StreetNet: Preference Learning with Convolutional Neural Network on Urban Crime Perception

Kaiqun Fu
Virginia Tech
7054 Haycock Road
Falls Church, Virginia 22043
fukaiqun@vt.edu

Zhiqian Chen
Virginia Tech
7054 Haycock Road
Falls Church, Virginia 22043
czq@vt.edu

Chang-Tien Lu
Virginia Tech
7054 Haycock Road
Falls Church, Virginia 22043
ctlu@vt.edu

ABSTRACT

One can infer from the broken window theory that the perception of a city street's safety level relies significantly on the visual appearance of the street. Previous works have addressed the feasibility of using computer vision algorithms to classify urban scenes. Most of the existing urban perception predictions focus on binary outcomes such as safe or dangerous, wealthy or poor. However, binary predictions are not representative and cannot provide informative inferences such as the potential crime types in certain areas. In this paper, we explore the connection between urban perception and crime inferences. We propose a convolutional neural network (CNN) - *StreetNet* to learn crime rankings from street view images. The learning process is formulated on the basis of preference learning and label ranking settings. We design a street view images retrieval algorithm to improve the representation of urban perception. A data-driven, spatiotemporal algorithm is proposed to find unbiased label mappings between the street view images and the crime ranking records. Extensive evaluations conducted on images from different cities and comparisons with baselines demonstrate the effectiveness of our proposed method.

CCS CONCEPTS

•**Information systems** → *Geographic information systems; Data mining;* •**Computing methodologies** → *Learning to rank; Perception;*

KEYWORDS

preference learning, street view, convolutional neural networks, spatial analysis

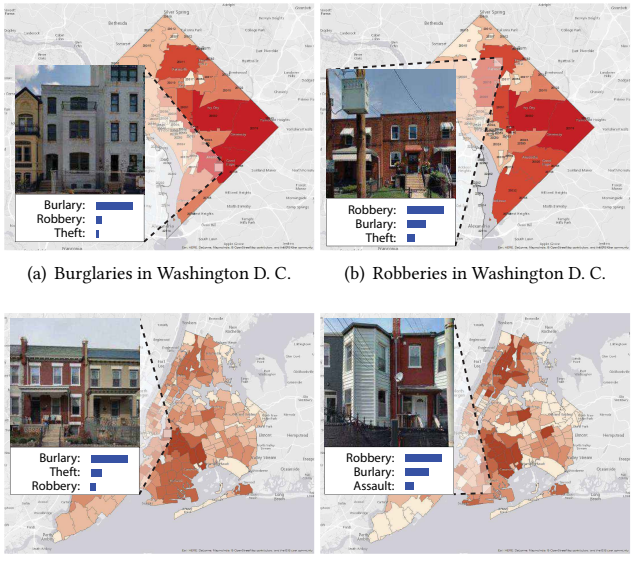
ACM Reference format:
Kaiqun Fu, Zhiqian Chen, and Chang-Tien Lu. 2018. StreetNet: Preference Learning with Convolutional Neural Network on Urban Crime Perception. In *Proceedings of 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, November 6–9, 2018 (SIGSPATIAL '18),* 10 pages.
DOI: 10.1145/3274895.3274975

SIGSPATIAL '18, Seattle, WA, USA

DOI: 10.1145/3274895.3274975

1 INTRODUCTION

The broken window theory is a criminological theory of the norm-setting and signaling effect of urban disorder and vandalism on additional crime and anti-social behavior. The theory was first proposed by *James Wilson* and *George Kelling* in *The Atlantic Monthly* in March 1982 [42]; quotes: *Consider a building with a few broken windows. If the windows are not repaired, the tendency is for vandals to break a few more windows. Eventually, they may even break into the building, and if it's unoccupied, perhaps become squatters or light fires inside.* Similar to the rapid development of the idea that social network's justify the *six degrees of separation* theory in sociology, the *broken window theory* in criminology may find its endorsement in our era of big data. Previous studies on urban crime analysis [22, 36, 37] have addressed significant associations between the locations of crime offenses and the categories of the offenses. However, all those works neglect the impact of street view images on urban safety perception problems.



(a) Burglaries in Washington D. C. (b) Robberies in Washington D. C.

(c) Burglaries in New York City (d) Robberies in New York City

Figure 1: Spatial Distributions of Different Categories of Crime

With the advent of image-based crowdsourcing services such as *Flickr* and *Instagram*, users can easily generate image data. Panoramic

image services such as Google Street View are also ubiquitously accessible from the Internet. Previous studies have presented spatial correlations between crime levels and residences of offenders due to the fact that most offenders prefer to commit illegal activities close by, and offenders will follow the same criminal patterns while they are committing illegal activities. For example, a person with burglary records is likely to commit burglary in the future [3]. Further assumptions are made for our exploration that different types of crimes will affect the urban visual appearance in a variety of ways. For instance, convenience stores located in suburban areas with high robbery rates will be equipped with substantial barriers or even bulletproof armor; a lot of graffiti will be witnessed in places with inadequate law enforcement. Figure 1 illustrates crime rate heat maps for different offense categories for both Washington D. C. and New York City. Figure 1 also shows that different areas of the cities are represented by different urban perceptions, and such distinctions can be utilized to infer hidden crime rankings. This is the main focus of our paper.

Learning crime rankings from urban perception or street view images can be challenging. This poses the following three issues: 1) **Features of images for crime ranking are not explicit**. Representations of urban appearances from street view images vary substantially due to changes in camera direction and imaging and lighting conditions. Previous studies [11, 43] in image classification extract features from clustered bag-of-visual-words (BOVW) methods, but the extracted features are not interpretable. 2) **Prior geographical knowledge should be considered for street view images retrieval**. Learning hidden knowledge from street view image datasets is different from traditional image classification problems. The selection of camera directions significantly affects the prediction results. Optimal camera directions are ones that are perpendicular to the streets' direction because this direction can minimize the noise introduced by recorded vehicles or pedestrians. 3) **Lack of labeled datasets**. To learn crime rankings from street view images, reasonable and unbiasedly labeled training data is required. However, there are no existing labeled street view image data available for crime ranking tasks. In addition, the techniques of feature engineering request tedious labor, and the human-annotated corpora are insufficient for training practicable models to identify crime rankings. 4) **Insufficient urban perception study on multi-label analysis**. Previous studies on urban perceptions only focus on binary classifications of income and safety levels for neighborhoods [16, 32, 33], which provide less informative inspections than the multi-label learning for residents and law enforcement agencies.

The methods proposed in this paper effectively address the above-mentioned issues. The proposed convolutional neural network extracts hidden features from street view images, and we formulate the crime rank learning problem under the preference learning framework. Improving upon previous works on safety level prediction [32], we demonstrate the feasibility of inferring crime ranking knowledge from cities' visual appearances.According to the previous assumption, we utilize spatiotemporal correlations between street view images and criminal offenses to construct crime ranking labels. The major contributions of this paper can be summarized as follows:

•**Propose a convolutional neural network (CNN) based preference learning approach for crime ranking inference from street view images**: A convolutional neural network is proposed and trained on street view images labeled with crime rankings from multiple cities. We formulate the problem under the settings of preference learning and label ranking.

