) approach was proposed to deal with large-scale imbalanced attribute classification tasks. With PCA appearance features and keypoint

features, Multi-task Restricted Boltzmann Machine (RBM) was adopted in [32] for estimating 40 face attributes on the CelebA database, and gender and smile classifications on the FotW database [44].

2.2 Multi-Task Learning in Deep Networks

As summarized in Table 1, approaches using hand-crafted and deep learning features can be grouped into two categories: (i) single-task learning (STL) of per attribute classifier [15], [16], [23], [26]–[28], [30], [34]; and (ii) multi-task learning (MTL) of a joint attribute classifier [32]. Compared with STL based methods, where each attribute is estimated separately, ignoring any correlations between the tasks, MTL based methods learn multiple models for multi-attribute estimation using a shared representation [45]. Deep models are well suited for MTL; therefore, a number of approaches seek to combine MTL with deep learning. Besides the MTL networks for face attribute estimation [20], [32], [33], MTL networks have been proposed for human pose estimation [46], human attribute prediction [47], face alignment [22], [48], etc.

The proposed approach falls under the MTL approach with CNNs, but with several differences compared with existing methods [22], [23], [32], [33], [46], [47].

- Unlike existing methods that have focused on face alignment, human pose estimation, and human attribute estimation [22], [46], [47], the proposed approach focuses on joint estimation of multiple attributes from a face image.
- Unlike the MTL in [22] which utilizes the auxiliary tasks to assist in the main task, we aim to boost the estimation accuracies of all the face attributes through utilizing attribute correlations and handling attribute heterogeneities;
- Unlike the methods in [23], [29] which utilized a two-step pipeline of CNN features followed by attribute classifiers, the proposed DMTL is an end-to-end learning approach;
- The proposed approach considers a number of practical scenarios for heterogeneous attribute estimation, single attribute estimation, and cross-database testing.

3 PROPOSED APPROACH

3.1 Deep Multi-task Learning

Our aim is to simultaneously estimate a large number of face attributes via a joint estimation model. While a large number of face attributes pose challenges to the feature learning efficiency, they also provide opportunities for leveraging the attribute inter-correlations to obtain informative and robust feature representation. For example, as shown in Fig. 3, a number of attributes in the CelebA database [23] have strong
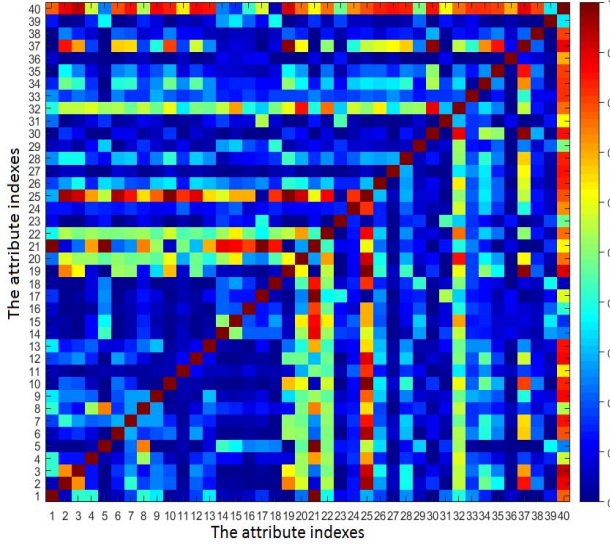
Fig. 3. Pair-wise co-occurrence matrix of the 40 face attributes (see Table 2) provided with the CelebA database[3]. Examples of attributes with a strong positive correlation include: #1 (5 O'Clock Shadow) and attribute #21 (Male), and attribute #19 (Heavy Makeup) and #37 (Wear Lipstick).

pair-wise correlations (elements with red color). MTL methods are naturally suited for this joint estimation problem. However, presence of appearance variations in facial images and the heterogeneity of individual attributes, the mapping from the face image space to the attribute space is typically nonlinear. Therefore, the joint attribute estimation model should also be able to capture the complex and compositional non-linear transformation between its input and output. CNN model is an effective approach for handling both MTL and such a nonlinear transformation learning. A good overview of MTL in neural network can be found in [45]. Following the success of MTL in neural networks [22], [46], [47], we choose to use Deep Multi-task Learning (DMTL) for estimating multiple attributes from a single face image.

We assume a training dataset with N face images, each with M attributes. The dataset is denoted as $\mathbf{D} = \{\mathbf{X}, \mathbf{Y}\}$, where $\mathbf{X} = \{X_i\}_{i=1}^N$, and $\mathbf{Y} = \left\{\{y_i^j\}_{j=1}^M\right\}_{i=1}^N$. A traditional DMTL model for joint attribute estimation can be formulated by minimizing the regularization error function

$$\underset{\{W^j\}_{j=1}^M}{\arg\min} \sum_{j=1}^M \sum_{i=1}^N \mathcal{L}\big(y_i^j, \mathcal{F}(X_i, W^j)\big) + \gamma \Phi(W^j), \quad (1)$$

where $\mathcal{F}(\cdot, \cdot)$ is an attribute prediction function of the input X_i and weight vector W^j; $\mathcal{L}(\cdot, \cdot)$ is a prescribed loss function (*e.g.,* empirical error) between estimated values by $\mathcal{F}$ and the corresponding ground-truth values y_i^j; $\Phi(\cdot)$ is a regularization term which penalizes the complexity of weights, and γ is a regularization
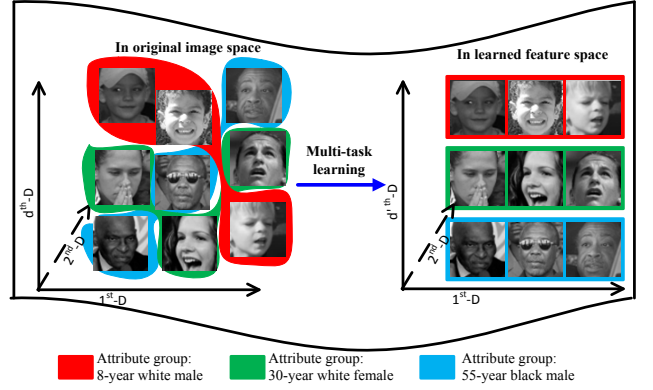
3. http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Fig. 4. The benefit of using MTL is that individual attribute groups which are not well separable from each other in the original image space could become separable in the feature space learned by MTL, leading to improved multi-attribute estimation accuracy.

parameter ($\gamma > 0$).

