Predicting and Understanding Urban Perception with Convolutional Neural Networks

Lorenzo Porzi
FBK, Trento, Italy
University of Perugia
Perugia, Italy

Samuel Rota Buló
FBK, Trento, Italy

Bruno Lepri
FBK, Trento, Italy

Elisa Ricci
FBK, Trento, Italy
University of Perugia
Perugia, Italy

ABSTRACT

Cities' visual appearance plays a central role in shaping human perception and response to the surrounding urban environment. For example, the visual qualities of urban spaces affect the psychological states of their inhabitants and can induce negative social outcomes. Hence, it becomes critically important to understand people's perceptions and evaluations of urban spaces. Previous works have demonstrated that algorithms can be used to predict high level attributes of urban scenes (*e.g.* safety, attractiveness, uniqueness), accurately emulating human perception. In this paper we propose a novel approach for predicting the perceived safety of a scene from Google Street View Images. Opposite to previous works, we formulate the problem of learning to predict high level judgments as a ranking task and we employ a Convolutional Neural Network (CNN), significantly improving the accuracy of predictions over previous methods. Interestingly, the proposed CNN architecture relies on a novel pooling layer, which permits to automatically discover the most important areas of the images for predicting the concept of perceived safety. An extensive experimental evaluation, conducted on the publicly available Place Pulse dataset, demonstrates the advantages of the proposed approach over state-of-the-art methods.

1. INTRODUCTION

Cities are shaped by and affects the life of their inhabitants. Several studies have shown that cities' visual appearance plays a central role in human perception and response to the surrounding environment. A notable example is the *Broken Windows Theory* suggesting that visual signs of environmental disorder, such as broken windows, abandoned cars, litter and graffiti, can induce negative social outcomes and increase crime levels [17]. Interestingly, the *Broken Windows Theory* has greatly influenced public policy strategies leading to aggressive police tactics to control the manifestations of social and physical disorder. Moreover, the visual qualities of urban spaces affect the psychological states of their inhabitants [22]. For example, urban scenes with vegetation tend to produce positive feelings [37], while urban disorder induces psychological distress [33]. Hence, it becomes critically important to understand people's perceptions and evaluations of urban spaces.

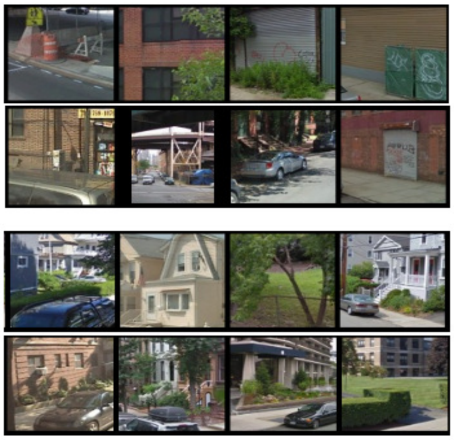

Figure 1: Given the images above showing specific details extracted from pictures of urban scenes, a human observer can easy choose which group corresponds to safe areas and which to unsafe ones.

In *The Image of the City* [23], Kevin Lynch introduced the city's mental map, indicating the city elements that are distinguished among hundreds, thousands, or millions of other city artifacts by their unique shapes, sizes, colors, etc. Traditionally, researchers have studied the city's mental map by interviewing city residents and manually reviewing photographs and videotapes: a tedious process, involving considerable collective efforts [26]. While this was unavoidable at the time of Lynch's study, today, modern information technology permits to analyze the image of the city in a quantitative manner. In particular, recent works have shown that modern crowdsourcing platforms can be used to collect millions of users' opinions about places and advanced machine learning approaches are very accurate in predicting human judgements of urban scenes, even in case

MM'15, October 26–30, 2015, Brisbane, Australia.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806273 .

of previously unseen locations. Moreover, geo-tagged images publicly available from Google Street View and from social network platforms, such as Flickr or Foursquare, represent an invaluable resource to study human perceptions of places. Recently, few works [31, 25, 2, 27] have proposed computational methods to automatically infer high level perceptual attributes from geo-referenced images of urban spaces.

In this paper we propose a novel approach for predicting human opinions about places. Specifically, confirming the recent findings in [27, 25], we demonstrate that a computer vision algorithm can emulate human perception and reliably predict judgments of safety from pictures of urban scenes. For a human observer it is generally hard to quantify the absolute degree of safety of a scene, while relative judgements (*e.g.* "*this looks safer than that*") are more natural [7]. Indeed, most crowdsourcing platforms collecting annotations of high levels human judgments[1] operate on image pairs. Previous works presented solutions based on a two-steps approach: first the pairwise annotations are used to calculate a set of absolute safety scores, then a regression model is learned to predict these scores. We believe that this indirect approach rises some fundamental issues, which are thoroughly discussed in Subsection 4.2.1. For these reasons we propose to adopt a ranking framework, which directly operates on relative judgements. The prediction algorithm we propose relies on a Convolutional Neural Network specifically designed for ranking tasks. Experiments conducted on the Place Pulse dataset [32] demonstrate that our approach is superior to state-of-the-art regression methods.

Previous works [27, 25] presented computational models capable of answering the question "*does this place look safe?*". In this paper we also tackle a second fundamental aspect about the way humans judge a place from a picture: "*what makes this place look safe?*". Our proposed CNN is based on a set of latent detectors, which are optimally combined for predicting urban safety. Each detector is asked to select specific parts of the images, correlating them with the concepts of perceived safety/unsafety. Indeed, a human observer can guess the degree of safety of a specific urban scene by simply looking at some parts of it (Figure 1). Importantly, the proposed detectors are designed in order to capture complementary information: some aim to discover very localized areas, thus focusing on discriminative objects (*e.g.* residential windows and trees, graffiti), while others are meant to capture spread areas, focusing on large objects and more diffuse patterns (*e.g.* roads and vegetation).

In summary, the main contributions of this study are: (i) we propose a ranking approach to predict human perceptions of safety from geo-tagged images of urban spaces. Our analysis and experimental evaluation demonstrates the advantages of a ranking-based framework over previous regression methods; (ii) To our knowledge, this is the first work based on CNN for predicting high level judgements of urban spaces. Previous works have considered traditional feature representations or descriptors derived from a pre-trained CNN. Differently, in this paper we propose a novel CNN architecture, demonstrating improved performance over state-of-the art methods; (iii) Our approach permits to automatically discover the parts of the image, which correlate with the concept of perceived safety. No previous works have proposed computational models to address this issue.

The rest of the paper is structured as follows. Related work for analysing patterns of human perception are discussed Section 2. The proposed approach for predicting the perceived safety from geo-tagged images of urban scenes is presented in Section 3. Extensive evaluations and comparisons with state-of-the-art methods are reported in Section 4. Finally, Section 5 is devoted to conclusions and future works.

2. RELATED WORK

In this section, we review key works from two research fields: urban perception and automatic scene recognition.