•**Develop a street view image retrieval algorithm with improved abilities in representing actual urban perceptions**: An efficient street view image retrieval algorithm is designed and implemented while generating the image datasets. The retrieved image datasets provide better urban perception representations than the previous datasets for those places. Such improvement assists our model in achieving a better prediction performance.

•**Design a data-driven spatiotemporal street view image and crime ranking labeling strategy**: A spatiotemporal based street view images and criminal offense record mapping algorithm is designed for labeling the images. The proposed labeling scheme is more representative and efficient than previous methods because the process is unbiased and systematic.

•**Extensive experiments and make comparisons to validate the effectiveness and efficiency of the proposed techniques**: We compare our proposed convolutional neural network with various methods. Conventional methods for learning label rankings are selected for comparisons. Evaluations of various metrics and detailed case study analysis are presented illustrating the effectiveness of the proposed method. Interesting discoveries for street view images' perception radius (in feet) are also presented and discussed.

The rest of our paper is structured as follows. Related works are reviewed in section 2. In section 3, we describe the problem setup of our work. In section 4, we present a detailed discussion of our proposed methods for predicting potential crime rankings from street view images. In section 5, extensive experiment evaluations and comparisons are presented. In the last section, we discuss our conclusion and directions for future work.

2 RELATED WORKS

In this section, we provide a detailed review of the current state of research for urban crime perception problem. There are several threads of related work of this paper: urban perception from street view imagery data, scene recognition and classification, and preference learning on multi-label learning.

Urban Perception. The earliest studies on urban perception [3, 4] indicate strong spatial coherence between the locations of illegal offenses and the residences of the offenders; these studies confirm that offenders who commit robberies, residential burglaries, thefts from vehicles, and assaults are more likely to target their current and former residential area than similar areas they never lived in. Previous works [10, 13, 20, 40] addressed problems of regional public safety and urban appearance perception. For example, correlations between a high initial level of homicide and losses in total population are observed [30] in suburban areas adjacent to a large city like Chicago. However, without street view images under a city-wide coverage, these previous works drew conclusions based on experiments with small datasets (160 manually taken photographs), which is insufficient for mining latent patterns for the majority of the urban appearances. In contrast, the method proposed in

this paper is trained on 44,694 street view images from two cities: Washington, D.C., and New York City.

Recent branches of works in urban perception applied computer vision and deep learning techniques to improve the resolution, precision, and scale. Ordonez et al. [34] proposed a regression model to predict the perceptual characteristics of places for wealth, uniqueness, and safety. The proposed model utilized features such as Gist, SIFT and Fisher Vectors. Such hand-craft features were not representative enough for large street view datasets and were outperformed by deep learning-based algorithms. Dubey et al. [12] proposed a convolutional neural network to quantify the urban perception along six perceptual attributes: safe, lively, boring, wealthy, depressing and beautiful. Andersson et al. [1] proposed a novel 4-Cardinal Siamese convolutional neural network to predict urban crime rates. However, this model applied four pre-trained VGG-16 architecture, which is not representative of the urban perception tasks. Liu et al. [26] also proposed a convolutional neural network for urban safety perception based on the crime dataset. Most of the deep learning based urban perception methods for safety inference focus on crime rate prediction and safety level comparison. Subjective labels are inevitably introduced in the previous works as most of the evaluations of the studies are performed by humans. In this paper, we inspirationally address the correlations between the urban appearance and the crime types. We also objectively labeled and evaluated based on official crime records as the golden standard.

Scene Recognition and Classification. Previous works have demonstrated the feasibility of considering images of the appearance of city streets as an indicator of hidden urban inferences such as safety, wealth, and aesthetics [2, 14, 49]. Several previous works have proposed computer vision techniques based on supervised classification algorithms such as SVM or convolutional neural networks (CNN) for predicting the safety level of a specific urban area. Although the question, "Does this place look safe?" has been resolved, previously proposed works only consider binary classes of safety levels or solving the safety index regression problem. Various research ideas on street view images have been proposed in recent years. Zamir et al. [45, 46] proposed a street view image location retrieval approach with SIFT vocabulary trees and generalized minimum clique graphs. Similar research problems of recognizing objects such as street numbers [17], storefronts [31], and other object recognition [15, 44, 48] also were addressed recently. Other works focusing on 3-D reconstruction and city modeling based on street level imagery have been proposed [7, 29].

Preference Learning. Preference learning algorithm for ranking was previously proposed in [19] for multi-label learning problems. Previous researchers utilized constraints derived from multi-label instances to enforce that the ranking of relevant classes is higher than the irrelevant ones. Based on the proposed preference learning structure, further applications of multiple-object detection and image tag ranking problems [23, 24] have been studied under such a problem setting. Most of the previous works in multi-label ranking applied the pairwise model [6]. However, the pairwise model for learning label preferences often suffers from the expensive computation. We formulate crime preference learning using a convolutional neural network. Such a design exploits convolutional layer's advantages for image feature extraction and deep neural network's learning ability for multi-label tasks.

3 PROBLEM STATEMENT

With machine learning algorithms on a huge street view image dataset, it may be feasible for a human to perceive or inference the types of criminal acts that are mostly likely to be committed to him in a certain area.

Consider a setting where potential crime rankings are inferred based on the given street view image in a certain area. We name this procedure a perceptional crime rank inference.

In the problem setting, we are given a street view image space $\mathcal{I}$ and a finite set of crime labels $C = \{c_1, c_2, ..., c_k\}$. The assumption has been made that there is a hidden correlation between the physical appearances of the city areas and the crime rankings in those areas. We denote the training dataset of n inferences as $\mathcal{D}_n \subseteq \mathcal{I} \times C$.

The general goal is to learn a "crime ranker" in the form of a $\mathcal{I} \rightarrow S_C$ mapping, where the output space S_C is given by the set of all permutations of the set of crimes C. Thus, label ranking can be seen as a generalization of conventional classification, a complete ranking is associated with a street view image I:

$$c_{\pi_I^{-1}(1)} \succ_I c_{\pi_I^{-1}(2)} \succ_I \ldots \succ_I c_{\pi_I^{-1}(k)} \tag{1}$$

where π_I is a permutation of $\{c_1, c_2, ..., c_k\}$ such that $\pi_I^{-1}(i)$ is the position of crime c_i i in the ranking associated with the given street view I.

