Deep Multi-Task Learning for Joint Prediction of Heterogeneous Face Attributes

Fang Wang[1,2], Hu Han[1], Shiguang Shan[1,2,3], and Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]CAS Center for Excellence in Brain Science and Intelligence Technology
fang.wang14@vipl.ict.ac.cn, {hanhu, sgshan, xlchen}@ict.ac.cn

Abstract— **Face attribute prediction has important applications in video surveillance, face retrieval, and social media. While a number of methods have been proposed for face attribute prediction, most of them did not explicitly consider the attribute correlation and heterogeneity during feature learning. In this paper, we propose a *Deep Multi-Task Learning (DMTL)* network to jointly learn multiple models; each addresses the prediction of one category of homogenous attributes. Specifically, we group the heterogeneous face attributes into two categories (*i.e.,* nominal and ordinal), and design corresponding prediction models. At the same time, we use a convolutional neural network (CNN) for early stage feature learning, which is shared by all the attributes. Experiments on the public-domain MORPH II, CelebA, and LFWA databases show that the proposed approach outperforms the state of the art in joint face attribute prediction, and has good generalization ability.**

I. INTRODUCTION

Face attribute prediction has wide applications in video surveillance [1] [2], face retrieval [3] [4] [5], and social media [6] [7]. Predicting face attributes from images in the wild is very challenging, because of the complicated face appearance variations: pose, lighting, and occlusion. A number of methods have been proposed for face attribute prediction in the past few years; however, most of them either did not explicitly consider the attribute correlations [8] or only estimate very few attributes [9] [8] [10] [4].

As shown in Fig. 1, when a person has long hair or wears lipstick, the person is more likely to be a woman. Such observations indicate that individual attributes should not be handled separately. Instead, it is more reasonable to jointly learn prediction models by leveraging the attribute correlations [11]. Face attribute correlation was exploited in [10] for face landmark detection, where auxiliary tasks such as gender, pose, smiling and wearing glasses were used to assist in facial landmark detection. Therefore, the goal is not to optimize the feature learning of all the tasks. In face attribute prediction, it is necessary to learn feature representations that are discriminative for all the attributes.

This study aims to jointly learn multiple attribute prediction models under a multi-task learning (MTL) framework. However, this task is non-trivial: (1) individual attributes are heterogeneous, requiring different handling strategies, such as nominal attribute like race group and ordinal attribute like age or age groups, (2) the number of attributes can be large in real applications requiring efficient framework to handle individual attributes simultaneously.

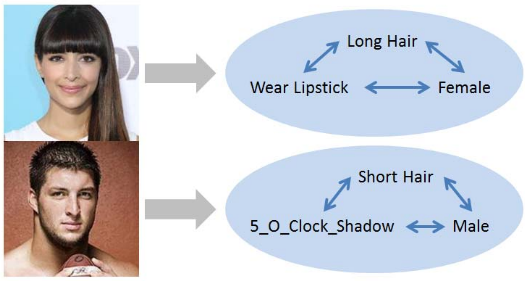


Fig. 1. Individual face attributes are often heterogeneous in semantic concept, but they also have potential correlations indicting their co-occurrence in a face image. Such co-occurrence can be utilized in learning face attribute prediction models.

In this paper, we propose a *Deep Multi-Task Learning (DMTL)* network to jointly optimize the entire face attribute prediction. Specially, we design different loss functions for different task categories to allow the errors of related tasks to be back-propagated jointly for shared feature learning. Multi-task learning is new, but to our knowledge, this is the first attempt to optimize heterogeneous tasks together in a single network by adopting different schemes.

The contributions of this work are: (1) a deep MTL approach for jointly handling various attributes by considering both attribute correlation and heterogeneity; (2) shared feature learning for individual attributes, which provides a good balance between speed and accuracy; (3) promising generalization ability under cross-database testing scenarios.

II. RELATED WORK

This work is related to face attribute prediction and multi-task learning. In this section, we briefly review the most recent literature on these two topics.

A. Face Attribute Prediction

Face attribute prediction methods can be categorized into two main groups: global and local attribute prediction. Global facial attributes (*e.g.,* gender) are often independent of local facial components. The features used for global facial attribute prediction are usually extracted from the whole

H. Han is the corresponding author.

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.30

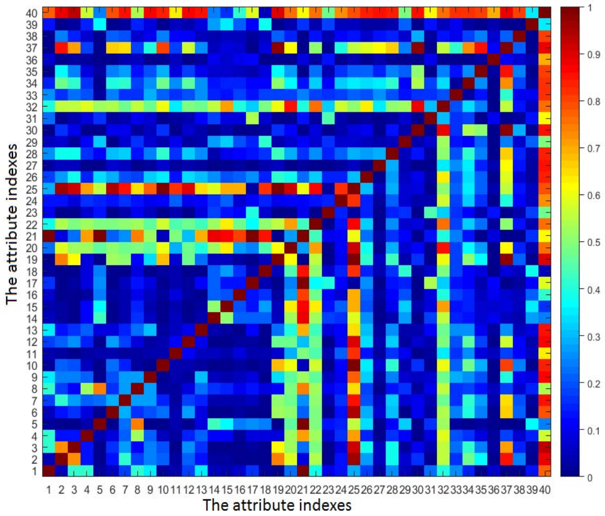


Fig. 2. Pair-wise co-occurrence matrix of the 40 attributes in the CelebA database [9].

