Multi-Task Deep Relative Attribute Learning for Visual Urban Perception

Weiqing Min, *Member, IEEE*, Shuhuan Mei, Linhu Liu, Yi Wang, and Shuqiang Jiang, *Senior Member, IEEE*

***Abstract*—Visual urban perception aims to quantify perceptual attributes (e.g., safe and depressing attributes) of physical urban environment from crowd-sourced street-view images and their pairwise comparisons. It has been receiving more and more attention in computer vision for various applications, such as perceptive attribute learning and urban scene understanding. Most existing methods adopt either (i) a regression model trained using image features and ranked scores converted from pairwise comparisons for perceptual attribute prediction or (ii) a pairwise ranking algorithm to independently learn each perceptual attribute. However, the former fails to directly exploit pairwise comparisons while the latter ignores the relationship among different attributes. To address them, we propose a Multi-Task Deep Relative Attribute Learning Network (MTDRALN) to learn all the relative attributes simultaneously via multi-task Siamese networks, where each Siamese network will predict one relative attribute. Combined with deep relative attribute learning, we utilize the structured sparsity to exploit the prior from natural attribute grouping, where all the attributes are divided into different groups based on semantic relatedness in advance. As a result, MTDRALN is capable of learning all the perceptual attributes simultaneously via multi-task learning. Besides the ranking sub-network, MTDRALN further introduces the classification sub-network, and these two types of losses from two sub-networks jointly constrain parameters of the deep network to make the network learn more discriminative visual features for relative attribute learning. In addition, our network can be trained in an end-to-end way to make deep feature learning and multi-task relative attribute learning reinforce each other. Extensive experiments on the large-scale Place Pulse 2.0 dataset validate the advantage of our proposed network. Our qualitative results along with visualization of saliency maps also show that the proposed network is able to learn effective features for perceptual attributes.**

I. Introduction

The urban visual appearance plays a central role in shaping human perception to their surrounding urban environment. Recently, more studies [1], [2], [3] resorted to computer vision methods to build the correlation between visual appearance of urban environment and urban attributes (e.g., safety, education and wealth), namely visual urban perception. For example, Naik *et al.* [1] developed a computer vision method including image region segmentation, feature extraction and attribute prediction to measure changes in the physical appearance of neighborhoods from street scene images. Gebrua *et al.* [4] utilized deep learning to estimate the demographic makeup of neighborhoods from street images. Because of growing attention of visual urban perception in computer vision and its various applications, such as urban scene understanding [5], [6], [7], in this paper, we focus on visual perceptual attribute learning from urban scenes, and then exploit it for urban representation, analysis and understanding.

There are mainly two types of methods for visual urban perceptual attribute learning. The first type is based on binary attribute learning. For example, Arietta *et al.* [8] adopted a support vector regression model to predict the relationship between visual urban appearance and urban attributes (e.g., housing prices and population density). Recently, many crowd-sourcing platforms collect annotations of human judgments about urban perceptual attributes on image pairs (e.g., “Which place looks safer?”), such as Place Pulse 2.0 [2], as it is more natural to quantify the degree of urban perceptual attributes using relative judgements. Therefore, as the second type of methods, some works considered perceptual attribute learning as relative attribute learning, and adopted pairwise ranking methods, such as RankSVM [9], [2] and RankNet [10] for perceptual attribute learning.For example, Dubey *et al.* [2] proposed a Siamese-like neural network with rankSVM to directly predict pairwise comparisons for relative attributes.

However, existing efforts on attribute learning for visual urban perception are still limited in three aspects. First, most existing works learn each perceptual attribute separately, and do not consider the relationship among these perceptual attributes. However, in real-world scenarios, if multiple urban perceptual attributes are involved, they usually exhibit correlation among them. For example, for some perceptual attributes from the Place Pulse 2.0 dataset, “beautiful” is very relevant to “lively” while “boring” and “depressing” often have positive correlation. Quercia *et al.* [11] have found that the strongest affiliation happens to be between beauty and happiness attributes for urban scene images. The correlation among different urban perceptual attributes should also be modeled to enable more effective and accurate relative attribute learning via appropriate sharing of visual knowledge. Multi-task learning [12] utilizes the relatedness among different tasks to learns several tasks simultaneously for potential performance gain. The combina-

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018 and 61602437, in part by Beijing Natural Science Foundation under Grant L182054, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals. W. Min, S. Jiang and L.Liu is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China, and also with University of Chinese Academy of Sciences, Beijing, 100049, China email: sqjiang@ict.ac.cn, linhu.liu@vipl.ict.ac.cn. Y. Wang are with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China email:minweiqing@ict.ac.cn, yi.wang@vipl.ict.ac.cn. S. Mei is with Shandong University of Science and Technology, Shandong, 266590, China, and also an intern with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China shuhuan.mei@vipl.ict.ac.cn.

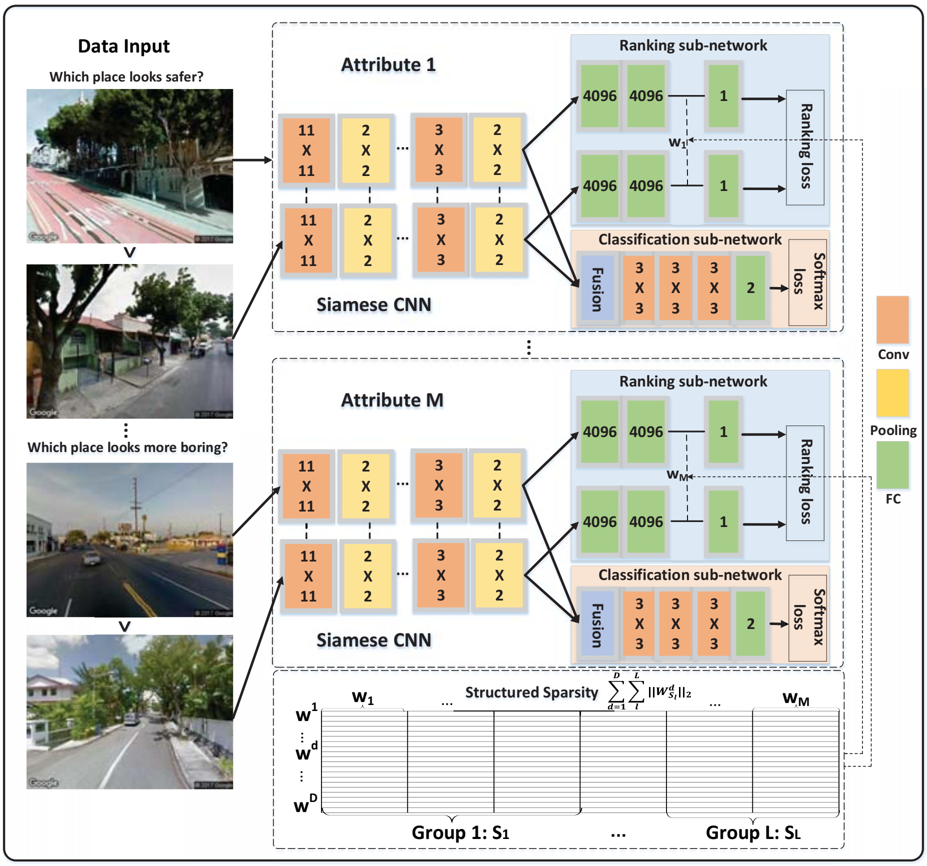


Fig. 1: The proposed multi-task deep relative attribute learning network: The input image pair with attribute comparison information is fed into the model. Each Siamese networks consists of two sub-networks, namely ranking sub-network and classification sub-network, and will predict one relative attribute. The relationship among different relative attributes are exploited through the structured sparsity to enable multi-task relative attribute learning.

tion between multi-task learning and relative attribute learning should be enabled for modeling such attribute correlation. Second, deep visual feature learning and multi-task relative attribute learning are neither used together nor tightly coupled. Deep learning's revolutionary advances in image representation have gained significant attention because of its powerful expressive capacity. The interaction between deep visual feature learning and multi-task relative attribute learning can help capture the most effective features for describing perceptual attributes of urban scene images, which in turn can promote relative attribute learning. Third, when combining deep learning with multi-task relative attribute learning for relative attributes, the ranking loss from the ranking network is generally adopted. However, existing deep learning models are pre-trained on certain large-scale datasets, and thus there still exists the difference between these datasets and task-specific datasets. In order to relieve this problem, introducing different types of losses can adjust the deep network from different aspects to boost its attribute learning capability for specific relative attribute learning task.