We formulate the problem of crime ranking from street view images as a pointwise preference learning problem on different crime types. The goal is to learn a relevance score $f_i(I) = rel_i$ prediction function for each crime type c_i from the street view images, and a set of pairwise preferences of the form $c_i \succ_I c_j$ from the training data $\mathcal{D}_n$. Such an outcome suggests that for street view image I, c_i is preferred to c_j. For each rank judgment on crime pairs c_i and c_j, the goal is to estimate a function $f \in \mathcal{I} \rightarrow \mathbb{R}$ and $\mathcal{F} = \{f | f_i(I) > f_j(I) \Leftrightarrow c_i \succ_I c_j; (i \neq j)\}$, where f_i represents a prediction function for crime type i. To generalize the proposed problem, we present the following:

$$f^* \in \arg\min_{f \in \mathcal{F}} \sum_{\mathcal{D}_n} R^*(f) + \Omega(f) \tag{2}$$

where $R^*(f)$ corresponds to the empirical risk whose performance is controlled by the selection of the loss function. A general representation of the empirical loss is given by:

$$R^*(f) = \frac{1}{|\mathcal{D}_n|} \sum_{(I,C) \in \mathcal{D}_n} L(y, f(I)) \tag{3}$$

To compare with the baseline algorithms, we discuss loss function selections for the conventional rank learning settings. The loss function $L(y, f(I))$ in the empirical risk determines the descending direction of the learning process. Note that y is the true relevance score of an image I for a given crime type. Under the pairwise preference learning setting, various loss functions can be chosen. In this paper, two loss functions are considered: 1) the logistic loss/cross entropy loss and 2) the squared Hinge loss for the SVMs. Both loss functions are smooth and convex. Consequentially, squared hinge loss and logistic loss are formed respectively:

$$L_{hinge^2} = \sum_{\mathcal{D}_n} max^2(1 - \phi(\mathbf{w}^T \mathcal{F}_I + b), 0) \tag{4}$$

$$\Omega_{l_2}(f) = \lambda \|\mathbf{w}\|^2 \tag{5}$$

In Equation 5, $\Omega(f)$ is the regularization term for controlling the complicity of the model. For the SVM classifiers, only l_2 norm regularization is utilized, shown in Equation 5, where λ is the trade-off parameter controlling the complexity of the model.

4 METHODOLOGY

In this section, we discuss the design of the proposed convolutional neural network and its training and solution processes. We also provide detailed discussions of the direction-based, street view image retrieval algorithm.

4.1 StreetNet

In conventional image classification tasks, performance is greatly dependent on feature selection. However, information loss is inevitably introduced to the classifier with such feature extraction mechanisms. In contrast, convolutional neural networks significantly keep complete image information. We propose an convolutional neural network - *StreetNet* for crime type inference from street view images. The structure of the proposed network is presented in Figure 2. The first several layers of the neural network are convolutional layers, and they can be considered as feature extraction operators on the images globally. The difference between our convolutional neural network based rank learning and other pointwise rank learning algorithms is that we can learn the relevance score simultaneously for different crime types. This advantage is introduced by the structures of the fully connected layers and output layer of our convolutional neural network.

4.1.1 Latent Features Extraction. Convolutional layers are implemented for extracting latent features of street view images. A Convolutional layer performs a convolution operation with a filter size of $k \times k$ on the output of its previous layer. The convolutional layer is represented:

$$\mathbf{I}_j^n = f\left(\sum_{i=1}^{L^{n-1}} \mathbf{I}_i^{n-1} * \mathbf{W}_{ij}^n + b_j^n\right) \tag{6}$$

where I is the image feature matrix, n represents the n^{th} layer of the convolutional neural network; $\mathbf{W}$ is the flattened filter with a size of $k \times k$; b_j^n is the bias of the feature filter $\mathbf{W}$; f is the specified activation function; and $*$ is the 2D convolution operation. The max-pooling layer calculates the maximum activation on the areas that are not overlapping with the filter $\mathbf{W}$. The max-pooling layer down- samples the street view images by the size of the filter.

4.1.2 Hidden Features Classification. Fully connected layers are utilized for inferring relevance scores from the extracted latent features. For each crime type, our goal is to learn a regression of the relevance score for the given street view image. A linear operation with weight matrix $\mathbf{w}$ and bias $\mathbf{b}$ is performed on the output features of the last convolutional layer. The result of this linear operation is fed into a rectified linear unit (*ReLU*) activation function. For each hidden node in the fully connected layer, *ReLU* outputs an activation. In the last output layer, we sum the activations and multiply the sum of the activations by a vector of 1s. While training, the root-means- squared-error (*RMSE*) is selected as the loss function for the fully connected layers. The design of our convolutional neural network is shown in Figure 2.

4.1.3 Parameter Optimization. Various selections of optimization methods are available to optimize the empirical risk minimization problem in convolutional neural networks. In our experiment, we use AdaDelta [47], a variation of gradient descent, for optimizing the neural network.

The AdaDelta on the other hand restricts the window of accumulated past gradients to some fixed size $\mathbf{w}$. This method reduces the aggressively decreasing learning rate compared to the previous methods. For representation simplicity, we define: $g_t = \nabla_{\mathbf{w}} R^*(\mathbf{w})$. The updating expectation $E[g^2]_t$ at time t depends on the previous expectation and the current gradient:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) {g_t}^2 \tag{7}$$

where γ is similar to the momentum term. In our settings, we set γ to 0.9, and we set the learning rate η to 0.05. We can rewrite the parameter update vector term:

$$\Delta \mathbf{w}_t = -\frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t \tag{8}$$

where ϵ is a smoothing term that avoids division by zero. As the denominator is just the root mean squared error criterion of the gradient.

The $RMS[\Delta \mathbf{w}]_t$ is approximated with the root mean squared error of parameter updates until the previous time step. Then the final AdaDelta updated rule is:

$$\Delta \mathbf{w}_t = -\frac{RMS[\Delta \mathbf{w}]_{t-1}}{RMS[g]_t} g_t \tag{9}$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \Delta \mathbf{w}_t \tag{10}$$

By using the AdaDelta method, our model is less dependent on the learning rate determination, since it is diminished from the update rule.

4.2 Direction based Street View Retrieval

To reduce street view image noise introduced by recorded vehicles or pedestrians, we select camera directions perpendicular to the streets' directions. The street view image retrieval process considers urban roadway structures as geographical prior knowledge. Under such consideration, the camera directions for the retrieved street view images are always perpendicular to the direction of the roadway. Compared to existing street view image datasets with fixed compass directions (UCF Google Street View Dataset [46], SUN dataset [35]), our dataset preserves a better representation of the real urban perception. Such improvement can be quantified explicitly from the experiment results in the following section of this paper.