Attr. Idx.	Attr. Def.	Attr. Idx.	Attr. Def.
1	5 O'ClockShadow	2	ArchedEyebrows
3	BushyEyebrows	4	Attractive
5	BagsUnderEyes	6	Bald
7	Bangs	8	BlackHair
9	BlondHair	10	BrownHair
11	GrayHair	12	BigLips
13	BigNose	14	Blurry
15	Chubby	16	DoubleChin
17	Eyeglasses	18	Goatee
19	HeavyMakeup	20	HighCheekbones
21	Male	22	OuthSlightlyOpen
23	Mustache	24	NarrowEyes
25	No Beard	26	OvalFace
27	PaleSkin	28	PointyNose
29	RecedingHairline	30	RosyCheeks
31	Sideburns	32	Smiling
33	StraightHair	34	WavyHair
35	Wear Earrings	36	Wear Hat
37	Wear Lipstick	38	Wear Necklace
39	Wear Necktie	40	Young

TABLE I

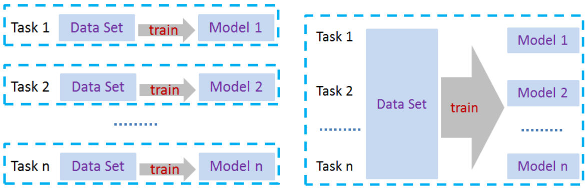
SUMMARY OF THE 40 ATTRIBUTES IN THE CELEBA DATABASE [9].

image. Sharif *et al.* [12] used OverFeat network to obtain a generic image representation for attribute recognition; Han *et al.* [13] extracted demographic informative features via a boosting algorithm, and then employed a hierarchical approach consisting of between-group classification and within-group regression for demographic estimation from face images. By contrast, local facial attribute prediction detect facial parts and extract features from these parts [14] [15] [4]. Kumar *et al.* [4] extracted hand-crafted features from ten face parts to predict face attributes; Zhang *et al.* [16] employed hundreds of poselets [17] to align human body parts to recognize human attributes. However, complicated facial appearance variations in the wild make the face detection and component localization difficult, leading to inaccurate attribute prediction.

B. Multi-Task Learning

Compared with single-task learning, where each task is performed separately ignoring any correlations with other tasks, multi-task learning (MTL) is used for sharing knowledge while solving multiple correlated tasks simultaneously [18]. As a result, MTL often leads to better generalization ability than single-task learning approaches.

Existing MTL methods can be categorized into two groups: discovering the relationship between tasks and mining the related features between tasks. For example, Zhang and Schneider [19] aimed at discovering task correlations during MTL. Yang *et al.* [20] focused on exploring the feature correlations. Li *et al.* [21] considered both task and feature correlations, combining the discriminative power of max-margin and flexibility of Bayesian modeling. All of these approaches rely on applying regularization on the model's parameters during feature learning. Thus, choosing the appropriate regularization method becomes important for the final system accuracy.



(a) Single-task Learning (b) Multi-task Learning

Fig. 3. Illustration of the differences between (a) single-task learning, and (b) multi-task learning.

Deep learning recently achieved great success in attribute prediction, due to their ability to learn compact and discriminative features [9] [22] [21]. Deep model is well suited for multi-task learning because features that are learned by convolutional neural networks (CNNs) are reported to be robust and generic in many computer vision tasks. As a result, a number of approaches seek to combine MTL with deep learning. Abdulnab *et al.* [8] constructed CNN for each binary attribute to generate attribute-specific feature representations, and then applied multi-task learning on the features to predict their attributes. Liu *et al.* [9] cascaded two CNNs, *i.e.,* LNet and ANet, to predict face attributes in the wild. Hwang *et al.* [22] designed a method for feature sharing between object and attribute prediction tasks. Li *et al.* [21] combined max-margin and Bayesian methods in multi-task learning with data augmentation. Alternatively to these MTL approaches, multi-label learning methods were recently employed for classification of images with multiple label [23]. Although multi-label learning utilizes the attribute correlations in classification, but it still did not consider the heterogeneous properties of individual attributes.

