Place-centric Visual Urban Perception with Deep Multi-instance Regression*

Xiaobai Liu
Dept. Computer Science, San Diego State University (SDSU), San Diego, California, U.S.A.

Qi Chen
Dept. Computer Science, San Diego State University (SDSU), San Diego, California, U.S.A.

Lei Zhu
School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia

Yuanlu Xu
Dept. Computer Science, University of California at Los Angeles (UCLA), Los Angeles, CA

Liang Lin
School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China

ABSTRACT

This paper presents a unified framework to learn to quantify perceptual attributes (e.g., safety, attractiveness) of physical urban environments using crowd-sourced street-view photos without human annotations. The efforts of this work include two folds. First, we collect a large-scale urban image dataset in multiple major cities in U.S.A., which consists of multiple street-view photos for every place. Instead of using subjective annotations as in previous works, which are neither accurate nor consistent, we collect for every place the safety score from government's crime event records as objective safety indicators. Second, we observe that the place-centric perception task is by nature a multi-instance regression problem since the labels are only available for places (bags), rather than images or image regions (instances). We thus introduce a deep convolutional neural network (CNN) to parameterize the instance-level scoring function, and develop an EM algorithm to alternatively estimate the primary instances (images or image regions) which affect the safety scores and train the proposed network. Our method is capable of localizing interesting images and image regions for each place. We evaluate the proposed method on a newly created dataset and a public dataset. Results with comparisons showed that our method can clearly outperform the alternative perception methods and more importantly, is capable of generating region-level safety scores to facilitate interpretations of the perception process.

1 INTRODUCTION

1.1 Background

The task of visual urban perception [9] [12] [13] [18] aims to quantify the connections between the physical appearance of urban environment and perceptual attributes (e.g., safety, attractiveness, wealth). The well known broken windows theory [14] [27] suggests that visual signs of environmental disorder, such as broken windows, abandoned cars, litter and graffiti, can induce negative social outcomes of and increase crime levels. Other studies [4] [15] [21] [23] also found the associations between environment disorder and criminal behaviors, health outcomes, incidences of obesity, and rates of female alcoholism. More recently, researchers in multi-media computing and Artificial Intelligence take as inputs geo-tagged street-view photos [8, 31], and develop learning-based methods using human-annotated street-view photos. The popularity of high throughput data collection methods (e.g., AMAZON MTurk) has increased the availability of urban

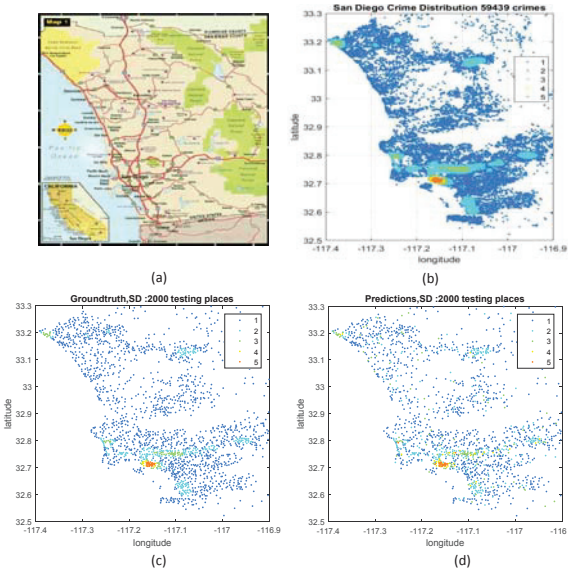

Figure 1: Place-centric visual urban perception. (a) The map of San Diego, U.S.A. (b) The density map of crime events for San Diego. The crime events were collected by various government agencies. The color of a place represents its densities. (c) The ground-truth densities (or safety ranks) for the 2000 testing places. (d) The density map estimated by the proposed method.

*The first two authors contributed equally to this work. All correspondences should be addressed to Dr. Xiaobai Liu (EMAIL: xiaobai.liu@mail.sdsu.edu).

MM'17, October 23–27, 2017, Mountain View, CA, USA.

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
http://doi.org/10.1145/3123266.3123271

perception data, and thus enables the learning based systems for automatic predictions.

However, existing efforts on visual urban perception are still limited in two aspects. First, the perception labels are usually crowdsourced by human annotators and there are huge variances or inconsistent annotations across subjects. For example, it is difficult for a human to consistently label the safety level of a scene from street-view photos. The labels might vary over time or across psychological statuses. Moreover, different annotators might give different ratings over the same photo. *Second*, most existing perception methods aim to learn scoring functions from individual images while the perception labels are actually associated with places. Figure 2 shows eight street-view photos of the same place (highlighted on the center map). These photos together provide a panoramic imagery description of the place, which is much more informative than a single street-view photo. For example, the safety level of a place might be mainly affected by one of the photos, which includes broken windows or other disorder visual patterns, but not other photos. Therefore, there is a demand for an urban perception method to develop a capability of learning from place-centric visual data.

Figure 2: A place with eight street-view photos of varying viewpoints. The perception labels (i.e. safety level) for these street-view photos should be collectively assigned.

1.2 Overview of this work

The *goal* of this work is to address the above issues with two innovative efforts. The first one is to collect a large-scale image-based urban perception dataset, which distinguishes itself from existing datasets by two aspects: (i) **place-centric**: providing multiple photos for every place, as shown in Figure 2; and (ii) **objectiveness**: mining the perceived level of safety using publicly available crime events records. The dataset includes about 20, 000 places in 5 major cities of U.S.A., and 8 street-view photos per place. To obtain the perceived safety level for every place, we collect crime event records from the government agencies in the past ten years and perform clustering analysis to obtain the crime density map for each city. Figures 1 (a) and (b) show the city map and crime density map for the city of San Diego, CA, U.S.A. To our best knowledge, the proposed objective and place-centric urban perception dataset is the first one in its catalog.

