Discovering Scenic Roads Using Neural Networks

Bogdan State
Stanford University
Stanford, California
bstate@stanford.edu

Abstract

Convolutional Neural Networks (CNNs) coupled with novel large image datasets offer ample hope for building a computational understanding of the relationship between humans and environment. CNNs are also particularly effective at learning complex sets of rules, such as the ones governing aesthetics. Here I present one potential application of computational aesthetics to the detection of scenic byways. To do so I use a dataset of Google Streetview images collected alongside California State Highways, with labels derived from the California Department of Transportation official "scenic byway" designation. A CNN is trained on a dataset of order-10^5 *Streetview images, achieving moderate accuracy against a validation set. The CNN is then integrated in an ensemble model, where training data is augmented with data from two additional sources, the YFCC 100M dataset, as well as a dataset compiled from traditional GIS sources. With suitable refinement the approach shows promise in quantifying the degree to which roadways are aesthetically pleasing, a measure with potential applications in tourism, transportation, and urban planning.*

1. Introduction

The modern human experience of nature takes place on pathways. Whether roads, trails or boardwalks, paths structure the landscape to create distinctive experiences. The *Camino de Santiago*, Appalachian Trail, the Pacific Coast Highway are but a few of the paths famous for their beauty. Most roads are more ... pedestrian[1], but their beauty, or lack thereof is not inconsequential. American commuters spend an average of 52 minutes daily traveling to and from work[4], much of it on heavily-congested roadways, the character of which is likely to lower well-being. Conversely, research has established a link between nature drives [12] and lower levels of stress. These considerations led me to the development of an algorithm for quantifying the scenic nature of roads, combining new advances in data collection (Streetview imagery) and in artificial intelligence (convolutional neural networks). If successful, such an algorithm could be used to optimize the everyday experience of nature and improve well-being.

2. Related Work

Although this work is concerned primarily with the beauty of rural locations, the computational literature it draws on is squarely anchored in the urban environment. Trivially, aesthetic consideration play a prominent role in urban design and planning. A first attempt at the systematic study of "the precise laws and specific effects of the geographic environment ... on the emotions and behavior of individuals" was articulated by Debord as early as 1955 [1]. The term used by Debord, *psychogeography* [14, 1], received renewed interest in the 1990s, primarily in urban studies.

Digital data has shown enormous promise for improving scientists' understanding of human-environment interaction from an aesthetic perspective. By eliciting user ratings of photographed street scenes [13] establish a relationship between image features and the perceived "aesthetic capital" of city locations. [14] expand on this line of work by showing how metadata associated with digital photographs can be used to predict an urban location's perceived degree of beauty.

A revolutionary computational approach to environmental aesthetics has become a possibility with the proliferation of digital geolocated data. New techniques in image recognition promise to help improve the methods and findings of landscape aesthetics, which has already established the existence of a positive relationship between human well-being and the perceived beauty of the natural environment [21, 16].

Of particular interest to the landscape aesthetic problem are approaches using deep neural networks, a method which in recent years has proven extremely effective in multiple areas of machine learning research, and which has seen extraordinary results in the field of image understanding [8].

[1] I hope the reader will forgive my pun.

Satellite imagery has been successfully used in the estimation of development indicators collected in the multinational DHS program [6].

"Streetview" images, captured in a drive-by fashion by a camera placed on a vehicle, have been another particularly promising source of visual information to be used for the understanding of human experience through artificial neural networks. Multiple studies have demonstrated the effectiveness with which Streetview imagery can be accurately geolocated [9, 25, 10, 11]. More recently, Streetview imagery has been used to explore social dimensions of local communities. [3] established how such imagery can be combined with existing geographically-specified data from the American Community Survey to predict the characteristics of urban communities. The model presented here further draws on this approach of combining geospatial metadata and photographic imagery to estimate the degree to which roadways are aesthetically pleasing.

The topic of image scenicness has likewise been of interest to the image recognition community. [23] combine data from the ScenicOrNot dataset, as well as aerial imagery to improve on scenic image classification.

[15] provide the most direct evidence that Streetview imagery can be tagged with scenicness through a transfer learning procedure leveraging tags provided by the GoogleNet model [18] to predict the scenicness of Streetview imagery and recommend scenic routes to drivers.