Most of the existing methods learned a separate model for each attribute, which leads to high computational cost and do not take into account the attribute correlations. In fact, many facial attributes are often correlated with each other. For example, if one person has long hair and wears lipsticks, then the person is more likely to be a woman and vise versa. Therefore, considering the correlation between attributes would benefit the model learning. Figure 2 shows the co-

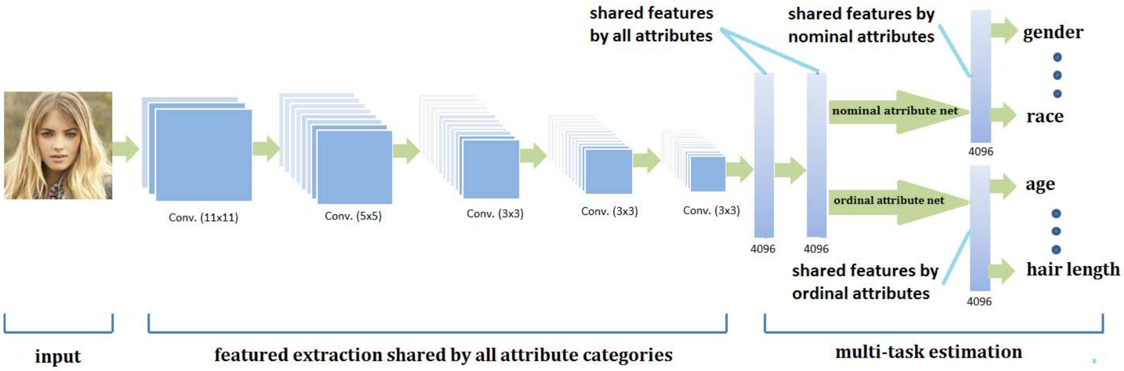


Fig. 4. Overview of the proposed deep multi-task learning network. The feature extraction stage shared by all attribute categories contains five convolutional layers and two fully connected (FC) layers. The multi-task estimation stage contains two sub-networks, each is designed to 'fine-tune' the shared features for attribute category-specific prediction.

occurrence matrix of the 40 facial attributes in the CelebA database [9]. From the co-occurrence matrix, We can clearly see that a number of attributes have strong correlations (elements with red color in Fig. 2). For example, attribute #1 (5 O'Clock Shadow) and attribute #21 (Male), as well as attribute #19 (Heavy Makeup) and #37 (Wear Lipstick) have strong correlations. These observations are consistent with the actual situation. Based on these observations, in this work we propose to simultaneously predict multiple attributes by exploring attribute correlation and also taking into account the attribute heterogeneity. Considering the success of CNNs in many vision tasks, and their natural advantage of shared feature learning, we design our MTL attribute prediction model based on CNNs.

III. Proposed Approach

A. Overview

Suppose we have a total of T facial attributes and the training data for the t-th attribute is denoted as (x^t, y^t), where $t = 1, 2, \ldots, T$. An attribute prediction model based on traditional MTL methods and CNNs can be formulated to minimize:

$$\arg\min_{w^t} \sum_{t=1}^{T} \Big(\sum_{i=1}^{N} l\big(y_i^{*t}, f(x_i^t; w^t)\big) + \phi(w^t) \Big), \qquad (1)$$

where N is the number of training samples, f is an attribute predict function, $f(x_i^t; w^t)$ is the prediction result given an input x_i^t and the parameter w_t; y_i^{*t} is the ground-truth attribute for x_i^t. $l(\cdot)$ is the loss function that is designed in particular to handle heterogeneous attributes. $\phi(w^t)$ is the regularization term to penalize the complexity of weights.

Through the aforementioned formulation, we can see that face attribute prediction model based on traditional MTL methods does not take into account the heterogeneous properties of individual attributes. Instead, all the attributes are treated equally. Such a scheme is not optimal because attributes like age and hair color have different characteristics. While age has ordinal values, hair color has nominal categories. Therefore, it is reasonable to design different models to handle different categories of attributes. Based on this analysis, we extend the attribute prediction model based on (1), and obtain the following formulation

$$\arg\min_{w^{g,j}} \sum_{g=1}^{G} \sum_{j=1}^{T_g} \Big(\sum_{i=1}^{N} l(y_i^{*g,j}, f(x_i^{g,j}; w^{g,j})) + \phi(w^{g,j}) \Big), \qquad (2)$$

where G is the total number of categories for the heterogeneous attributes, T_g is the number of attribute in the g-th attribute category. Thus, $G \times T_g = T$, where T is mentioned above. In this work, we group various face attributes into two main categories: nominal attribute and ordinal attribute.

Nominal attribute. Nominal attribute is one that has two or more categories, but there is no intrinsic ordering among the categories. For example, gender is a nominal attribute having two classes (male vs. female), and there is no intrinsic order. Hair color is also a nominal attribute having a number of classes, such as black, brown, gray, and red, and there is no agreed way to order these classes. Therefore, it is more reasonable to handle nominal attribute in a classification scheme. Under the CNN framework, we use softmax as the classification loss.

