Finding Common Image Semantics for Urban Perceived Safety Based on Pairwise Comparisons

Gabriel Costa
Institute for Systems and Robotics
Instituto Superior Tecnico
Lisbon, Portugal
gabrielcosta@tecnico.ulisboa.pt

Claudia Soares
Institute for Systems and Robotics
Instituto Superior Tecnico
Lisbon, Portugal
csoares@isr.tecnico.ulisboa.pt

Manuel Marques
Institute for Systems and Robotics
Instituto Superior Tecnico
Lisbon, Portugal
manuel@isr.ist.utl.pt

***Abstract*—What influences people's perception of safety in an urban environment? Does everyone perceive safety the same way or do different people look for different contents in an image, safety-wise? We present a user analysis on a crowdsourced dataset that contains pairwise comparisons regarding the perceived safety of street imagery from different municipalities in the greater Lisbon area, Portugal. We use state-of-the-art semantic segmentation to extract the contents of images and cluster different people according to what they perceive as safe. Then, we study semantic classes and analyze clusters of users for semantic elements appearing in images classified as safer (or more dangerous). The results show that clusters share a lot of similarities. Our analysis evidences that, for users with more pairwise comparisons, there is only one group, while spurious groupings appear when users contribute less. This result emphasizes that a pairwise image comparison dataset potentiates agreement of users in perceptual tasks, for moderate comparison data size.**

***Index Terms*—Urban perceived safety, Semantic urban segmentation, Crowdsourced perceptual dataset, Pairwise comparisons.**

I. Introduction

Understanding what makes people feel safer allows policy-makers to better shape neighborhoods, improving residents' well being, and enables citizens to monitor and draw conclusions on the collective perception of safety.

Quantifying perception is a difficult problem mainly due to the need for covering a full range of different levels of perception with judgments of much more than one person, and to the subjectivity and variability of perception judgments themselves. One of the most popular approaches for perception assessment is through lengthy surveys [1], covering a small sample and prone to biases [2]. Side by side pairwise comparisons address these shortcomings but transfer complexity to the data processing side. This method is used to approach the very relevant question of whether different profiles of safety perception can be found among participants or not, and how can we describe those profiles. This paper shows that, for urban perceived safety, our data in the form of pairwise comparisons on street imagery has, indeed, a strong inter-observer agreement. They also show that human-related contents are linked with safer perception while vehicle-related contents are seen as unsafe. From the specificity of our data, we developed this conclusion based on the analysis of the semantic contents and individual preferences using state-of-the-art semantic segmentation methods. Our data does not comply with the assumptions of repeated observations because pairs of images are randomly generated from a very large pool of images, and so, there are not, in general, repeated individual comparisons as assumed in the work stemmed from reference [3].

II. Related Work

Estimating the urban safety perception and computing its relationship with socio-economic indicators are longstanding challenges for the scientific community. Social scientists have been studying links with, for instance, neighborhood disorder [4], physical activity among the elderly [5] and academic achievement [6]. More recently, the increasing availability and diversity of data sources concerning the physical city, such as street imagery from Google Street View, combined with new computer vision techniques have allowed researchers to conduct studies both with more resolution and on a larger scale [7]. These resources have been used to assess urban greenery [8], estimate multiple demographic indicators [9] and study what drives physical change in neighborhoods [10].

The Place Pulse project collected crowdsourced pairwise comparisons of images from all over the world to estimate perceptions of safety, wealth or beauty on a global scale [11]. These data were then used to estimate safety perception from street imagery alone using both generic image features (only concerning safety in Boston and New York) [12] and deep learning [13]. Computer vision has seen big improvements in recent years due to deep neural networks, which were enabled by the availability of very large datasets. Semantic segmentation has been a fundamental area of research as it provides visual understanding. Street imagery semantics is especially relevant for many urban applications, for example, autonomous driving. Researchers have, thus, developed multiple datasets that focus on urban scene contents like COCO-Stuff [14], Mapillary Vistas [15] and Cityscapes [16]. Due to the large amount of data required by deep learning, the use of artificial scene datasets in conjunction with real imagery has also shown good results in semantic segmentation [17]. There has also been significant work in developing better feature extraction networks [18]–[21] to better capture objects

978-9-0827-9703-9/19/$31.00

at multiple scales and improve feature resolution, which are the two main challenges in semantic segmentation.