Details of direction based street view retrieval are presented in Figure 3, the red dots represent the crime point locations reported from the crime record datasets; the arrows represent the directions; and the dashed blue lines represent the roadway networks. This

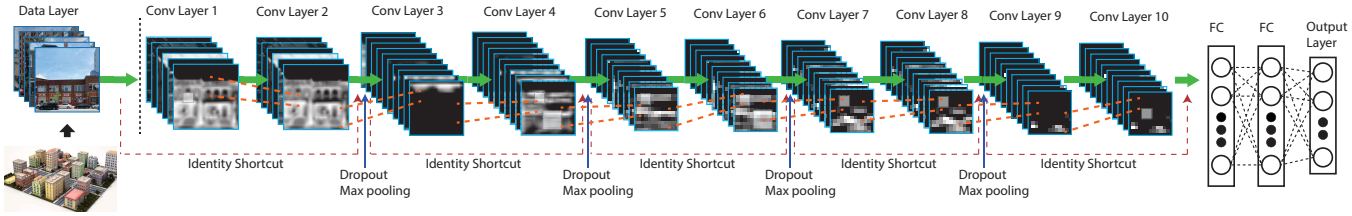


Figure 2: Representation of the Convolutional Neural Network Structure

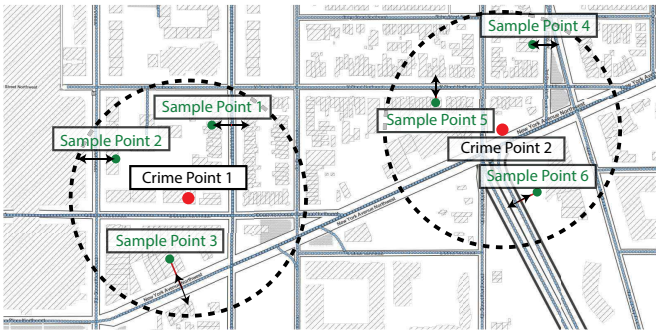


Figure 3: Street View Points Sampling from Crime Records

algorithm preserves the urban surroundings with better representations; most of the previous works did not consider the directions of the street view images [38]. Directions that are perpendicular to the roadway are calculated based on the topological structure information of the target city. *CycloMedia GlobalSpotter API*[1] takes the directions as queries retrieving the street view images.

Based on the given topological structures of the target cities' roadways, the structure can be represented by a shapefile[2] or a spatial database: $Shp = \{r_1, r_2, r_3, ..., r_n\}$ where r_i represents one road in the target city. The procedure of identifying the directions perpendicular to the roads is presented in Algorithm 1, where the operations $<P_s, r_i>$, $proj(r_i, P_s)$, and $perp(P_v, r_i)$ are spatial functions. $<P_s, r_i>$ calculates the spatial distance between the sample point P_s and the road r_i; the function $proj(r_i, P_s)$ finds the projection point of P_s on the road r_i; the function $perp(P_v, r_i)$ returns the directions that are perpendicular to the tangent line of the road r_i at the tangency location P_v [25].

4.3 Crime Rank Labeling

While calibrating ground truth street view images dataset with crime rankings, we build spatiotemporal associations between official crime records datasets of two cities and the street view images. We utilize crime records datasets of Washington D. C. and New York City. In the official crime records datasets, key information of a crime record such as reported time, offense type, and geolocation specified by latitude and longitude is provided. Street view images with timestamps and geolocations are labeled with a localized crime density ranking. For a given street view image $I_{s_i}^{t_i}$ with a timestamp t_i and geolocation pair $s_i = \{lat, lon\}$, we define

[1] *https://globespotter.cyclomedia.com/us/*
[2] *http://doc.arcgis.com/en/arcgis-online/reference*

Algorithm 1: *Direction based Street View Retrieval*

R_c: Crime Records;
Shp: Topological Structure of the Roadways;
P_s: Location of a Sampled Point;
Function *Directions*(shp, P_s)
  for $P_s \in Crime\ Range : R_c$ **do**
    $Closest\ Distance : D_c \leftarrow \infty$;
    for *all* $r_i \in Shp$ **do**
      if $<P_s, r_i> < D_c$ **then**
        $D_c \leftarrow <P_s, r_i>$; // update closest road
      end
    end
    $P_v \leftarrow proj(r_i, P_s)$; // project direction to road
  end
  $direction1, direction2 \leftarrow perp(P_v, r_i)$;
  return $direction1, direction2$
for $C_i \in Crime\ Records : R_c$ **do**
  $\{P_{s_1}, P_{s_2}, ..., P_{s_n}\} \leftarrow Gaussian(C_i.geom, Std)$;
  for $P_{s_i} \in \{P_{s_1}, P_{s_2}, ..., P_{s_n}\}$ **do**
    $direction1, direction2 \leftarrow Directions(Shp, P_{s_i})$;
    $Img\_Retr(direction1, direction2)$
  end
end

a time window τ and a radius r. The local crime set is defined: $C = \{C_{s_c}^{t_c} | t_i - \tau < t_c \leq t_i + \tau \text{ and } dist(S_c, S_i) \leq r\}$, where $C_{s_c}^{t_c}$ represents the crime record with a report time at t_c and a location at s_c; the function $dist()$ returns the distance between two points. Then the crime types are ranked based on the local crime density with a descending order. The density for crime type k is calculated by: $D_k = |C_k|/|C|$. The labeling process is presented in Figure 4. We manipulate the radius parameter r and generate street view crime ranking datasets under multiple levels of resolutions. The radius selected are 1 thousand feet and 2 thousand feet, which generates four datasets for two cities: *DC-1k*, *DC-2k*, *NYC-1k*, and *NYC-2k*.

5 EXPERIMENT

In this section, we present the experiment environment, dataset introduction, evaluation metrics and comparison methods, extensive experimental analysis, and discussions of case studies.

5.1 Experimental Environment and Datasets

The convolutional neural network model is implemented utilizing both Caffe and Keras frameworks respectively. All convolutional neural network experiments were proceeded on a NVIDIA Tesla