Ordinal attribute. The difference between ordinal attribute and nominal attribute is that ordinal attribute has a clear ordering of its variables. For example, suppose you have an attribute, hair length, with three categories (short, medium and long). For example, ages are ordered, and typically ranges from 0 to 100. Therefore, it is more reasonable to handle ordinal attribute in a regression scheme. We utilize Euclidean distance as the regression loss in our deep multi-task network.

Based on the above analysis, our MTL face attribute prediction model based on (2) can be written as:

$$\begin{aligned} \arg\min_{w^r, w^c} & \sum_{r=1}^{T_r} \sum_{i=1}^{N} \lambda_r l_r\big(y_i^{*r}, f(x_i^r; w^r)\big) \\ & + \sum_{c=1}^{T_c} \sum_{i=1}^{N} \lambda_c l_c\big(y_i^{*c}, f(x_i^c; w^c)\big) \end{aligned} \qquad (3)$$

where λ_r and λ_c balance the importance of different attribute categories (here, nominal and ordinal), and by default we use equal importance, *i.e.,* $\lambda_r = \lambda_c = 1$. In fact, we tried different values of λ_r and λ_c. Experiments show that different values have little effect on the results, thus we choose the same value between λ_r and λ_c.

By replacing the loss functions in (3) with their specific definitions, we get

$$\begin{aligned}\arg\min_{w^r, w^c} \frac{1}{2} \sum_{r=1}^{T_r} \sum_{i=1}^{N} \lambda_r \| y_i^{*r} - f(x_i^r; w^r) \|^2 \\ - \sum_{c=1}^{T_c} \sum_{i=1}^{N} \sum_{k=1}^{M} \lambda_c \mathrm{B}\{y_i^{kc} == y_i^{*c}\} log\big(p(y_i^{kc} | x_i^c; w^c)\big)\end{aligned} \tag{4}$$

where the first term denotes the Euclidean loss, and the second term denotes the cross-entropy loss. In the second term, M is the number of classes of each attribute, *e.g.,* gender has two classes: male and female; $\mathrm{B}\{\cdot\}$ is the indicator function, so that $\mathrm{B}\{\text{a true statement}\} = 1$, and $\mathrm{B}\{\text{a false statement}\} = 0$.

However, learning a deep convolutional network for each attribute category may lead to high computational cost. Therefore, as shown in Fig. 4, we design an early-stage feature learning network that is shared by all the attribute categories. Attribute category-specific sub-networks are connected after the last FC layer of the shared network for feature fine-tuning. This way, we are able to optimize the feature learning of heterogeneous attributes and avoid high computational cost.

B. Implementation Details

As shown in Fig. 4, the proposed Deep Multi-task Learning (DMTL) network has five convolutional layers and two fully connected (FC) layers at the stage of featured extraction shared by all attribute categories. Each convolutional layer is followed by a max pooling layer. The DMTL prediction stage has two sub-network, one is designed for nominal attribute prediction and the other is designed for ordinal attribute prediction. The attribute category specific sub-networks are connected to the last FC layer of the shared feature extraction network.

By learning a sequence of non-linear mappings, the feature leaning network shared by all the attributes projects a given face image x^0 to higher-level representation x^l gradually.

$$x^0 \xrightarrow{\sigma((w_1^s)^{\mathrm{T}} x^0)} x^1 \xrightarrow{\sigma((w_2^s)^{\mathrm{T}} x^1)} \dots \xrightarrow{\sigma((w_l^s)^{\mathrm{T}} x^{l-1})} x^l \tag{5}$$

Here, $\sigma(\cdot)$ and w^{s_l} represent the non-linear activation function and weight parameters.

We perform stochastic gradient descent to update the weights both the shared network and the category-specific sub-networks. For example, the weight matrix of regression task is updated by $\triangle w^r = \eta \frac{\partial l_r}{\partial w^r}$ and $\triangle w^c = \eta \frac{\partial l_c}{\partial w^c}$, where η is the learning rate ($\eta = 0.0001$ in our implementation). In summary, the update of the parameters can be performed as follows:

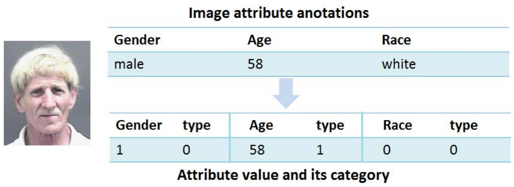


Image attribute anotations

Gender	Age	Race
male	58	white

Gender	type	Age	type	Race	type
1	0	58	1	0	0

Attribute value and its category

Fig. 5. Revised network input where each attribute has two fields: one for the attribute value and the other for attribute category (*e.g.,* 0 for nominal and 1 for ordinal). Here, 0 and 1 represent nominal and ordinal attributes, respectively.

$$\frac{\partial l_r}{\partial w^r} = \big(y_i^r - (w^r)^{\mathrm{T}} x_i^r\big) x_i^r \tag{6}$$

$$\frac{\partial l_c}{\partial w^c} = \big(h(x_i^c; w^c) - y_i^{c'}\big)(x_i^c)^{\mathrm{T}} \tag{7}$$

where $h(x_i^c; w^c)$ is the output of softmax function and $y_i^{c'}$ is a column vector with the y_i^{*c}-th element being 1, and other elements being 0.