2.1 Urban Perception

Since the seminal study of Lynch [23], several works in urban studies and environmental psychology have investigated people's preferences for certain environments and their aesthetic judgments of urban scenes, such as streets, parks, buildings [38]. In particular, the urban activist Jane Jacobs discussed extensively, in *The Death and Life of Great American Cities*, the role played by streets as principal visual scenes in a city [14]. In particular, research has shown seven environmental features as prominent in human evaluation of places: naturalness, complexity, order, novelty, openness, historical significance, and upkeep [26]. More specifically, people prefer vegetation and dislike obtrusive signs, intense land uses and traffic. Weber and colleagues identified uniformity in architectural style, symmetry, scale and presence of vegetation as primary factors driving the aesthetic judgments of urban spaces [38]. Again, several studies highlights people notice and prefer order. Dilapidation and disorder such as trash, boarded up buildings, abandoned property and cars, and litter, which researchers refer to as *physical incivilities*, also contribute to a perception of the breakdown of social controls, fear of crime, and crime [36, 17]. For example, Schroeder and Anderson found vacant buildings and graffiti associated with judgements of low safety [35]. Based on videotaping and systematic rating of more than 23,000 street segments in Chicago, Sampson and Radenbausch [34] constructed scales of social and physical disorder for 196 neighborhoods. In particular, graffiti, abandoned cars, garbage or litter on street or sidewalk were recognized by raters as visual signs of physical disorder. Another set of environmental cues evoking the possibility that an area is unsafe are related with *entrapment*, referring to the difficulty a person would have escaping, and *physical concealment*, referring to a visual occlusion of space big enough to hide potential offenders. Thus, people usually consider safer places offering open vistas [26]. The studies described above were mostly based on interviews or visual surveys where people were asked to rate images of streets and neighborhoods on a 1-10 scale. Nowadays, the large amount of geo-tagged images publicly available on Google Street View and in social network platforms such as Flickr or Foursquare opens a new way to study human perceptions of places. Our work represents one of the first attempts in this novel research direction.

2.2 Automatic Scene Recognition

In the last few years there has been significant progress in the area of automatic scene recognition [21, 39, 40]. However, previous research on scene understanding has mainly focused on traditional problems such as place classification or analysis of scene composition (*i.e.* the detection and

[1] http://pulse.media.mit.edu/,http://urbangems.org/

recognition of the objects present in the scene). This paper tackles a different problem: we are interested in learning to infer human high level (safety) judgments of places. Perceptual prediction from images has received an increased interest recently in the vision and multimedia community. However, previous works have focused mainly on tasks such as assessing automatically the interestingness and the aesthetic quality of images [10, 24] or discovering the style of a city or an object [8]. These tasks are substantially different from ours, as we focus on predicting high level judgments of urban areas from geo-tagged images and specifically from pictures depicting outdoor street level scenes.

Geo-tagged images are an invaluable resource for researchers in multimedia analysis and have been exploited for many tasks, such as for automatic localization of non geo-referenced pictures [12], for photo trip pattern mining [1], for augmented reality [30, 29], for developing personalized travel recommendations [6]. Prediction of human judgments from geo-tagged images depicting urban spaces has been attempted only very recently. Quercia *et al.* [31] used crowdsourcing to collect a dataset of street level images and the associated perceptual attributes, corresponding to the concepts of attractiveness, happiness and quiet. They also investigated the role of different visual features (*i.e.* color, texture and compositional features) for automatic high level judgments prediction. Naik *et al.* [25] proposed an approach based on support vector regression to predict three perceptual characteristics of scenes, *i.e.* safety, wealth and uniqueness). Their analysis is performed on the Place Pulse dataset [32], consisting of about 4K images downloaded from Google StreetView and corresponding to four different cities (New York, Boston, Linz and Salzburg). Human annotations have been collected with crowdsourcing, asking users to decide, among two scene images, which picture corresponds to the most safe/upper-class/unique place. The Place Pulse dataset has been used in [27], where deep convolutional activation features [20] have been shown superior to traditional descriptors (*e.g.* GIST, Dense and Sparse SIFT, HOG) for high level judgments prediction. In this paper we also consider the Place Pulse dataset, but differently from previous works based on regression [25, 27], we propose a ranking framework, showing that our approach guarantees more accurate estimates of human perceptions.

The problem of automatically identifying visual elements (*i.e.* small image patches) that correlate with high level non-visual attributes has been rarely considered. Notable exceptions are the works in [2, 18]. Arietta *et al.* [2] proposed a computational model, based on support vector regression, to predict non-visual attributes (crime rate, housing prices) from images and to automatically discover visual elements correlated with the predicted attributes. However, they did not consider human annotations obtained with crowdsourcing and the annotated data are derived from public databases of city municipalities. Khosla *et al.* [18] presented an approach to *look beyond the visible scene* and used deep learning features extracted from visual data to infer properties about the environment (*e.g.* crime rate). They also proposed an approach based on sparse coding to analyze the importance of different image regions for predicting high level attributes. To our knowledge, no previous work have considered the problem of discovering mid-level visual elements responsible for human perceptual judgments in the context of Convolutional Neural Networks.

3. PREDICTING THE PERCEIVED SAFETY OF URBAN SPACES

In this section, we first describe the Place Pulse dataset, then we present the proposed ranking method for learning the perceived safety of urban places. Specifically, we first show an approach based on Support Vector Machines (SVM) and precomputed feature descriptors, then we introduce a novel method based on CNNs. Additionally, we discuss the challenges of learning from human judgments, analyzing the noise of users annotations in the Place Pulse dataset.

3.1 Place Pulse 1.0

We consider the publicly available Place Pulse 1.0 dataset [32]. This dataset contains 4,136 geo-tagged images for the cities of New York (including Manhattan and parts of Queens, Brooklyn and The Bronx), Boston (including parts of Cambridge), Linz and Salzburg. As typical for images downloaded from Google Street View, these are captured from a vehicle in the early morning, thus depicting mostly empty sidewalks, little traffic and limited human activity.

For each image, the scores for perceived safety, wealth and uniqueness are provided. These scores were computed from user preferences using the Microsoft Trueskill algorithm [11]. User preferences were collected with crowdsourcing through an online website[2]. Participants were shown two randomly chosen images from the dataset and were asked to choose one of the two in response to the question: *Which place looks safer/more upper-class/more unique?*. In total, user annotations corresponding to 208,738 votes were obtained between August and November 2011. 7,872 unique users from 91 countries took part to the survey. The user preferences, *i.e.* the votes corresponding to pairwise comparisons, are publicly available only for the attribute of safety. Therefore, in this work we only consider this attribute. However, it is worth nothing that our approach is general and can be used for predicting arbitrary high level judgments.

3.2 Learning to Predict the Perceived Safety

It is hard for a human observer to quantify the degree of safety of a place given a single picture, while it may be relatively easy to give judgments over image pairs. In confirmation of this, in the Place Pulse dataset, as well as in other studies on urban perception [31], annotations are collected in the form of votes on couples of images. Therefore, in this paper we propose to formulate the problem of learning to predict the perceived safety of places as a ranking task.