The other effort of this work is to develop a weakly supervised method to regress the perceived safety label for each place. We cast the learning of such a regression function in the multi-instance setting. Our method employs a multi-layer bag-instance representation: each place is described with a bag of street-view photos and each photo (at a particular viewpoint) with a bag of image regions. An image region can recursively decompose into a bag of sub-regions. This Place-Photo-Region-Subregion hierarchy forms a rich and redundant representation, and in the training stage, only place-level labels are available. We employ a deep convolutional neural network (DCNN) to parameterize the scoring function, which maps an input image or image region to a real-valued safety label. Our method follows the traditional rules of multi-instance regression [22]:(i) only a portion of the instances in a bag will affect the regression function to be learned; (ii) the bag-level label specifies the upper-bound of the predictions over all instances in the bag, i.e. that instance-level predictions should be less than or equal to the bag-level real-valued label. We divide all instances of a bag into a primary subset and a non-primary subset, and use the instances in the primary subset to estimate the scoring function. The primary/nonprimary labels of parent-children instances should be collectively assigned so that (i) all children nodes of non-primary node are non-primary instances as well; (ii) for any primary-instance, there is at least one primary child node. We refer our method to *hierarchical deep multi-instance regression* (HDMiR), which is different from the conventional multi-instance methods with two-layer bag-instance representation.

We develop a novel Expectation-Maximization (EM) [30] method for training HDMiR from weakly supervised street-view photos. Our method alternates between estimating primary instances for each bag (E-step), and optimizing the CNN parameters (M-step). In the E-step, we apply the current DCNN to score every instance (images or regions) and introduce a clustering method to robustly assign the highly scored instances to be primary instances and the others to be non-primary ones. We also introduce two bottom-up/top-down label propagation steps in order to preserve the above parent-children consistencies. In the M-step, we use the stochastic gradient descent (SGD) method to train the DCNN network in order to maximize the complete data likelihood defined over primary instances.

We apply the proposed EM method over a public dataset and a newly created dataset, and compare it to the alternative urban perception methods. Results show that the proposed HDMiR method can significantly improve perception accuracy on both datasets. This is a considerably improvement since the proposed method only requires place-level labels, instead of instance-level annotations. More importantly, the proposed method is capable of discovering the most influenced instances (e.g., images or regions) which might

affect the safety score of a place, which results in an explainable urban perception method.

1.3 Contributions

The two major contributions of this work include (i) a new large-scale place-centric urban street-view dataset with objective safety scores; (ii) a hierarchical multi-instance regression method for urban perception which can learn a deep scoring function from weakly supervised data. The proposed EM method can be used for learning multi-instance regression functions, and has great potentials in a wide variety of image-based applications.

2 RELATIONSHIPS TO PREVIOUS WORKS

The proposed urban perception method is closely related to three research streams in the multimedia community.

Visual Urban Perception aims to predict perceptual responses to scene images and play critical roles in public health and other socio-economic activities. In the past literature, researchers already developed a wide variety of perception models for predicting aesthetics [13], memorability [12], interestingness [6], virality [5]. In particular, Naik et al. [20] contributed a visual perception dataset, known as Place Pulse 1.0, and proposed to use various image features and support vector regression for predicting the perceived safety of street-view images. Dubey et al. [8] made efforts to a large-scale urban perception dataset, i.e. Place Pulse 2.0, through crowd-sourcing human annotations from online tools. While being remarkable successful, these datasets are restricted to the variances of human perceptions as well as mistakes or errors made during labeling. Moreover, human annotations are provided for images, instead of places, and might impose inconsistent labels for different street-view photos of the same place. In this work, we propose to create a large-scale place-centric dataset with place-level objective safety scores which are mined from crime event records.

Learning based perception methods can take advantages of large-scale labeled data and become much more feasible with the availability of high throughput online survey websites. These methods can be divided into two categories. The first category aims to directly learn a regression function to assign a real-valued label to the input images [22] [2] [10]. Such methods are limited since human annotators are not good at providing annotations in continuous space [8]. The second category, in contrast, aims to learn to rank pairs of images according to their perception scores. For example, Kiapour et al.[16] employed image features and manual annotations to learn to rank images according to their clothing styles. Zhu et al. [33] ranked facial images for attractiveness, for generating better portrait images. Wang et al. [25] introduced a deep ranking method for image similarity metric computation. Zagoruyko and Komodakis [29] developed a Siamese architecture for computing image patch similarity for applications like wide-baseline stereo. The predicated ranks of scene images are often convented to real-valued label in post-processing steps, e.g.,using the Microsft Trueskill algorithm [11], which might bring considerable errors.

Multi-instance Perception Learning We identify that visual urban perception is essentially a multi-instance learning (MIL) problem since the perceived labels, e.g., safety, are by its nature associated with places, instead of images (of specific viewing angles). Traditional MIL classification methods [7] [19] [1] [3] assume that the aggregation function over instance labels is an OR function, i.e. that a positive bag contains at least one positive instance and a negative bag contains only negative instances. A number of approaches relax the assumption and propose other forms of aggregation. Weidmann et al. [26] considers a generalization where the presence of a combination of instance types determines the label of the group. Xu and Frank [28] assume that all instances contribute equally and independently to a bag's class label. Zhou et al. [32] build a model that solves MIL through semi-supervised learning techniques by considering a negative label for every instance in a negative group. These solutions are typically tailored to handle specific assumptions about the whole-part relationships between groups and instances. The focus of this work is on the predictions of real-valued labels, e.g., safety scores, which is different from the discrete labels. Ray and Page [22] proposed an alternative procedure to find the primary instances and estimate regression functions. Our proposed approach generalizes such two-level bag-instance relationships through exploring multi-level place-image-region-subregion relationships in a hierarchy of instance representation.