Dataset	Method	*nDCG@3*	*nDCG@5*	*nDCG@7*	*P@3*	*P@5*	*P@7*	*MAP*
DC-1k	rSVM-HOG	0.6026	0.5620	0.6951	0.3421	0.6058	0.8549	0.4433
	rSVM-SIFT	0.5882	0.6465	0.7098	0.3084	0.6752	**0.9433**	0.4880
	RLS-HOG	0.5154	0.5312	0.6178	0.2865	0.5475	0.8322	0.4672
	RLS-SIFT	0.6052	0.7054	0.7984	0.3612	**0.7097**	0.8579	0.5896
	AlexNet	0.5598	0.5713	0.7823	**0.4476**	0.6884	0.8831	0.4568
	VGGNet	0.6119	0.6546	0.6802	0.3782	0.6790	0.9217	0.5337
	PlacesNet	0.6251	0.6619	0.7682	0.4146	0.6853	0.8964	0.6209
	StreetNet	**0.6809**	**0.7530**	**0.8210**	0.4353	0.7079	0.9393	**0.6340**
NYC-1k	rSVM-HOG	0.6286	0.7051	0.8105	0.3315	0.6049	**0.9213**	0.4684
	rSVM-SIFT	0.6569	**0.8177**	0.7290	0.3086	0.6639	0.8775	0.5181
	RLS-HOG	0.4691	0.5138	0.6271	0.3822	0.6050	0.7849	0.4650
	RLS-SIFT	0.6333	0.7114	0.7522	0.4071	0.6781	0.8797	0.5388
	AlexNet	0.5092	0.6389	0.7256	0.4133	0.7233	0.8673	0.5376
	VGGNet	0.6007	0.5873	0.7145	0.3997	0.6980	0.9103	0.6231
	PlacesNet	0.6182	0.7378	0.7953	**0.4586**	0.7360	0.9353	**0.6315**
	StreetNet	**0.6793**	0.7512	**0.8226**	0.4297	**0.7438**	0.9206	0.6245
DC-2k	rSVM-HOG	0.5469	0.5972	0.7356	0.3724	0.6294	0.7136	0.5006
	rSVM-SIFT	0.3777	0.6093	0.6694	0.3469	0.5778	0.8483	0.5062
	RLS-HOG	0.3940	0.4644	0.6376	0.3730	0.4956	0.7185	0.3624
	RLS-SIFT	0.5511	0.6372	0.6856	**0.3992**	0.6730	0.8612	0.5493
	AlexNet	0.5440	0.5891	0.6936	0.3522	0.6358	0.7983	0.498
	VGGNet	0.5880	0.6780	0.7008	0.3208	0.5984	0.8439	0.5594
	PlacesNet	0.6081	0.6300	0.7149	0.3722	**0.7123**	0.8280	0.5368
	StreetNet	**0.6116**	**0.6769**	**0.7583**	0.3695	0.6728	**0.9124**	**0.5637**
NYC-2k	rSVM-HOG	**0.6313**	0.4937	0.7659	0.2797	0.4386	0.8337	0.4538
	rSVM-SIFT	0.4390	0.5397	0.6922	0.2729	0.5947	0.7587	0.5195
	RLS-HOG	0.4364	0.4139	0.6261	0.3359	0.5371	0.6932	0.4166
	RLS-SIFT	0.5698	0.5725	0.6987	0.2745	0.6569	0.8947	0.5277
	AlexNet	0.4793	0.5839	0.6704	**0.3792**	0.6002	0.8576	0.4796
	VGGNet	0.5338	0.6193	0.7423	0.2860	0.5784	0.8233	0.5207
	PlacesNet	0.6100	0.6455	**0.7804**	0.3557	0.6507	**0.9204**	**0.5781**
	StreetNet	0.6139	**0.6771**	0.7602	0.3645	**0.6718**	0.9120	0.5516

Table 1: Crime Ranking Performance

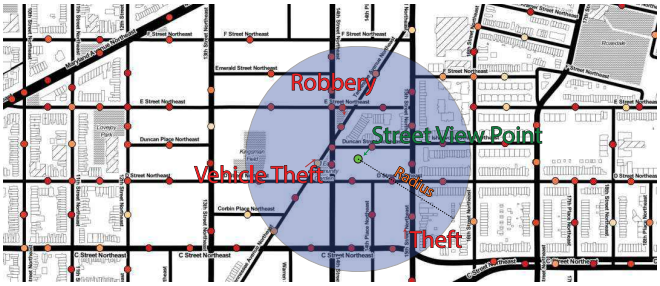


Figure 4: Image Label Strategies

K20 GPU. For support vector machine and regression models, we run the experiments on an Intel Core i7-4790 3.60GHz CPU with 32 GB memory. Standard libraries such as LibSVM and LibLINEAR are utilized as baseline methods.

The experiments are conducted on street view datasets of two major locations: Washington, D.C., and New York City. We trained our proposed models on a set of 44,694 images for the physical appearances for the street view, which is significantly more images than in previous works [18, 39]. The street view images for the Washington, D.C., area are obtained from the *CycloMedia GlobalSpotter API*. The CycloMedia GlobalSpotter is an interactive web-based application that provides access to CycloMedia' s panoramic street level images. The Atlas PanoramaRendering Service of the CycloMedia GlobalSpotter API provides a controllable RESTful API for requesting street view images. The retrieved street view images are directed horizontally and vertically after being given geo-location and a spatial reference index.

Street view images for New York City were extracted from the Google Street View data set [46]. Total of 23,764 images are provided by the New York City Google Street View data set. There are 5,941 unique location points contained in this data set, each location consists of 4 directions, and each direction represents one view. Each image from the data set is geo-tagged with latitude and longitude. Note that the image quality of the New York City Google Street View data set is lower than the Cyclomedia GlobalSpotter generated street view data set, and the camera view compass directions for the New York City Street View dataset are fixed to 0°, 90°, 180°, and 270°. As we will show in the later sections, this camera direction mismatch to the street direction shows its insufficiency in representing the actual street view.

Crime record datasets for Washington, D.C.,[3] and New York City[4] are utilized for extracting the spatiotemporal correlations between the street view images and the crime types. Nine types of common crimes are considered as ranking labels: *theft*, *theft from auto*, *robbery*, *motor vehicle theft*, *burglary*, *assault with dangerous weapon*, *sex abuse*, *homicide*, and *arson*[5]. 36,484 cases of criminal offenses in Washington, D.C., and 102,327 cases in New York City are collected.

5.2 Baseline Methods

We compare the proposed method to the two major branches of methods in urban perception and scene recognition areas. Firstly, we implement hand-craft feature extraction methods on traditional supervised learning methods. These methods include ranking-SVM with HOG features (*rSVM-HOG*), ranking-SVM with SIFT features (*rSVM-SIFT*), regularized least squares with HOG features (*RLS-HOG*), and regularized least squares with SIFT features (*RLS-SIFT*). Note that we utilize RBF kernel while solving the ranking-SVM. HOG [9] is a popular feature descriptor used in computer vision and image processing for multiple purposes of learning tasks. In our experiments, descriptor blocks with a size of 8 by 8 are utilized for HOG feature generation. Using scale invariant feature transform (SIFT) [27] as a key point extraction mechanism has become increasingly popular in recent years. variousus previous works have justified its effectiveness [5, 28]. In this paper, SIFT key points are extracted for each street view image to construct a bag of key points; this method is referred to as the bag of words paradigm [8]. The other branch of baseline methods is deep regression networks for urban perception problems. Baseline method of this branch includes AlexNet [21], VGGNet [41], and the PlacesNet [49]. We used the pre-trained models of the deep regression networks and fine-tuned all these baseline methods on our street view image dataset separately.

5.3 Evaluation Metrics

The effectivenesses of the *nDCG@k*, *Precision@k*, and *MAP* are analyzed for all the comparison methods and the proposed method.

5.3.1 nDCG@k. The first metric is normalized discounted cumulative gain at top k (*nDCG@k*) to evaluate the accuracy of the crime ranking produced by a given crime ranking prediction model. *nDCG@k* was first defined as an information retrieval (IR) evaluation metric to consider the degree of relevance in retrieved results. The more relevant results retrieved at top positions in the rank would accumulate higher score to the top k gain. This metric is chosen because it is suited for crime rankings that have multiple levels of assessment. For a given ground truth crime rank $\{c_1, c_2, ..., c_k\}$ and its prediction $\{\hat{c_1}, \hat{c_2}, ..., \hat{c_k}\}$, the relevance scores $\{\hat{rel}_1, \hat{rel}_2, ..., \hat{rel}_k\}$ of the prediction ranking are firstly permutated by the indexes of the ground truth; then *nDCG@k* is measured on the permutation in the form of:

$$nDCG@k = \frac{1}{Z} \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log(1+i)} \tag{11}$$

The term Z is a normalization factor derived from a perfect ranking of top k articles so that it would yield a *nDCG@k* of 1.