More formally, we consider a setting, where we observe judgments about the relative safeness of pairs of images. Each judgment consists of a triplet $(I, J, y) \in \mathcal{I} \times \mathcal{I} \times \mathcal{Y}$, where $I, J \in \mathcal{I}$ are images and $y \in \mathcal{Y} = \{+1, -1\}$ is a label that indicates whether image I is perceived by a human as safer ($y = +1$) or unsafer ($y = -1$) than image J.[3] We assume the judgments to follow an unknown distribution P, which is in general noisy due to inconsistencies that exist among humans' perceptions. We denote by $\mathcal{D}_\mathsf{n} \subseteq \mathcal{I} \times \mathcal{I} \times \mathcal{Y}$ a training set of n judgments, being *i.i.d.* samples from P.

The goal of the learning task is to estimate from the training set $\mathcal{D}_\mathsf{n}$ a function $f \in \mathcal{I} \to \mathbb{R}$ that provides for each im-

[2] `http://pulse.media.mit.edu/`

[3] To facilitate comparison with previous methods we have excluded ties in our analysis, but in principle they could be added as a third class.

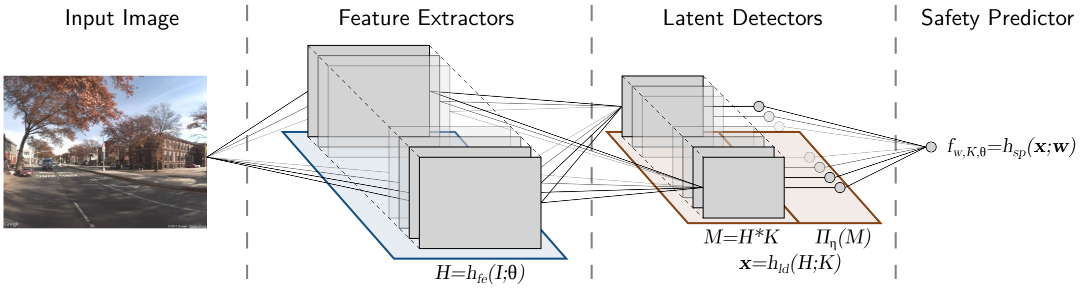

Figure 2: Schematic representation of the three blocks composing our rCNN architecture.

age $I \in \mathcal{I}$ a real value $f(I)$ that reflects the degree of safety perceived by a human. The prediction function f is sought within a subset of feasible functions $\mathcal{F} \subseteq \mathcal{I} \to \mathbb{R}$ (*a.k.a.* hypothesis space) following the regularized, empirical risk minimization principle. This yields the following optimization problem

$$f^\star \in \arg\min_{f\in\mathcal{F}} \lambda\,\Omega(f) + R_{\text{emp}}(f)\,, \tag{1}$$

where $\Omega(f)$ is a regularization term penalizing complex functions to prevent overfitting, λ is a nonnegative trade-off parameter and $R_{\text{emp}}(f)$ is the empirical risk term, which is given by:

$$R_{\text{emp}}(f) = \frac{1}{|\mathcal{D}_{\mathsf{n}}|}\sum_{(I,J,y)\in\mathcal{D}_{\mathsf{n}}} L(y, f(I), f(J))\,.$$

In the empirical risk, the quality of a prediction function with respect to the observed data is assessed via a *loss function*. Each loss term $L(y, z_I, z_J)$ represents the cost one incurs by assigning image I and J a safety degree of z_I and z_J, respectively, given that y is the true, relative safety label.

The loss function that is typically regarded as the reference loss, and that is used in practice at test time to measure the errors committed by the learning algorithm, is the *binary loss*, which is defined as

$$L_{\text{bin}}(y, z_I, z_J) = \begin{cases} 1 & \text{if } y(z_I - z_J) \le 0 \\ 0 & \text{otherwise.} \end{cases}$$

At training time, however, the binary loss is typically replaced with some other convex, surrogate loss that renders the optimization of (1) easier. Later in this section, we will consider differentiable, convex losses such as the squared hinge loss and the logistic-loss.

In the following we present two different learning settings for the problem at hand: in Subsection 3.2.1 we present a solution based on the well-known SVM; in Subsection 3.2.2 we propose a Convolutional Neural Network (CNN) architecture, which includes a new type of pooling unit. Finally, in Subsection 3.3 we discuss the problem of optimizing the empirical risk in the case of a binary loss and unconstrained function space, with the purpose of determine the intrinsic noise level of a given dataset.

3.2.1 RankingSVM

We describe a learning machine for our task that is known as RankingSVM [13, 5]. In this setting, the hypothesis space $\mathcal{F}_{\text{SVM}}$ comprises linear, generalized decision functions of the following form:

$$\mathcal{F}_{\text{SVM}} = \{f_{\boldsymbol{w}} = I \mapsto \boldsymbol{w}^\top\phi(I)\,:\,\boldsymbol{w}\in\mathbb{R}^{\mathsf{m}}\} \subset \mathcal{I}\to\mathbb{R}\,.$$

where $\phi \in \mathcal{I} \to \mathbb{R}^{\mathsf{m}}$ is a pre-defined *feature map* that provides a vector-valued feature abstraction for image I. Functions $f_{\boldsymbol{w}} \in \mathcal{F}_{\text{SVM}}$ are penalized by means of an ℓ_2-regularization of the parameter vector $\boldsymbol{w}$, *i.e.*

$$\Omega_{\ell_2}(f_{\boldsymbol{w}}) = \|\boldsymbol{w}\|_2^2\,,$$

and the loss function we consider is given in terms of the squared hinge loss as follows:

$$L_{\text{hinge}^2}(y, z_I, z_J) = |1 - y(z_I - z_J)|_+^2\,,$$

where $|x|_+ = \max(x, 0)$. With this loss, one incurs no penalization if $y = +1$ and $z_I > z_J + 1$, or if $y = -1$ and $z_J > z_I + 1$, *i.e.* when the image deemed as safer has indeed a safety degree that is at least 1 point better than the one of the other image.

Under this setting, the optimization in (1) is convex in $\boldsymbol{w}$ and a global solution can be recovered using the algorithm in [5]. Once the learned parameters $\boldsymbol{w}$ are obtained, the degree of safety for a novel image $\hat{I}$ can be computed as $f_{\boldsymbol{w}}(\hat{I})$.

3.2.2 Convolutional Neural Networks

A drawback of the SVM-based approach is its dependence on a pre-defined feature mapping ϕ. To overcome this limitation, we consider a richer hypothesis space in which functions admit a compositional factorization of the type $f = h_{\mathsf{k}} \circ \cdots \circ h_1$, such as those implemented by multilayer neural networks. In the specific, the architecture we propose can be organized into three compositional blocks (Figure 2):

Feature extractor

$$h_{fe}(\cdot;\theta) \in \mathcal{I} \to \mathbb{R}^{\mathsf{r}\times\mathsf{s}\times\mathsf{t}}$$

This block maps an input image I to an intermediate $\mathsf{r}\times\mathsf{s}\times\mathsf{t}$-dimensional representation. It consists of the first 2/3 layers (we tried both settings) of the deep convolutional network proposed in [20] and known as AlexNet. All the parameters of the block are held by θ;

Latent detectors

$$h_{ld}(\cdot;K) \in \mathbb{R}^{\mathsf{r}\times\mathsf{s}\times\mathsf{t}} \to \mathbb{R}^{\mathsf{m}}\,, \quad K = (K_1, \dots, K_{\mathsf{m}})$$