Given the objective function in (1), a straightforward approach is to learn multiple CNNs in parallel, one per attribute. Such an approach is not optimal because individual face attribute estimation tasks may share some common features. This is supported by the fact that off-the-shelf CNN features learned for face recognition were directly used for face attribute estimation [34]. However, the formulation in (1) does not explicitly enforce a large portion of feature sharing during MTL. To this end, we reformulate the DMTL for multi-attribute estimation as

$$\underset{W_c, \{W^j\}_{j=1}^M}{\arg\min} \sum_{j=1}^M \sum_{i=1}^N \mathcal{L}\big(y_i^j, \mathcal{F}(X_i, W^j \circ W_c)\big), \quad (2)$$
$$+ \gamma_1 \Phi(W_c) + \gamma_2 \Phi(W^j)$$

where W_c controls feature sharing among the face attributes, and W^j controls update of the shared features w.r.t. each face attribute. Specifically, as shown in Fig. 2, a face image is first projected to a high-level representation through a shared deep network (W_c) consisting of a cascade of complex non-linear mappings, and then refined by shallow subnetworks ($\{W^j\}_{j=1}^M$) towards individual attribute estimation tasks. The formulation in (2) makes it possible to explore the attribute correlations and learn a compact representation shared by various attributes. Figure 4 explains the benefit of jointly estimating multiple face attributes via MTL.

3.2 Heterogeneous Face Attribute Estimation

Although the above formulation of DMTL utilizes the attribute correlations in feature learning, the attribute heterogeneity still needs to be considered. Hetero-geneity of individual face attribute is ever present, but has not received sufficient attention. The reasons are two-fold: (i) many of the public-domain face databases are labeled with a single attribute,

the requirement of designing corresponding models becomes no longer urgent. (ii) many of the published methods choose to learn a separate model for each face attribute; model learning for individual attributes does not face the attribute heterogeneity problem.

We treat each of the heterogeneous attribute categories separately, but attributes within each category are expected to share feature learning and classification model to a larger extent. To accomplish this, the objective function in (2) is rewritten as

$$\arg\min_{W_c,\{W^j\}_{j=1}^M} \sum_{g=1}^{G}\sum_{j=1}^{M^g}\sum_{i=1}^{N} \lambda^g \mathcal{L}^g\left(y_i^j, \mathcal{F}(X_i, W^g \circ W_c)\right), \quad (3)$$
$$+\gamma_1\Phi(W_c)+\gamma_2\Phi(W^g)$$

where G is the number of heterogeneous attribute categories, and M^g is the number of attributes within each attribute category; λ^g balances the importance of each attribute category ($\lambda^g = 1$ by default); W^g refines the shared features w.r.t. each of the heterogeneous attribute categories. $\mathcal{L}^g(\cdot,\cdot)$ is a prescribed loss function for each of the heterogeneous attribute categories, given the estimated values by $\mathcal{F}$ and the corresponding ground-truth y_i^j.

Grouping a large number of attributes into a few heterogeneous categories depends on prior knowledge. Here, we consider face attribute heterogeneities in terms of *data type and scale* (*i.e.,* ordinal vs. nominal) [24] and *semantic meaning* (*i.e.,* holistic vs. local) [25], and explain our category-specific modeling for these heterogeneous attribute categories.

Nominal vs. ordinal attributes. Nominal attributes have two or more classes (values), but there is no intrinsic ordering among the categories [24]. For example, race is a nominal attribute having multiple classes, such as Black, White, Asian, etc., and there is no intrinsic ordering of these values (classes). We handle nominal attributes in a classification scheme, and choose to use the cross-entropy loss [49]

$$\mathcal{L}^{g_{\mathbb{N}}} = -\sum_{j=1}^{M^{\mathbb{N}}}\sum_{i=1}^{N}\sum_{k=1}^{C^j}\mathbf{1}(y_i^j,\hat{y}_i^{j,k})\log p(\hat{y}_i^{j,k}), \quad (4)$$

where

$$p(\hat{y}_i^{j,k}) = \frac{e^{\hat{y}_i^{j,k}}}{\sum_{k=1}^{C^j}e^{\hat{y}_i^{j,k}}} \quad (5)$$

is the softmax function, $\hat{y}_i^{j,k}$ is the k-th element (C^j elements in total) of the prediction by $\mathcal{F}(X_i, W^{g_{\mathbb{N}}} \circ W_c)$ for the estimation of the j-th nominal attribute; y_i^j is the ground-truth attribute; and $\mathbf{1}(a,b)$ outputs 1 when $a = b$, and 0 otherwise.

The difference between ordinal attribute and nominal attribute is that ordinal attribute has a clear ordering of its variables. For example, age of a person, typically ranging from 0 to 100, is an ordinal attribute.. Actually, age is not only ordinal but also interval [24].

We handle ordinal attributes in a regression scheme, and choose to use the Euclidean loss

$$\mathcal{L}^{g_{\mathbb{O}}} = \sum_{j=1}^{M^{\mathbb{O}}}\sum_{i=1}^{N}\|y_i^j - \hat{y}_i^j\|_2^2, \quad (6)$$

where $\hat{y}_i^{j,k}$ is the prediction by $\mathcal{F}(X_i, W^{g_{\mathbb{O}}} \circ W_c)$.

Holistic vs. local attributes. While attributes such as age, gender, and race describe the characteristics of the whole face, attributes such as pointy nose and big lips, mainly describe the characteristics of local facial components. Therefore, the optimal features used for estimating holistic and local attributes could be different. Such attribute heterogeneities in semantic meaning can also be modeled by the proposed DMTL, *e.g.,* using multiple holistic attribute subnetworks and multiple per-component attribute subnetworks. Both holistic and local attributes could further consist of nominal and ordinal categories. So, a joint consideration of nominal vs. ordinal and holistic vs. local heterogeneities leads to four types of subnetworks: holistic-nominal, holistic-ordinal, local-nominal, and local-ordinal. The choice of the loss function for each type of subnetwork is still determined by whether the subnetwork is nominal or ordinal.

The proposed category-specific modeling differs the proposed approach from [33], which manually classifies the binary attributes into 9 groups based on the attribute location (*e.g.,* eyes, nose, and mouth), but does not consider the heterogeneity in terms of data type and scale. In addition, each of the 9 attribute groups in [33] was handled equally via regression.

3.3 Implementation Details

As shown in Fig. 2, the proposed DMTL network mainly consists of a deep network for shared feature learning, and variable number of shallow subnetworks for category-specific feature learning.

Network structure. For the shared feature learning network, we use a modified AlexNet network (5 convolutional (Conv.) layers, 5 pooling layers, 2 fully connected (FC) layers [11]) with a batch normalization (BN) layer inserted after each of the Conv. layers. Each of the category-specific feature learning networks contains two FC layers, and is connected to the last FC layer of the shared network.

Network input. Since the proposed DMTL network is designed to handle heterogeneous attribute categories, we revise the network input format, and use two fields to represent each attribute label, *i.e.,* $y_i^j = (val, cat)$, where val and cat denote the attribute value and category, respectively (see Fig. 5). After we introduced an attribute category field, the order of the input attributes no longer matters; the corresponding attribute values used for computing individual losses (cross-entropy and Euclidean) can be easily determined based on the attribute category fields, e.g.,

Fig. 5. Revised network input for the label information, with each attribute taking two fields: one for the attribute value and the other for attribute category. Here, 'N' and 'O' represent the nominal and ordinal attributes, respectively.