As far as the authors know, the present work is the first agreement analysis of image semantics from a large perception dataset. Due to the low probability in forming a specific image pair, our agreement study follows a new and specific methodology. Our analysis evidences strong agreement on image semantics for hundreds of different observers.

III. The City-SAFE dataset

Our City-SAFE dataset, inspired by the Place Pulse project [11], comprises pairwise comparisons for perceived safety on street images from the greater Lisbon area (2056 from Lisbon, 1008 from Amadora and 2034 from Cascais). Images were retrieved from the Google Street View API, where the coordinates of each image were obtained using a grid bounded by polygons of the municipalities' limits. The available images for collection were taken on either sunny or cloudy days, from 2009 to 2018.

We chose to collect pairwise comparisons as it makes the participants' task easier and provides more reliable data when compared to directly providing a score [2]. Place Pulse has also shown that safety perception is not related to their gender, age or location. Through a website created by us, two random images from the image pool were presented side by side and we asked "Which place looks safer?". The possible answers were *left*, *right* or *equal*. A unique ID was attributed to each participant by setting a cookie and recorded with each comparison.

A total of 439 random people have participated in our ongoing crowdsourced data collection so far, having generated over 19k pairwise comparisons from November of 2018 until May of 2019. This results in an average of 44 comparisons per participant and about 25% of all comparisons were ties (which will not be considered in this study). There were slightly more *left* than *right* choices but both are fairly similar. Images were compared, on average 7 times with some images being compared as many as 19 times. Fig. 1 depicts the expected long tail behavior of the number of comparisons by each user. Most users concentrate on the < 20 comparisons group. We have identified different patterns in users as some label image pairs as *ties* much more often than others (some never do while others have up to 70% of ties). Fig. 2 exemplifies how preferences are distributed in a random group of users.

IV. Agreement assessment procedure

A. Semantic segmentation

We use the state-of-the-art Inplace-ABN (DeepLab3+WideResNet38) semantic segmentation implementation, trained on the Mapillary Vistas dataset of street imagery [19], to segment each image from our dataset into $N = 52$ semantic classes (this configuration provides a mean IoU of 82%). Fig. 3 depicts two examples of Google Street View images and the respective semantic segmentation results. In Fig. 4 we can see how often each class appeared on the comparisons. Classes like *sky*, *building*,

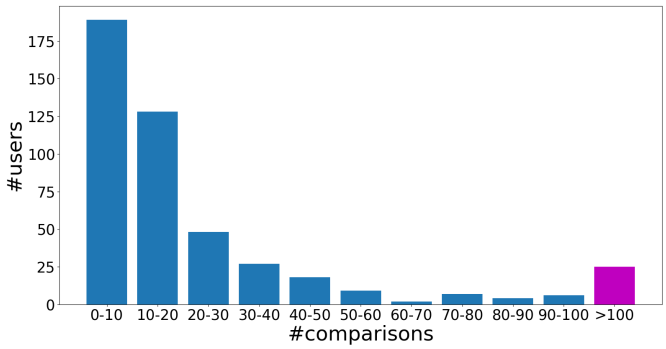

Fig. 1: Breakdown of participants by the number of comparisons they have provided. Most users provide less than 20 comparisons but there is still a significant number of users that provide over 100 comparisons. This is a natural example of a power law.

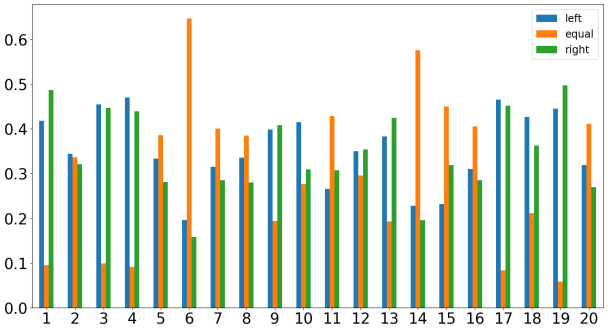

Fig. 2: The choices of the 20 most active users. Different profiles can be found where some users answer *equal* much more often than others. There are not significant *left*/*right* biases among the 20 most active users.