This block takes the output H of the feature extractor and returns the scalar responses of m detectors of latent visual concepts. The i-th detector performs the convolution of the input H with a linear filter $K_i \in \mathbb{R}^{\mathsf{u}\times\mathsf{v}\times\mathsf{t}}$, followed by a ReLU non-linearity, obtaining a matrix. This matrix is then fed to a pooling operator $\Pi_{\eta_i}(\cdot)$, parametrized by $0 \le \eta_i \le 1$, to obtain a single scalar. The pooling parameters $(\eta_1, \dots, \eta_{\mathsf{m}})$ are model choices to be fixed a priori. In our experiments we use linear filters of spatial dimensions $\mathsf{u} = \mathsf{v} = 3$, corresponding to a receptive window of 99×99

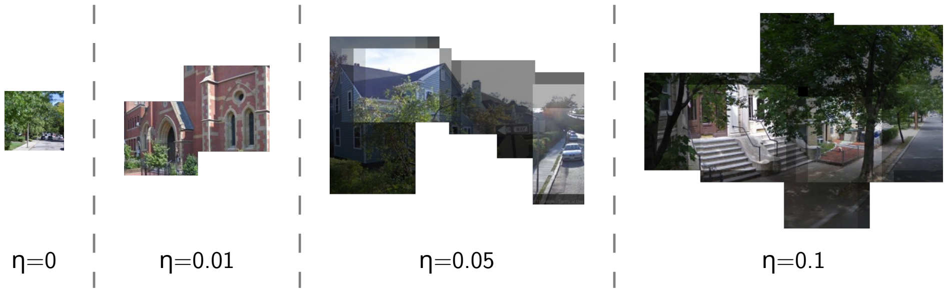

Figure 3: Visual representation of our pooling method. Given different values of the η parameter we show the areas of four images that concur to the output value of the pooling function, *i.e.* those corresponding to the $1 + \lceil\eta(\mathsf{wz}-1)\rceil$ largest elements in its input M. Each pixel's intensity is scaled according to the magnitude of its corresponding element in M.

pixels in the input image. The pooling operator $\Pi_\eta(M)$ that we propose generalizes the well-known max-pooling and average-pooling operators. It takes as input a matrix $M \in \mathbb{R}^{\mathsf{w}\times\mathsf{z}}$ and depends on a parameter $0 \leq \eta \leq 1$. It computes the average of the $1 + \lceil\eta(\mathsf{wz}-1)\rceil$ largest elements in M. If $\eta = 0$ then it returns the largest element as the max-pooling operator, while if $\eta = 1$ it returns the average of the elements in M as the average-pooling operator. Intuitively, this pooling operator selects a variable-sized portion of the input image as the most relevant for the detector's response. In other words, the proposed structure for latent detectors permits to discover the regions of the image which are discriminative with respect to the perceived safety/unsafety. A graphical representation of this intuition is shown in Figure 3.

In summary, this block takes the following form:

$$h_{ld}(H;K) = (\Pi_{\eta_1}(|H * K_1|_+), \ldots, \Pi_{\eta_\mathsf{m}}(|H * K_\mathsf{m}|_+)) \ ,$$

where $|M|_+$ returns M with negative elements set to 0.

Previous works [4, 19] have investigated similar parametric pooling methods, which allow a smooth transition between average and max-pooling. However, none of these works have focused specifically on CNN architectures for automatic scene analysis.

Safety predictor

$$h_{sp}(\cdot;\boldsymbol{w}) \in \mathbb{R}^\mathsf{m} \to \mathbb{R}$$

This block is a linear decision function with parameter $\boldsymbol{w} \in \mathbb{R}^\mathsf{m}$ applied to the responses $\boldsymbol{x} \in \mathbb{R}^\mathsf{m}$ of the latent detectors, *i.e.*

$$h_{sp}(\boldsymbol{x};\boldsymbol{w}) = \boldsymbol{w}^\top\boldsymbol{x} \, .$$

The final form of the safety prediction function is

$$f_{\boldsymbol{w},K,\theta} = h_{sp}(\cdot;\boldsymbol{w}) \circ h_{ld}(\cdot;K) \circ h_{fe}(\cdot;\theta) \, ,$$

where $\boldsymbol{w} \in \mathbb{R}^\mathsf{m}$, $K \in \mathbb{R}^{\mathsf{u}\times\mathsf{v}\times\mathsf{t}\times\mathsf{m}}$ and $\theta \in \Theta$, Θ being the set of all possible parametrizations of the h_{fe}. The corresponding hypothesis space is thus given by

$$\mathcal{F}_{\text{rCNN}} = \left\{ f_{\boldsymbol{w},K,\theta} \, : \, \boldsymbol{w} \in \mathbb{R}^\mathsf{m}, K \in \mathbb{R}^{\mathsf{u}\times\mathsf{v}\times\mathsf{t}\times\mathsf{m}}, \theta \in \Theta \right\} \subset \mathcal{I} \to \mathbb{R} \, .$$

Functions in $\mathcal{F}_{\text{rCNN}}$ are regularized as usually done for neural networks through an ℓ_2-penalization of the parameters in each layer. As for the loss function, we adopt the so-called logistic loss, which is given by

$$L_{\text{logistic}}(y, z_I, z_J) = \log(1 + \exp(z_J - z_I)) \, .$$

As opposed to the SVM case, the optimization of (1) in this scenario is non-convex and we rely on Stochastic Gradient Descent (SGD). Specifically, we adopt an SGD solver with momentum $\mu = 0.9$ and learning rate $\alpha = 0.01$ (or $\alpha = 0.001$ for fine-tuned layers), scheduling a tenfold decrease of the learning rate every 6 training epochs.

As a final remark, it is worth noting that, comparing our CNN with recent deep learning approaches popular in the vision community [20, 40], the proposed architecture is relatively shallow. Our choice is motivated by the fact that, with the proposed latent detectors, we aim to discover mid-level visual representations (*i.e.* small image regions), which are discriminative with respect to the perceived safety. Importantly, our experimental results demonstrate that this choice does not imply a decrease in terms of prediction accuracy.

3.3 Determining the training set noise

The learning task that we address has a data distribution that is by nature affected by noise, because humans, to some degree, tend to provide inconsistent judgements regarding the relative safety of image pairs. For this reason, it is interesting to quantify the smallest error that can be achieved for some data $\mathcal{D}_\mathsf{n}$. To this end, we setup a learning task under an unconstrained hypothesis space $\mathcal{F}$ consisting of explicit mappings of images in $\mathcal{D}_\mathsf{n}$ to real values, without regularization (*i.e.* $\lambda = 0$) and with a binary loss. By doing so, the solution to (1) yields an error, which corresponds to the intrinsic noise of the dataset.

Unfortunately, the ERM optimization under this setting is NP-hard, as it is substantially equivalent to the *Minimal Feedback Arc Set* (MFAS) problem [16], which asks for the minimum set of edges that should be removed from a directed graph to obtain an acyclic graph. The directed graph $G = (\mathcal{V}, \mathcal{E})$ we refer to can be constructed from the training set by considering the set of images in $\mathcal{D}_\mathsf{n}$ as the vertex set $\mathcal{V}$, and for each judgement $(I, J, y) \in \mathcal{D}_\mathsf{n}$ we have an edge $(I, J) \in \mathcal{E}$ if $y = +1$, or $(J, I) \in \mathcal{E}$ if $y = -1$.