$cat = N$ for choosing the nominal attribute values, and $cat = O$ for choosing the ordinal attribute values. This is an advantage of the proposed approach over existing methods like [33], [46].

Network training. We perform stochastic gradient descent (SGD) [11] with weight decay [50] to jointly optimize the weights of both the shared network and the category-specific subnetworks in an end-to-end way. Specifically, given two types of loss functions (for ordinal and nominal attributes), the derivatives used for updating $W^{g_{\mathbb{N}}}$ and $W^{g_{\mathbb{O}}}$ can be calculated as

$$\frac{\partial \mathcal{L}^{g_{\mathbb{N}}}}{\partial W^{g_{\mathbb{N}}}} = \left(y_i^j - p(\hat{y}_i^{j,k})\right) X_i^{\mathrm{T}}, \qquad (7)$$

and

$$\frac{\partial \mathcal{L}^{g_{\mathbb{O}}}}{\partial W^{g_{\mathbb{O}}}} = \left(y_i^j - (W^{g_{\mathbb{O}}})^{\mathrm{T}} X_i\right) X_i^{\mathrm{T}}. \qquad (8)$$

The sum of (7) and (9) is used for updating W_c. Finally, the network weights are updated as

$$\begin{aligned} \Delta W^{g_{\mathbb{N}}} &= \eta \frac{\partial \mathcal{L}^{g_{\mathbb{N}}}}{\partial W^{g_{\mathbb{N}}}}, \\ \Delta W^{g_{\mathbb{O}}} &= \eta \frac{\partial \mathcal{L}^{g_{\mathbb{O}}}}{\partial W^{g_{\mathbb{O}}}}, \\ \Delta W^c &= \eta \left(\frac{\partial \mathcal{L}^{g_{\mathbb{N}}}}{\partial W^{g_{\mathbb{N}}}} + \frac{\partial \mathcal{L}^{g_{\mathbb{O}}}}{\partial W^{g_{\mathbb{O}}}}\right), \end{aligned} \qquad (9)$$

where η is the learning rate. Random initialization is used for all the weights in network pre-training.

4 EXPERIMENTAL RESULTS

4.1 Databases

As summarized in Section 2, the widely used public-domain face database for attribute estimation include: MORPH II [36], CelebA [23], LFWA [23], and ChaLearn LAP [51] and FotW [44]. Besides these databases, we also constructed the LFW+ database (LFW augmented by $2,466$ images of children) with three heterogeneous attributes, labeled for each face image via the Amazon Mechanical Turk (MTurk) crowdsourcing[4].

MOROH II. MORPH is a large database of mugshot images, each with associated metadata containing three heterogeneous attributes: age (ordinal),

TABLE 2
Summary of the 40 face attributes provided with the CelebA database [23].

Attr. Idx.	Attr. Def.	Attr. Idx.	Attr. Def.
1	5 O'ClockShadow	21	Male
2	ArchedEyebrows	22	MouthSlightlyOpen
3	BushyEyebrows	23	Mustache
4	Attractive	24	NarrowEyes
5	BagsUnderEyes	25	NoBeard
6	Bald	26	OvalFace
7	Bangs	27	PaleSkin
8	BlackHair	28	PointyNose
9	BlondHair	29	RecedingHairline
10	BrownHair	30	RosyCheeks
11	GrayHair	31	Sideburns
12	BigLips	32	Smiling
13	BigNose	33	StraightHair
14	Blurry	34	WavyHair
15	Chubby	35	WearEarrings
16	DoubleChin	36	WearHat
17	Eyeglasses	37	WearLipstick
18	Goatee	38	WearNecklace
19	HeavyMakeup	39	WearNecktie
20	HighCheekbones	40	Young

gender (nominal), and race (nominal). We investigate all the three attribute estimation tasks on MORPH Album2 (MORPH II) containing about $78K$ images of more than $20K$ subjects. Results on MORPH II are reported with a five-fold, subject-exclusive cross-validation protocol [16], [17].

CelebA. CelebA is a large-scale face attribute database [23] with more than $200K$ celebrity images of more than $10K$ identities, each with 40 attribute annotations (see Table 2). The images in this dataset contain large variations in pose, expression, race, background, etc., making it challenging for face attribute estimation. Additionally, since there are 40 attribute annotations, the CelebA database poses challenges to joint attribute estimation algorithms in terms of feature learning efficiency. Results on CelebA are reported following the protocol provided in [23].

LFWA. LFWA is another unconstrained face attributes database [23] with face images from the LFW database ($13,233$ images of $5,749$ subjects) [52], and the same 40 attribute annotations as in the CelebA database. Results on LFWA are reported following the protocol provided in [23].

ChaLearn LAP and FotW. The ChaLearn challenge series, started in 2011, has been very successful in promoting advances in visual or multi-modal analysis of people [53]. LAPAge2015 is an unconstrained face database for apparent age estimation released at ICCV 2015.[5] This database contains $4,699$ face images, each with an average age of the estimates by at least 10 different users. The database was split into $2,476$ images for training, $1,136$ images for validation, and $1,087$ images for testing [51]. Since the age information for the testing set was not available, we follow the protocol in [17], and report the results on the valida-

---

4. https://www.mturk.com

5. http://gesture.chalearn.org/2015-looking-at-people-iccv-challenge

tion set. The FotW database was created by collecting publicly-available images from the Internet, which contains two datasets, one for accessary classification, and the other for gender and smile classification. The FotW accessary dataset contains $5,651$, $2,826$, and $4,086$ face images for training, validation, and testing, respectively; each is annotated with seven binary accessory attributes (see Table 5 (a)). The FotW gender and smile dataset is composed of $6,171$, $3,086$, and $8,505$ face images for training, validation, and testing, respectively; each is annotated with ternary gender (male, female, and not sure) and binary smile attributes. We following the same testing protocols to report the results on FotW.

LFW+. We extended the LFW database [52] to study the joint attribute estimation (age, gender, and race) from unconstrained face images. Since the number of young subjects (*e.g.,* in the age group 0–20) in the LFW database is very small (only 209 subjects among the $5,749$ subjects according to the labels provided by MTurk workers), the LFW database was extended by collecting $2,466$ unconstrained face images of subjects in the age range 0–20 years using Google Images search service. Specifically, we first used the keywords such as "baby", "kid", and "teenager" to find about $5,000$ images of interest from Google Images. The Viola-Jones [54] face detector was then applied to generate a set of candidate faces. Finally, we manually removed false face detections as well as most of the subjects that appeared to be older than 20. The extended LFW database (LFW+) contains $15,699$ unconstrained face images of about $8,000$ subjects. For each face image, three MTurk workers were asked to provide their estimates of age, gender, and race. The apparent age is determined as the average of the three estimates, and the gender and race are determined by the majority vote rule. Results on LFW+ are reported with a five-fold, subject-exclusive cross-validation protocol.