3. Data and Methods

To understand roadway aesthetics I leveraged a publicly-available source of labels, the list of scenic state highways made available as an ArcGIS shapefile by the California Department of Transportation[2]. I used the *spatialEco* R package [2] to obtain a sample of points spaced 100 meters apart alongside all California state highways. Spatial sampling code is provided in the *streetviewSampler* R package, released at `github.com/bogdanstate/streetview-sampler`. A sampled point was considered to be "scenic" if covered by the California DoT scenic byways shapefile and non-scenic otherwise.

This label-generation procedure makes a very coarse simplifying assumption, namely that all points on a scenic byways are aesthetically pleasing, while all points on a non-scenic state highway are not. Despite the inherent limitations of this assumption, which translate into both noisier model training and lower maximal accuracy, the initial results appear sufficiently promising to justify this source of labels.

Google Streetview images (at resolution 640x640) were obtained for every sample point, through structured API calls.[3] About 177,000 images were obtained through this technique and were used in model training in a 17:1:2 training, validation, and test split. A convolutional neural network was trained on the dataset, with images scaled to to 256x256 resolutions. The network used the following architecture, implemented using the PyTorch library:

1. 2d convolutional layer, kernel size 3, depth 10, and stride 1.
2. Leaky ReLu activation layer ($p = 0.01$).
3. 2d batch normalization layer. [5]
4. 2d max-pooling layer, kernel size 7 and stride 2.
5. 2d convolutional layer, kernel size 3, depth 3, stride 1 and dilation=3.
6. 2d dropout layer (p=0.5). [17]
7. 2d max-pool layer, kernel size 7 and stride 2.
8. ReLu activation layer.
9. 2d max-pool layer, kernel size 7 and stride 2.
10. Linear layer w/ 128 output neurons.
11. Linear layer w/ 2 output neurons.

The network thus specified was implemented using starter code available in the PyTorch Github Examples [4]. An exponential learning rate scheduler was implemented, following the model provided in the PyTorch Transfer Learning Tutorial.[5] The network was trained using the Adam stochastic optimization algorithm [7], with a cross-entropy loss function. Model training was performed on an NVIDIA 1080-Ti GPU.

The best accuracy achieved on the validation accuracy was 0.6173, after training for 11 epochs, time at which whereas the training set recorded an accuracy of .5888. The small, albeit consistent, training-validation discrepancy is likely due to the use of an aggressive dropout layer which removes half the weights after the 2nd convolutional layer.

[2] Data is available at `http://www.dot.ca.gov/hq/tsip/gis/datalibrary/`

[3] I used the scripts available at `https://github.com/robolyst/streetview`.

[4] Available at `https://github.com/jcjohnson/pytorch-examples`.

[5] Available at `https://github.com/pytorch/tutorials/blob/master/beginner_source/transfer_learning_tutorial.py`.

4. Results

4.1. Visual Analysis

Despite the low accuracy achieved in training the simplistic model presented before, results show good face validity. Batches of 16 images, labeled as most and least scenic by the model are shown in Figures 2 and 1. Images identified as least scenic appear to be pictures of flat landscapes, likely from California's inland Central Valley, a region known for industrialized agriculture. Images identified as most scenic present uphill climbs, vegetation and narrower roadways.

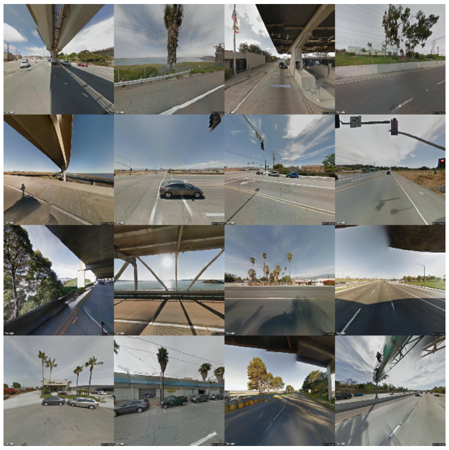

Figure 1. Least Scenic Images

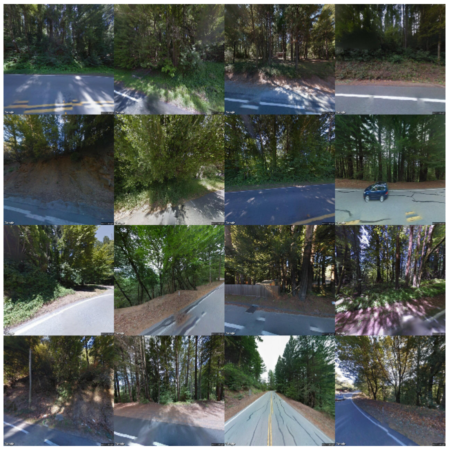

Figure 2. Most Scenic Images

The images ranked as most scenic by the model appear to concur at least partially with existing research into the aesthetic perception of landscapes. In a 1986 synthesis, Ulrich [21] summarizes the established characteristics of pleasing landscapes:

"1. complexity, or the number of independently perceived elements in the scene.

"2. the complexity is structured to establish a focal point, or other kind of patterning is also present.

"3. a moderate to high level of depth that is clearly defined.

"4. the ground surface has even or uniform length textures that are relatively smooth, and the observer judges that the surface is favorable to movement.

"5. a deflected or curving sightline is present, conveying a sense that new landscape information lies immediately beyond the observer's visual bounds.

"6. judged threat is negligible or absent

Of these criteria, Streetview images rated as scenic appear distinct in their complexity, or criterion (1). They present higher contrast between roadway, vegetation and sky. There is no focal point to speak of besides the roadway, but the vegetation does provide texture. Indeed, the model seems to have converged on the mottled shadows of leaves as a characteristic feature of scenic roadways. Similarly, the front-facing images rated as scenic show a higher level of depth, and reveal a curving sightline, a result of the hilly landscape in which the scenic roads are situated. In contrast, non-scenic images are suggestive of flat urban landscapes with very large roads.

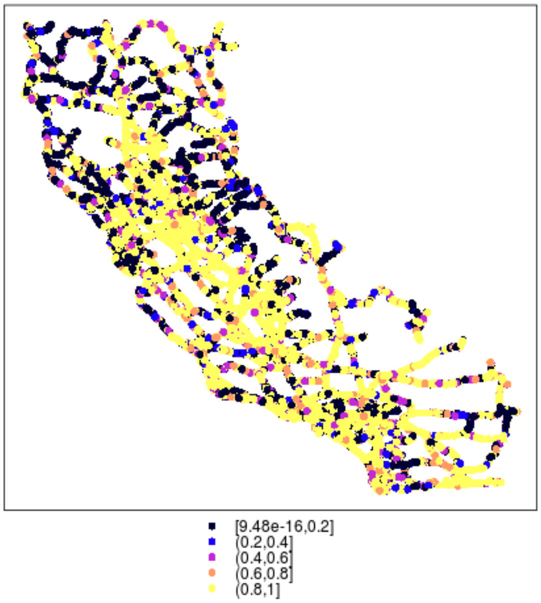

Figure 3. Predictions on Sampled Images

Predictions on sampled images are shown in Figure 3, ranging from yellow (least scenic) to black (most scenic).

The model correctly identifies known scenic roads in the northern part of the state, but very conspicuously misses California's coastal highways, traditionally known for their scenicness.

4.2. Data Augmentation

Some limitations to the image-recognition approach have become apparent in the previous section. While the model correctly learns that most scenic roads are in mountainous, forested areas, it does less well at developing a notion of coastal scenic roads. A reasonable solution meant to address these shortcomings involves data augmentation with information about the geographic context, as suggested by [19].

4.2.1 Baseline GIS model

Given that the visual analysis revealed that altitude, rurality, as well as coastal locations are likely to play a role in determining the baseline scenicness of a location, a model was developed to predict baseline scenicness using a number of simple but powerful GIS features[6]:

- Elevation (based on DEM30 data).
- Distance to coast.
- Distance to nearest urban area.
- Distance to nearest wild or scenic river.
- Distances to nearest national, state, and county parks.
- Distance to nearest lake.
- Distance to nearest major stream.

The baseline GIS model was trained in PyTorch using the same dataset, cross entropy loss function and optimization strategy used for the image recognition model presented before. The GIS model featured a fully-connected architecture, more specifically a linear layer with 10 input features and 20 output features, followed by a hyperbolic tangent activation and a linear layer with 20 input features and 2 outputs. The baseline GIS model achieved an accuracy of .5709 against the validation set after 5 epochs of training.