From the MFAS of G, denoted by $\mathcal{A} \subset \mathcal{E}$, we can directly recover the optimal empirical error as $R_{\text{emp}}(f^\star) = |\mathcal{A}|/|\mathcal{E}|$. As for the minimizer $f^\star$, there are uncountably many possible solutions that belong to the set

$$\mathcal{F}^\star = \{ f \in \mathcal{F} \, : \, f(I) > f(J) \, , \forall (I, J) \in \mathcal{E} \setminus \mathcal{A} \} \ .$$

To construct an arbitrary function $f \in \mathcal{F}^\star$, we initialize f_I for all $I \in \mathcal{V}$ with a random real number. Then, we visit the directed acyclic graph $G' = (\mathcal{V}, \mathcal{E} \setminus \mathcal{A})$ in breadth-first order starting from the source vertices, *i.e.* those ones having no incoming edge. Each time a new vertex $I \in \mathcal{V}$ is visited, we set $f_J \leftarrow \max(f_J, f_I + r)$ with r a random positive number for each vertex $J \in \mathcal{V}$ that is adjacent to I, *i.e.* such that (I, J) is an edge of G'. At the end of the graph visit, f represents an arbitrary function in $\mathcal{F}^\star$ and, therefore, a minimizer of (1). Since solving the MFAS problem exactly is in general NP-hard, we rely on the heuristic algorithm of Berger and Shor [3] to find a possibly good solution. In the following section we show the results of this analysis on the Place Pulse dataset.

4. EXPERIMENTAL RESULTS

4.1 Experimental setup

As mentioned above, our experimental evaluation is conducted on the Place Pulse 1.0 dataset. We implemented Convolutional Neural Networks approaches using the Caffe package [15] and run them using a NVIDIA Tesla K40 GPU. For RankingSVM, we adopted a publicly available implementation[4], while for Support Vector Regression (SVR) we used LIBLINEAR [9]. In all our experiments, the regularization parameters of RankingSVM and SVR have been chosen with a five-fold cross-validation. The performance of the tested methods has been evaluated in terms of the share of correctly predicted users' votes on pairs of images.

4.2 Prediction of Perceived Urban Safety

In this section we demonstrate that a ranking approach can be successfully used to learn a function that automatically predicts the perceived safety of places from images.

4.2.1 Ranking vs Regression

In Section 3 we have shown that we can learn from the users' judgements a real-valued function that assigns a score to each image. This function induces a total ordering (*a.k.a.* ranking) on the set of images, which should be as consistent as possible with the pairwise relations gathered from the users. Although the true source of information for the Place Pulse dataset is given by the users' pairwise judgements, previous works [25, 27] have focused on training their algorithms in a way to mimic a ”ground-truth” image scoring function, which has been computed by the Place Pulse creators from the users' annotations through the Microsoft Trueskill algorithm. In this way, the attention has been shifted from the original ranking problem based on the users' judgements to a regression problem relying on the provided ”ground-truth” scoring function.

In our opinion, this drift raises two important issues. First of all, the ability of an algorithm to mimic the provided scoring function *weakly* correlates with the ability of the algorithm to explain the true data distribution, which consists of the users' pairwise annotations, being the only, true, physical observations one can rely on. Indeed, there are uncountably many possible scoring functions (see, Subsection 3.3) performing equally well, but being arbitrarily different, that could potentially have been taken as a ”ground-truth” scoring function. Moreover, errors measured for the regression task on the scoring function do not properly map to errors relative to the true data distribution (*i.e.* violations of the users' annotations). A second issues is that, according to the previous point, the only reliable way to measure how well a trained algorithm performs for the safety prediction task on the Place Pulse dataset is via an error measurement on the true data distribution. However, any attempt in this direction will be intrinsically biased, if the algorithm has been trained using the ”ground-truth” scoring function. This is because the ”ground-truth” scoring function has been constructed using all the users' annotations and, hence, any algorithm trained for regression on those scores is a function of the whole dataset, thus leaving no independent piece of information for a proper error assessment.

Table 1: RankingSVM: Performance with different features representations.

Features	Average Accuracy (std)
GIST	65.42% (±0.95)
HOG	65.82% (±0.79)
ImageNet	62.02% (±1.15)
PLACES	66.34% (±1.16)
SSIM	64.48% (±1.03)
SUN	65.97% (±1.04)

We also found out that the ”ground-truth” scoring function based on Microsoft Trueskill, which is used in [25, 27], is not the best representative for the users' annotations. Indeed, it yields an accuracy of 72.87%, which is smaller than what we obtain with a scoring function constructed using the procedure described in Subsection 3.3, namely $76,9\%$.

For the aforementioned reasons, we do not find reliable to train an algorithm for regression on the ”ground-truth” scoring function provided with the Place Pulse dataset: it will hinder the possibility of a more principled error assessment via the users' annotations, *i.e.* the true data distribution. We consider instead more appropriate to train our algorithm directly on the users' votes.

4.2.2 RankingSVM

In this section we analyze the performance of a ranking approach, *i.e.* RankingSVM, when different types of features are considered: **GIST**, Histogram of Oriented Gradients (**HOG**) and Self-similarity descriptors (**SSIM**), as described in [28]; features extracted from the sixth layer of the Caffe reference network, trained on the 1.2 million images of **ImageNet** (ILSVRC 2012) [15]. These features have been used in previous related works [25, 27]. We also propose two additional feature representations: (i) features extracted from the sixth layer of a CNN having the same architecture as the Caffe reference network, trained on the recent **PLACES** [40] dataset; and (ii) features derived from the **SUN** Attribute dataset [28]. Specifically, we used as high-level features the scores computed by predicting the presence of each of the 102 crowd-sourced scene attributes available in the SUN Attribute dataset [28]. The 102 attributes are quite heterogeneous and span materials, surface properties, functions or affordances, and spatial envelope properties. The prediction of attributes is obtained following the approach described in [28] using a publicly available code.[5]

The results shown in Table 1 demonstrate that, despite the many challenges inherent to the task, a ranking approach can be used to learn to predict human judgements of safety. Table 1 also shows that the best performance is obtained using the features derived from the pre-trained Places-CNN. This is somehow expected, as these features represent state-of-the-art descriptors for scene recognition problems [40]. However, confirming the findings of previous works on the

[4] http://olivier.chapelle.cc/primal/

[5] http://cs.brown.edu/~gen/sunattributes.html

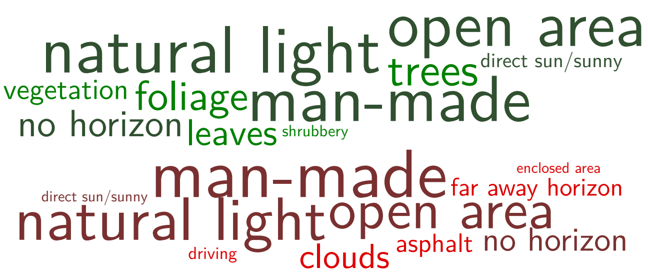

Figure 4: Word clouds of SUN attributes. Top, green: safe, bottom, red: unsafe.