Taking these factors into consideration, given pairwise street view images and their comparisons, we propose a Multi-Task Deep Relative Attribute Learning Network (MTDRALN) (Fig. 1) to jointly learn all the perceptual attributes (e.g., safe and beautiful attributes) from crowdsourced street view images. Particularly, MTDRALN adopts multi-task Siamese networks with two types of sub-networks, namely classification sub-network and ranking sub-network for relative attribute learning, where each Siamese network will predict one relative attribute. In order to leverage grouping information of attributes, we utilize the structured sparsity to encourage feature sharing among related attributes and feature competition among unrelated attributes, leading to relative attribute learning in a multi-task way. In addition, combined with the ranking sub-network, the introduced classification sub-network can fine-tune the deep network from different aspect to boost its relative attribute learning capability. Our network can be trained in an end-to-end way to make visual feature learning and multi-task learning benefit each other. The proposed network is general, and we have demonstrated the detailed instantiation of MTDRALN to two frequently used learning-to-rank algorithms, i.e., RankNet [13] and RankSVM [14]. We then utilize learned attributes to identify region-level visual elements for perceptual attribute analysis.

Our main contributions in this paper can be summarized as follows:

- We propose a Multi-Task Deep Relative Attribute Learning Network (MTDRALN) to explore relative attribute correlation to learn all the relative attributes simultaneously via multi-task Siamese networks, where each Siamese network is responsible for each relative attribute learning.

- Our proposed MTDRALN consists of both classification sub-network and ranking sub-network and can be trained in an end-to-end fashion to enable positive interaction between deep feature learning and multi-task relative attribute learning, where multi-task relative attribute learning can leverage attribute grouping information to encourage relative attributes in the same group to share more knowledge and less knowledge among different groups.
- We conduct the experiments on a large perceptual attribute dataset Place Pulse 2.0, and quantitative as well as qualitative evaluation results have validated the effectiveness of our proposed perceptual attribute learning approach.

II. Related Work

Since our work is related to visual urban perception, attribute learning and multi-task learning, we will mainly review works related to these three parts in this section.

A. Visual Urban Perception

Visual urban perception aims to predict perceptual responses to scene images and play critical roles in urban scene understanding. Researchers have already developed a wide variety of visual perception models for predicting aesthetics [15], memorability [16], etc. For visual urban perception, in the earlier literature, Lynch *et al.* [17] investigated people's preference for certain environments and their aesthetic judgments of urban scenes, such as streets and buildings. More recently, some studies [18], [19], [2], [1], [3], [20] used human-annotated street view images to model the relation between visual elements of one city and high-level attributes. For example, Ordonez *et al.* [19] utilized classification and regression models to predict human perceptive attributes, such as uniqueness and wealth. Naik *et al.* [1] adopted the scene understanding algorithm Streetscore [21] to quantify the physical urban change. Gebru *et al.* [22], [4] used deep learning based methods to estimate socio-economic characteristics of regions from US cities using 50 million images of street scenes. In addition, our work is also relevant to scene recognition [23], location recognition [24] ,landmark mining [25] and location visualization [26]. Similar to [1], [2], in this work, we focus on outdoor street level urban images for perceptual attribute learning, representation and analysis. We further propose a multi-task deep relative attribute learning network to learn all the relative perception attributes simultaneously, and provide some insights based on learned perceptual attributes.

B. Attribute Learning

There are two kinds of attributes: binary attributes and relative attributes. Binary attributes are a visual property that appear or disappear in an image. Earlier works relied on hand-crafted features to predict binary attributes [27], [28]. Recently, more and more works used deep Convolutional Neural Networks (CNN) to learn the attribute representation [29], [30], [31]. For example, Zhang *et al.* [29] combined part-based models and deep learning for attribute classification. Abdulnabi *et al.* [30] proposed a multi-task learning method to simultaneously learn all the binary attributes using CNN.

In addition to binary attribute learning, there are also a lot of related works using hand-crafted features [32], [33], [34] or deep features [9], [10], [35] for relative attribute learning [36], which can capture richer semantic relationships. For example, You *et al.* [32] proposed an active-learning method for relative attribute learning using GIST and color histogram features. Chen *et al.* [33] proposed a multi-task relative attribute learning method, where hand-crafted features are first extracted from images for the model training. Benefiting from powerful expressive capability of CNN, more works combined CNN and ranking function learning for deep relative attribute learning [9], [10]. For example, Yang *et al.* [9] combined relative attribute learning and CNN in a unified framework to learn each attribute separately. Different from [9] using the RankSVM model, Souri *et al.* [10] proposed a deep relative learning method with RankNet for relative attribute learning. He *et al.* [37] considered attribute learning as a multi-label prediction problem to learn multiple relative attributes. Recently, Cruz *et al.* [35] proposed a DeepPermNet to learn ordered image sequences, which can be also used for relative attribute learning. Meng *et al.* [38] adopted the graph neural network for relative attribute learning. They both needed additional relative ordering of all classes for each attribute to build the sequence of images, where the length of image sequence is generally greater than or equals 4. However, such information is not always provided, such as the Place Pulse 2.0. In this dataset, only pairwise comparison is provided. There are also some works [39], [40], which focused on learning to both rank and localize relative attributes. Similar to [37], our work also uses CNN to learn relative attributes. However, combined with deep relative attribute learning, we further consider the correlation among different attributes based on natural attribute grouping to enable multi-task deep relative attribute learning. Furthermore, the proposed network is trained in an end-to-end way to make deep visual feature learning and multi-task relative attribute learning reinforce each other.

C. Multi-Task Learning

Multi-Task Learning (MTL) aims to improve the generalization performance of multiple tasks by utilizing the relativeness among them [12]. In attribute learning, most studies [33], [41], [30], [42], [43], [44] focused on binary attribute learning in a multi-task way via exploiting attribute relationships. Traditional methods adopt hand-crafted visual features for model learning. For example, Chen *et al.* [33] proposed a multi-task learning method to learn different attributes. Jaya *et al.* [41] modeled the relativeness of attributes for binary attribute learning through the prior of attribute grouping. Recently, CNNs are used for multi-task learning [43], [44], [42]. For example, Lu *et al.* [44] proposed a multi-task deep network architecture through a dynamic branching procedure for person attribute classification. Han *et al.* [42] proposed a deep multi-task face attribute learning network consisting of shared feature learning for all the attributes, and category-specific feature learning for heterogeneous attributes. Inspired by [41], we leverage the natural grouping of perceptual attributes in a multi-task fashion for deep relative attribute learning, and apply it to the visual urban perception task.

III. Our Method

A. The General Framework

Before we introduce our proposed network, we first give some notations. The attribute set $A = \{a_m\}_{m=1}^M$. For each attribute m, each tuple denotes $(\mathbf{x}_i, \mathbf{x}_j, y_{ij})$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $\mathbf{x}_j \in \mathbb{R}^D$ belong to one image pair from the training set $\{\mathbf{x}_i\}_{i=1}^N$, and D is the length of visual features. There are two kinds of image pairs for each attribute m: a set of ordered pairs $O_m = \{(i,j)\}$ and a set of un-ordered pairs $S_m = \{(i,j)\}$. $(i,j) \in O_m$ means that image i has a higher relative value than image j while $(i,j) \in S_m$ means that image i and j have similar relative value. $P_m = O_m \cup S_m$. $y_{ij} \in \{1, 0\}$. $y_{ij} = 1$ means that $(i,j) \in O_m$ while $y_{ij} = 0$ means that $(i,j) \in S_m$.

Some key ideas are exploited in our proposed network. First, for each pairwise comparison on two images from Place Pulse 2.0 in our experiment, users have made the judgment on one of 6 questions, such as "which place looks safer?" and 'which place looks more lively?" Based on judgments over these image pair comparisons, we can formulate visual urban perception as a relative attribute learning task. Correspondingly, the goal is to directly estimate a ranking function to provide a real value that reflects the degree of each perceived attribute for each image. Second, CNNs have been successfully applied to various visual perception tasks, such as image style, aesthetics and quality estimation [45]. Therefore, a deep network should be introduced into the framework for urban visual feature learning. Deep visual feature learning and ranking function learning should be jointly conducted to make them benefit each other. Third, multiple perceptual attributes are often correlated. For example, "lively" is very relevant to "beautiful" while "depressing" and "boring" often have higher correlation. Such attribute correlation should be exploited for better attribute learning via the prior of natural attribute grouping. Particularly, the structured sparsity is exploited based on the prior of attribute grouping to make it learn in a multi-task fashion. Furthermore, existing CNN models are pre-trained on certain large-scale datasets, such as ImageNet [46] and Places [47]. However, there still exists the difference between these datasets and task-specific datasets, such as Place Pulse 2.0 dataset. In order to release this problem, we can introduce different types of losses. Particularly, the proposed network consists of two types of sub-networks, namely ranking sub-network and classification sub-network. Two types of loss functions from these two sub-networks can fine-tune the deep network from two different aspects to boost its attribute learning capability for specific relative attribute learning task.