Since the two sub-networks are designed to handle nominal and ordinal attributes, respectively, we revise the network input format, and use two fields to represent each attribute, *i.e.,* $y^t = (y^t, g^t)$, where y^t and g^t denote the attribute value and category, respectively (see Fig. 5). Such an input format makes it possible to flexibly handle various attributes without concerning the attribute input order.

Different from the multi-task CNN model in [8], which constructed a deep convolutional model for each attribute, the proposed approach shared a deep convolutional model by all the attributes, followed by two category-specific sub-networks. Apparently, the proposed approach has much lower computational cost than the model in [8]. The proposed approach also differs from [24] in that while [24] clustered the attributes into groups based on feature similarity, such groups did not consider the heterogeneous characteristics, *e.g.,* between nominal and ordinal attributes.

IV. Experimental Results

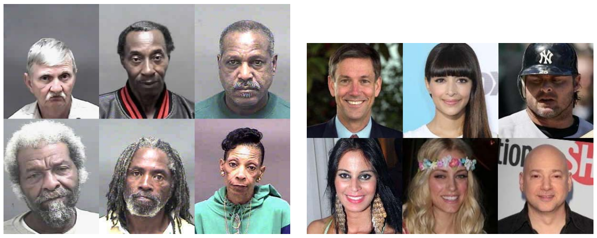
A. Experimental settings

The input RGB face image of the network is normalized to 256×256 after face and landmark detection using an open source SeetaFaceEngine[1]. In order to get a better initial parameter, we pre-train our DMTL network on the ImageNet 2012 database, and then fine-tune the initial model using the training set of each attribute database. We use a small base learning rate of 0.0001 and reduce the learning rate by 0.1 every 100,000 iterations. All the training and testing are performed on a Titan X GPU.

B. Selection of the Shared Network

For the early stage feature extraction shared by all the attributes (see Fig. 4), we tried different networks of varying depths, *e.g.,* AlexNet (7 layers) [25] and GoogLeNet (22

[1] `https://github.com/seetaface`



(a) MORPH II

(b) CelebA

Fig. 6. Example face images from the MORPH II and CelebA databases.

layers) [26] for attribute prediction on the CelebA database [9]. The CelebA database has about 200,000 face images, each is annotated with up to 40 binary (with or without) attributes (see Table I). Figure 6(b) shows some example face images of CelebA. The average prediction accuracies for 40 attributes on the CelebA database by AlexNet and GoogLeNet are 91.98% and 92.05%, respectively. The performance difference between AlexNet and GoogLeNet is minor. Therefore, we choose to use AlexNet for the early-stage shared feature learning, which is much simpler and faster than GoogLeNet for real applications.

C. Demographic Estimation on MORPH II

We first evaluate the proposed approach on the MORPH II database [27], which is a longitudinal face database containing about 75,000 unique images of more than 13,000 individuals. Each face image in MORPH II is provided with two types of heterogeneous attributes: nominal attributes such as gender and race, and ordinal attribute such as age. The chronological ages of the entire database range from 16 to 77 with a median age of 33. Gender has two classes of male and female. Race has three classes, *i.e.,* Black, White, and other races. We use a five-fold cross validation protocol following the state of the art methods, such as [13]. Figure 6(a) shows some example images of the MORPH II database.

As shown in Table II, the proposed approach achieves 85%, 97.9%, and 96.4% accuracies for the estimations of age (at a 5-year absolute error), gender, and race, respectively. As a comparison, the state of the art method used three separate SVM models for age, gender, and race estimation. The accuracies for age, gender, and race estimation reported by [13] are 78%, 97.6%, and 99.1%, respectively. We can see that by using multi-task learning with attribute category-specific prediction models, the proposed approach improves the accuracy of the most challenging age estimation task (see Fig. 7), while achieving almost the same accuracies for gender and race estimation. Considering the attribute correlations and heterogeneous properties leads to good performance of the proposed method.

D. Attribute Prediction on CelebA

We also evaluate the proposed approach on another public-domain face attribute database named CelebA [9]. The CelebA database was used in a number of the state of the art methods such as [9], we use the same training and testing protocols of [9]. Images of 80% in CelebA (about 160,000 images) are used to fine-tune, images of 10% are used as validation data, and another 10% are used as testing data. All of these images are randomly selected.

Method	Estimation accuracy (in %)		
	Age (at 5-yr AE)	Gender	Race
Han *et al.* [13]	75.0	97.6	99.1 (2 classes)
DMTL	**85.3**±0.6	**98**±0.044	96.6±0.28

TABLE II

AGE, GENDER AND RACE ESTIMATION PERFORMANCE ON MORPH II BY THE PROPOSED APPROACH AND A STATE OF THE ART METHOD [13].