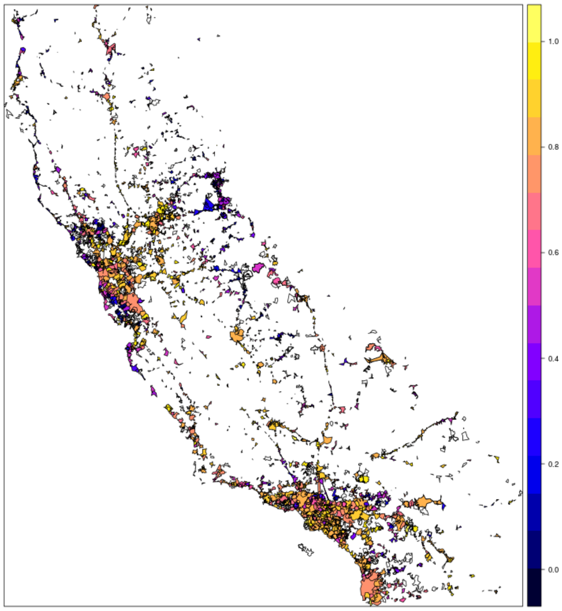

Figure 4. Average Flickr WOEID polygon scenicness. (Lower values translate to higher scenicness.)

4.2.2 Flickr YFCC100M model

Another potential source of information regarding the scenicness of a particular location comes from social media. Given the wealth of geotagged information available on social media platforms, it is quite plausible that the scenicness of a location may be inferred from geotagged posts. To this end I use data from the Flickr YFCC100M dataset [20]. This dataset provides a set of "deep tags," complex image features detected using a model trained to predict Flickr photo captions given image features. The dataset has been used successfully by other teams to improve the accuracy of Google Streetview geolocation [19]. Similar approaches using photo captions (or in this case, features derived from photo captions) have been used successfully to predict the beauty of locations [24, 13, 14].

About half of the YFCC100M data also contains unique geographic identifiers, in the form of WOEIDs (Where on Earth IDentifier), and 1.5m of these examples are located within the boundaries of California. It is possible to map arbitrary geographic coordinates to WOEIDs using the Flickr Shapefiles 2.0 dataset[7]. Thus we can compute an average scenicness per WOEID polygon, shown in Figure 4.

The results in Figure 4 are suggestive of a further improvement. Given that each image in the Flickr 100M dataset is embedded using a 1570-dimensional tag-space, we can compute an average embedding at the WOEID level, which can be mapped to Streetview images. Thus, each

[6] DEM data was obtained from the USGS National Elevation Dataset, urban areas were obtained from the 2015 TIGER/Line shapefiles, while all other features were downloaded from the GIS Clearinghouse at the California Department of Fish and Wildlife, https://www.wildlife.ca.gov/Data/GIS/Clearinghouse.

[7] Released at `http://code.flickr.net/2011/01/08/flickr-shapefiles-public-dataset-2-0/`

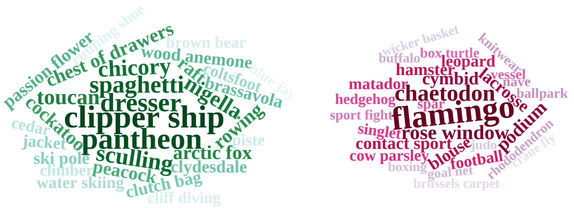

Figure 5. Scenic and non-Scenic Flickr 100m tags.

image receives a tag typically associated with Flickr photos taken in the proximity. The most discriminative such tags w/r to scenicness are shown in Figure **??**. The tags are quite revealing: scenic tags correspond to flowers (clipper ship, nigella, chicory, passion flower, coltsfoot, brassavola), outdoor sports (ski pole, climber, water skiing, jacket, cliff diving, running shoe, rowing), or to animals (toucan, arctic fox, cockatoo, brown bear, blue jay). Non-scenic tags correspond to lawn decor (flamingo, cow parsely, rhododendron), and indoor or stadium sports (spar, contact sport, football, goal net, boxing, sport fight, lacrosse, ballpark, judo, podium).

With this model setup, it is possible to predict whether a Streetview image is located along a scenic or non-scenic California state highway given the average Flickr tag embedding of its vicinity. About 125k of the 177k training images are covered by such Flickr vicinities and can be used in training. A very simple linear model, with only 2 linear layers, a batch normalization layer and a linear layer with 1570 inputs was trained and achieved an accuracy of 64.51% after 2 epochs of training.