To realize these ideas, we propose a Multi-Task Deep Relative Attribute Learning Network (MTDRALN) for all tuples from all attributes. Compared with existing methods, MTDRALN is capable of jointly utilizing deep learning, multi-task learning and relative attribute learning to learn all the attributes simultaneously, and can be trained in an end-to-end fashion to benefit each other. MTDRALN is formulated as follows:

$$\begin{aligned}\min_{\boldsymbol{\Theta},\mathbf{W}}\{&\sum_m [L_r(f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_i), g_{\boldsymbol{\Theta}}(\mathbf{x}_j)), y_{ij}) \\ &+ \lambda L_c(\sigma_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_i), g_{\boldsymbol{\Theta}}(\mathbf{x}_j)))] + \mu\Omega(\mathbf{W})\}\end{aligned} \tag{1}$$

where $L_r(\cdot)$ denotes the ranking loss from ranking sub-network for each attribute a_m, e.g., the hinge loss in RankSVM [14] or the cross-entropy in RankNet [13]. $L_c(\cdot)$ denotes the classification loss from the classification sub-network. $g_{\boldsymbol{\Theta}}(\cdot)$ is the transformed representation through the Siamese deep network with parameters $\boldsymbol{\Theta}$. $f_m(\cdot)$ is a ranking function for certain attribute m. We can consider $f_m(\cdot)$ as a fully connected neural network layer with linear activation function in the deep network, namely $f_m(\mathbf{x}_i) = \mathbf{w}_m^\top \mathbf{x}_i$. $\sigma_m(\cdot)$ is the softmax of final layer activations from another subnetwork. $\Omega(\mathbf{W})$ is a regularization function defined on the matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_m, ..., \mathbf{w}_M] \in \mathbb{R}^{D\times M}$. It is used to model the correlation among different attributes/tasks. Both λ and μ are trade-off constants.

Next, we will introduce how to design each term in our objective function in details.

B. Ranking Loss $L_r(f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_i), g_{\boldsymbol{\Theta}}(\mathbf{x}_j)), y_{ij})$

As shown in Fig. 1, we introduce one ranking sub-network into our network with the ranking loss. Various pairwise ranking methods can be used as the ranking loss. We adopt two common ranking losses in our framework, namely RankSVM [14] and RankNet [13].

RankSVM is a classic pairwise learning-to-rank algorithm and has been used in relative attribute learning [36], and it is formalized as follows:

$$\begin{aligned}\min_{\mathbf{w}_m}\{&\sum_{(i,j)\in P_m} [y_{ij} L_o(\mathbf{x}_i, \mathbf{x}_j) + (1 - y_{ij}) L_s(\mathbf{x}_i, \mathbf{x}_j)] \\ &+ \lambda||\mathbf{w}_m||_2^2\}\end{aligned} \tag{2}$$

where $L_o(\mathbf{x}_i, \mathbf{x}_j)$ and $L_s(\mathbf{x}_i, \mathbf{x}_j)$ denote the contrastive constraint for ordered image pairs and the similar constraint for un-ordered image pairs, respectively. When incorporating deep feature learning into Eqn.2, they are denoted as:

$$\begin{aligned}L_o(\mathbf{x}_i, \mathbf{x}_j) &= \frac{1}{2}(\max(0, 1 - (f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_i)) - f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_j)))))^2 \\ L_s(\mathbf{x}_i, \mathbf{x}_j) &= \frac{1}{2}(f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_i)) - f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_j)))^2\end{aligned} \tag{3}$$

λ is a trade-off constant.

Therefore, the ranking loss of RankSVM is

$$\begin{aligned}&L(f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_i), g_{\boldsymbol{\Theta}}(\mathbf{x}_j))), y_{ij}) \\ &= \sum_{(i,j)\in P_m} [y_{ij} L_o(\mathbf{x}_i, \mathbf{x}_j) + (1 - y_{ij}) L_s(\mathbf{x}_i, \mathbf{x}_j)] + \lambda||\mathbf{w}_m||_2^2\end{aligned} \tag{4}$$

RankNet is another learning-to-ranking method. For each pair of feature vectors from deep learning $(g_{\boldsymbol{\Theta}}(\mathbf{x}_i), g_{\boldsymbol{\Theta}}(\mathbf{x}_j))$, corresponding scores are $f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_i))$ and $f_m(g_{\boldsymbol{\Theta}}(\mathbf{x}_j))$, respectively. They are then mapped to probabilities using a logistic function

$$p_{ij} = \frac{1}{1 + \exp[-(f_m(g_{\mathbf{\Theta}}(\mathbf{x}_i)) - f_m(g_{\mathbf{\Theta}}(\mathbf{x}_j)))]} \tag{5}$$

The model applies the cross entropy as its cost function and is defined as follows:

$$\min_{\mathbf{w}_m} \sum_{(i,j)\in P_m} [-t_{ij}\log(p_{ij}) - (1-t_{ij})\log(1-p_{ij})] \tag{6}$$

where the true probability obtained from the labels is t_{ij}.

Similar to [10], if $(i,j) \in O_m$, there are two cases: if the attribute of image i is stronger than image j, then t_{ij} is expected to be larger than 0.5; if image i exhibits less of the attribute than image j, then t_{ij} is expected to be smaller than 0.5. If $(i,j) \in S_m$, t_{ij} is expected to be 0.5. That is

$$t_{ij} = \begin{cases} 0 & i \prec j \\ 0.5 & i = j \\ 1 & i \succ j \end{cases}$$

Therefore, the ranking loss of RankNet is formulated as the cross-entropy loss:

$$L(f_m(g_{\mathbf{\Theta}}(\mathbf{x}))) = \sum_{(i,j)\in P_m} [-t_{ij}\log(p_{ij}) - (1-t_{ij})\log(1-p_{ij})] \tag{7}$$

C. Classification Loss $L_c(\sigma_m(g_{\mathbf{\Theta}}(\mathbf{x}_i), g_{\mathbf{\Theta}}(\mathbf{x}_j)))$

Training based on ranking sub-network may not be sufficient to learn the deep network due to the difference between ImageNet and special task oriented datasets, and the complex deep network. Therefore, we further introduce the classification sub-network, where the input is concatenated feature maps from each image pair and it consists of some convolutional layers and fully-connected layers with softmax loss $L_c(\cdot)$.

We use the standard softmax classification loss and perform binary classification for each image pair given certain attribute m. We introduce an indicator variable $z_{ij} \in \{1, 0\}$. Given an image pair $(i,j) \in O_m$, $z_{ij} = 1$ denotes a win for image i and $z_{ij} = 0$ denotes a win for image j. The softmax loss is

$$\begin{aligned} & L_c(\sigma_m(g_{\mathbf{\Theta}}(\mathbf{x}_i), g_{\mathbf{\Theta}}(\mathbf{x}_j))) \\ & = \sum_{(i,j)\in O_m} -[\mathcal{I}[z_{ij}=1]\log(\sigma_m(g_{\mathbf{\Theta}}(\mathbf{x}_i), g_{\mathbf{\Theta}}(\mathbf{x}_j)) \\ & + \mathcal{I}[z_{ij}=0]\log(1-\sigma_m(g_{\mathbf{\Theta}}(\mathbf{x}_i), g_{\mathbf{\Theta}}(\mathbf{x}_j))] \end{aligned} \tag{8}$$

where $\mathcal{I}[\cdot]$ is an indicator function. Note that in our network, similar to [2], when we consider the classification sub-network, only ordered pairs are used for the network training.

D. $\Omega(\mathbf{W})$

Regularization is one critical key to balance feature sharing of intra-group relative attribute learning and feature competition between inter-group relative attribute learning. When applying the ℓ_1 norm $\sum_{m=1}^{M} \|\mathbf{W}_m\|_1$, it will encourage the sparsity on each feature row of $\mathbf{W}$ and can generate a competition between tasks. Additionally, when applying the ℓ_{21} norm $\sum_{d=1}^{D} \|\mathbf{W}^d\|_2$. This can be seen as applying the ℓ_1 norm on collapsed column-wise output of ℓ_{21}, which forces tasks to select only dimensions that are sharable by other tasks as a way to encourage feature sharing.

As a middle solution, if there are L groups for all the attributes based on the relativeness among different attributes, the competition can be applied on the groups; meanwhile, the sharing can be applied inside each group. Particularly, each group $S_l = \{m_1, m_2, m_3, ...\}$ contains indices of specific attributes in that group, and $1 \le m_i \le M$. Each attribute appears in one group only. For example, in Place Pulse 2.0, there are $M = 6$ attributes in this dataset, including Safe, Lively, Beautiful, Wealthy, Depressing and Boring. It is more reasonable to be divided into two groups: positive group $\{\text{Safe}, \text{Lively}, \text{Beautiful}, \text{Wealthy}\}$ and negative group $\{\text{Depressing}, \text{Boring}\}$. In order to leverage the information of natural grouping, we also adopt the ℓ_{21}-type sharing regularizer within every semantic group into our loss function:

$$\Omega(\mathbf{W}) = \sum_{l=1}^{L}\sum_{d=1}^{D} \|\mathbf{W}_{S_l}^d\|_2 \tag{9}$$

where $\mathbf{W}_{S_l}^d$ is the d^{th} row vector containing a subset of the entries in row $\mathbf{w}^d$, specified by indices in the semantic group S_l.