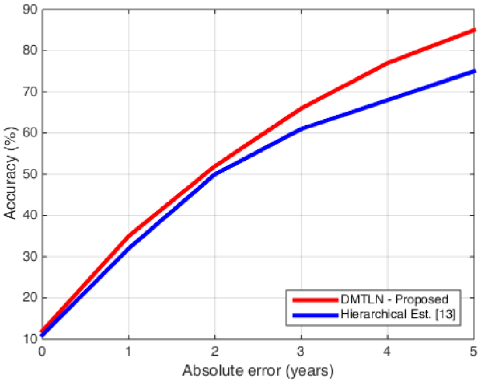


Fig. 7. Comparison with the state of the art method [13] for age estimation performance on MORPH II in terms of cumulative score.

We compare the proposed approach with a state of the art method, LNets+ANet [9] on CelebA. The attribute prediction performance by the proposed approach and [9] is reported in Fig. 8 (attributes #1-40). The correspondence between the attribute index and the attribute description can be found in Table I.

The average accuracy of all the 40 attributes achieved by our method is 92% which outperforms the average accuracy (87%) by LNets+ANet [9]. If we look at the individual attribute accuracies, the proposed approach outperforms L-Nets+ANet [9] in 36 of the 40 attributes. For example, the prediction accuracies of bangs (#7 in Fig. 8), gray hair (#11 in Fig. 8), oval face (#26 in Fig. 8), and wearing necklace (#38 in Fig. 8) of LNets+ANet [9] are 68%, 84%, 66%, and 71%, respectively. Be contract, the proposed method achieves 79%, 96%, 78%, and 89% accuracies, respectively. Our method outperforms LNets+ANet [9] by nearly 10% for these attributes. This result can support the importance of considering the attribute correlations and heterogeneous properties.

We also notice that several other methods such as FaceTracer [28], PANDA-w [16], and PANDA-l [16] also reported results on face images in the wild. Since LNets+ANet [9] reported better result than all of these methods on the CelebA database, we compare our method with [9]. These comparisons show that the proposed DMTL with shared feature learning and category-specific learning can handle heterogeneous attributes more effectively.

Another straightforward baseline is to train a separate CNN model for each attribute. Since there are up to 40 attributes, we simply choose eight common attributes and train eight separate CNN models. At the same time, we

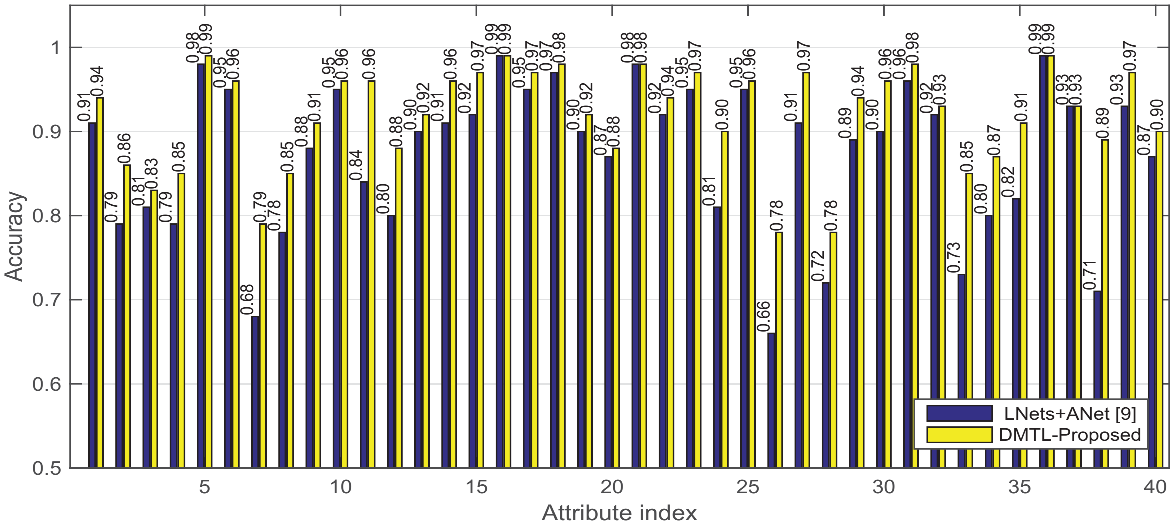


Fig. 8. Prediction accuracies for all the 40 attributes on the CelebA database by the proposed DMTL and a state of the art method, LNets+ANet [9].