5.3.2 Precision@k. Precision measures in (IR) consider the number of relevant documents among the top k documents. In our evaluation, relevant crime types in the predicted crime ranking refer to the crime types that are also presented in the ground truth crime ranking at cutting off point k. However, unlike *nDCG@k*, the *Precision@k* measurement is incapable of capturing the order of information within the top k rankings. The *Precision@k* is measured in the form of:

$$P@k = \frac{|Predicted\_Crimes@k \cap Ground\_Truth\_Crimes@k|}{k}$$

5.3.3 MAP. Mean average precision (*MAP*) for a set of street view images is the mean of the average precision scores for each street view image. *MAP* has been shown to have especially good discrimination and stability. The *MAP* of a given set of rankings is calculated:

$$MAP = \frac{1}{|\mathcal{I}|} \sum_{j=1}^{|\mathcal{I}|} \frac{1}{k} \sum_{i=1}^{k} Precision@i \tag{12}$$

where $\mathcal{I}$ is the complete set of the street view images for validation.

5.4 Experimental Analysis

In this section, we demonstrate the results of the crime type prediction from street view urban perceptions. Experimental evaluations of our proposed methods and extensive comparisons to the baseline methods are conducted.

5.4.1 Crime Ranking Prediction. As shown in Table 1, our proposed *StreetNet* outperforms the baseline methods in general. Such performance increase is even more significant when the parameter k is relatively small for both *nDCG@k* and *precision@k*. This result also implies that the process of feature selection and extraction is critical for image label ranking tasks, and the convolutional layers in the convolutional neural networks achieve better feature extraction.

We compare the performance of our proposed convolutional neural network and competing methods for crime ranking prediction on different datasets. We generate four datasets out of two major cities, Washington, D.C., and New York City, with two levels of street view perception radius: 1,000 feet and 2,000 feet.

Table 1 shows that crime ranking prediction performance generally decreases as the street view perception radius becomes larger. For example, comparing the *DC-1k* and *DC-2k* datasets, the *nDCG@k* score of the *DC-1k* is 4% greater than the score of *DC-2k*; for the metric *precision@k*, the prediction results of the *DC-1k* also outperform the results of *DC-2k* by 2% in general; the *MAP* of the *DC-1k* also exceed *DC-2k* by 6%. From the previous experimental observations, we find that the increase of the *precision@k* metric is not as significant as the increase of the *nDCG@k* metric. This

[3] *http://data.octo.dc.gov/*
[4] *https://data.cityofnewyork.us/Public-Safety/Historical-New-York-City-Crime-Data/hqhv-9zeg*
[5] *http://crimemap.dc.gov/CrimeDefinitions.aspx*