4.3. Ensemble model

The three models presented so far underscore the fuzziness of the "scenic" concept, which can be triangulated using data from sources as diverse as Google Streetview, Flickr, and traditional GIS systems. To strengthen inferences regarding this concept an ensemble model can allow different models to build on each other's strengths. To set up this model I use the pre-trained Streetview, GIS and Flickr models. They each contribute equally to the predictions (for the Flickr model, only where data is available). Additionally, the loss function for the ensemble model introduces a disagreement penalty in the form of the KL divergence between any pair of two models making a prediction.

4.4. Test Set

A test set comprised of 10% randomly-selected examples from the collected data was held out from model training. The held-out test set achieved 61.4% accuracy. An ROC curve against the held-out set is shown in Figure 6. The ensemble model achieved 0.66 AUC.

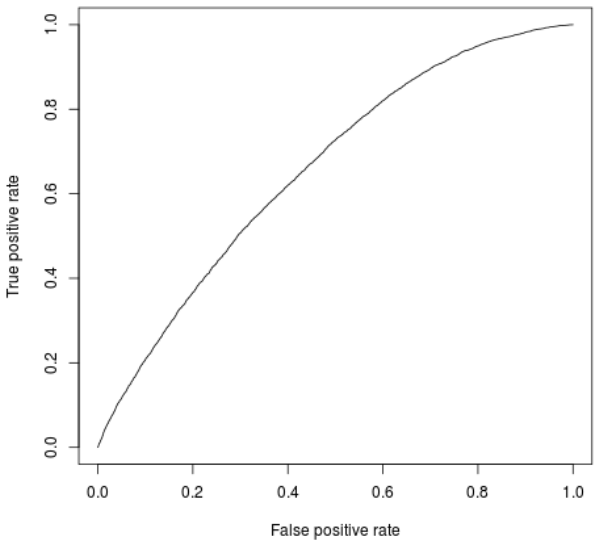

Figure 6. ROC curve for ensemble model (AUC=0.66)

It is important to re-emphasize an earlier point made in the Data section. The intent of this paper is to show how GIS data can be used as a source of labels for the scenicness classification task. To leverage official scenicness designations as a source of labels, a very strong assumption is made, namely that all points along a scenic route are truly scenic, whereas all other points among all (eligible) routes designated as non-scenic are truly non-scenic. This assumption arguably holds for a sufficient number of points to be useful from a statistical perspective. Sizeably more points along scenic routes are truly scenic than points along non-scenic routes. But this assumption is by no means always true. Truck stops and quarries – not exactly pleasing road features – do exist along routes designated as scenic, while areas of natural beauty are also found alongside roads lacking the non-scenic designation. Thus it is reasonable to believe that the maximal accuracy achieved by the model is going to be less than 100%, since the training data does not represent a real "ground truth" with respect to aesthetic pleasingness, but rather a noisy version thereof.

4.5. Discussion

Geolocated digital data promise to radically change the ways we understand the relationship between humans and environment. The method presented here shows how expert knowledge, distilled into painstakingly-developed datasets can be generalized to new previously unexplored datasets. Scenic state highways follow an official definition in California, established through a cumbersome political process. Pretty roads, however, can be anywhere, and we can bootstrap their discovery through image recognition.

4.6. Out-of-sample prediction

To illustrate a potential application of scenicness prediction (also explored by [15]) I sampled points every 30 meters along all county-maintained roads in Santa Cruz County[8]. This location was chosen for its varied landscape – coastal, mountainous and urban, which make it a good candidate for out-of-sample generalization of the model. An additional set of sample points was obtained for primary and secondary roads in the San Francisco Bay Area from the US Census Bureau's TIGER database.[9] The previously-developed model was used to make predictions on the "scenicness" of roads in Santa Cruz County and environs.

Results are computed using the Streetview-only model and are shown in Figure 7. The results are roughly consistent with local knowledge of the area. Scenic road scenes (blue points) predominate in the Santa Cruz mountains, and are sparse in the urbanized areas of Santa Cruz and Watsonville, colored red to indicate "unscenic" road scenes. The mostly-suburban roadscape of the San Jose urban area, shown to the north, is likewise judged to be low in aesthetic pleasingness.