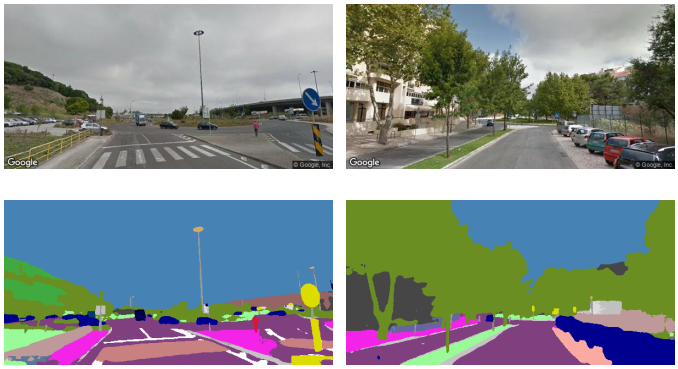

Fig. 3: Example of the semantic segmentation of street imagery rich in content including *cars* (dark blue), *sidewalks* (pink), *person* (red), *road* (purple) *poles* (grey), *street signs* (yellow) and *lane markings* (white). The segmentation procedure provides a mean IoU of 82%.

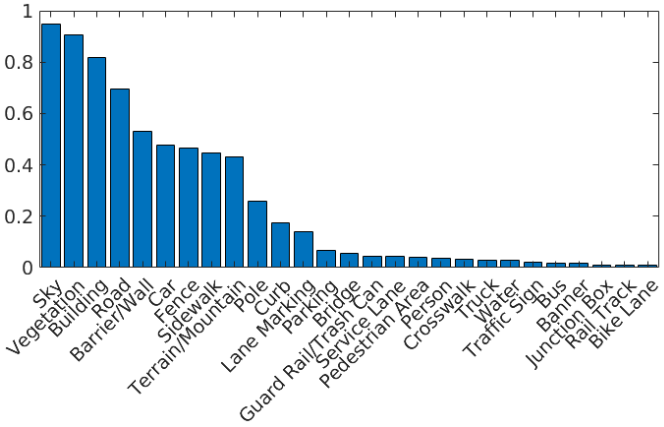

Fig. 4: Ratio of images where each semantic class has been detected. More than half of the classes appear in less than 10% of images. *Sky* and *vegetation* naturally appear in almost every image.

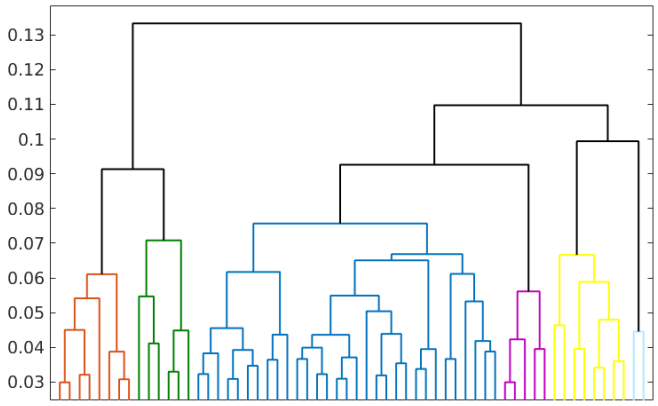

Fig. 5: Dendrogram of the agglomerative hierarchical clustering. Only 60 leaf nodes are shown (each is a cluster of one or more different users). Due to complete linkage being used no significant chaining effects are visible.

vegetation, *road* and *car* are extremely common while *rail track* and *water* are not very typical of street imagery. *Junction box* and *bus* do not get detected very often in our dataset because of their nature (there are not that many buses) although they are more typical of an urban scene than of a not urban scene. Using the segmentation, we derived a binary representation $x \in \mathbb{R}^N$ for every image where $x_n = 1$ if the semantic class i is present, and $x_n = 0$ otherwise. We empirically set a minimum of 1000 pixels (which is 0.5% of the size of the image) from a semantic class to consider it present in an image. We considered that instances smaller than 1000 pixels were not noticed when comparing two images.

Some classes were found to be redundant, as they are very similar and would often be detected interchangeably (for instance, there were no apparent differences between *curb* and *curb cut* detections) and for that reason these were merged. There were also some semantic classes that generated many misclassifications and were for this reason discarded.