Table 2: Comparison between different ranking methods.

	Method	Accuracy
CNN	rCNN$_2$[$\mathsf{m}=24,\eta_A$]	70.25%
	rCNN$_3$[$\mathsf{m}=24,\max$]	69.12%
	AlexNet-noinit	66.85%
	AlexNet-ImageNet	67.15%
	AlexNet-PLACES	70.65%
SVM	GIST	66.18%
	HOG	66.37%
	ImageNet	63.01%
	PLACES	65.93%
	SSIM	64.56%

Place Pulse dataset [27], we did not observe a significant increase in accuracy when using pre-trained CNN-derived features with respect to more traditional descriptors. It is worth noting that, discarding CNN-based features, the best performance is obtained considering SUN attributes.

Looking at SUN features, it is also interesting to try to interpret these data, by analyzing which of the SUN attributes correlate with the concept of perceived safety/unsafety. To this aim we performed a simple experiment. Given the ranking scores computed with the algorithm described in Section 3.3, we took the 100 images corresponding to places perceived as safe and 100 pictures associated with the most unsafe locations. For each of the SUN attributes we analyzed the distribution of the scores computed as in [28] and we selected the 10 most represented attributes in both sets. Figure 4 shows these attributes in the form of a word cloud, where the size of each attribute's name is proportional to the number of images in which it was detected. It is interesting to see that, among attributes being discriminative of a specific set, those corresponding to vegetation are associated to safe areas, while concepts like clouds or asphalt correlate with non-safe places. As expected, common attributes involve aspects like natural light, man-made structures and open area, indicating general concepts related to outdoor scenes.

4.2.3 Ranking with CNN

In this section we analyze the performance of the proposed CNN architectures for ranking (rCNN). Specifically we consider the following configurations (*cfr.* Section 3.2.2):

Table 3: Comparison between different regression methods using the same split as Table 2.

Features	Accuracy
GIST [25, 27]	65.29%
HOG [25]	66.04%
ImageNet [27]	66.79%
PLACES	67.93%
SSIM [25]	63.93%

Table 4: Comparison among the proposed CNNs.

Method	Accuracy
rCNN$_2$[$\mathsf{m}=24,\eta_A$]	70.25%
rCNN$_2$[$\mathsf{m}=32,\eta_A$]	69.44%
rCNN$_2$[$\mathsf{m}=24,\eta_B$]	69.70%
rCNN$_2$[$\mathsf{m}=32,\eta_B$]	69.73%
rCNN$_2$[$\mathsf{m}=24,\max$]	66.87%
rCNN$_2$[$\mathsf{m}=32,\max$]	68.39%
rCNN$_3$[$\mathsf{m}=24,\max$]	69.12%
rCNN$_3$[$\mathsf{m}=32,\max$]	68.98%

- rCNN$_2$[$\mathsf{m}=24,\eta_A$], rCNN$_2$[$\mathsf{m}=32,\eta_A$]: 2-layers feature extractor, respectively 24 and 32 latent detectors split into four groups with pooling factors $\eta_A = (0, 0.01, 0.05, 0.1)$;
- rCNN$_2$[$\mathsf{m}=24,\eta_B$], rCNN$_2$[$\mathsf{m}=32,\eta_B$]: 2-layers feature extractor, respectively 24 and 32 latent detectors split into four groups with pooling factors $\eta_B = (0, 0.1, 0.25, 0.5)$;
- rCNN$_2$[$\mathsf{m}=24,\max$], rCNN$_2$[$\mathsf{m}=32,\max$]: 2-layers feature extractor, respectively 24 and 32 latent detectors with max pooling ($\eta = 0$);
- rCNN$_3$[$\mathsf{m}=24,\max$], rCNN$_3$[$\mathsf{m}=32,\max$]: 3-layers feature extractor, respectively 24 and 32 latent detectors with max pooling ($\eta = 0$).

The feature extractor layers of all configurations are fine-tuned from AlexNet trained on the Places dataset [40], while all other layers are learned from scratch. Due to the long training time of CNNs, the results presented in this section refer to a single random split of the data, where 80% of the images and their votes are used for training and 20% for testing. We do not expect significant differences in the results when a different split is chosen.

In a first series of experiments we compare the two best performing configurations of our CNN architecture with RankingSVM and state-of-the-art deep learning methods. In particular, we adapt the well-known AlexNet architecture [20] to our task: we keep the original topology for the bottom 6 layers (5 convolutional, 1 fully connected) and attach a final fully connected layer that outputs the ranking score. The resulting network is trained both from scratch (denoted as AlexNet-noinit) and by fine-tuning the bottom six layers from two publicly available models learned on object recognition [20] (AlexNet-ImageNet) and scene recognition [40] (AlexNet-PLACES) tasks. Table 2 reports the results of this comparison. It is immediately clear that CNN approaches outperform SVM-based ones, confirming the advantages of deep learning for predicting human perception over previous works based on traditional features descriptors [25, 27]. By comparing the accuracy of our CNNs with the AlexNet-derived ones, we observe that AlexNet-PLACES slightly outperforms our approach. Indeed, a deeper network typically guarantees improved performance when sufficient training data is available. The requirement for considerable amounts of training data is reflected by the results obtained for AlexNet-noinit, which shows the worst performance among the CNN approaches. Finally, we note that AlexNet-ImageNet achieves a lower accuracy than both AlexNet-PLACES and our CNNs. We think that this depends on the scene recognition task on which AlexNet-PLACES was pre-trained begin closer to our safety-prediction task, as opposed to the object recognition task in ImageNet.

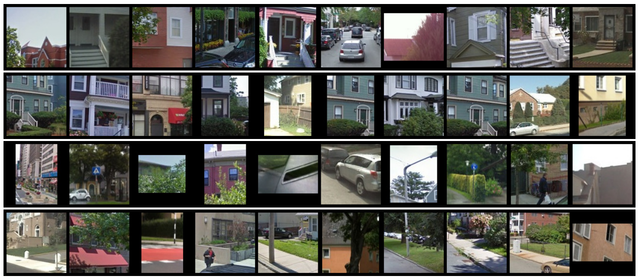

Figure 5: The most important patterns correlated with perception of safety discovered by the network rCNN$_2$[$\mathsf{m} = 24, \eta_B$]

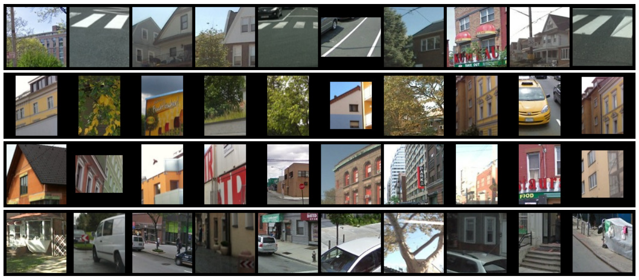

Figure 6: The most important patterns correlated with perception of safety discovered by the network rCNN$_2$[$\mathsf{m} = 32, \eta_A$]