3 OUR APPROACH

The goals of this work include two aspects: creating a place-centric urban perception dataset with objective safety scores, and developing deep machine perception algorithms which can leverage weak supervisions.

3.1 Community-Centric Urban Safety Dataset

An distinctive effort of this work is to collect an urban perception dataset, including both street-level photos and objective safety scores for each place. To do so, we employ crime event records (e.g., theft, fight, incidents, robbery etc.), which are publicly available in websites of government agencies, e.g., county police departments. The records include various details of crime events in the past decades (usually 10-15 years), e.g., date, time, place (longitude and latitude), and event types. Figure 1 (c) maps crime events in San Diego City, where each point represents a crime event. In this dataset, we obtain a total of $1,056,533$ crime events for the five major cites in U.S.A., including San Diego, Chicago, Seattle, San Francisco and New York City. Given places of these events, we employ the Parzen Window method to estimate the density of each place, and quantize a density label into five levels: 1 to 5. Places with lower density level is safer. These scores are automatically mined from historical community data and serve as objective measurements of place-wise safety indicators, which are much more accurate than human annotated labels used in existing datasets [8] [20].

These original place data, though informative, are obviously redundant since, for example, two places might be physically close to each other. In order to create a compact place dataset, we perform cluster analysis over the place clouds to group adjacent places together. In particular, we represent each place using its longitude/latitude coordinates, i.e. a two-dimensional feature vector, and run the K-means method over these vectors. We empirically set K and calculate the center location of each cluster of places. We consider each cluster as a place and set its safety level to be the

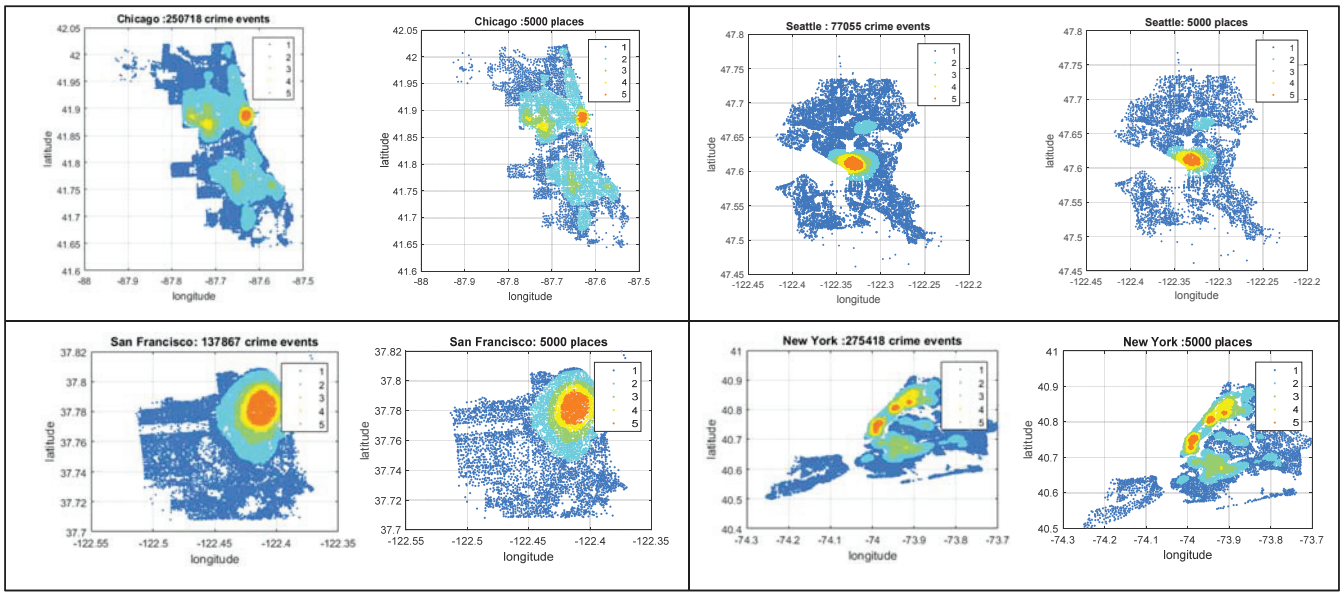

Figure 3: Visual places used in four major cities in U.S.A, including Chicago, San Francisco, Seattle, New York City. For each city, we visualize the locations of crime events in the left and the discovered places using cluster analysis in the right. The crime events are collected from local government agencies.

average safety level of its membership places. In this work, we set $K = 5000$ for all cities. Figure 3 shows the places generated for four cities: Chicago, Seattle, San Francisco, and New York City. For each cell, we show the original places (longitude/latitude) with crime densities (colored) in the left figure, and the clustered 5000 places in the right figure. The selected places are well distributed over the urban areas of the cities.

We employ the Google Street-View API to retrieve scene photos for each place in the dataset. The API provides access to the panoramic 360 degree views of a place using two parameters: location and point-of-view (or heading angles w.r.t. the true north). We evenly quantize heading angles into 8 bins and retrieve one street-level image for each bin at each location. Figure 2 shows the 8 street-view images downloaded for a place in San Diego. Thus, we obtain a place-centric urban image dataset, where safety scores are available for places only, rather than images or image regions. In contrast, most existing urban datasets utilize image-level perception labels while training machine learning models.