These databases can be divided into three group based on the type of annotation method used: (i) databases with nominal and ordinal attributes (MORPH II and LFW+), (ii) databases with binary attributes (CelebA, LFWA and FotW), and (iii) databases with a single attribute (LAPAge2015). Example face images from the six databases are shown in Fig. 6. We can see that except for the MORPH II database, the other five databases mainly contain unconstrained face images. Evaluations of attribute estimation on such databases could provide insights of the system's performance under real application scenarios. In addition, we also evaluate the generalization ability of the proposed approach under *cross-database testing* scenarios[6].

6. In a cross-database testing, the attribute estimation method is trained on one face database, and tested on a different face database.

(a) Face images from the MORPH II database

(b) Face images from the LFW+ database

(c) Face images from the CelebA database

(d) Face images from the LFWA database

(e) Face images from the ChaLearn LAPAge15 and FotW databases

Fig. 6. Examples of face images with nominal and ordinal attributes from (a) MORPH database (total of $78K$ face images) [36], and (b) LFW+ database (total of $15K$ face images); face images with 40 binary attributes from (c) CelebA database (total of $200K$ face images) [23], and (d) LFWA database (total of $13K$ images) [23]; and face images from (e) ChaLearn LAPAge2015 and FotW databases (total of $4K$ and $30K$ face images) [51]. M/F and B/W in (a–b) denote the gender (male, female) and race (black, white) information, respectively.

4.2 Experimental Settings

For all the face images, we perform face and landmark detection using an open source SeetaFaceEngine[7], and normalize the face images into $256 \times 256 \times 3$ (height $\times$ width $\times$ channels) based on five facial landmarks (*i.e.,* two eye centers, nose tip, and two mouth corners). Unless otherwise stated, we pre-train our DMTL network on the CASIA-WebFace database [55], and then fine-tune this model on the training set of each individual database. We use a base learning rate of 0.0001, and reduce the learning rate to 10% every $100,000$ iterations. All the training and testing (except for our prototype system) are performed on a Nvidia Titan X GPU. For the baseline methods used in Sections 4.3, 4.4, and 4.5 for which the code is not available in the public domain, we directly report the results in their publications.

There is no constraint in the network architecture for the shared feature learning in our DMTL. We tried two networks (AlexNet [11] and GoogLeNet [56]) with varying depths for attribute estimation on

7. https://github.com/seetaface

TABLE 3
Estimation accuracies of the three heterogeneous
attributes (age, gender, and race) on the MORPH II
and LFW+ databases (in %).

Approach	MORPH II			LFW+		
	Age[2]	Gender	Race	Age[2]	Gender	Race
Guo and Mu [19]	3.92/70.0	98.5	99.0	NA	NA	NA
Yi et al. [20]	3.63/NA	98.0	99.1	NA	NA	NA
DIF [16]	3.8/75.0	97.6	99.1[3]	7.8/42.5[4]	94[4]	90[3,4]
DEX [17]	3.25/NA	NA	NA	NA	NA	NA
DEX [17][1]	2.68/NA	NA	NA	NA	NA	NA
Proposed	3.0/85.3	98.0	98.6	4.5/75.0	96.7	94.9

[1]The IMDB-WIKI database [17] was used for network pre-training. [2]Age estimation results are reported in terms of both mean absolute error (MAE) and the accuracy with a 5-year absolute error. [3]Only two race classes (White vs. other) were used in [16], but the proposed approach used three classes (Black, White, and Other). [4]Only the frontal face images in LFW were used in [16].

CelebA. The average accuracies of all the 40 attributes by AlexNet and GoogLeNet are 91.98% and 92.05%, respectively. The performance difference is minor, but AlexNet is much faster. Therefore, we choose to use AlexNet (with a few modifications as described in Section 3.3) for shared feature learning in our DMTL.

4.3 Nominal and Ordinal Face Attributes

The MORPH II and LFW+ databases, which contain age, gender, and race annotations, represent the scenario with heterogeneous attributes of nominal and ordinal. Table 3 lists the performance by the proposed approach and the state-of-the-art methods [16], [19], [20] on the MORPH II and LFW+ databases. Methods in [19] and [20] provided a joint estimation of three face attributes, but both methods used multi-label regression. Since the performance of [19] and [20] is not available on the LFW+ database, we only compare the proposed approach with [19] and [20] on the MORPH II database. While our results on gender and race estimations are comparable with [19] and [20], the proposed approach performs much better than [19] and [20] on the more challenging age estimation task (3.0 years MAE by the proposed approach vs. 3.92 and 3.63 years MAE by [19] and [20], respectively). The possible reason is that multi-label learning in [19] and [20] utilizes the same features for estimating different attributes, which may not be optimal. By contrast, the subnetworks in our approach can fine-tune the shared features to obtain better feature representation for individual attributes.

Another baseline we considered is DEX [17], which is not a multi-attribute estimation method, but reported the best known age estimation accuracy on MORPH II (3.25 years MAE). Under the same settings, our approach performs better than DEX [17], which suggests that by leveraging attribute correlations via MTL, our simple network can be as effective as a very deep VGG-16 network. This also indicates that MTL
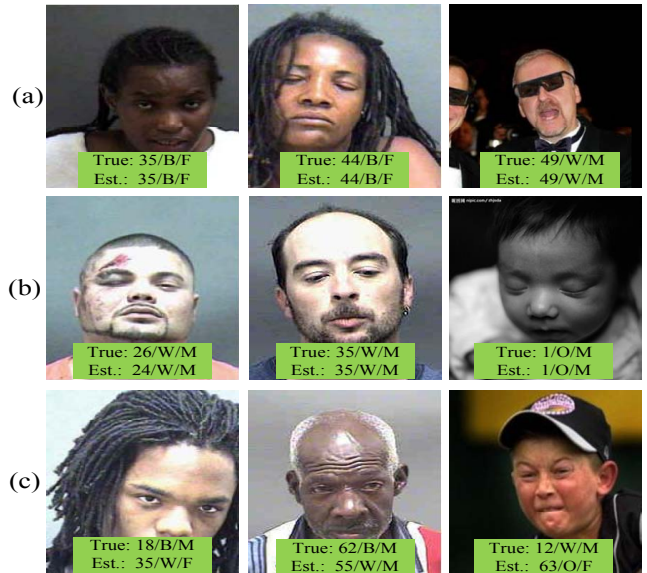

Fig. 7. Examples of (a,b) good and (c) poor estimates for age, gender, and race by the proposed approach on the MORPH II and LFW+ databases. 'm/n/l' denotes the age/race/gender information of each image, with 'M/F' denoting male/female, and 'W/O' denoting white/other, respectively.

could be a better choice than STL when multiple face attributes need to be jointly estimated.