Each pairwise comparison is represented by $c = (x_{\text{winner}}, x_{\text{loser}})$, which is the Cartesian product (or concatenation) between the semantic representations of the winning and losing image, respectively. Each comparison is, thus, represented in the nonnegative orthant of $\mathbb{R}^{2N}$. Ties were discarded for this user study. Since all pairwise comparisons have associated a user ID, for a user u, we generated a descriptor $\tilde{c}_u$ of all comparisons, C_u, made by this user

$$\tilde{c}_u = \sum_{c \in C_u} c. \tag{1}$$

Since some users had made very few comparisons, their votes were not considered faithful representatives of how the user safety perception translates into the semantic classes. Furthermore, since some semantic classes appear very rarely on the compared images, whether they are found in winning or losing images may provide inaccurate information as these results may be biased by other semantic classes in the same image or by the pairwise match. For these reasons, semantic classes with less than 200 total appearances were ignored and users with less than 10 comparisons were discarded. The final number of users was 197 and of semantic classes was 27.

B. Clustering users

Using the chosen representation, we conducted an agglomerative hierarchical clustering with complete linkage among all users where cosine distances were used. The cosine distance metric captures profile similarities, regardless of the number of times each user voted, which affects our representation and should not be relevant for clustering. The farthest distances to measure cluster proximity were chosen to avoid chaining, which is typical of clustering by the shortest distance (especially if the data are not separable with clear gaps) and thus fails to provide useful insights when searching for user preferences. Using complete linkage generates more compact clusters since it minimizes the largest distance within each cluster, despite being more sensitive to outliers. Since we already control for users with very few comparisons and undesired semantic classes to remove outliers, using complete linkage should provide satisfactory and meaningful results.

Fig. 5 shows the dendrogram of the agglomerative clustering of our representations $\tilde{c}_u$, as detailed in the previous Section. We prune to 60 clusters, and so each leaf node may represent one or more users. We cut the dendrogram into 6 clusters:

- Cluster 1 (light blue): 2 users with 22 comparisons;
- Cluster 2 (yellow): 13 users with 209 comparisons;
- Cluster 3 (purple): 6 users with 73 comparisons;
- Cluster 4 (dark blue): 169 users with 11812 comparisons;
- Cluster 5 (orange): 12 users with 192 comparisons;
- Cluster 6 (green): 6 users with 79 comparisons.

Clusters 1, 3 and 6 were not considered in the following analysis for representing few users. We notice that cluster 4 features most of our dataset and includes all of the most active

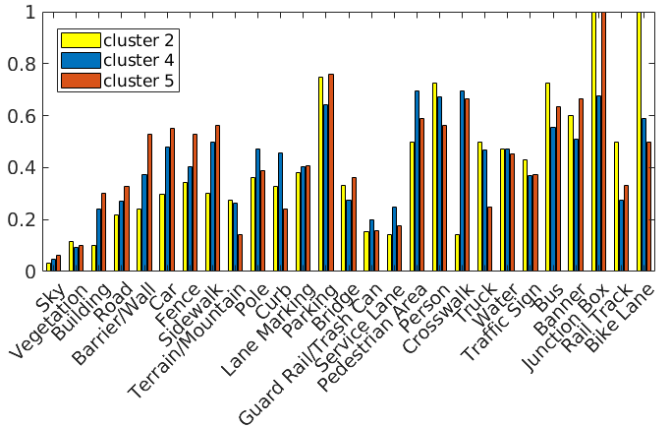

Fig. 6: How often each semantic class appears in winning images while not appearing in the losing image. Classes are ordered from the most common to the least common in our dataset with the former being not very discriminative. In general, all clusters follow a similar profile in which classes appear more often on winning images only.

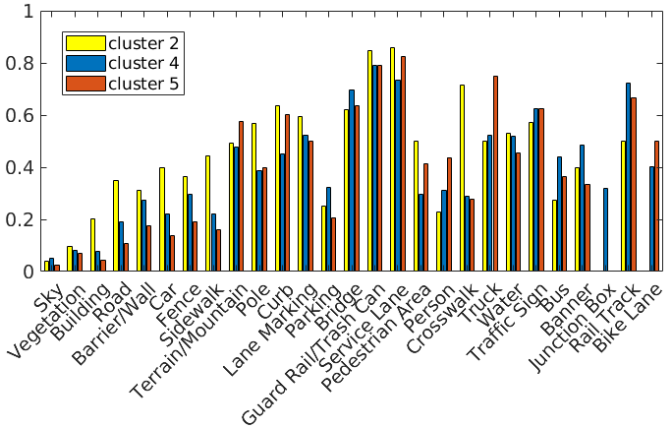

Fig. 7: How often each semantic class appears in losing images while not appearing in the winning image. In general, all clusters follow a similar profile in which classes appear more often on losing images only.

users, while the remaining clusters are small and of less active users.