3.2 Learning Deep Multi-instance Regression for Urban Perception

We formulate the learning of visual urban perception models in the multi-instance setting. We consider a place as a bag of instances, where each instance represents a street-view image or an image region. In the training set, there are a set of n bags. Each bag consists of multiple instances, and a real valued safety label. Our goal is to learn a scoring function $f(\cdot)$ that returns a real value for each instance. Classical multi-instance regression methods [22] often divide all the instances of a bag into two categories: primary instances and non-primary instances, and use only primary instances to estimate the prediction function $f(\cdot)$. In this work, for example, an image full of facade surfaces or corners of an intersection might not be useful for determining the scoring function. We thus propose to identify primary instances and use these instances to train the proposed deep regression network.

In order to harness the recent technical breakthrough in deep feature learning, we employ deep convolutional network (DCNN) [17] to parameterize the scoring function $f(\cdot)$. Figure 4 illustrates the network architecture, which includes five convolution layers, three pooling layer sand three fully connected layers. For any instance $x \in R^d$, we perform forward propagation to get its activations of the last fully connected layer $fc8$ which can be viewed as high level features of the input image. We connect $fc8$ to an output unit, i.e., $f(\cdot)$. We take $f(x)$ as the real-valued prediction of the instance x. The outputs $f(x)$ of all instances in the same bag are connected to an aggregate layer in order to get the bag-level prediction. In this work, we use max() as the aggregate function.

Formally, let $\mathbf{x}_k = \{x_{ki}, z_{ki}\}, i = 1, 2, \ldots, n$ denote a bag of n instances, where each instance x_{ki} represents an image or image region, and $z_{ki} = 1$ if z_{ki} is a primary instance and 0 otherwise. Let $f(x_{ki})$ denote the prediction on the instance x_{ki}. Let $\mathbf{z}_k = [z_{ki}]$ pool over the latent variables for the bag k. The prediction over a bag $\mathbf{x}_k$ is defined as

$$f(\mathbf{x}_k; \theta, \mathbf{z}_k) = \phi\Big(f(x_{k1}), \ldots, f(x_{ki}), \ldots,\Big), \tag{1}$$

where ϕ represents the aggregate function, θ represents the DCNN parameters. Only primary instances, i.e., $z_{ki} = 1$, are used in the

above aggregate function and thus the non-primary instances do not affect the training of the deep scoring function $f()$.

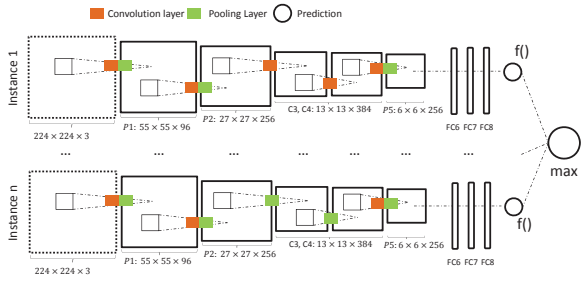

Figure 4: Deep Multi-instance Regression network. There are five convolution layers, three pooling layers, and three fully connected layers. All these layers are shared across the instances of the same bag.

When the hidden labels $\mathbf{z}$ are available, the objective function of the proposed multi-instance regression method is defined as follows,

$$J(\theta; \mathbf{z}) = -\log P(\mathbf{y}|\mathbf{x}; \theta, \mathbf{z}) = \sum_{k=1}^{M} \mathcal{L}\Big(f(x_k; \theta, \mathbf{z}_k), y_k\Big) \tag{2}$$

where $\mathbf{z}_k = [z_{ki}]$ pools over the hidden variables for the instances of the bag k, M is the number of bags, $\mathcal{L}(\cdots)$ is the loss function. We use the least square method, i.e. $\mathcal{L} = [y_k - f(x_k; \theta, \mathbf{z}_k)]^2$.

We develop an Expectation-Maximization (EM) method to jointly estimate the hidden labels$\mathbf{z}$ and learn the neural network from weakly supervised data. Let $\mathbf{x} = [x_k], \mathbf{y} = [y_k], k = 1, 2, \ldots, M$. We use the following probabilistic graphical model:

$$P(\mathbf{x}, \mathbf{y}, \mathbf{z}; \theta) = \prod_{x \in \mathbf{x}} P(x) \Big(\prod_{k,i} P(z_{ki}|x; \theta) \Big) P(\mathbf{y}|\mathbf{z}, \mathbf{x}) \tag{3}$$

The model $P(x)$ is a background reference distribution, $P(\mathbf{y}|\mathbf{z}, \mathbf{x})$ is the observation model.