This regularizer restricts the column-collapsing effect of the l_2 norm to within the semantic groups. This encourages (1) in-group latent feature sharing and (2) between-group competition for latent features.

E. Optimization

We introduce two settings for MTDRALN. In the first setting, we first fine-tune the Siamese deep network for each attribute to obtain attribute-specific deep visual features. After feature extraction, we use the output of the FC7 layer to obtain deep visual features for each image $\{g_{\mathbf{\Theta}}(\mathbf{x}_i)\}_{i=1}^{N}$, and then apply deep learning features for the proposed multi-task relative attribute learning task. Such setting is Loosely Combined, and named as MTDRALN-LC. In the second setting, we train our model in an end-to-end fashion, where each Siamese deep model is dedicated to relative attribute learning for one attribute. The loss function with the group structure prior is used to all the Siamese models. This setting is Tightly Combined and named as MTDRALN-TC.

1) MTDRALN-LC: In this setting, we should solve Eqn. 9 to estimate parameters $\mathbf{W}$. Similar to [48], we can change Eqn. 9 to iteratively optimize the following problem:

$$\begin{aligned} & \min_{\mathbf{W}} \{ \sum_m [L_r(f_m(g_{\mathbf{\Theta}}(\mathbf{x}_i), g_{\mathbf{\Theta}}(\mathbf{x}_j)), y_{ij}) \\ & + L_c(g_{\mathbf{\Theta}}(\mathbf{x}_i), g_{\mathbf{\Theta}}(\mathbf{x}_j), z_{ij})] + \mu \sum_d \sum_l \frac{\|\mathbf{W}_{S_l}^d\|_2^2}{q_{d,l}} \} \\ & \text{s.t.} \sum_d \sum_l q_{d,l} = 1, \ q_{d,l} \ge 0, \ \forall d, l \end{aligned} \tag{10}$$

and

$$q_{d,l} = \frac{\|\mathbf{W}_{S_l}^d\|_2}{\sum_d \sum_l \|\mathbf{W}_{S_l}^d\|_2} \tag{11}$$

We can use the smoothing proximal gradient descent to estimate $\mathbf{W}$ using (10) and $q_{d,l}$ using (11) alternately [41].

2) MTDRALN-TC: In this setting, we should jointly learn parameters $\mathbf{\Theta}$ from the deep network and $\mathbf{W}$ from MTL. Similarly, we can change Eqn. 9 to iteratively optimize the following two problems:

$$\begin{aligned}\min_{\mathbf{\Theta},\mathbf{W}}\{\sum_m[L_r(f_m(g_{\mathbf{\Theta}}(\mathbf{x}_i),g_{\mathbf{\Theta}}(\mathbf{x}_j)),y_{ij})\\ +L_c(g_{\mathbf{\Theta}}(\mathbf{x}_i),g_{\mathbf{\Theta}}(\mathbf{x}_j),z_{ij})]+\mu\sum_d\sum_l\frac{\|\mathbf{W}^d_{S_l}\|_2^2}{q_{d,l}}\} \quad (12)\\ \text{s.t.}\sum_d\sum_l q_{d,l}=1,\ q_{d,l}\geq 0,\ \forall d,l\end{aligned}$$

and

$$q_{d,l}=\frac{\|\mathbf{W}^d_{S_l}\|_2}{\sum_d\sum_l\|\mathbf{W}^d_{S_l}\|_2} \quad (13)$$

Note that both parameters $\mathbf{\Theta}$ from the deep network and $\mathbf{W}$ from MTL are jointly learned. During a training epoch, the forward pass will generate the input $g_{\mathbf{\Theta}}(\mathbf{x})$ for the multi-task loss layer from all the Siamese models with learned parameters $\mathbf{\Theta}$. We then optimize Eqn. 12 and Eqn. 13 to obtain updated parameters $\mathbf{W}$. Then they are taken back in the backward pass alongside the gradients with respect to its input to update $\mathbf{\Theta}$.

IV. Experiment

A. Experimental Setting

Dataset. We utilize the Place Pulse 2.0 (PP2) dataset with 111,390 images and 1,223,649 pairwise comparisons from 56 cities[1]. These comparisons come from the response of six questions based on six perceptual attributes, namely Safe, Lively, Beautiful, Wealthy, Boring and Depressing, such as "Which place looks safer?" and "Which place looks more depressing?" The left part of Fig. 1 shows some samples from PP2. We use the Google Street View Image API[2] to download street images based on provided GPS information. After discarding images, where URLs are unavailable, we finally obtain 110,688 images and 1,208,808 pairwise comparisons, including 1,046,926 unequal ones and 161,882 equal ones[3]. In PP2, a natural group splitting is to divide attributes into 2 groups: the positive group $\{\text{Safe},\text{Lively},\text{Beautiful},\text{Wealthy}\}$ and the negative one $\{\text{Depressing},\text{Boring}\}$.

Implementation Details. All the deep models are implemented based on the Caffe [49] platform, and the Siamese network architecture is adopted. We adopt PlacesNet in each branch of the Siamese network in the following experiment, which is one network with 7 layers and pre-trained on the scene dataset Places205 [47], and thus suitable for our task, which consists of street view images. In addition, we remove the last fully connected layers, and substitute these layers with our own objective loss.

In the Siamese network, we set the weight decay to $5e-4$. The momentum, learning rate and mini-batch is 0.9, 0.01 and 64, respectively. For model fine-tuning, we set the maximum number of iterations to 20 epoches. Empirically, $\lambda=\mu=1.0$.

For unequal comparisons, similar to [2], we split the set for each attribute randomly in the ratio 65%, 5% and 30% for training, validation and testing, respectively. The set of pairwise equal comparisons can be also added into the training set for methods without introducing the classification sub-network.

B. Attribute Prediction

Similar to [36], the performance is evaluated in terms of the ranking accuracy, which is the share of correctly predicted users' votes on pairs of images. Since our task is to learn relative attribute for urban perception, we use the following relative attribute learning methods for comparison:

- RA-CNN [36]. This method first uses the CNN to extract visual features, and then adopts RankSVM to learn each attribute separately.
- RN-CNN [13]. This method first uses the CNN to extract visual features, and then adopts RankNet to learn each attribute separately.
- DRA [9]. Compared with RA-CNN, this method learns deep visual features and ranking function in an end-to-end way, where RankSVM is adopted.
- DRN [10]. This method learns deep visual features and ranking function in an end-to-end way, where RankNet is adopted.
- STN-RN [40]. STN-RN adopts the spatial transformer network to discover the most informative image region for the attribute. The image and its local region are both used by the ranking module to learn a ranking model of the attribute,where RankNet is adopted.
- RSS-CNN [2]. This method uses a Siamese-like convolutional neural architecture with classification sub-network and ranking sub-network to jointly learn the classification and ranking loss.
- RSS-CNN-RN. Similar to RSS-CNN, RSS-CNN-RN also consists of classification sub-network and ranking sub-network, but RankNet-style ranking loss is adopted.

For our MTDRALN-LC, there are two network settings:

- MTDRALN-LC(RankSVM): we use the fine-tuned PlacesNet to extract visual features based on jointly classification loss and ranking loss, and optimize Eqn.10 for optimization based on the visual features, where RankSVM is adopted.
- MTDRALN-LC(RankNet): this method is similar to MTDRALN-LC(RankSVM), but adopts RankNet for relative attribute learning.

Similarly, MTDRALN-TC(RankSVM) and MTDRALN-TC(RankNet) are trained in an end-to-end way for simultaneous deep feature learning and relative attributes learning. For fair comparison, the CNN network adopts the PlacesNet architecture for all the methods.