Among the multi-attribute estimation methods, only DIF [16] reported their results on a subset of LFW with frontal face images. On this frontal subset, DIF [16] achieved 42.5% (@ 5-year AE), 94%, and 90% accuracies for age, gender, and race estimations, respectively. The proposed DMTL achieves 75.0% (@ 5-year AE), 96.7%, and 94.9% accuracies for age, gender, and race estimations, on the much larger LFW+ database with unconstrained face images.

Examples of correct and incorrect age, gender, and race estimates by the proposed approach on the MORPH II and LFW+ databases are shown in Figure 7. We find that the proposed approach is quite robust to pose and illumination variations. However, we also notice that the small number of young and old subjects in both the MORPH II and LFW+ databases can make the age and race estimation difficult.

4.4 Binary Face Attributes

In practice, it is relatively easy to annotate the presence of each attribute (binary attribute) than fine-grained annotations (e.g., nominal and ordinal). The CelebA, LFWA and FotW databases represent the scenario of joint estimation for multiple binary attributes. Binary attributes could be heterogeneous in terms of holistic vs. local (e.g., in CelebA and LFWA), but no longer heterogeneous in terms of nominal vs. ordinal. Therefore, we can handle binary attributes through holistic and local subnetworks with the same loss. Specifically, for the CelebA and LFWA databases, we use one holistic nominal subnetwork (for attributes:

TABLE 4

Attribute estimation accuracies (in %) for the 40 binary attributes (see Table 2) on the CelebA and LFWA databases by the proposed approach and the state-of-the-art methods [23], [27], [33], [34], [57]. The average accuracies of [27], [57], [23], [34], [33], and the proposed approach are 81%, 85%, 87%, 86.6%, 91% and 93%, respectively, on CelebA, and 74%, 81%, 84%, 84.7%, 86.0%, and 86.0%, respectively, on LFWA.

	Approach	Attribute index																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
CelebA	FaceTracker [27]	85	76	80	78	76	89	88	70	80	60	90	64	74	81	86	88	98	93	85	84
	PANDA [57]	88	78	86	81	79	96	92	85	93	77	94	67	75	86	86	88	98	93	90	86
	LNets+ANet [23]	91	79	90	81	79	98	95	88	95	80	97	68	78	84	91	92	99	95	90	87
	CTS-CNN [34]	89	83	87	82	79	96	94	87	93	79	95	70	79	87	88	89	99	94	91	87
	MCNN-AUX [33]	**95**	83	**93**	83	85	**99**	**96**	90	96	89	**98**	71	85	**96**	96	96	**100**	97	**92**	88
	Proposed	**95**	86	85	85	99	99	96	85	91	96	96	88	92	96	97	99	99	98	92	88
LFWA	FaceTracker [27]	70	67	67	71	65	77	72	76	88	62	78	68	73	73	67	70	90	69	88	77
	PANDA [57]	84	79	79	81	80	84	84	87	94	74	81	73	79	74	69	75	89	75	93	86
	LNets+ANet [23]	84	82	82	83	83	88	88	90	97	77	84	75	81	74	73	78	**95**	78	95	88
	CTS-CNN [34]	77	83	83	79	83	91	**91**	90	97	76	87	78	83	**88**	75	80	91	83	95	88
	MCNN-AUX [33]	77	82	**85**	80	83	92	90	**93**	97	81	**89**	79	**85**	85	77	82	91	83	**96**	88
	Proposed	80	**86**	82	**84**	92	93	77	83	92	**97**	89	**81**	80	75	**78**	92	86	**88**	95	89

	Approach	Attribute index																			
		21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
CelebA	FaceTracker [27]	91	87	91	82	90	64	83	68	76	84	94	89	63	73	73	89	89	68	86	80
	PANDA [57]	97	93	93	84	93	65	91	71	85	87	93	92	69	77	78	96	93	67	91	84
	LNets+ANet [23]	98	92	95	81	95	66	91	72	89	90	96	92	73	80	82	**99**	93	71	93	87
	CTS-CNN [34]	**99**	92	93	78	94	67	85	73	87	88	95	92	73	79	82	96	93	73	91	86
	MCNN-AUX [33]	98	**94**	**97**	87	96	76	**97**	77	**94**	95	**98**	93	84	84	90	**99**	**94**	87	**97**	88
	Proposed	98	**94**	**97**	90	97	78	**97**	78	**94**	96	**98**	94	85	87	91	**99**	93	89	**97**	90
LFWA	FaceTracker [27]	84	77	83	73	69	66	70	74	63	70	71	78	67	62	88	75	87	81	71	80
	PANDA [57]	92	78	87	73	75	72	84	76	84	73	76	89	73	75	92	82	93	86	79	82
	LNets+ANet [23]	**94**	82	92	81	79	74	84	80	85	78	77	91	76	76	94	88	**95**	88	79	86
	CTS-CNN [34]	**94**	81	94	81	80	75	73	83	**86**	82	82	90	77	77	94	90	**95**	90	**81**	86
	MCNN-AUX [33]	**94**	84	93	**83**	**82**	77	**93**	84	**86**	88	83	92	79	82	95	90	**95**	90	**81**	86
	Proposed	93	**86**	95	82	81	75	91	**84**	85	86	80	92	79	80	94	92	93	**91**	81	87

#4, 14, 15, 19, 21, 26, 27, 32, and 40 in Table 2) and seven local nominal subnetworks (subnet1 for attributes #6, 7, 8, 9, 10, 11, 29, 33, 34, 36; subnet2 for attributes #2, 3, 5, 17, 24; subnet3 for attributes: #13, 28; subnet4 for attributes: #20, 30, 31, 35; subnet5 for attributes: #1, 12, 22, 23, 37; subnet6 for attributes: #16, 18, 25; subnet7 for attributes: #38, 39).

The results on CelebA and LFWA by the proposed approach and several state-of-the-art methods [23], [27], [33], [34], [57] are reported in Table 4. The proposed approach outperforms [23], [27], [34], [57] for most of the 40 face attributes on both the CelebA and LFWA databases. Comparisons with [23], [27], which used per attribute SVM classifiers, show superior performance of the proposed DMTL in jointly estimating multiple attributes. Our approach achieves similar accuracies to MCNN-AUX [33] on LFWA. The possible reason is that both methods tend to show overfitting on such a small training set of LFWA (6K images), leading to unsatisfactory results on the testing set. Given a larger training set such as CelebA (160K images), both methods are improved, but our method performs better than [33]. Figure 8 shows examples of good and poor attribute estimates by our approach on the CelebA database. Some of the poor estimates by the proposed approach are due to the inconsistencies in the provided attributes. For example, the first image in Fig. 8 (c) was labeled with both attribute #1 '5 o'Clock Shadow' and attribute #25 'No Beard'.