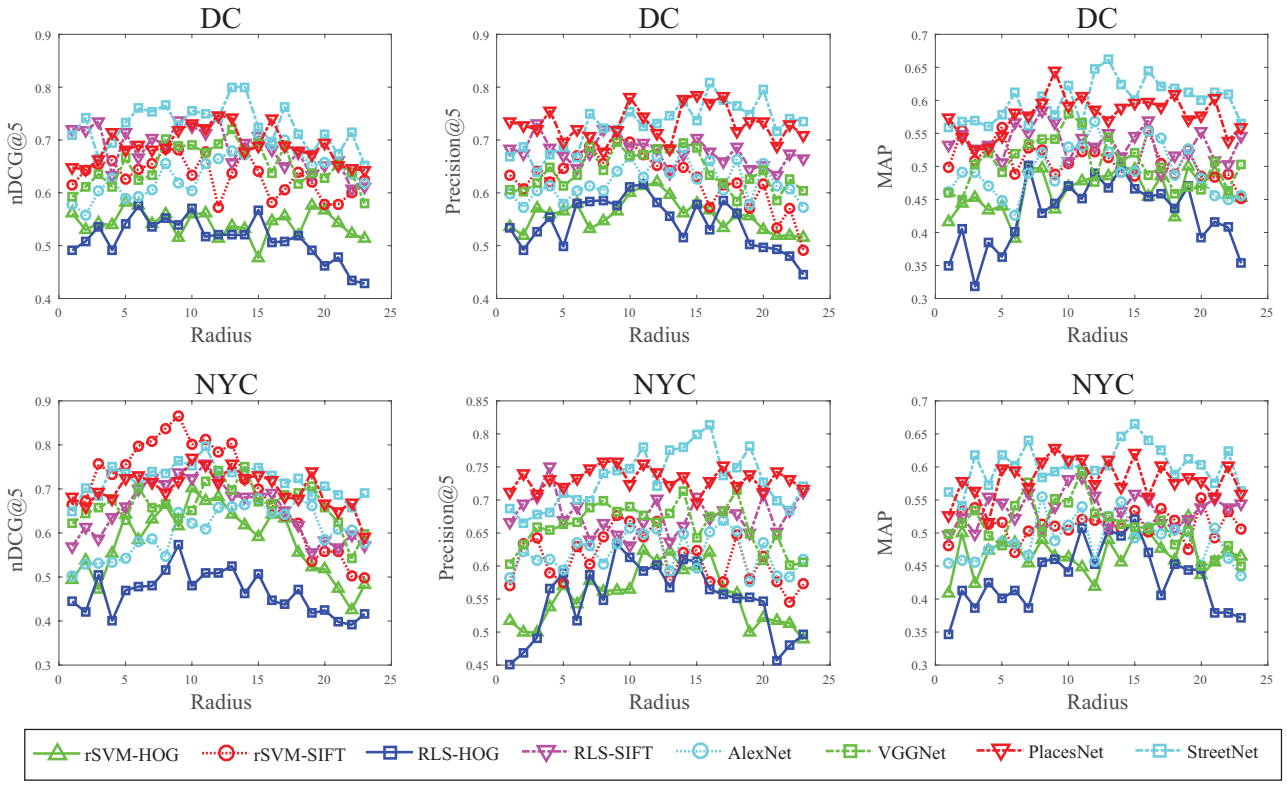


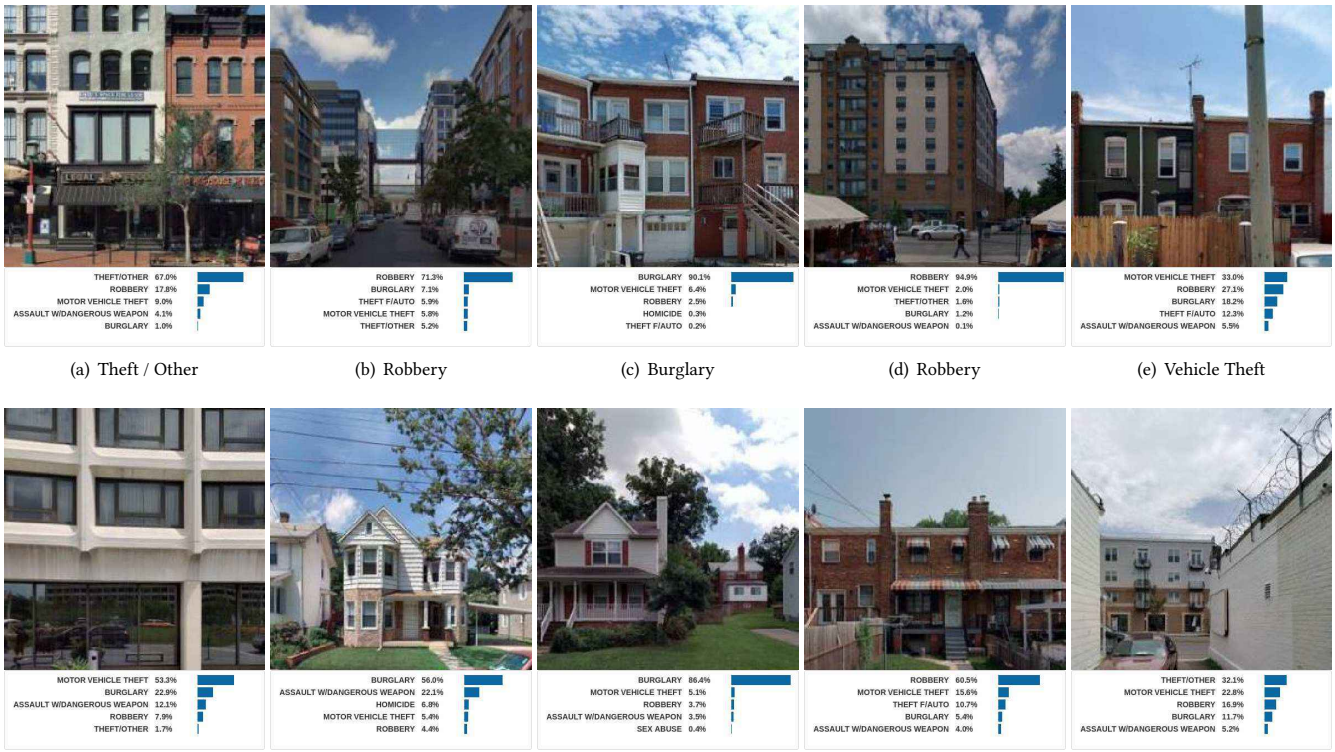
Figure 5: Street View Perception Radius Analysis

may be caused by the different properties of the evaluation metrics: *nDCG@k* considers ordering of the crime relevance scores, while *precision@k* is calculated based on the number of crime intersected with the true crime set.

From the crime type ranking prediction results, interesting performance patterns can be observed. Firstly, for metrics *nDCG@k* and *Precision@k*, when the ranking parameter k is relatively small ($k = 3$ or $k = 5$), some of the hand-craft features based methods can outperform the deep neural networks (AlexNet, VGG-16, and PlacesNet) . On the other hand, when the ranking parameter k is set to be relatively large ($k = 7$), the deep neural networks can outperform the hand-craft features based methods. Secondly, the PlacesNet achieves better performance than other baseline methods when trained on the NYC-1k and NYC-2k datasets. This is because the pre-trained PlacesNet model was trained on an imagery dataset with higher diversity. While handling street view images on urban perception task, the PlacesNet will converge faster.

5.4.2 Street View Perception Radius Analysis. Further analysis of the correlations between the selection of the radius and the evaluation metrics are studied. The results help us to learn, in an empirical way, the best crime rank representation area (resolution) of a given street view image. Such a finding is highly practical in the study of urban perception. For example, given a street view image with geo-location, one is always interested in questions such as: Can the street view represent the crime rank for the whole city? Or can the street view only represent the crime rank for a small neighborhood? What is the resolution? Figure 5 shows the evaluation metrics results by varying the selections of the street view perception radius. We find that for different learning methods, the optimal radius varies. For *nDCG@5*, our proposed convolutional neural network outperforms other comparison methods, and the optimal radius for our method locates at 1,200 feet; ranking-SVM with SIFT features locates its optimal selection of radius at 900 feet for the same metric. In order to achieve the best of *precision@5*, our method locates the optimal radius selection at 1,500 feet; and ranking-SVM at 1,000 feet.

5.4.3 Direction-based Street View Image Retrieval Analysis. As proposed in the methodology section, the direction-based street view image retrieval algorithm is applied for retrieving the street views with higher- quality urban representation. The street view image datasets *DC-1k* and *DC-2k* are retrieved by our algorithm; on the other hand, the other two street view image datasets are retrieved with fixed compass directions of 0°, 90°, 180°, and 270°. As in Table 1, we find that the performances of our method on the *NYC-1k* and *NYC-2k* datasets is not as stable. This result is intuitive, because a tremendous amount of noise can be introduced to street view images with camera directions not perpendicular to the streets. For example, if the camera direction is always the same as the street's direction, the retrieved street view image will always present the street surface or the sky. In other words, the



(a) Theft / Other (b) Robbery (c) Burglary (d) Robbery (e) Vehicle Theft

(f) Vehicle Theft (g) Burglary (h) Burglary (i) Robbery (j) Theft / Other

Figure 6: Crime Type Inferences from Street Views

resulting street view images will not be representative enough for the real street view; image content like the front of a store or the appearance of a building will be neglected.

5.5 Case Studies Discussions

In this section, a number of interesting crime ranking prediction patterns are observed discovered by the proposed convolutional neural network. The top 5 crime types with the highest relevance scores learned are listed for each input street view image in Figure 6. The corresponding relevance scores are also presented. From Figure 6, we can find interesting correlations between urban appearance and the predicted crime rankings. For example, crime types such as robbery and motor vehicle theft are more likely to be inferred from the street views of downtown areas. Such findings are presented in Figures 6(a), 6(b), and 6(f). On the other hand, from the street view images of residential areas or suburbs, crime types such as burglary and theft are more likely to be predicted by our approach; the results are shown in Figures 6(c), 6(g), and 6(h). As presented in Figure 1, the predicted crime ranking results for both downtown areas and the suburbs fit the crime distributions in those places. The consistency with official crime records indicates the feasibility of inferring crime rankings or other safety information from street view images or other forms of urban perception.

In order to test the performance of our model, we manually extract street view images from *Google Maps*, and selected areas with no public crime records accessible. The results are shown in Figures 6(d), 6(e), 6(i), and 6(j). Similar crime ranking prediction patterns can be witnessed from these results. From these tests, we show that our model is highly practical for application scenarios such as 1) areas and cities with no easy access to public crime records data and 2) end users traveling to an unfamiliar area with no idea how safe it is.