Table I shows the experimental results on predicting pairwise comparisons. We can see that (1) The performance of

[1] http://pulse.media.mit.edu/data/

[2] https://developers.google.com/maps/documentation/streetview/

[3] We began crawling street images in July, 2017 and it took us about one month to complete the image collecting.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2932502, IEEE Transactions on Image Processing

TABLE I: Performance comparison of the attribute prediction between our method and others in the PP2

Method	Safe	Lively	Beautiful	Wealthy	Boring	Depressing	Avg.
RA-CNN [36]	59.98%	67.46%	65.14%	58.65%	61.60%	62.31%	62.52%
RN-CNN [13]	60.18%	67.23%	65.53%	58.82%	60.80%	61.92%	62.41%
DRA [9]	60.23%	69.23%	66.23%	59.54%	63.41%	64.00%	63.77%
DRN [10]	60.56%	69.58%	66.82%	61.42%	63.54%	**64.84%**	64.46%
STN-RN [40]	60.52%	60.62%	67.89%	**66.27%**	60.00%	62.81%	63.02%
RSS-CNN [2]	64.10%	72.10%	70.90%	64.60%	65.40%	61.20%	66.38%
RSS-CNN-RN	63.93%	71.92%	70.83%	64.24%	65.62%	61.50%	66.34%
MTDRALN-LC(RankNet)	64.19%	72.12%	70.95%	64.64%	65.83%	61.72%	66.58%
MTDRALN-TC(RankNet)	64.96%	73.12%	71.65%	65.42%	66.78%	62.45%	67.40%
MTDRALN-LC(RankSVM)	65.07%	73.90%	71.50%	64.92%	66.79%	63.21%	67.57%
MTDRALN-TC(RankSVM)	**65.82%**	**74.58%**	**72.32%**	65.84%	**67.20%**	64.05%	**68.30%**

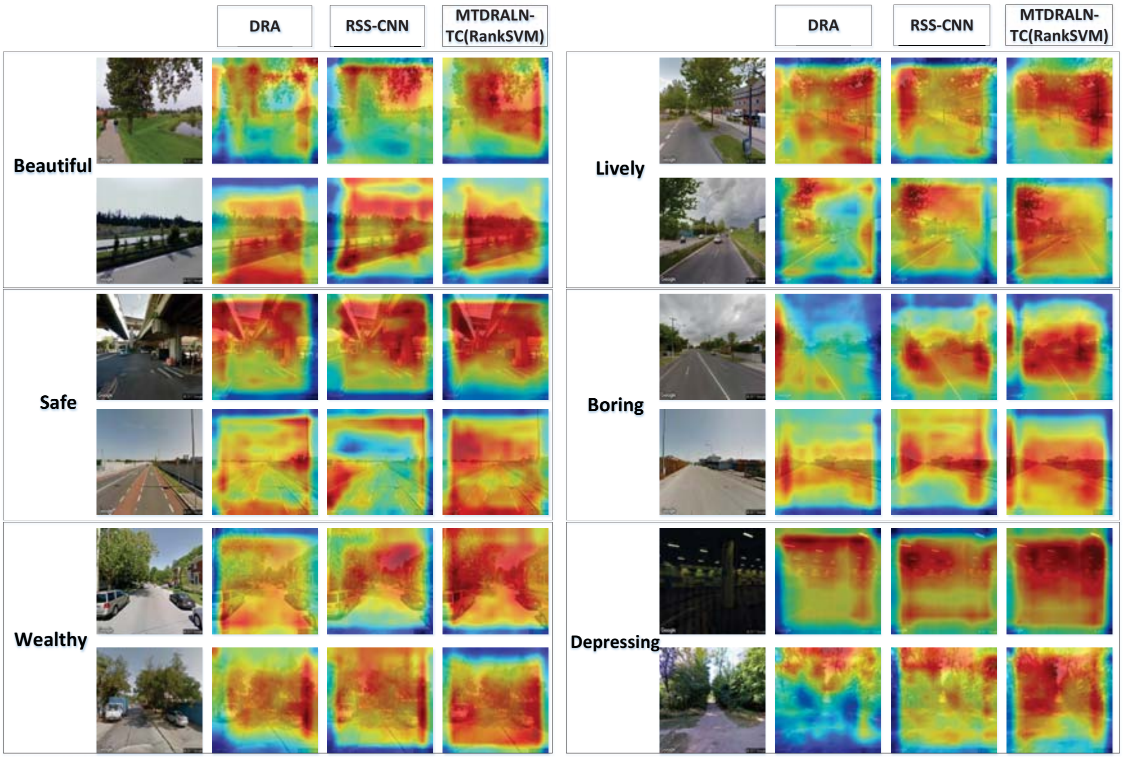


Fig. 2: Discriminative localization maps comparison from some images among three methods: (1) DRA, (2) RSS-CNN and (3) MTDRALN-TC(RankSVM). Grad-CAM [50] is adopted to implement discriminative localization region visualization(the warmer the color of the overlay image, the more discriminative that pixel is).

DRA is better than RA-CNN. This is because DRA is trained to learn each attribute in an end-to-end way. Deep visual feature learning and ranking function learning benefit each other. Similarly, the performance of DRN is better than RN-CNN. (2) The performance of RSS-CNN is better than DRA on the whole. Because RSS-CNN learns joint classification and ranking loss. Combined with the ranking sub-network, the introduced classification sub-network can further reduce the domain gap when fine-tuning basic network. Note that there is a small difference between the performance of RSS-CNN and the one reported in [2] with the fine-tuned PlacesNet. The probable reason is that the used dataset is a little different from theirs. Similarly, we can also see that the performance of RSS-CNN-RN is better than DRN. (3) After introducing the attribute grouping information, our method MTDRALN-LC on different fine-tuned networks both consistently improves the performance of their original networks on most attributes. This verifies the effectiveness of our method in utilizing the prior of attribute grouping to model the relativeness among different attributes. (4) the performance of MTDRALN-TC is higher than MTDRALN-LC in all the six attribute learning tasks under two different settings: RankSVM and RankNet. These results suggest that MTDRALN-TC can jointly utilize visual feature learning and multi-task relative attribute learning in an end-to-end fashion to achieve better performance.

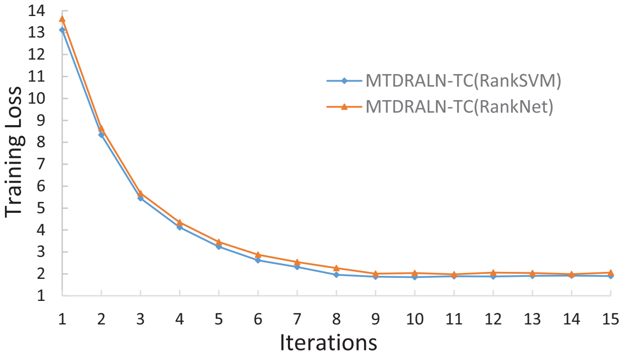


Fig. 3: Training losses with different iterations for MTDRALN-TC(RankSVM)and MTDRALN-TC(RankNet) on the PP2 dataset.

C. Qualitative Evaluation

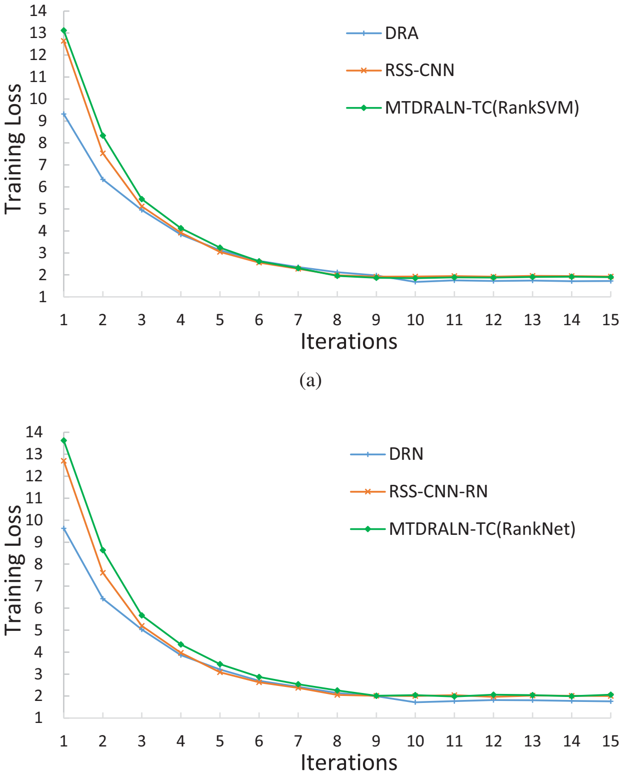
In this section, we further qualitatively study the contribution of the classification sub-network and attribute grouping. Without loss of generality, we compare the following three methods, where RankSVM is adopted for the ranking function: (1) DRA, (2) RSS-CNN and (3) MTDRALN-TC (RankSVM).

For the qualitative evaluation, we use Grad-CAM [50] for region localization. Particularly, we visualize the significance for each attribute. Using two images as input to the network, we extract the feature map of the last convolutional layer, find the weighted sum of their corresponding feature maps, then upsample it and superimpose it on the original image. The pixels in these saliency map visualizations contribute most to the ranking of network predictions. It can learn to find areas of great contribution and then generate heat maps based on the contribution.

Fig. 2 shows some examples with saliency maps obtained from these networks. From these results, we can see that (1) After introducing the classification sub-network, RSS-CNN localized more interpretable regions than DRA. For example, for the safe attribute, the second example shows that RSS-CNN successfully localizes the fence in the left of the image than DRA, where the fence can be considered as the safe attribute. For the boring attribute, RSS-CNN focused more on the region with the bare road in this image than DRA, where bare road can be considered as a reasonable explanation of boring attribute. (2) After further introducing the group prior, MTDRALN-TC makes localized regions more comprehensive and accurate than these two methods on the whole. For example, for wealthy attribute, compared with other two methods, MTDRALN-TC(RankSVM) can accurately localize the regions of car, an indication of wealth. For boring attribute, MTDRALN-TC(RankSVM) can locate comprehensively the bare road compared with RSS-CNN. These qualitative results further verify the effectiveness of our framework.