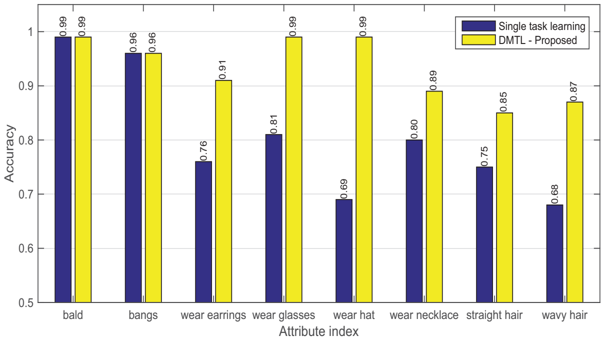


Fig. 9. Attribute prediction accuracies by the proposed DTML approach and the baseline single-task learning method for eight common attributes from the CelebA database.

also train DMTL method for these eight attributes. Again, as shown in Fig. 9, the proposed DMTL method outperforms the separate CNN models. Thus, considering attribute correlation and heterogeneity leads to better face attribute prediction accuracy. To better understand how the proposed DMTL network learns attribute correlation, we visualize the responses of the last FC layer of the shared AlexNet network. We randomly choose about 2,000 images from CelebA, each containing both attribute #1 (5 O'Clock Shadow) and attribute #21 (Male). Figure 9 shows the response values of the 4,096 nodes in the last FC layer of AlexNet. We can see that a few nodes have much stronger responses than the other nodes indicating that some nodes are learned to better capture the correlation between '5 O'Clock Shadow' and 'Male' attributes.

E. Generalization Ability

Generalization ability of an approach is one of the key factors in real applications. We evaluate the generalization ability of the proposed approach using the CelebA database and another face attribute database named LFWA [9], which has about 13,000 images, and is also annotated with the same 40 attributes.

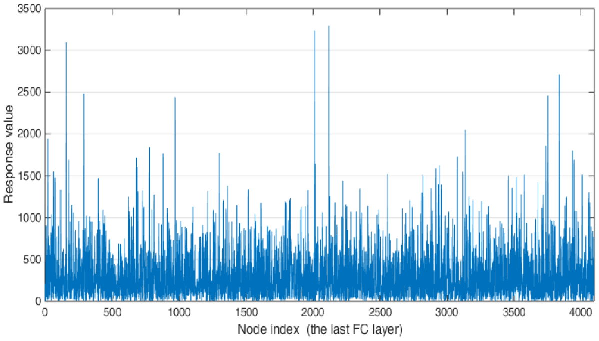


Fig. 10. Response values by the noes of the last FC layer of the shared network show that a few nodes have much stronger responses than the other nodes for images with both '5 O'Clock Shadow' and 'Male' attributes.

We first train our model using CelebA alone, and test our model on the testing set of LFWA defined in [9]. We denote this model as 'DMTL-CrossDB'. We also build another model by fine-tuning 'DMTL-CrossDB' model using the training set of LFWA defined by [9], and denote this model as 'DMTL-IntraDB'. We compare the accuracies of these to models with the intra-database performance ('LNets+ANet-IntraDB [9]') of [9] on the testing set of LFWA. The results are shown in Table III. We can see that although cross-database testing is much more challenging than intra-database testing, the proposed approach still achieves very promising accuracies (DMTL-CrossDB) for the 40 attributes on the LFWA database. These results suggest that the proposed DMTL network is able to learn more general and representative features than the methods that do not consider the attribute heterogeneous property and correlations.

Given an aligned face image, the proposed DMTL takes 5ms to extract features and predict 40 attributes on a Titan X GPU. By contrast, the LNets+ANet approach [9] requies 14ms to extract features. The proposed approach has large potential in real-world applications.

Attr. Idx.	LNets+ ANet [9]	DMTL Cross/IntraDB	Attr. Idx.	LNets+ ANet [9]	DMTL Cross/IntraDB
1	84%	60%/75%	21	94%	91%/93%
2	82%	76%/82%	22	82%	74%/82%
3	83%	66%/79%	23	82%	90%/93%
4	83%	59%/81%	24	81%	41%/79%
5	88%	91%/92%	25	79%	78%/78%
6	88%	87%/91%	26	74%	49%/75%
7	75%	66%/77%	27	84%	49%/91%
8	81%	56%/83%	28	80%	31%/81%
9	90%	86%/ 92%	29	85%	45%/85%
10	97%	97%/97%	30	78%	80%/86%
11	74%	84%/86%	31	77%	70%/80%
12	77%	68%/81%	32	91%	83%/90%
13	82%	48%/77%	33	76%	60%/79%
14	73%	65%/75%	34	76%	57%/79%
15	78%	64%/78%	35	94%	89%/94%
16	95%	91%/92%	36	88%	87%/90%
17	78%	77%/82%	37	95%	93%/94%
18	84%	85%/88%	38	88%	82%/89%
19	95%	94%/95%	39	79%	72%/81%
20	88%	83%/87%	40	86%	73%/85%

TABLE III

Generalization ability test of the proposed approach under a cross-database (CrossDB) testing scenario. As a comparison, the intra-database (IntraDB) testing results by a state-of-the-art method (LNets+ANet [9]) and our method are also provided.