Our EM method includes two iterative steps. In the M-step, with the estimated hidden variables $\hat{\mathbf{z}}$, we aim to estimate the network parameters θ through maximizing the following expected complete-data likelihood model,

$$Q(\theta; \theta') = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}; \theta') \log P(\mathbf{z}|\mathbf{x}; \theta) \approx \log P(\hat{\mathbf{z}}|\mathbf{x}; \theta) \tag{4}$$

In the E-step, with the estimated network parameters θ', we aim to solve the latent variables $\mathbf{z}$ by

$$\hat{z} = \arg\max_{\mathbf{z}} \prod_{k,i} P(z_{ki}|x_{ki}; \theta') P(\mathbf{y}|\mathbf{z}, \mathbf{x}) \tag{5}$$

$$= \arg\max_{z} \sum_{k,i} \log P(z_{ki}|x_{ki}; \theta') + \log P(\mathbf{y}|\mathbf{z}, \mathbf{x}) \tag{6}$$

The model $\log P(z_{ki}|x_{ki}; \theta')$ is defined over the outputs of the CNN regression network with the newly updated parameters θ'. In particular, we specify $P(z_{ki} = 1|x_{ki}; \theta') \propto f(x_{ki}; \theta')$, i.e. the predicated safety score for the instance x_{ki}. We set the observation model $\log P(\mathbf{y}|\mathbf{z}, \mathbf{x})$ to be a constant which allows us to estimate the hidden variables for each bag separately. In particular, we employ clustering methods, e.g., K-Means, to group all instances of a bag into two clusters according to their real-valued predictions $f(x_{ki}; \theta')$ with the current network parameters θ', and consider the higher-scored cluster as the primary subset. The predictions y_{ki}, however, might not be accurate at early stages of the EM iterations. To address this issue, we represent each instance x_{ki} using the activations $h(x_{ki})$ of the last layer of the CNN network, and develop algorithms to solve the two center features $c_{kp}, p = 1, 2$ and z_{ki} iteratively. For the bag k, we aim to optimize the following objective function.

$$\min_{c_{kp}, z_{ki}} \sum_{p=1,2} \sum_{z_{ki}=p} \|c_p - h(x_{ki})\|^2 + \alpha \sum_{z_{ki}=1} \sum_{z_{kj}=0} \mathbf{1}\Big(f(x_{ki}) \geq f(x_{kj})\Big) \tag{7}$$

which is an integer optimization problem. The first team is a classical data objective for clustering and the second term encourages the ordering of the two desired clusters.

We develop a constrained clustering method to solve Eq. (7). Our method, like K-Means, randomly selects two instances and use their activations as the center vectors c_1 and c_2. Then, we employ three alternative steps: (i) to assign every instance into one of the two groups, which has closer center vector; (ii) to swap instances between the two clusters so that all instances in the first group have higher safety scores than the other group; (iii) to calculate the new center vector for each group. We alternate these three steps multiple times until convergence. The swapping step is used to enforce the constrains defined in Eq. (7).

We further extend the above constrained clustering method to explore the hierarchical relationships between places, street-view photos and image regions. An image region can further decompose into multiple sub-regions in a recursive fashion. In such a hierarchical setting, it is convenient to denote the number of layers to be L. The nodes of top layer (i.e., places) are provided with real valued safety labels, while the other nodes are not available. An node of layer l includes multiple children nodes of layer $l + 1$, which are labeled as primary or non-primary instances. There are two multi-instance assumptions: (i) all the children nodes of a non-primary node are non-primary as well; (ii) a primary node should have at least one primary child node.

Algorithm 1 summarizes the sketch of identifying primary instances in the hierarchy. We employ two propagation steps to ensure consistencies of parent-child label assignments. The top-down propagation is used to set the children nodes of a non-primary node to be non-primary and the bottom-up propagation is used to set the ancestor nodes of an primary node to be primary. These two steps are performed once for each iteration.

3.3 Sketch of HDMiR

Algorithm 2 summarizes the proposed HDMiR method. In the E-step, we use Algorithm 1 to estimate the latent variables; in the M-step, we employ the stochastic gradient descent method to optimize the network parameters θ. Note that regions of varying resolutions are resized and used to train the same deep network. Using a single deep network for multiple purposes have been proved to be successful in the past efforts [24].

Algorithm 1 Hierarchical Primary Instance Recognition

Input: streetview photos of a place and their features;
Output: Latent variables z_{ki} for individual instance nodes in the hierarchy.
Initialize two centroids: c_{k1}, c_{k2};
Iterate until convergence,
- Assign every instance to one of the two groups with closer centroid;
- Swap instances between the two clusters if they violate the constraints in Eq. (7);
- *Top-down Propagation*: for any non-primary node, set its offspring nodes to be non-primary nodes;
- *Bottom-up Propagation*: for any primary node, set its ancestor nodes (including parent nodes) to be primary nodes;
- Re-calculate the centroids for each cluster;

Algorithm 2 Sketch of the proposed EM algorithm

Input: multiple places with their street-view photos and place-level safety scores;
Output: deep multi-instance regression network parameters θ
Initialize θ with pretrained network models;
Iterate until convergence,
- E-step: estimate the latent variables z by Algorithm 1;
- M-step: train the network parameters θ by the stochastic gradient descent method;

The proposed HDMiR method can be also applied over the traditional image-centric urban perception dataset, considering each image as a hierarchy of region-subregions. In experiments, we will show that such a deep network with alternative optimization techniques can significantly improve urban perception quality on both traditional image-centric public datasets and the newly created place-centric image datasets.

4 EXPERIMENTS

In this section, we apply the proposed hierarchical deep multi-instance regression (HDMiR) method over both public datasets and the newly created urban perception dataset and evaluate it in both qualitative and quantitative ways.

4.1 Evaluation Protocols

Datasets

We use two datasets to evaluate the proposed regression method. The first one is a newly created dataset, which includes safety ranks (or scores) for 20,000 places in five major U.S.A. cities and 8 photos for each place. We retrieve these 160,000 images using the Google Street-view API and mine their safety scores from the historical crime records maintained by local government agencies. We divide each photo into 5 subregions to enable the proposed hierarchical multi-instance regression method. For evaluation purposes, we split the places of each city into two subsets: 3000 places for training and validation, and 2000 for testing. More details of this dataset can be found in Section 3.1.