We treat the generated clusters as possessing all of the comparisons made by the users that were grouped in it. Thus, when picking a random comparison from our dataset, we can analyze any event concerning the following random variables:

- S — semantic class;
- U — user (or cluster);
- W — winning image;
- L — losing image.

For instance, $P(\neg W, L, S = car | U = u_i)$ describes the proportion that for all comparisons of user u_i, the class *car* appears on the losing image but not on the winning image. Such proportions are frequentist proxies to probabilities.

To analyze which features better separate group preferences, we decided to focus on $P(W, \neg L | C, U)$, how often a semantic class appears on the winning image but not on the losing image for each cluster and for each semantic class, and on $P(\neg W, L | C, U)$, if the class appears on the losing image but not on the winning image. These proportions seek to capture which classes are the most relevant for safety perception by appearing mostly on the winning image only or losing image only. When comparing the different clusters (in Fig. 6 for $P(W, \neg L | C, U)$ and in Fig. 7 for $P(\neg W, L | C, U)$) we can see that all clusters share a seemingly common profile. There are, though, some notorious differences, for instance in *crosswalk* for cluster 2, and at the least common classes for both the smaller clusters, but in general we see that the same classes have higher appearances on winning images (while not being on the losing ones) for all clusters (and vice-versa).

Looking at the clustering results and their profiles, we hypothesize that there might be a general agreement on which semantic classes are perceived as safer for all clusters (making them, in fact, a single one) if all clusters had enough comparisons.

C. Agreement hypothesis

Following our hypothesis, we used all of the comparisons to create a general profile of safety perception $\tilde{c}^\star$, the centroid of the overall dataset. Then, we took the most active member from each cluster and measured its cosine distance to $\tilde{c}^\star$ as we consider a greater sample of comparisons. Fig. 8 shows that all the users' distance to the generic profile decreases towards zero as a higher number of comparisons is considered and at a similar rate, supporting our hypothesis. The multiple clusters earlier defined are indeed groups of users who have not provided enough comparisons for their profile to converge to the general semantic description of perceived safety.

Finally, after establishing empirical evidence of inter-observer agreement, we do a final exploration of the data to try to understand which semantic classes appear more frequently in winning and losing images. Fig. 9 shows that human-related classes, like *Pedestrian Area*, *Crosswalk* and *Person*, are found more frequently in winning images. Classes like *Bridge*, *Service Lane*, *Rail Track* and *Guard Rail* are more common in losing images while not being in winning images and are more associated with motor vehicle presence. Fig. 9 also evidences which classes appear frequently on both images. Thus, such classes are not as relevant for safety perception (*Sky* and *Vegetation*) as the others. This is a paramount conclusion regarding how clean the data is, because images tend to be mostly sky, for example, and the sky is not, in our results, a confounding factor for perceived safety.

V. Conclusions

We presented an agreement analysis using image semantics from pairwise comparisons. We showed that grouping users according to their preferences leads to a seemingly common semantic profile. When analyzing which semantic classes are

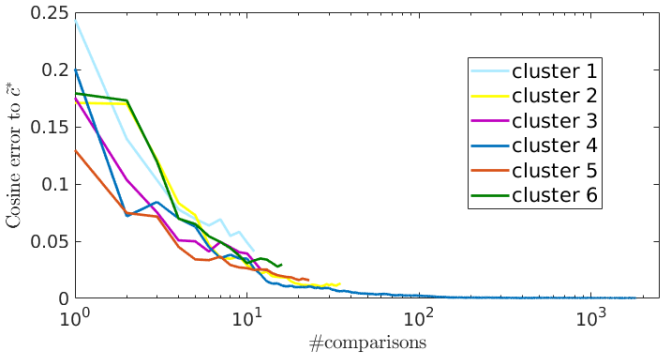

Fig. 8: Evolution of the cosine distance between each user $\tilde{c}_u$ and the general profile $\tilde{c}^{\star}$ from all comparisons, when increasing the number of comparisons considered. All clusters approach the general profile as more comparisons are considered.