We also provide the results by STL, *i.e.*, training a separate AlexNet model for each face attribute. Since there are up to 40 face attributes in CelebA, we simply chose eight common attributes. Figure 9 shows that while STL may work well for a few attributes, overall the proposed DMTL performs much better than STL. It is not clear to what degree the attribute correlations were utilized in the published methods, but we checked the incorrect estimation results for attribute #38 ('WearNecklace') by our approach, and find that the number of males (attribute #21) satisfying this attribute is very small. This makes sense because males wear necklace much less often than females do.

For the two FotW datasets, since there is no clear attribute heterogeneity, either nominal vs. ordinal or holistic vs. local, we simply use a nominal subnetwork in our DMTL. Results by our approach and the state-of-the-art methods (reported in [44]) for accessory classification, and smile and gender classification on FotW are shown in Table 5. Our approach achieves an average accuracy of 94.0% for accessory classification, which is better than the best result (93.5%) by SIAT MMLAB [58]. For smile and gender classification, our approach achieves an average accuracy of 86.1%, which is lower than the top-2 methods (SIAT MMLAB [58] and IVA NLPR [59]) reported in [44]. However, while methods in [58], [59] used very deep networks like VGG [31], our approach only uses a network with complexity similar to AlexNet.

These results indicate that our DMTL can make

Fig. 8. Examples of (a,b) good and (c) poor esti-mates for the 40 binary face attributes by the proposed approach on the CelebA databases. 'm/n' denotes (the number of correct estimates)/(total number of at-tributes) for each face image.

use of attribute correlations to achieve better attribute estimation results. In addition, our DMTL is effective in handling attribute heterogeneities, *e.g.,* nominal vs. ordinal and holistic vs. local, by using different num-ber and different types of subnetworks for category-specific feature learning.

4.5 Single Face Attribute

Some application scenarios may require the esti-mate of a single attribute, *e.g.,* age estimation used for preventing minors from purchasing alcohol or cigarette from camera-enabled vending machines.[8] The LAPAge2015 database represents such a scenario with age estimation from unconstrained face images. Following [17], we train our DMTL network without and with pre-training on the IMDB-WIKI database[9]. Both the MAE and ϵ-error ($\varepsilon = 1 - \exp(-\frac{(y-\mu)^2}{2\sigma^2})$) are used to measure the performance. When the proposed DMTL network is trained from scratch using only the training set of the LAPAge2015 database, it achieves an ϵ-error of 0.449, and 5.2 years MAE. This result is comparable to the 8-th best method among all the 115 participants of LAPAge2015 [51]. If we pre-train our DMTL approach using the IMDB-WIKI database, and then fine-tune the model on the training set of the LAPAge2015 database, the proposed approach achieves an ϵ-error of 0.289. This result is comparable to the best age estimation result (an ϵ-error of 0.265) on LAPAge2015, which was reported by DEX in [17]. However, while DEX [17] is an ensemble of 20 VGG-16

8. http://newsfeed.time.com/2011/12/27/scram-kids-new-ven ding-machine-dispenses-pudding-to-adults-only

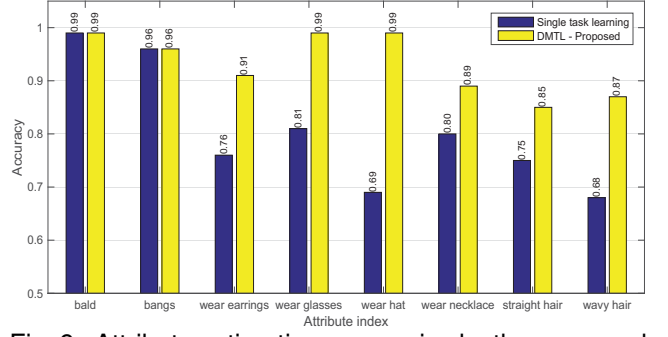9. https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki

Fig. 9. Attribute estimation accuracies by the proposed DMTL approach and the baseline single-task learning (STL) method for eight common attributes from the CelebA database. On average, MTL works much better than STL using networks with a similar depth.

TABLE 5
Accuracies (in %) of the proposed approach and the state-of-the-art methods (reported in [44]) for (a) accessory classification, and (b) smile and gender classification on the FotW datasets.

Method	Hat	Headband	Glasses	Earrings	Necklace	Tie	Scarf	Avg.
SIAT MMLAB	94.7	94.9	94.7	91.0	88.2	97.3	93.7	93.5
IVA NLPR	92.2	95.1	93.9	85.3	87.4	96.1	94.0	92.0
Proposed	94.7	96.1	96.1	89.1	89.5	97.4	95.1	94.0

(a) FotW - accessory classification

Method	Smile	Gender	Avg.
SIAT MMLAB	92.7	85.8	89.3
IVA NLPR	91.5	82.5	87.0
VISI.CRIM	90.2	82.1	86.1
SMILELAB NEU	90.0	81.5	85.7
Proposed	84.9	87.3	86.1

(b) FotW - smile and gender classification

networks, the proposed approach is a single network with complexity similar to AlexNet.

Figure 10 shows examples of good and poor age estimates by our approach for age estimation on the LAPAge2015 database. Loss of face details due to overexposure of the image is responsible for some poor age estimates (see Fig. 10 (c)).

4.6 Generalization Ability

The data distribution in the system deployment en-vironment can be different from that during model development. We evaluate the generalization ability of the proposed approach with cross-database testing on the MORPH II, LFW+, CelebA, and LFWA databases.

Specifically, cross-database testing of age, gender, and race estimation between the MORPH II and LFW+ databases is performed by training our ap-proach on LFW+ and testing it on MORPH II, and vice versa. Similarly, cross-database testing of 40 face attribute estimation is performed between the CelebA and LFWA databases. The attribute estimation results with cross-database testing are shown in Table 6. As expected, cross-database testing performance is lower

Fig. 10. Examples of (a,b) good and (c) poor age estimations by the proposed approach on the LAPAge2015 database. 'm/n' denotes the (estimated age)/(ground-truth apparent age) respectively, for each face image.

than intra-database testing. But, we believe these accuracies (not reported in other published studies) are still quite good. Image distribution (age, gender, race, pose, expression, occlusion, and illumination) differences between the MORPH II and LFW+ databases are responsible for the drop in performance. For example, there are more males than females in the MORPH II (84%) and LFW+ (74%) databases, and the race distributions in MORPH II and LFW+ are significantly biased towards black (75%) and white (79%), respectively. The reasons for the drop in performance of the cross-database testing between CelebA and LFWA are similar. In addition, although both CelebA and LFWA contain face images of individuals such as celebrities, public figures, etc., face images in LFWA were selected by using the Viola-Jones face detector [54]. Thus, face images in LFWA have relatively small variations in pose, expression, occlusion, etc. Finally, the LFWA database only contains $13,233$ face images, making it difficult to train a robust CNN model.