D. Convergence Analysis

To explore the convergence of the proposed MTDRALN-TC, we show the training losses of MTDRALN-TC trained with different numbers of iterations under two settings: RankSVM and RankNet in Fig. 3. We can see that our proposed method has very fast convergence. The training loss of MTDRALN-TC has been steady after the 10th iteration.



(a)

(b)

Fig. 4: Convergence analysis for (a) RankSVM based methods and (b) RankNet based methods.

In addition, in order to evaluate the convergence of introducing the classification sub-network, we further showed the training losses on three end-to-end deep learning methods trained with different numbers of iterations under two settings: RankSVM Fig. 4 (a) and RankNet Fig. 4 (b). From Fig. 4, we can see that although introducing the classification sub-network makes the network more complicated, the ranking loss and classification loss work together to guarantee faster convergence speed of this network, especially in the first five iterations. As a result, all the methods achieve the convergence after about 10 iterations.

E. Predicting Urban Perceptual Attributes across Cities

In order to use our model for visual urban perception at a global scale, similar to [21], [2], we need to verify the generalization performance of our model in terms of the geographical distance. Therefore, we conduct the experiment by training on one city and testing on another city. We use the data from 4 cities, namely, New York, Berlin, Tokyo and Moscow. These cities are from different countries. For each city, as mentioned before, we also split the set randomly in the ratio 65%, 5% and 30% for training, validation and testing, respectively for each attribute. The test data from the intra-city and other cities are used to test the trained model. Table II shows the experimental results on different perceptual attributes. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2932502, IEEE Transactions on Image Processing

TABLE II: Ranking accuracy from one city to another on attribute prediction,where the first column denotes the model learned from the training data from corresponding city, and each row denotes predicted results based on test data from different cities.

(a) Safe

Test Training	New York	Berlin	Tokyo	Moscow
New York	**62.94%**	62.35%	59.56%	61.54%
Berlin	61.23%	**63.22%**	60.52%	61.82%
Tokyo	58.52%	57.62%	**62.51%**	59.23%
Moscow	59.50%	60.62%	59.84%	**61.23%**

(b) Lively

Test Training	New York	Berlin	Tokyo	Moscow
New York	**69.52%**	69.39%	67.52%	67.00%
Berlin	70.43%	**71.22%**	68.83%	69.52%
Tokyo	66.52%	65.85%	**68.38%**	65.25%
Moscow	64.72%	64.85%	63.25%	**66.72%**

(c) Depressing

Test Training	New York	Berlin	Tokyo	Moscow
New York	**60.20%**	59.70%	58.10%	59.60%
Berlin	58.80%	**60.80%**	57.20%	60.20%
Tokyo	56.80%	55.90%	57.80%	**58.90%**
Moscow	54.83%	55.42%	54.23%	**56.90%**

(d) Boring

Test Training	New York	Berlin	Tokyo	Moscow
New York	**62.29%**	60.05%	59.20%	59.50%
Berlin	62.43%	**64.73%**	60.42%	62.82%
Tokyo	59.60%	58.50%	**61.22%**	60.02%
Moscow	58.52%	59.25%	57.52%	**60.56%**

(e) Beautiful

Test Training	New York	Berlin	Tokyo	Moscow
New York	**66.20%**	65.89%	63.52%	64.85%
Berlin	66.52%	**68.51%**	64.83%	66.94%
Tokyo	62.34%	62.82%	**64.52%**	63.54%
Moscow	62.13%	61.23%	60.20%	**63.60%**

(f) Wealthy

Test Training	New York	Berlin	Tokyo	Moscow
New York	**64.83%**	64.44%	62.52%	63.51%
Berlin	62.73%	**63.82%**	61.92%	63.32%
Tokyo	61.20%	60.50%	**63.52%**	61.52%
Moscow	62.53%	61.56%	61.00%	**62.95%**

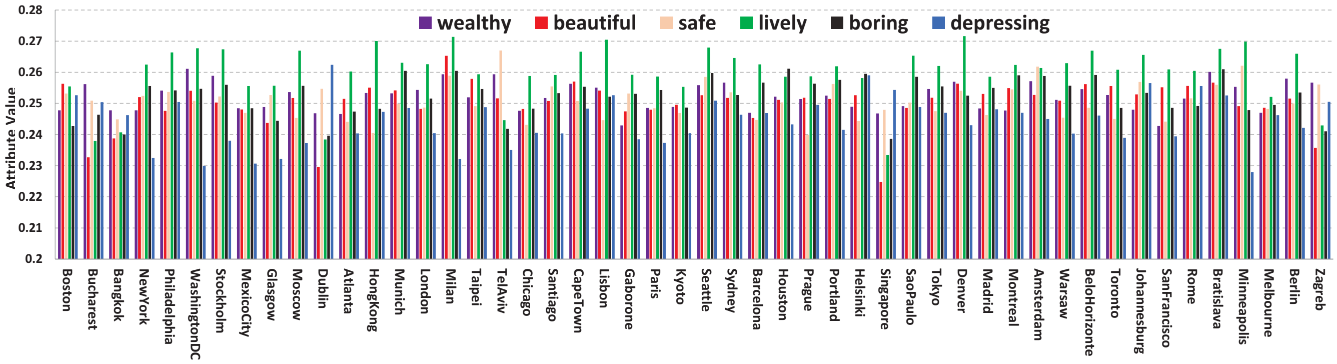


Fig. 5: The attribute distribution for each city (Best view in enlargement).

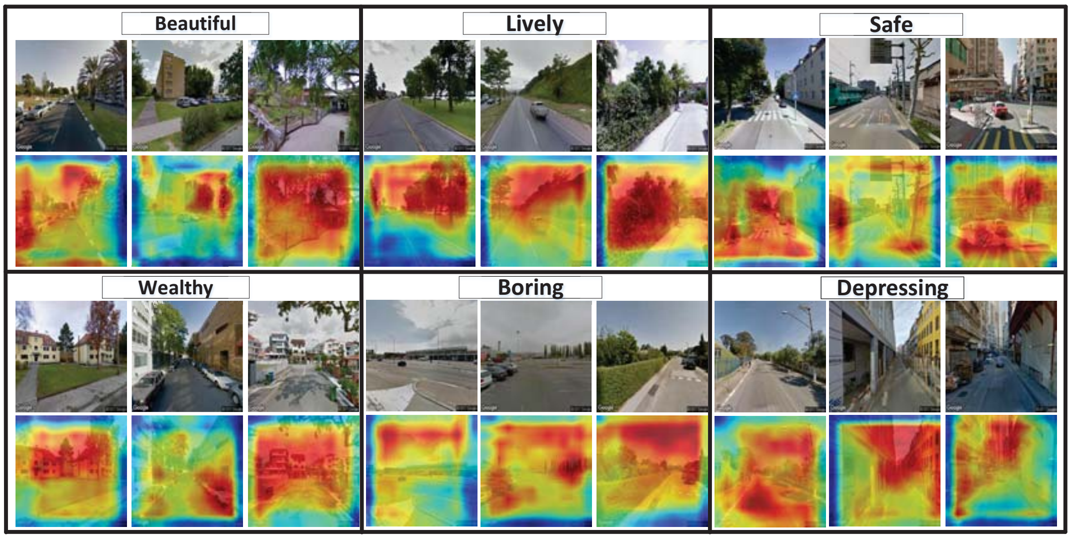
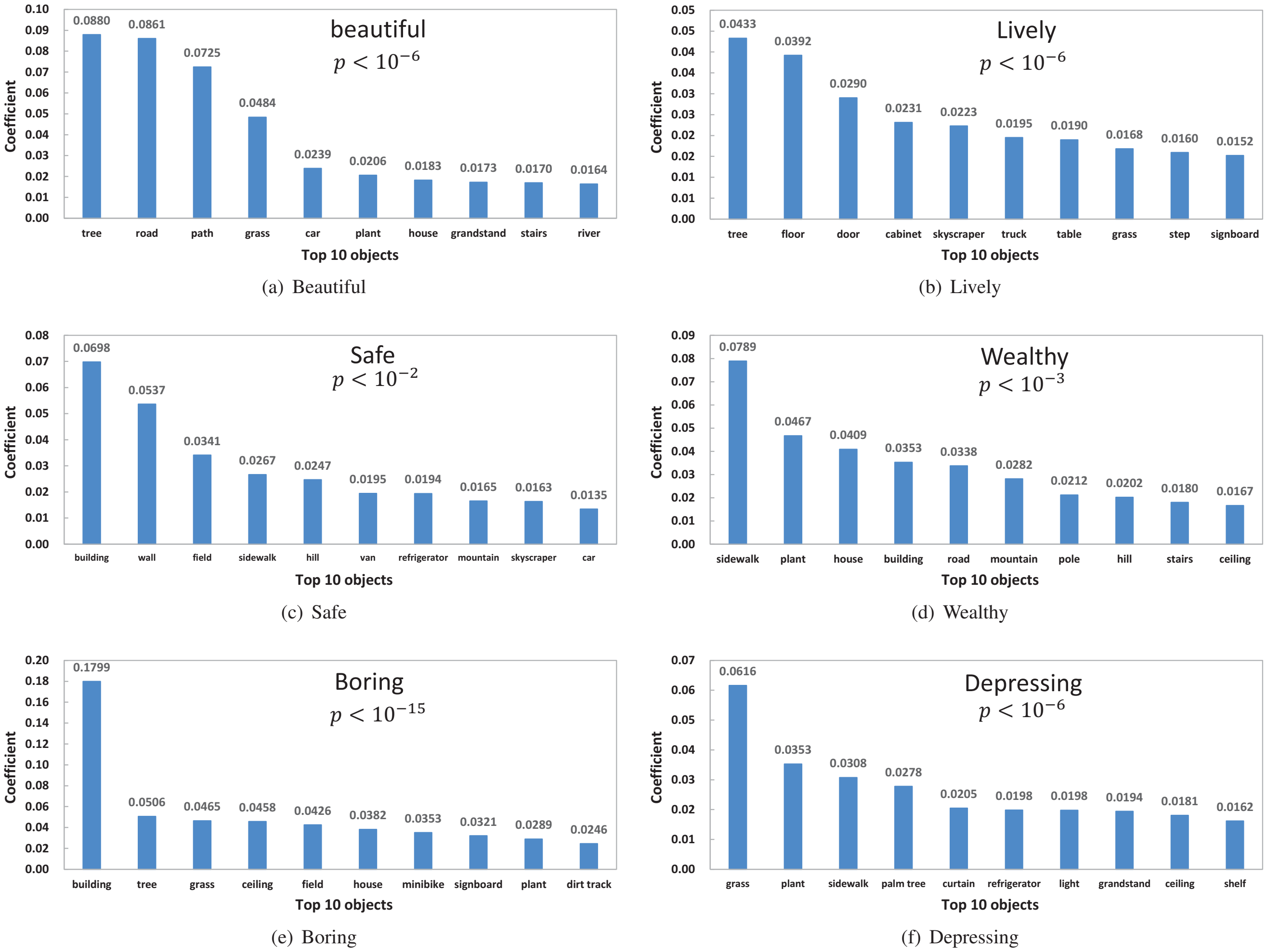