For the sake of comparison, we also report the results obtained on the same data split using SVR (Table 3). Note that the values in Table 3 suffer from the score bias discussed in Section 4.2.1. Nevertheless, Table 3 clearly shows that the proposed CNN outperforms previous regression-based methods [25, 27], independently from the feature representation used. Table 4 reports a comparison among the aforementioned configurations of our CNN architecture. We observe an advantage in terms of accuracy for the mixed-pooling configurations compared to the max-only ones. In particular, rCNN$_2$[$\mathsf{m} = 24, \eta_A$] achieves the best overall result with an accuracy of 70.25%, while rCNN$_2$[$\mathsf{m} = 24, \max$] shows the worst with an accuracy of 66.87%, suggesting that both localized and diffused features are more informative for the prediction of perceived safety. Looking at the effects of changing the number of latent detectors, only slight differences in accuracy are observed between $\mathsf{m} = 24$ and $\mathsf{m} = 32$. We did not report results for $\mathsf{m} > 32$ as they typically correspond to a decrease in accuracy as well as a less clear separation between safe/unsafe visual patterns.

4.3 Visualizing Patterns of Safety/Unsafety

In this section we analyze the ability of the proposed CNN approaches to automatically discover visual patterns associated with the perception of safety. With reference to Section 3.2.2, we use the learned weights $\boldsymbol{w}$ to characterize the latent detectors. Intuitively, since a latent detector's output is always greater than or equal to zero, its positive or negative contribution to the predicted image safety only depends on the sign of the corresponding entry of $\boldsymbol{w}$. Thus, we mark as "safe" the detectors associated with learned positive weights and "unsafe" the others, while we use the weight's magnitude to determine how discriminative the detector is.

As a first visualization approach, we select the four most discriminative "safe" and "unsafe" latent detectors from the two best performing CNNs of Table 4: rCNN$_2$[$\mathsf{m} = 24, \eta_A$] and rCNN$_2$[$\mathsf{m} = 24, \eta_B$]. Then, we extract from the dataset the 10 patches showing the highest response on each detector and collect them (arranged row-wise) in Figures 5-8. Looking at Figures 5 and 6, we can see a prevalence of residential single houses, and, in particular, windows of residential houses. Another interesting aspect is the presence of trees and portions of gardens. Interestingly, Quercia *et al.* [31] found residential trees and residential windows as visual words associated with quiet images. The prevalence of residential trees and residential gardens is also in line with the results obtained in environmental psychology [38], highlighting the positive influence of vegetation, particularly gardens, parks and residential trees, on judgements about urban spaces. Finally, in some of the patches are represented crosswalks. Turning now our attention to the four most discriminative "unsafe" latent detectors (Figures 7 and 8), we can see the presence of gates, graffiti and other signs of vandalism. As pointed out by the *Broken Windows Theory*, graffiti contribute to a perception of breakdown of social order [17] and usually are associated with judgments of low safety [35]. Moreover, *Broken Windows Theory* suggested a positive relationship not only between signs of disorder and perception of unsafety, but also between disorder and crime rates. Another emerging sign from the images associated with judgments of unsafety is the presence of empty roads, electricity and communication pylons. Finally, the buildings are mainly council houses and industrial buildings and their selected portions represent usually the roofs.

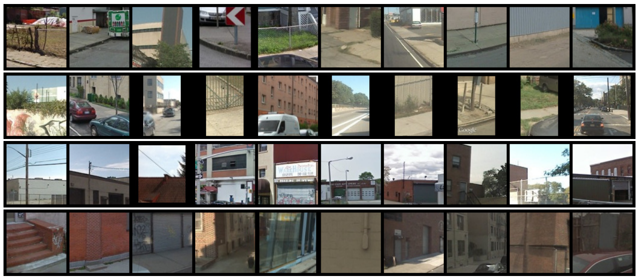

Figure 7: The most important patterns correlated with perception of unsafety discovered by the network rCNN$_2$[$\mathsf{m} = 24, \eta_B$]

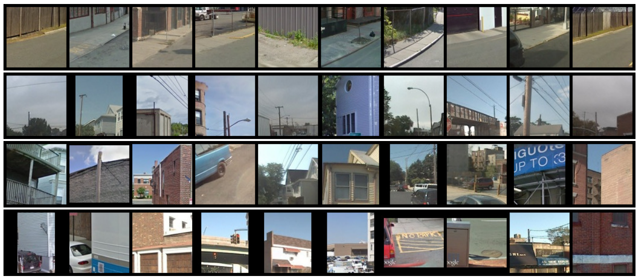

Figure 8: The most important patterns correlated with perception of unsafety discovered by the network rCNN$_2$[$\mathsf{m} = 32, \eta_A$]

As mentioned in Section 3.2.2, the output of the pooling operator adopted in our rCNN networks only depends on a portion of the input image. Depending on the η parameter the size of this portion varies from a single patch to the whole image. This observation suggests an effective way to visualize the internal representation learned by each of our latent detectors, which is exploited to generate the images in Figure 9. Here we chose three "safe" and three "unsafe" detectors from $\mathrm{rCNN}_2[\mathsf{m} = 24, \eta_A]$, each having a different pooling parameter, and study their output on some of the images that exhibit the strongest response on each detector. For each image we show the portions that concurred to the detectors' output and scale the pixel intensities according to the local detector response. As expected, the images corresponding to a small value of the pooling parameter depict very localized features of the environment, like roofs and windows. Increasing values of η shift the detectors' attention towards bigger areas, comprising sidewalks, roads, facades or entire buildings. The same observations concerning the nature of the features associated with safety and unsafety reported in the previous paragraphs also apply here (presence of residential gardens and trees, and residential single houses for safety, while council houses and industrial buildings for unsafety). It is also interesting to note how the aspect ratio of the image regions selected by the "unsafe" detectors tends to be wider, highlighting horizontal structures typical of industrialized open areas.

5. CONCLUSIONS AND FUTURE WORK

We presented a novel approach for predicting the perceived safety of urban scenes from Google Street View images. Our extensive experimental evaluation, conducted on the publicly available Place Pulse dataset, demonstrates that the proposed method, combining a ranking framework with the representational power of CNNs, is more accurate than state of the art methods. Moreover, our CNN-based approach permits to discover automatically mid-level visual patterns correlated with urban perception. To our knowledge, this is the first work that introduces a computational model for addressing the issue: *What makes a place look safe?*. Future works include extending the proposed approach to prediction of other high level attributes and analyzing the difference among scenes of various geographic areas (Europe vs USA).

Acknowledgments

We would like to thank Cesar Hidalgo and Nikhil Naik for the insightful discussion, as well as for sharing the data of Place Pulse 1.0. We also thank NVIDIA Corporation for providing us a K40 GPU.

6. REFERENCES

[1] Y. Arase, X. Xie, T. Hara, and S. Nishio. Mining people's trips from large scale geo-tagged photos. In *ACM Multimedia*, 2010.

[2] S. M. Arietta, A. A. Efros, R. Ramamoorthi, and M. Agrawala. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Trans. on Visualization and Computer Graphics*, 2014.