The second dataset *StreetScore* includes 4109 images in two cities: Boston and New York. Each image is provided with a safety score (between 0 and10). The scores are derived from pair-wise rankings of photo pairs in response to the question 'which place looks safer?'. A total of 208,738 comparisons were collected from an online game, which are converted to ranked scores using the Microsoft Trueskill algorithm [11]. It is noteworthy that the upgraded version of this dataset, known as Place Pulse 2.0 [8], provides pair-wise ranking annotations only and is not applicable for the regression purposes. In particular, the TrueSkill algorithm usually needs $24-36$ comparisons per image for obtaining stable ranking, whereas there are only 3.35 comparisons per image in the Pulse 2.0. We evenly split the images of *StreetScore* dataset into two subsets and use one for training and the other for testing. Note that this dataset is image-centric, and only individual street-view photos are provided with safety scores. To apply multi-instance methods, we consider each street-view photo as a bag of instances, and each instance represents an image region.

Implementation

We implement the proposed hierarchical deep multi-instance regression (HDMiR) method as follows. For the E-Step, we set the maximal iteration of the Algorithm 1 to be 20. For the M-step, we use the stochastic gradient descent method with mini-batches to train the network [17]. Each mini-batch contains 30 bags. The initial learning rate is 0.001 and is decreased by a factor of 0.1 after every 2000 iterations. We set the momentum to be 0.9 and the weight decay to be 0.0005. The maximal iteration is set to be 120,000. The network parameters θ are pre-trained on the ImageNet for classification purpose. Fine-tuning the proposed network on the newly created dataset takes about 96 hours on a NVIDIA Telsa K40 GPU. The average inference time for one image is about 0.1 seconds. We resize all images or image regions in the instance hierarchy to be 224 by 224 pixels, and use them as inputs to to the deep neural network.

Baseline Methods

We compare the proposed method to alternative regression methods in both supervised setting or multi-instance setting. In the supervised setting, we directly assign bag-level labels to instances and train various machine learning models using either manually engineered features or deep learned features. These methods include. (I) *Streetscore* [20], which uses support vector regression (SVR) and a group of appearance features, including GIST, Texton histogram, color histogram, HOG, Dense SIFT, LBP etc. (II) SVR with deep features over pre-trained networks. In particular, for each image or image region, we feed it to the pretrained CNN network (VGG Network [24]), and use the activations (of 4096 dimensions) as features. (III) Deep regression networks, including the AlexNet [17], the VGGNet [24] and the PlacesNet [31], which are trained in an end-to-end fashion. For these deep models, we used their pre-trained models publicly available in the Caffe framework, and fine-tuned them on the two datasets separately for regression purposes. We use least square loss for all these three regression networks.

In the multi-instance setting, we use the multi-instance regression (MIR) [22] method, which employs an alternative method to jointly discover a single primary instance for each bag, and to solve the regression problem. The original work used linear regression

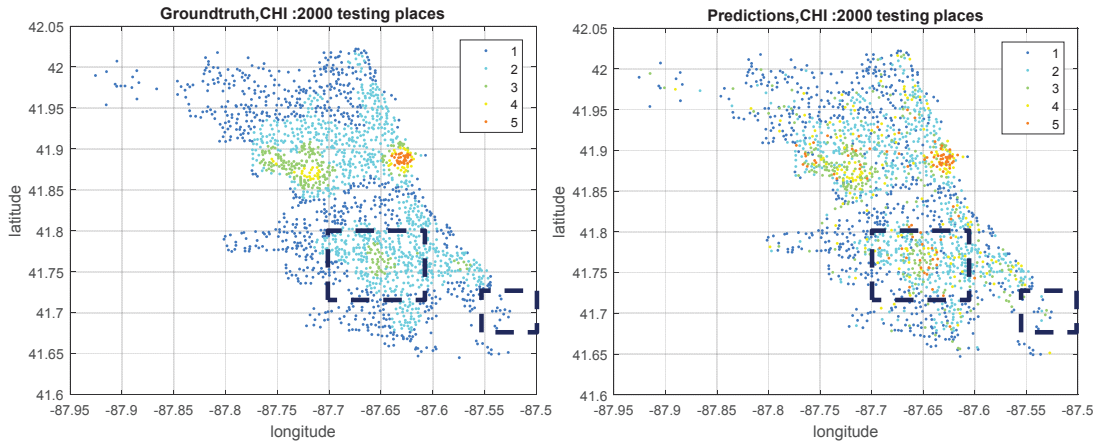

Figure 5: Results for the places in Chicago. Left: groundtruth safety levels for the 2000 testing places; right: the safety levels predicated by the proposed algorithm. Highlighted areas include failure places.