Fig. 6: Sample images associated with six types of perceptual attributes, where saliency areas correspond to the ones that mostly contribute to perceptual attributes.



(a) Beautiful

(b) Lively

(c) Safe

(d) Wealthy

(e) Boring

(f) Depressing

Fig. 7: The results of multivariate regression analysis between scene elements and perception scores. The top 10 objects that contribute to each of the 6 perception attributes are shown.

find that our model is able to reasonably predict perceptual attributes well across cities, with the range of the ranking accuracy $[0.60, 0.70]$, indicating the generalization ability of our model to new places. Note that for the depressing attribute in the Tokyo city, training on this city different from the test city actually improves performance as compared to training on the same city. This might occur because of the unbalanced test data. In addition, in order to utilize learned attributes for urban representation, we use the Gaussian distribution with $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}$ to represent attribute distribution of each city. We select 50 cities with more street images and show their attribute distribution on 6 perceptual attributes. Fig. 5 shows the results. We can have some reasonable observations. For example, the scores of the wealthy attribute from some cities in the USA are very high, such as Seattle, New York and Minneapolis because of their better natural scenes and living environment in the USA. Meanwhile, their scores of the safe attribute are also high. Many cities have lower scores of both depressing and boring attributes because of their higher scores for positive attributes. Note that because of sample bias, the demonstrated results have variance to represent one

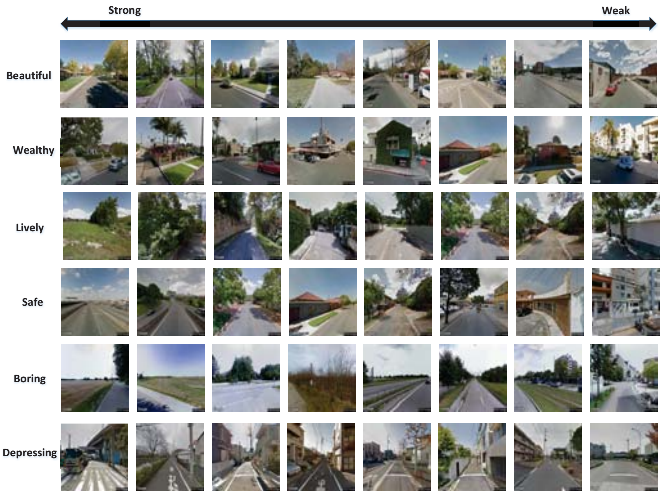


Fig. 8: Sample images from PP2, ordered according to the predicted value of their respective attribute.

city. In order to more accurately reveal the final role of urban perception attribute for describing one city,we should collect more comprehensive data, which is beyond the scope of this paper. In this work, we highlight the perceptive representation for city can be considered as mid-level representation for great potentials in urban computing. For example, we can use perceptive attribute representation to calculate the city similarity. We can also combine visual features and perceptive attribute features to improve the recognition performance of city identity. Also, we can combine other urban attributes, such as water coverage and transportation [51] for comprehensive visual urban representation from different aspects.

F. Attribute Region Localization

As mentioned before, we use the method [50] to visualize the saliency of each attribute. Fig. 6 shows some examples, where visual regions are highlighted in the image. We can see that visual elements corresponding to safe attributes are pedestrian crosswalks and the fence between apartment and road. The green, such as residential trees and public gardens enjoyed stronger association with both beautiful and lively attributes. High buildings, villa and cars indicate the wealthy attribute. Large bare roads indicate more boring attributes, and are also very relevant to depressing attributes. The region from the end of two buildings also indicates depressing attributes. In addition, such discovered results provide potentials for business opportunities. For example, knowing what makes people lively, and then designing for such 'lively' could increase profits. It could guide investment in urban landscaping.

In order to identify visual elements, which are highly contributed to perceptual attributes, similar to [20], we first use the scene parsing model PSPNet [52] to recognize and segment object instances from each street view image, and the area ratio of each visual element in one image was calculated by counting the pixel numbers in the segmentation mask. For each perceptual attribute, the perceptual scores are calculated from each image in PP2.0 via our model. We finally model the correlation between perceptive attribute score and the area ratio of each visual element via multivariate regression analysis, where 150 categories of objects are employed. Fig. 7 present the results of multivariate regression analysis between the perceptual attributes and the presence of visual elements, where top 10 objects that contributed to each perceptual attribute are ranked and listed.

Like other ranking based networks, our network can achieve a total ordering of images, given a set of pairwise orderings. Figure 8 shows some images ordered according to their value of the respective attribute.

V. Conclusions

In this paper, we have proposed a novel Multi-Task Deep Relative Attribute Learning Network (MTDRALN), which is capable of incorporating deep learning, multi-task learning and relative attribute learning into a unified framework to jointly predict all the perceptual attributes. The introduced classification sub-network and ranking sub-network jointly reduce the domain gap when training the neural network, leading to more discriminative attribute learning ability. The grouping information among different attributes has been leveraged to improve the performance of relative attribute learning. Quantitative and qualitative experimental results have demonstrated its effectiveness of our proposed framework in visual urban perception. We finally found some interesting visual associations with perceptual attributes. These findings further the agenda of perceiving and understanding the urban scenes.

Visual urban perception is a typically multidisciplinary field. Few works focuses on this topic, especially in the large-scale dataset. Therefore, there is a lot of work to do in the future. We first plan to analyze the difference among urban scenes from various geographic areas with different granularity [53]. For example, we can conduct the analysis on certain neighborhoods and buildings from one city. Second, although our results show that it is possible to locate image regions to understand why people find certain urban scenes to be safe or lively, there are limitations to this analysis because of subjective annotations on street view photos. Users' votes do not depend on visual characteristics alone, but might be influenced by other factors, such as the picture quality and culture. Therefore, we also plan to find differences of various factors. Third, the method itself is independent of underlying tasks, and thus can be applied to other relative attribute learning problems. In addition, we can also plan to use deeper architectures (e.g., VGG-16 [54]) or other types of deep architectures such as ResNet [55] and Densenet [56] in our framework to continue improving the performance of our model. Fourth, in some cases, especially multi-task learning for many attributes, a manual split for these attributes is hard to operate. Therefore, automatically learning the relation among different relative attributes with deep visual feature learning is also necessary. Finally, the images used for online rating in Place Pulse are taken in 2014 or even earlier. The visual environment may change during the gap period to some extent for our collected food images in 2017. Therefore, it is necessary to evaluate the potential bias in the future work, since such visual change may probably change the user's rating. We can also plan to consider the temporal dimension and study the spatio-temporal urban patterns based on street-level images [57].