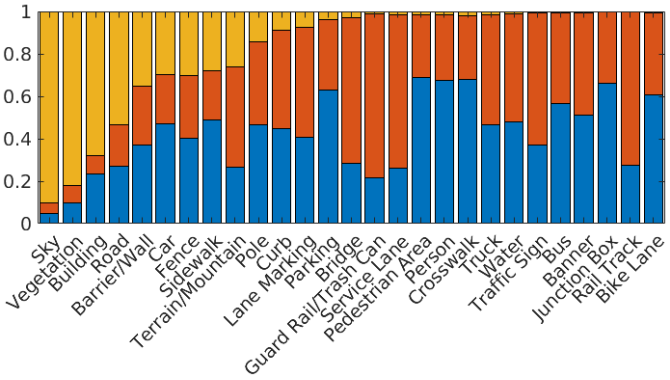

Fig. 9: How often semantic classes appear in the winning image only (blue), the losing image only (orange) or both (yellow). The most common classes also appear most of the times on both images. Human-related classes are more typical of winning images and vehicle-related classes are more typical of losing images.

more common on both winning and losing images only, as well as how often classes appear on both we reinforce our conclusion and observe that human-related semantic classes are perceived as safer while vehicle-related classes are perceived as less safe. We show how clustering is actually a result of each cluster profile having not yet converged to the general safety perception profile due to insufficient data from these clusters. The City-SAFE dataset and source code can be found in https://github.com/gabrielcosta1995/City-SAFE.

Acknowledgements

This research was partially supported by Fundação para a Ciência e a Tecnologia (projects UID/EEA/50009/2013, DSAIPA/AI/0087/2018), and the EC through EU2020-MSCA-ITN BIGMATH (812912) and H2020-ICT-2018-2 AI4EU (825619), and by the Smart City Sense project (LISBOA-01-0247-FEDER-01796).

References

[1] B. Jiang, C. N. S. Mak, H. Zhong, L. Larsen, and C. J. Webster, "From broken windows to perceived routine activities: Examining impacts of environmental interventions on perceived safety of urban alleys," *Frontiers in Psychology*, vol. 9, p. 2450, 2018.

[2] N. Stewart, G. D. Brown, and N. Chater, "Absolute identification by relative judgment." *Psychological review*, vol. 112, no. 4, p. 881, 2005.

[3] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[4] J. Q. Wilson and G. L. Kelling, "Broken Windows: The Police and Neighborhood Safety," *Atlantic Monthly*, 1982.

[5] F. N. Piro, Ø. Nœss, and B. Claussen, "Physical activity among elderly people in a city population: the influence of neighbourhood level violence and self perceived safety," *Journal of Epidemiology & Community Health*, vol. 60, no. 7, pp. 626–632, jul 2006.

[6] A. J. Milam, C. D. M. Furr-Holden, and P. J. Leaf, "Perceived School and Neighborhood Safety, Neighborhood Violence and Academic Achievement in Urban School Children," *The Urban Review*, vol. 42, no. 5, pp. 458–467, dec 2010.

[7] E. L. Glaeser, S. D. Kominers, M. Luca, and N. Naik, "Big Data and Big Cities: the Promises and Limitations of Improved Measures of Urban Life," *Economic Inquiry*, 2018.

[8] X. Li, C. Zhang, W. Li, R. Ricard, Q. Meng, and W. Zhang, "Assessing street-level urban greenery using Google Street View and a modified green view index," *Urban Forestry & Urban Greening*, vol. 14, no. 3, pp. 675–685, 2015.

[9] E. L. Aiden, Y. Wang, J. Krause, J. Deng, L. Fei-Fei, T. Gebru, and D. Chen, "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.

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

[11] P. Salesses, K. Schechtner, and C. A. Hidalgo, "The Collaborative Image of The City: Mapping the Inequality of Urban Perception," *PLoS ONE*, vol. 8, no. 7, p. e68400, jul 2013.

[12] N. Naik, J. Philipoom, R. Raskar, and C. Hidalgo, "Streetscore – predicting the perceived safety of one million streetscapes," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, ser. CVPRW '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 793–799.

[13] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo, "Deep learning the city: Quantifying urban perception at a global scale," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 196–212, 2016.

[14] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-Stuff: Thing and Stuff Classes in Context," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[15] G. Neuhold, T. Ollmann, S. R. Bulo, and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1–16.

[16] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. Lopez, "The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes," 2016.

[18] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *Engineering Analysis with Boundary Elements*, vol. 100, pp. 246–255, feb 2018.

[19] S. R. Bulò, L. Porzi, and P. Kontschieder, "In-Place Activated BatchNorm for Memory-Optimized Training of DNNs," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 5639–5647.

[20] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *CoRR*, vol. abs/1809.00916, 2018.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *CoRR*, 2017.