[3] B. Berger and P. W. Shor. Approximation alogorithms for the maximum acyclic subgraph problem. In *ACM-SIAM symposium on Discrete algorithms*, 1990.

[4] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *International Conference on Machine Learning*, 2010.

[5] O. Chapelle and S. S. Keerthi. Efficient algorithm for ranking with svms. *Information Retrieval*, 13(3):201–215, 2010.

[6] Y.-Y. Chen, A.-J. Cheng, and W. H. Hsu. Travel recommendation by mining people attributes and travel group types from community-contributed photos. *IEEE Trans. on Multimedia*, 15(6):1283–1295, 2013.

[7] H. David. *The method of paired comparisons*. Hodder Arnold, 1988.

[8] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012.

[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 2008.

[10] H. Grabner, F. Nater, M. Druey, and L. Van Gool. Visual interestingness in image sequences. In *ACM Multimedia*, 2013.

[11] T. Graepel, T. Minka, and R. T. Herbrich. A bayesian skill rating system. *Neural Information Processing Systems*, 2007.

[12] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition*, 2008.

[13] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000.

[14] J. Jacobs. *The death and the life of great american cities*. Random House, 1961.

[15] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. *http://caffe.berkeleyvision.org*, 2013.

[16] R. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, The IBM Research Symposia Series, pages 85–103. 1972.

[17] G. Kelling and C. Coles. *Fixing broken windows: Restoring order and reducing crime in our community*. Free Press, 1998.

[18] A. Khosla, B. An, J. J. Lim, and A. Torralba. Looking beyond the visible scene. In *Computer Vision and Pattern Recognition*, 2014.

[19] P. Koniusz, F. Yan, and K. Mikolajczyk. Comparison of mid-level feature coding approaches and pooling strategies in visual concept detection. *Computer Vision and Image Understanding*, 117(5):479–492, 2013.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, 2012.

[21] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, 2006.

[22] P. Lindal and T. Hartig. Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology*, 2012.

[23] K. Lynch. *The image of the city*, volume 11. MIT press, 1960.

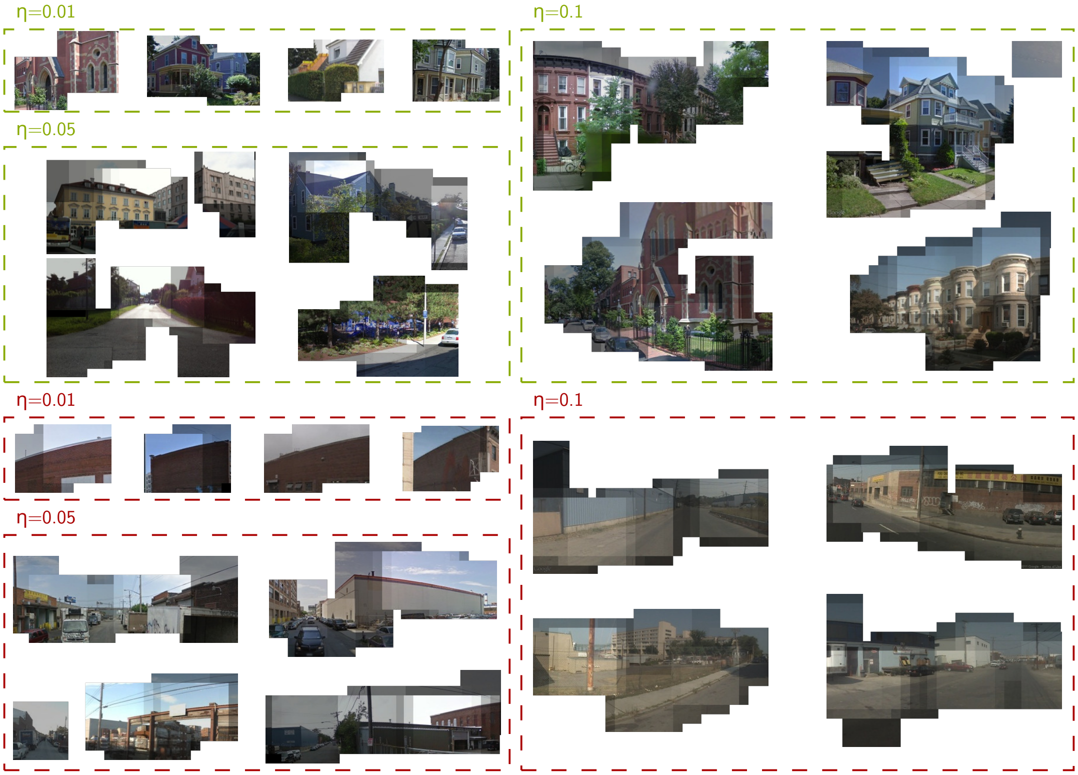

Figure 9: Visualization of the responses of latent detectors associated to safe (green boxes) and unsafe areas (red boxes) from our rCNN$_2$[m $= 24, \eta_A$] network. For each image we show the regions that concur to a detector's output after the pooling function with the shown η value is applied.

[24] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *International Conference on Computer Vision*, 2011.

[25] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo. Streetscore–predicting the perceived safety of one million streetscapes. In *Computer Vision and Pattern Recognition Workshops*, 2014.

[26] J. Nasar. *The evaluative image of the city*. Sage Publications, 1997.

[27] V. Ordonez and T. L. Berg. Learning high-level judgments of urban perception. In *European Conference on Computer Vision*. 2014.

[28] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[29] L. Porzi, S. R. Buló, P. Valigi, O. Lanz, and E. Ricci. Learning contours for automatic annotations of mountains pictures on a smartphone. In *International Conference on Distributed Smart Cameras*, 2014.

[30] L. Porzi, E. Ricci, T. Ciarfuglia, and M. Zanin. Visual-inertial tracking on android for augmented reality applications. In *IEEE Workshop on Environmental, Energy and Structural Monitoring Systems*, 2012.

[31] D. Quercia, N. K. O'Hare, and H. Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *ACM Computer supported cooperative work & social computing*, pages 945–955, 2014.

[32] P. Salesses, K. Schechtner, and C. A. Hidalgo. The collaborative image of the city: mapping the inequality of urban perception. *PloS one*, 8(7):e68400, 2013.

[33] R. J. Sampson, J. Morenoff, and T. Gannon-Rowley. Assessing 'neighborhood effects': Social processs and new directions in research. *Annual Review of Sociology*, 2002.

[34] R. J. Sampson and S. W. Raudenbush. Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. *American Journal of Sociology*, 105(3):603–651, 1999.

[35] H. Schroeder and L. Anderson. Perception of personal safety in urban recreation sites. *Journal of Leisure Research*, 16:177–194, 1983.

[36] W. Skogan. *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. University of California Press, 1990.

[37] R. Urlich. Visual landscapes and psychological well-being. *Landscape Research*, 4:17–23.

[38] R. Weber, J. Schnier, and T. Jacobsen. Aesthetics of streetscapes: Influence of fundamental properties on aesthetic judgments of urban space. *Perceptual and Motor Skills*, 106:128–146, 2008.

[39] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer Vision and Pattern Recognition*, 2010.

[40] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Neural Information Processing Systems*, 2014.