We also combine the age and race information in our LFW+ database with the 40 attributes in the LFWA database, leading to a new LFWA database (LFWA+) with 42 attributes.[10] Since LFW+ and LFWA were constructed independently, crowd-sourcing methods used in the two databases could be different. We evaluate the proposed approach using LFWA+ to see its effectiveness in handling both attribute heterogeneity and different annotation sources. We used a five-fold, subject-exclusive cross-validation protocol. The proposed approach using nominal and ordinal subnetworks achieves 4.8 years

10. The gender information is already provided with the LFWA database.

TABLE 6
Cross-database testing accuracies (in %) of the proposed approach using MORPH II and LFW+, as well as CelebA and LFWA.

Database		Accuracy		
Training	Testing	Age[1]	Gender	Race
MORPH II	MORPH II	3.0/85.3	98.0	98.6
LFW+	MORPH II	7.0/60.1	89.0	85.7
LFW+	LFW+	4.5/75.0	96.7	94.9
MORPH II	LFW+	9.4/52.6	77.4	70.5
		Avg. accuracy of 40 attributes		
CelebA	CelebA		93.0	
LFWA	CelebA		70.2	
LFWA	LFWA		86.0	
CelebA	LFWA		73.0	

[1] Age estimation results are reported using both mean absolute error (MAE) and the accuracy with a 5-year AE.

MAE for age estimation, 91% accuracy for race classification, and 83% accuracy for the average of the other 40 attributes. Compared with the results on the separate LFW+ and LFWA databases (see Tables 3 and 4), the accuracies on the combined LFWA+ database are slightly lower. This experiment indicates that different sources of annotations may pose additional challenges to face attribute estimation, but the proposed approach still achieves quite good results in such a challenging scenario.

4.7 Computational Cost

We summarize the computational cost of the proposed approach and several state-of-the-art methods on the MORPH II, CelebA, LFWA, and LAPAge2015 databases. For feature learning and joint attribute estimation, the proposed approach takes 8ms on a Titan X GPU, and 35ms on an Intel Core I7 3.6 GHz CPU. Only a few of the state-of-the-art methods reported their computational costs using machines with different GPUs and CPUs. We still report their computational costs for reference in Table 7. Compared with the methods that reported computational cost on GPU, the proposed approach is much faster than state-of-the-art methods except for MS-CNN [20]. However, our approach works much better than [20] for age estimation on MORPH II. Compared with the best method on LAPAge2015 (DEX [17]), the proposed approach is about 10 times faster than a single VGG-16 model used in [17]. For the computational cost on CPU, the proposed approach is faster than the rKCCA method in [19] and MS-CNN in [20]. A prototype implementation of the proposed approach is able to run in real-time (about 16 fps) on the CPU (Intel Core I7 3.6 GHz) of a commodity desktop machine (see a demo at: http://ddl.escience.cn/f/FOrq), which suggests that our approach can be used in wide application scenarios.

5 CONCLUSIONS

This paper presents a deep multi-task learning approach for joint estimation of multiple face at-

TABLE 7
Computational cost of different face attribute
estimation methods.

Method	Face Detection	Feature learning	Prediction
MS-CNN [20] (GPU)	N/A	2ms[1]	
LNet+ANet [23] (GPU)	35ms	14ms	N/A
DEX [17] (GPU)	N/A	$\sim$ 75ms[2] with VGG-16	
Proposed (GPU)	5ms[1]	8ms[2]	
rKCCA [19] (CPU)	N/A	N/A	1,600ms[3]
MS-CNN [20] (CPU)	N/A	200ms[4]	
Proposed (CPU)	25ms[5]	35ms[5]	

[1,2,3,4,5]The computational costs are profiled on a Tesla K20 GPU, Titan X GPU, Intel Core2 2.1 GHz CPU, Intel Core I3 2.4 GHz CPU, Intel Core I7 3.6 GHz CPU, respectively.

tributes. Compared to the existing approaches, the proposed approach models both attribute correlation and attribute heterogeneity in a single network, allowing shared feature learning for all the attributes, and category-specific feature learning for heterogeneous attributes. The LFW+ database was created by augmenting the LFW database with 2,466 images of subjects in 0-20 years of age. This helps evaluate the proposed approach on a wider age range. Our approach performs well on large and diverse databases (including MORPH II, LFW+, CelebA, LFWA, LAPAge2015, and FotW), which replicate several representative scenarios such as face databases with multiple heterogeneous attributes and a single attribute. Generalization ability of the proposed approach is studied under the cross-database testing scenarios. Experimental results show that the proposed approach generalizes well to the unseen scenarios. The cross-database testing highlights the importance of training database in real-world face attribute estimation systems. Additionally, the ambiguity of annotation for some attributes would be another issue that makes it difficult to learn efficient models. One possible solution to this issue could be integrating noisy label refining with deep multi-task learning.

ACKNOWLEDGMENTS

REFERENCES

[1] H. Han and A. K. Jain, "Age, gender and race estimation from unconstrained face images," Michigan State University, Tech. Rep. MSU-CSE-14-5, 2014.
[2] F. Wang, H. Han, S. Shan, and X. Chen, "Multi-task learning for joint prediction of heterogeneous face attributes," in *Proc. IEEE FG*, 2017, pp. 173–179.
[3] S. Z. Li and A. K. Jain, Eds., *Handbook of Face Recognition*, 2nd ed. New York: Springer, 2011.
[4] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Proc. IEEE WACV*, 2009, pp. 1–8.
[5] J. Kim and V. Pavlovic, "Attribute rating for classification of visual objects," in *Proc. IEEE ICPR*, 2012, pp. 1611–1614.
[6] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum, "Scalable face image retrieval with identity-based quantization and multireference reranking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1991–2001, Oct. 2011.
[7] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 1962–1977, Oct. 2011.
[8] S. Xia, M. Shao, and Y. Fu, "Toward kinship verification using visual attributes," in *Proc. IEEE ICPR*, 2012, pp. 549–552.
[9] G. Qi, C. Aggarwal, Q. Tian, H. Ji, and T. S. Huang, "Exploring context and content links in social media: A latent space method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 850–862, May 2012.
[10] G. Qi, X. Hua, and H. Zhang, "Learning semantic distance from community-tagged media collection," in *Proc. ACM MM*, 2009, pp. 243–252.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
[12] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Nov. 2010.
[13] B. Ni, Z. Song, and S. Yan, "Web image and video mining towards universal and robust age estimator," *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1217–1229, Dec. 2011.
[14] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.
[15] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014.
[16] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.
[17] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vision*, pp. 1–14, Aug. 2016.
[18] Y. Sun, M. Zhang, Z. Sun, and T. Tan, "Demographic analysis from biometric data: Achievements, challenges, and new frontiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017 (To appear).
[19] G. Guo and G. Mu, "A framework for joint estimation of age, gender and ethnicity on a large database," *Image Vision Comput.*, vol. 32, no. 10, pp. 761–770, Oct. 2014.
[20] D. Yi, Z. Lei, and S. Z. Li, "Age estimation by multi-scale convolutional network," in *Proc. ACCV*, 2014, pp. 144–158.
[21] P. Luo, X. Wang, and X. Tang, "A deep sum-product architecture for robust facial attributes analysis," in *Proc. IEEE ICCV*, 2013, pp. 2864–2871.
[22] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 918–930, May 2016.
[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE ICCV*, 2015, pp. 3730–3738.
[24] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.
[25] S. Samangooei, B. Guo, and M. S. Nixon, "The use of semantic human description as a soft biometric," in *Proc. BTAS*, 2008, pp. 1–7.
[26] G. W. Cottrell and J. Metcalfe, "EMPATH: Face, emotion, and gender recognition using holons," in *Proc. NIPS*, 1990, pp. 564–571.
[27] N. Kumar, P. N. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *Proc. ECCV*, 2008, pp. 340–353.
[28] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proc. IEEE CVPR Workshops*, 2015, pp. 1–9.