References

[1] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo, "Computer vision uncovers predictors of physical urban change," *Proceedings of the National Academy of Sciences*, vol. 114, no. 29, pp. 7571–7576, 2017.

[2] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo, "Deep learning the city: Quantifying urban perception at a global scale," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 196–212.

[3] X. Liu, Q. Chen, L. Zhu, Y. Xu, and L. Lin, "Place-centric visual urban perception with deep multi-instance regression," in *Proceedings of the ACM Multimedia Conference*, 2017, pp. 19–27.

[4] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and F. F. Li, "Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states," *Proceedings of the National Academy of Sciences*, vol. 114, no. 50, pp. 13 108–13 113, 2017.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[6] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6230–6239.

[7] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3309–3318.

[8] S. M. Arietta, A. A. Efros, R. Ramamoorthi, and M. Agrawala, "City forensics: Using visual elements to predict non-visual city attributes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2624–2633, 2014.

[9] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim, "Deep relative attributes," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1832–1842, 2016.

[10] Y. Souri, E. Noury, and E. Adeli, "Deep relative attributes," in *Proceedings of the Asian Conference on Computer Vision*, 2017, pp. 118–133.

[11] D. Quercia, N. K. O'Hare, and H. Cramer, "Aesthetic capital: What makes london look beautiful, quiet, and happy?" in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2014, pp. 945–955.

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

[13] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of International Conference on Machine Learning*, 2005, pp. 89–96.

[14] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 133–142.

[15] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[16] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 145–152.

[17] K. Lynch, "The image of the city," *MIT press*, vol. 11, 1960.

[18] P. Salesses, K. Schechtner, and C. A. Hidalgo, "The collaborative image of the city: Mapping the inequality of urban perception," *PLOS ONE*, vol. 8, no. 7, pp. 1–12, 2013.

[19] V. Ordonez and T. L. Berg, "Learning high-level judgments of urban perceptionlearning high-level judgments of urban perceptionlearning," in *Proceedings of European Conference on Computer Vision*. Springer, 2014, pp. 494–510.

[20] F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, and C. Ratti, "Measuring human perceptions of a large-scale urban region using machine learning," *Landscape and Urban Planning*, vol. 180, pp. 148 – 160, 2018.

[21] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo, "Streetscore – predicting the perceived safety of one million streetscapes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 793–799.

[22] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, and F. F. Li, "Fine-grained car detection for visual census estimation," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4502–4508.

[23] X. Song, S. Jiang, and L. Herranz, "Multi-scale multi-feature context modeling for scene recognition in the semantic manifold," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2721–2735, 2017.

[24] N. Vo, N. Jacobs, and J. Hays, "Revisiting im2gps in the deep learning era," in *IEEE International Conference on Computer Vision*, 2017, pp. 2640–2649.

[25] W. Min, B. Bao, and C. Xu, "Multimodal spatio-temporal theme modeling for landmark analysis," *IEEE Multimedia*, pp. 20–29, 2014.

[26] J. Sang, Q. Fang, and C. Xu, "Exploiting social-mobile information for location visualization," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 39:1–39:19, Jan. 2017.

[27] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1778–1785.

[28] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014.

[29] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.

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

[31] N. Murrugarra-Llerena and A. Kovashka, "Learning attributes from human gaze," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2017, pp. 510–519.

[32] X. You, R. Wang, and D. Tao, "Diverse expected gradient active learning for relative attributes," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3203–3217, 2014.

[33] L. Chen, Q. Zhang, and B. Li, "Predicting multiple attributes via relative multi-task learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1027–1034.

[34] Q. Buyue, W. Xiang, C. Nan, J. Yu-Gang, and D. Ian, "Learning multiple relative attributes with humans in the loop," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5573–5585, 2014.

[35] R. S. Cruz, B. Fernando, A. Cherian, and S. Gould, "Deeppermnet: Visual permutation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6044–6052.

[36] D. Parikh and K. Grauman, "Relative attributes," in *Proceedings of the International Conference on Computer Vision*, 2011, pp. 503–510.

[37] L. C. Yuhang He and J. Chen, "Multi-task relative attribute prediction by incorporating local context and global style information," in *Proceedings of the British Machine Vision Conference*, 2016, pp. 1–12.

[38] M. Zihang, A. Nagesh, J. K. Hyunwoo, F. Glenn, and S. Vikas, "Efficient relative attribute learning using graph neural networks," in *Proceedings of the European Conference on Computer Vision*. Springer, 2018, pp. 575–590.

[39] F. Xiao and Y. J. Lee, "Discovering the spatial extent of relative attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1458–1466.

[40] K. K. Singh and Y. J. Lee, "End-to-end localization and ranking for relative attributes," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 753–769.

[41] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1629–1636.

[42] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2597–2609, 2018.

[43] T. Gebru, J. Hoffman, and L. Fei-Fei, "Fine-grained recognition in the wild: A multi-task domain adaptation approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1358–1367.

[44] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1131–1140.

[45] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 990–998.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[47] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2014, pp. 487–495.

[48] S. Kim and E. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 543–550.

[49] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv:1408.5093*, 2014.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization." in *Proceedings of the IEEE International Conference on Computer Vision Workshop*, 2017, pp. 618–626.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2932502, IEEE Transactions on Image Processing

[51] B. Zhou, L. Liu, A. Oliva, and A. Torralba, "Recognizing city identity via attribute analysis of geo-tagged images," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 519–534.

[52] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6230–6239.

[53] M. Kafsi, H. Cramer, B. Thomee, and D. A. Shamma, "Describing and understanding neighborhood characteristics through online social media," in *Proceedings of International Conference on World Wide Web*, 2015, pp. 549–559.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[56] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

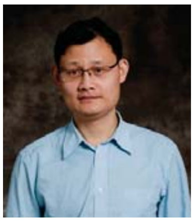
[57] F. Zhang, L. Wu, D. Zhu, and Y. Liu, "Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 153, pp. 48 – 58, 2019.

Weiqing Min received the B.E. degree from Shandong Normal University, Jinan, China, in 2008 and M.E. degree from Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2015, respectively. He is currently an associate professor at the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences. His current research interests include multimedia content analysis, understanding and applications, food computing and geo-multimedia computing. He has authored and co-authored more than 20 peer-referenced papers in relevant journals and conferences, including ACM Computing Surveys, IEEE Transactions on Multimedia, ACM TOMM, IEEE Multimedia Magazine, ACM Multimedia, IJCAI, etc. He is the reviewer of some international journals including IEEE Trans. on Multimedia, IEEE Trans. on Cybernetics, IEEE Trans. on Circuits and Systems for Video Technology, IEEE Trans. on Neural Network and Learning System, ACM TOMM, etc. He is the recipient of 2016 ACM TOMM Nicolas D. Georganas Best Paper Award and the 2017 IEEE Multimedia Magazine Best Paper Award.

Shuhuan Mei received the M.E. degree in Shandong University of Science and Technology, Shandong, China. His current research interests include multimedia content analysis, understanding and applications.

Linhu Liu is pursuing the M.E. degree in University of Chinese Academy of Sciences, Beijing, China. His current research interests include multimedia content analysis, understanding and applications.

Yi Wang received the M.E. degree in Beijing University of Posts and Telecommunications, China. His current research interests include multimedia content analysis and geo-multimedia computing.



Shuqiang Jiang (SM'08) is a professor with the Institute of Computing Technology, Chinese Academy of Sciences(CAS), Beijing and a professor in University of CAS. He is also with the Key Laboratory of Intelligent Information Processing, CAS. His research interests include multimedia processing and semantic understanding, pattern recognition, and computer vision. He has authored or coauthored more than 150 papers on the related research topics. He was supported by the New-Star program of Science and Technology of Beijing Metropolis in 2008, NSFC Excellent Young Scientists Fund in 2013, Young top-notch talent of Ten Thousand Talent Program in 2014. He won the Lu Jiaxi Young Talent Award from Chinese Academy of Sciences in 2012, and the CCF Award of Science and Technology in 2012. He is the senior member of IEEE and CCF, member of ACM, Associate Editor of IEEE Multimedia, Multimedia Tools and Applications. He is the vice chair of IEEE CASS Beijing Chapter, vice chair of ACM SIGMM China chapter. He is the general chair of ICIMCS 2015, program chair of ACM Multimedia Asia2019 and PCM2017. He has also served as a TPC member for more than 20 well-known conferences, including ACM Multimedia, CVPR, ICCV, IJCAI, AAAI, ICME, ICIP, and PCM.