6 CONCLUSION

This paper presents a novel convolutional neural network solution to the problem of inferring crime rankings from street view images of an area. The convolutional neural network model is designed based on the settings of a preference learning framework. By taking road structure data as prior knowledge, the proposed direction-based street view image retrieval method presents better preservation of urban perceptions. By exploiting the spatiotemporal correlations between the street view images and official crime records datasets, we generate labeled training data in a data-driven way, which greatly reduces bias. Comparisons with previous image feature extraction and ranking learning algorithms show that the proposed convolutional neural network approach outperforms the baseline methods in learning crime rankings from street view images. Extensive experiments based on multiple street view image datasets and crime records confirm the feasibility of inferring

hidden knowledge such as crime ranking from urban perception data.

REFERENCES

[1] Virginia O Andersson, Marco AF Birck, and Ricardo M Araujo. 2017. Investigating Crime Rate Prediction Using Street-Level Images and Siamese Convolutional Neural Networks. In *Latin American Workshop on Computational Neuroscience*. Springer, 81–93.

[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] Wim Bernasco. 2010. Modeling micro-level crime location choice: Application of the discrete choice framework to crime at places. *Journal of Quantitative Criminology* 26, 1 (2010), 113–138.

[4] Wim Bernasco. 2010. A sentimental journey to crime: Effects of residential history on crime location choice. *Criminology* 48, 2 (2010), 389–416.

[5] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. 2006. Scene classification via pLSA. In *European conference on computer vision*. Springer, 517–530.

[6] Gang Chen, Yangqiu Song, Fei Wang, and Changshui Zhang. 2008. Semi-supervised multi-label learning by solving a sylvester equation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 410–419.

[7] Nico Cornelis, Bastian Leibe, Kurt Cornelis, and Luc Van Gool. 2008. 3d urban scene modeling integrating recognition and reconstruction. *International Journal of Computer Vision* 78, 2-3 (2008), 121–141.

[8] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. 2004. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, Vol. 1. Prague, 1–2.

[9] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 886–893.

[10] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. 2016. Are safer looking neighborhoods more lively?: A multimodal investigation into urban life. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1127–1135.

[11] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. 2008. Bag-of-visual-words models for adult image classification and filtering. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 1–4.

[12] 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.

[13] Kaiqun Fu, Chang-Tien Lu, Rakesh Nune, and Jason Xianding Tao. 2015. Steds: Social Media Based Transportation Event Detection with Text Summarization.. In *ITSC*. 1952–1957.

[14] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. 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* (2017), 201700035.

[15] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, and Li Fei-Fei. 2017. Fine-Grained Car Detection for Visual Census Estimation.. In *AAAI*, Vol. 2. 6.

[16] Edward L Glaeser, Scott Duke Kominers, Michael Luca, and Nikhil Naik. 2018. Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry* 56, 1 (2018), 114–137.

[17] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. 2013. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082* (2013).

[18] Asaad Hakeem, Roberto Vezzani, Mubarak Shah, and Rita Cucchiara. 2006. Estimating geospatial trajectory of a moving camera. In *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 2. IEEE, 82–87.

[19] Sariel Har-Peled, Dan Roth, and Dav Zimak. 2002. Constraint classification for multiclass classification and ranking. *Urbana* 51 (2002), 61801.

[20] Taoran Ji, Kaiqun Fu, Nathan Self, Chang-Tien Lu, and Naren Ramakrishnan. 2018. Multi-task Learning for Transit Service Disruption Detection. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'18)*.

[21] 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.

[22] Ickjai Lee and Peter Phillips. 2008. Urban crime analysis through areal categorized multivariate associations mining. *Applied Artificial Intelligence* 22, 5 (2008), 483–499.

[23] Xirong Li, Cees GM Snoek, and Marcel Worring. 2010. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 10–17.

[24] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag ranking. In *Proceedings of the 18th international conference on World wide web*. ACM, 351–360.

[25] Meiling Liu, Kaiqun Fu, Chang-Tien Lu, Guangsheng Chen, and Huiqiang Wang. 2014. A search and summary application for traffic events detection based on twitter data. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 549–552.

[26] Xiaobai Liu, Qi Chen, Lei Zhu, Yuanlu Xu, and Liang Lin. 2017. Place-centric Visual Urban Perception with Deep Multi-instance Regression. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 19–27.

[27] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[28] Yong Luo, Dacheng Tao, Chang Xu, Dongchen Li, and Chao Xu. 2013. Vector-Valued Multi-View Semi-Supervsed Learning for Multi-Label Image Classification.. In *27th AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 647–653.

[29] Branislav Micusik and Jana Kosecka. 2009. Piecewise planar city 3D modeling from street view panoramic sequences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2906–2912.

[30] Jeffrey D Morenoff and Robert J Sampson. 1997. Violent crime and the spatial dynamics of neighborhood transition: Chicago, 1970–1990. *Social forces* 76, 1 (1997), 31–64.

[31] Yair Movshovitz-Attias, Qian Yu, Martin C Stumpe, Vinay Shet, Sacha Arnoud, and Liron Yatziv. 2015. Ontological supervision for fine grained classification of street view storefronts. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1693–1702.

[32] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore–Predicting the Perceived Safety of One Million Streetscapes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 793–799.

[33] 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. *American Economic Review* 106, 5 (2016), 128–32.

[34] Vicente Ordonez and Tamara L Berg. 2014. Learning high-level judgments of urban perception. In *European Conference on Computer Vision*. Springer, 494–510.

[35] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision* 108, 1-2 (2014), 59–81.

[36] Peter Phillips and Ickjai Lee. 2011. Crime analysis through spatial areal aggregated density patterns. *Geoinformatica* 15, 1 (2011), 49–74.

[37] Peter Phillips and Ickjai Lee. 2012. Mining co-distribution patterns for large crime datasets. *Expert Systems with Applications* 39, 14 (2012), 11556–11563.

[38] Lorenzo Porzi, Samuel Rota Bulò, Bruno Lepri, and Elisa Ricci. 2015. Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 139–148.

[39] Grant Schindler, Matthew Brown, and Richard Szeliski. 2007. City-scale location recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–7.

[40] Herbert W Schroeder and LM Anderson. 1984. Perception of personal safety in urban recreation sites. *Journal of Leisure Research* 16, 2 (1984), 178.

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

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

[43] Jun Yang, Yu-Gang Jiang, Alexander G Hauptmann, and Chong-Wah Ngo. 2007. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, 197–206.

[44] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In *30th AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 308–314.

[45] Amir Roshan Zamir and Mubarak Shah. 2010. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*. Springer, 255–268.

[46] Amir Roshan Zamir and Mubarak Shah. 2014. Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. *IEEE transactions on pattern analysis and machine intelligence* 36, 8 (2014), 1546–1558.

[47] Matthew D Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).

[48] Man Zhang, Ran He, Dong Cao, Zhenan Sun, and Tieniu Tan. 2016. Simultaneous Feature and Sample Reduction for Image-Set Classification. In *30th AAAI Conference on Artificial Intelligence (AAAI)*. AAAI, 1401–1407.

[49] 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.