V. Conclusions and Future Work

Face attribute prediction is challenging due to the nature of their heterogeneous characteristics. We solve this problem by proposing a deep multi-task learning network, which consists of an early stage shared feature learning for all the attributes, followed by per category feature learning, *e.g.,* nominal and ordinal attributes. Compared with the existing methods, the proposed approach considers both attribute correlation and heterogeneity, leading to balanced network speed and accuracy. Experiments on the public-domain databases (MORPH II, CelebA, and LFWA) show the effectiveness of the proposed approach in predicting heterogeneous attributes in a network.

In the future work, we would like to investigate illumination and pose normalization method [29], as well as automatic attribute category grouping method [24] for efficient attribute prediction. Additionally, attribute provides a middle-level representation independent of modalities, and thus has large potential in sketch to image matching [30], [31].

ACKNOWLEDGMENT

This research was partially supported by Natural Science Foundation of China (grant Nos. 61672496 and 61390511), and 973 Program (grant No. 2015CB351802). H. Han gratefully acknowledges the support of NVIDIA Corporation with the donation of the Titan X GPU used for his research.

References

[1] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk. Attribute-based people search in surveillance environments. In *Proc. WACV*, pages 1–8, 2009.

[2] J. Kim and V. Pavlovic. Attribute rating for classification of visual objects. In *Proc. ICPR*, pages 1611–1614, 2012.

[3] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum. Scalable face image retrieval with identity-based quantization and multireference reranking. *IEEE Trans. PAMI*, 33(10):1991–2001, Oct. 2011.

[4] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE Trans. PAMI*, 33(10):1962–1977, Oct. 2011.

[5] S. Xia, M. Shao, and Y. Fu. Toward kinship verification using visual attributes. In *Proc. ICPR*, pages 549–552, 2012.

[6] 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. PAMI*, 34(5):850–862, May 2012.

[7] G. Qi, X. Hua, and H. Zhang. Learning semantic distance from community-tagged media collection. In *Proc. ACM MM*, pages 243–252, 2009.

[8] A. Abdulnabi, G. Wang, J. Lu, and K. Jia. Multi-task CNN model for attribute prediction. *IEEE Trans. Multimedia*, 17(11):1949–1959, Nov. 2015.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. ICCV*, pages 3730–3738, 2015.

[10] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *Proc. ECCV*, pages 94–108, 2014.

[11] J. Roth and X. Liu. On the exploration of joint attribute learning for person re-identification. In *Proc. ACCV*, pages 673–688, 2014.

[12] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *Proc. CVPR Workshops*, pages 806–813, 2014.

[13] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Trans. PAMI*, 37(6):1148–1161, Jun. 2015.

[14] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proc. CVPR*, pages 955–962, 2013.

[15] J. Chung, D. Lee, Y. Seo, and C. D. Yoo. Deep attribute networks. In *Proc. NIPS Workshop*, 2012.

[16] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proc. CVPR*, pages 1637–1644, 2014.

[17] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *Proc. ICCV*, pages 1543–1550, 2011.

[18] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[19] Y. Zhang and J. G. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. In *Proc. NIPS*, pages 2550–2558, 2010.

[20] M. Yang, Y. Li, and Z. Zhang. Multi-task learning with gaussian matrix generalized inverse gaussian model. In *Proc. ICML*, pages 423–431, 2013.

[21] Chengtao Li, Jun Zhu, and Jianfei Chen. Bayesian max-margin multi-task learning with data augmentation. In *Proc. ICML*, pages 415–423, 2014.

[22] S. Hwang, F. Sha, and K. Grauman. Sharing features between objects and their attributes. In *Proc. CVPR*, pages 1761–1768, 2011.

[23] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino. Matrix completion for weakly-supervised multi-label image classification. *IEEE Trans. PAMI*, 37(1):121–135, Jan. 2015.

[24] N. Yousefi, M. Georgiopoulos, and G. C. Anagnostopoulos. Multi-task learning with group-specific feature space sharing. In *Proc. ECML PKDD*, pages 120–136, 2015.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, pages 1097–1105, 2012.

[26] 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. CVPR*, pages 1–9, 2015.

[27] K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Proc. FGR*, pages 341–345, 2006.

[28] N. Kumar, P. N. Belhumeur, and S. Nayar. Facetracer: A search engine for large collections of images with faces. In *Proc. ECCV*, pages 340–353, 2008.

[29] H. Han, S. Shan, X. Chen, and W. Gao. A comparative study on illumination preprocessing in face recognition. *Pattern Recognition*, 46(6):1691–1699, Jun. 2013.

[30] H. Han, B. F. Klare, K. Bonnen, and A. K. Jain. Matching composite sketches to face photos: A component-based approach. *IEEE Trans. IFS*, 8(1):191–204, Jan. 2013.

[31] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain. The FaceSketchID system: Matching facial composites to mugshots. *IEEE Trans. IFS*, 9(12):2248–2263, Dec. 2014.