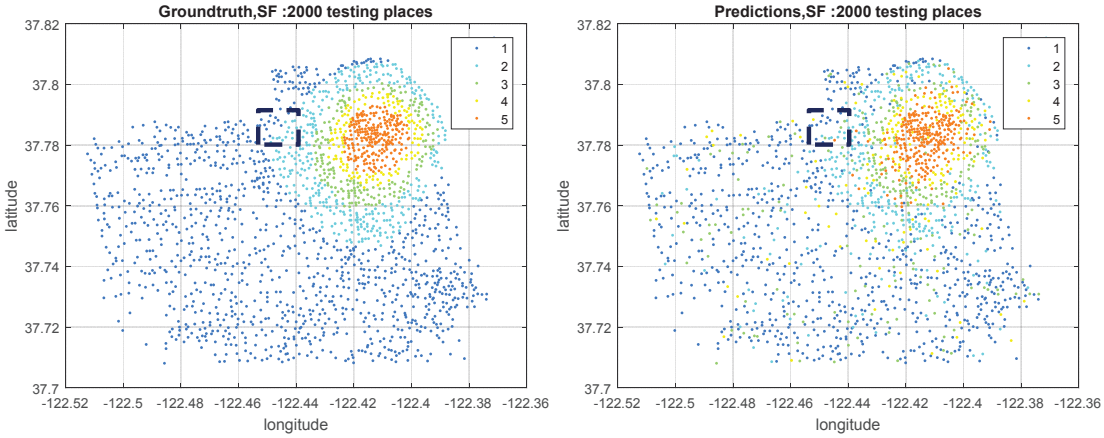

Figure 6: Results for the places in San Francisco. Left: groundtruth safety levels for the 2000 testing places; right: the safety levels predicated by the proposed algorithm. Highlighted areas include failure places.

which is not suitable for dealing with high-dimensional image data. Instead, we use the deep regression network VGGNet, as introduced before, to regress the primary instances.

Metrics

We use the coefficient of determination R^2 [20] between true scores y and predicated scores $\hat{y}$ to evaluate the accuracy of a regression model. R^2 is a quantitative measure for the proportion of total variance of true data explained by the prediction model. Let $\bar{y}$ denote the average predicated safety level of all testing samples. We defined $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})}$.

Results on the place-centric dataset

We apply the proposed HDMiR method over the newly created place-centric image data for learning to predict instance-level safety levels. Table 1 reports the quantitative results of various methods. In order to analyze the effects of the proposed instance hierarchy, we implement a variant of Algorithm 2 which uses two-layer bag-instance image representation, called Deep Multi-instance Regression (DMiR). From the table, we found that proposed weakly supervised deep regression method methods (DMiR and HDMiR) clearly outperform the classical regression method *Streetscore* [20] and the more recent end-to-end learning based deep methods. It is noteworthy that the three networks AlexNet, VGGNet and PlacesNet are popular deep learning models which achieved remarkable successes in multiple domains. These improvements are because of the proposed weakly supervised framework which can takes advantages of the bag-instance constraints. Moreover, from the comparisons between DMiR and HDMiR, we can observe that the proposed instance hierarchy can further enhance system accuracy.

Figures 5 and 6 visualize the predicated safety levels (1-5) for two cities, Chicago and San Francisco, respectively. Each point represents a testing place in longitude/latitude coordinates. The predictions are obtained by the proposed HDMiR method. For each

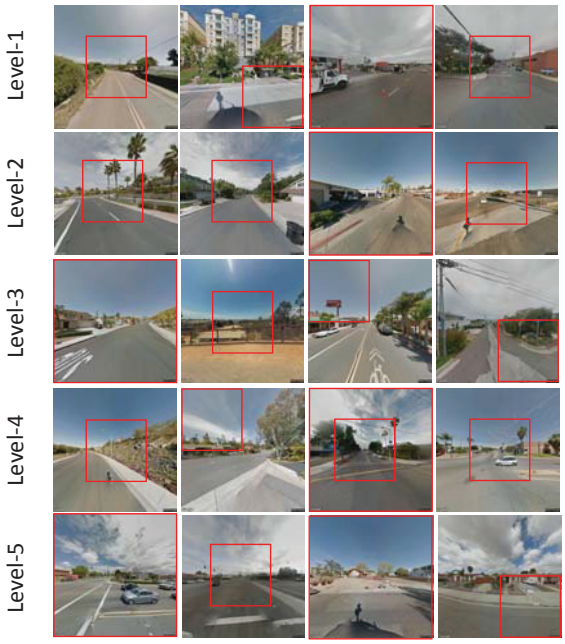

Figure 7: Exemplar results on places in San Diego. Each row shows four street-view photos of different places, which are successfully classified by the proposed HDMiR (within 0.5 distance from the true level) and are not classified correctly by the baseline methods. For each image, we show the regions with highest safety score.

figure, we show the ground-truth labels in the left column and the predicated labels in the right column. Figure 1 visualizes the predications for the places in San Diego. Note that we learn a single regression model for all the five cities. We can observe that the predications are consistent with the true safety levels. We also highlight a few areas which include obvious wrong ranks. Figure 7 visualizes four exemplar street-view photos (one per place), which are correctly predicated by the proposed HDMiR method, but are mis-predicated by the other baseline methods. We consider a prediction to be correct if it is 0.5 away from the ground-truth level. Each image is overlaid with the highest scored instance, either a subregion or a whole image. These exemplar street-view photos are very challenging because, for example, the photos of level 1 have similar appearance as the photos of level 2, and existing urban perception methods failed to predict the correct safety level. In contrast, the proposed method can work well through exploring the consistencies in the hierarchy of instances.

Results on the StreetScore dataset

Table 2 reports the quantitative results on the StreetScore dataset [20]. We consider each photo as a bag of five regions, and apply the proposed HDMiR method or other multi-instance regression methods. It is noteworthy that similar observations can be drawn from the comparisons between the proposed multi-instance regression methods and alternate methods. These results demonstrate that the proposed regression method for place-centric street-view data can be generalized to the image-centric image data, and can achieve equally promising perception accuracies.

Algorithms	R^2
Streetscore [20]	0.49
SVR+Deep Features	0.51
AlexNet [17]	0.53
VGGNet [24]	0.61
PlacesNet [31]	0.65
MIR [22]+VGGNet Features	0.69
DMiR	0.78
HDMiR	**0.81**

Table 1: Quantitative results (R^2)on the Place-centric Dataset.