[29] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proc CVPR*, 2016, pp. 5375–5384.

[30] M. Uricár, R. Timofte, R. Rothe, J. Matas, and L. V. Gool, "Structured output SVM prediction of apparent age, gender and smile from deep features," in *Proc. IEEE CVPR Workshops*, 2016, pp. 730–738.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv e-prints*, Sept. 2014.

[32] M. Ehrlich, T. J. Shields, T. Almaev, and M. R. Amer, "Facial attributes classification using multi-task representation learning," in *Proc. IEEE CVPR Workshops*, 2016, pp. 752–760.

[33] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network for attribute classification," *ArXiv e-prints*, Apr. 2016.

[34] Y. Zhong, J. Sullivan, and H. Li, "Face attribute prediction using off-the-shelf CNN features," in *Proc. ICB*, 2016, pp. 1–7.

[35] X. Geng, Z.-H. Zhou, and K. Smith-Miles, "Automatic age estimation based on facial aging patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2234–2240, Dec. 2007.

[36] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *IEEE Proc. FGR*, 2006, pp. 341–345.

[37] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. FG*, 2000, pp. 46–53.

[38] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[39] E. Mäkinen and R. Raisamo, "An experimental comparison of gender classification methods," *Pattern Recogn. Lett.*, vol. 29, no. 10, pp. 1544–1556, Jul. 2008.

[40] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. WACV*, 2015, pp. 534–541.

[41] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the FG-NET ageing database," *IET Biometrics*, vol. 5, pp. 37–46, May 2016.

[42] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[43] C. Li, J. Zhu, and J. Chen, "Bayesian max-margin multi-task learning with data augmentation." in *Proc. ICML*, 2014, pp. 415–423.

[44] S. Escalera, M. T. Torres, B. Martínez, X. Baró, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. Oliu, M. A. Bagheri, and M. Valstar, "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proc. CVPR Workshop*, 2016, pp. 706–713.

[45] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[46] S. Li, Z.-Q. Liu, and A. B. Chan, "Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network," *Int. J. Comput. Vision*, vol. 113, no. 1, pp. 19–36, May 2015.

[47] A. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov. 2015.

[48] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: database and results," *Image Vision Comput.*, vol. 47, pp. 3–18, Mar. 2016.

[49] P. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[50] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. NIPS*, 1991, pp. 950–957.

[51] S. Escalera, J. Fabian, P. Pardo, X. Baró, J. Gonzàlez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon, "ChaLearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proc. ICCV Workshops*, Dec. 2015, pp. 243–251.

[52] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Tech. Rep. 07-49, 2007.

[53] S. Escalera, X. Baró, H. J. Escalante, and I. Guyon, "Chalearn looking at people: A review of events and resources," *ArXiv e-prints*, Jan. 2017.

[54] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004.

[55] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *ArXiv e-prints*, 2014.

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE CVPR*, 2015, pp. 1–9.

[57] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. IEEE CVPR*, 2014, pp. 1637–1644.

[58] K. Zhang, L. Tan, Z. Li, and Y. Qiao, "Gender and smile classification using deep convolutional neural networks," in *Proc. CVPR Workshop*, 2016, pp. 739–743.

[59] C. Li, Q. Kang, G. Ge, Q. Song, H. Lu, and J. Cheng, "DeepBE: Learning deep binary encoding for multi-label classification," in *Proc. CVPR Workshop*, 2016, pp. 744–751.

Hu Han is an Associate Professor of the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). He received the B.S. degree from Shandong University, and the Ph.D. degree from ICT, CAS, in 2005 and 2011, respectively, both in computer science. He was a Research Associate in the Department of Computer Science and Engineering at Michigan State University, and a visiting researcher at Google in Mountain View from 2011 to 2015. His research interests include computer vision, pattern recognition, and image processing, with applications to biometrics, forensics, law enforcement, and security systems. He is a member of the IEEE.

Anil K. Jain is a University Distinguished Professor in the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence (1991-1994). He served as a member of the United States Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. He has received Fulbright, Guggenheim, Alexander von Humboldt, and IAPR King Sun Fu awards. He is a member of the National Academy of Engineering and foreign fellow of the Indian National Academy of Engineering. He is a Fellow of the AAAS, ACM, IAPR, SPIE, and IEEE.

Fang Wang received the B.S. degree from Tianjin University in 2014, and the M.S. degree from ICT, CAS in 2017. Her research interests include computer vision and pattern recognition.

Shiguang Shan is a Professor of ICT, CAS, and the Deputy Director with the Key Laboratory of Intelligent Information Processing, CAS. His research interests cover computer vision, pattern recognition, and machine learning. He has authored over 200 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition. He was a recipient of the China's State Natural Science Award in 2015, and the Chinas State S&T Progress Award in 2005 for his research work. He has served as the Area Chair for many international conferences, including ICCV'11, ICPR'12, ACCV'12, FG'13, ICPR'14, and ACCV'16. He is an Associate Editor of several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING, the Computer Vision and Image Understanding, the Neurocomputing, and the Pattern Recognition Letters. He is a Senior Member of IEEE.

Xilin Chen is a Professor of ICT, CAS. He has authored one book and over 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He served as an Organizing Committee/Program Committee member for over 50 conferences. He was a recipient of several awards, including the China's State Natural Science Award in 2015, the China's State S&T Progress Award in 2000, 2003, 2005, and 2012 for his research work. He is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, a Leading Editor of the Journal of Computer Science and Technology, and an Associate Editor-in-Chief of the Chinese Journal of Computers. He is a Fellow of the China Computer Federation (CCF), and IEEE.