Algorithms	R^2
Streetscore [20]	0.54
SVR+Deep Features	0.58
AlexNet [17]	0.62
VGGNet [24]	0.68
PlacesNet [31]	0.69
MIR [22]+VGGNet Features	0.72
HDMiR	**0.84**

Table 2: Quantitative results R^2 on the StreetScore Dataset.

5 CONCLUSIONS

This paper presented a hierarchical deep multi-instance regression (HDMiR) method for learning robust visual perception models from weakly supervised images. We used advanced neural networks as base models, and introduced a multi-instance regression network to predict safety scores for instances, i.e. street-view images or their image regions. We contributed an innovative hierarchical instance representation and developed an EM algorithm to jointly identify primary instances for each bag, and to learn the deep regression network. Another significant effort of this work is to create a novel place-centric urban perception dataset with objective safety scores, which is different from most existing datasets which crowd-sourced image-based safety scores from human beings. We apply the proposed regression method over both the newly created dataset and publicly available dataset. Results with comparisons to alternative methods clearly demonstrated the advantages of the proposed methods.

The proposed HDMiR method has great potentials in various regression problems, and the developed deep networks and EM techniques can be applied to solve other types of perception tasks, e.g., house pricing, market analysis, and transportation demand estimations.

ACKNOWLEDGEMENTS

This project was supported by the National Science Foundation (Grant No. 1657600), China State Key Development Program (Grant No. 2016YFB1001004), and Guangdong Natural Science Foundation (Grant No. 2015A030313129)

REFERENCES

[1] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Support vector machines for multiple-instance learning. *Advances in neural information processing systems* (2003), 577–584.

[2] Sean M Arietta, Alexei A Efros, Ravi Ramamoorthi, and Maneesh Agrawala. 2014. City forensics: Using visual elements to predict non-visual city attributes. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2624–2633.

[3] Olivier Chapelle and Alexander Zien. 2005. Semi-Supervised Classification by Low Density Separation.. In *AISTATS*. 57–64.

[4] Deborah A Cohen, Karen Mason, Ariane Bedimo, Richard Scribner, Victoria Basolo, and Thomas A Farley. 2003. Neighborhood physical conditions and health. *American journal of public health* 93, 3 (2003), 467–471.

[5] Arturo Deza and Devi Parikh. 2015. Understanding image virality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1818–1826.

[6] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1657–1664.

[7] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89, 1 (1997), 31–71.

[8] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. In *European Conference on Computer Vision*. Springer, 196–212.

[9] Akilah Dulin-Keita, Herpreet Kaur Thind, Olivia Affuso, and Monica L Baskin. 2013. The associations of perceived neighborhood disorder and physical activity with obesity among African American adolescents. *BMC public health* 13, 1 (2013), 440.

[10] Edward L Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik. 2016. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry* (2016).

[11] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkillâĎć: a Bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*. MIT Press, 569–576.

[12] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2011. What makes an image memorable?. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 145–152.

[13] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z Wang, Jia Li, and Jiebo Luo. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28, 5 (2011), 94–115.

[14] Kees Keizer, Siegwart Lindenberg, and Linda Steg. 2008. The spreading of disorder. *Science* 322, 5908 (2008), 1681–1685.

[15] Cheryl M Kelly, Jeffrey S Wilson, Elizabeth A Baker, Douglas K Miller, and Mario Schootman. 2013. Using Google Street View to audit the built environment: inter-rater reliability results. *Annals of Behavioral Medicine* 45, 1 (2013), 108–112.

[16] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. 2014. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*. Springer, 472–488.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[18] Mirte AG Kuipers, Mireille NM van Poppel, Wim van den Brink, Marleen Wingen, and Anton E Kunst. 2012. The association between neighborhood disorder, social cohesion and hazardous alcohol use: a national multilevel study. *Drug and alcohol dependence* 126, 1 (2012), 27–34.

[19] Oded Maron and Aparna Lakshmi Ratan. 1998. Multiple-Instance Learning for Natural Scene Classification.. In *ICML*, Vol. 98. 341–349.

[20] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 779–785.

[21] Nikhil Naik, Ramesh Raskar, and César A Hidalgo. 2016. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *The American Economic Review* 106, 5 (2016), 128–132.

[22] Soumya Ray and David Page. 2001. Multiple instance regression. In *ICML*, Vol. 1. 425–432.

[23] Philip Salesses, Katja Schechtner, and César A Hidalgo. 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS one* 8, 7 (2013), e68400.

[24] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[25] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1386–1393.

[26] Nils Weidmann, Eibe Frank, and Bernhard Pfahringer. 2003. A two-level learning method for generalized multi-instance problems. In *European Conference on Machine Learning*. Springer, 468–479.

[27] James Q Wilson and George L Kelling. 1982. Broken windows. *Critical issues in policing: Contemporary readings* (1982), 395–407.

[28] Xin Xu and Eibe Frank. 2004. Logistic regression and boosting for labeled bags of instances. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 272–281.

[29] Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4353–4361.

[30] Yongyue Zhang, Michael Brady, and Stephen Smith. 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging* 20, 1 (2001), 45–57.

[31] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[32] Zhi-Hua Zhou and Jun-Ming Xu. 2007. On the relation between multi-instance learning and semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*. ACM, 1167–1174.

[33] Jun-Yan Zhu, Aseem Agarwala, Alexei A Efros, Eli Shechtman, and Jue Wang. 2014. Mirror mirror: Crowdsourcing better portraits. *ACM Transactions on Graphics (TOG)* 33, 6 (2014), 234.