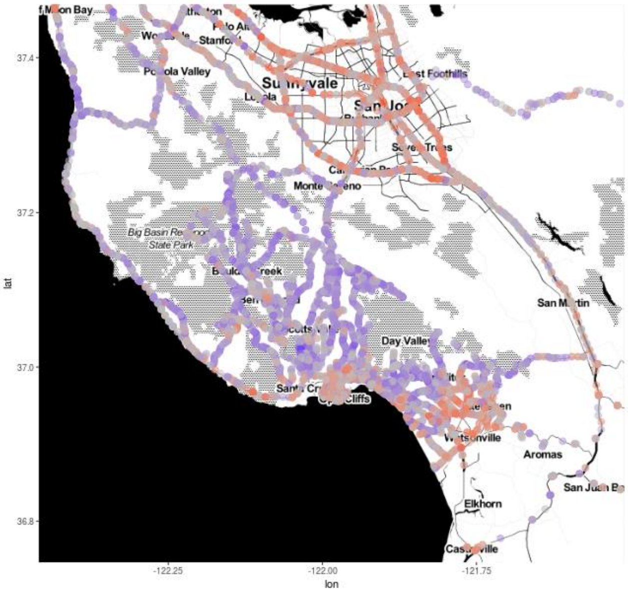

Figure 7. Predictions for Santa Cruz County and environs

A clear omission visible in the initial model concerns the supposed lack of scenicness of coastal regions. Highway 1, which skirts along the Pacific coast is shown to be low in "scenicness," despite being both a designated Scenic State Highway, and a destination for many weekend road trips. This arguable error is likely due to the model learning that scenic landscapes are primarily forested, whereas coastal landscapes often appear without tree cover.

This omission hints at the usefulness of data augmentation improvements for a convolutional model to become maximally useful at a task such as the detection of scenicness, which requires both local information given by a streetview image, as well as a more global context. Specifying a deeper neural network that can learn more complex relations beyond the presence of forested landscapes would be essential to improving the model's performance. Likewise, acquiring more image data can be useful, although this strategy is limited by API constraints in acuiring image data.

As the experiments performed in this paper suggested, a promising avenue involves data augmentation using geocoded tags. The YFCC100M dataset [20] contains about 50m geocoded photographs as well as their caption metadata. Especially when taken in rural settings such captions are likely to be informative with respect to landscape aesthetics. Understanding how to integrate them into the model is an important step to ensuring its usefulness. GIS features, such as altitude, distance to nearest body of water) likewise capture important information about scenicness.

Finally, deeper aesthetics may lie in sequences of contiguous streetview images which may be better addressed by more sophisticated architectures such as pointer networks [22].

5. Conclusion

The algorithm presented here shows the promise of artificial intelligence for the better understanding of both the natural environment, and of the human relationship with nature. With some refinement, the algorithm could be used to provide a map of all the scenic roads in the United States, and potentially asses the environmental aesthetics of any location on Earth where Streetview imagery is collected. This algorithm could be used not only to improve everyday experience, but could also provide a means to monitor changes in the natural environment, inasmuch as they appear on the road.

References

[1] G. Debord. Introduction to a critique of urban geography. *Critical Geographies A Collection of Readings*, 1955.

[2] J. Evans. spatialeco: an r package for spatial analysis and modeling. *R package version 0.1-1. http://cran. r-project. org/package= spatialEco*, 2015.

[3] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of the us. *arXiv preprint arXiv:1702.06683*, 2017.

[8] Available at http://www.co.santa-cruz.ca.us/Departments/GeographicInformationSystems(GIS).aspx

[9] Available at https://catalog.data.gov/dataset/tiger-line-shapefile-2013-state-%2D-california-primary-and-secondary-roads-state-based-shapefile

[4] C. Ingraham. The astonishing human potential wasted on commutes. *Washington Post*, February 25 2016.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[6] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

[7] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] M. Kroepfl, Y. Wexler, and E. Ofek. Efficiently locating photographs in many panoramas. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 119–128. ACM, 2010.

[10] X. Li, M. Larson, and A. Hanjalic. Geo-visual ranking for location prediction of social images. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 81–88. ACM, 2013.

[11] H. Liu, T. Mei, J. Luo, H. Li, and S. Li. Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 9–18. ACM, 2012.

[12] R. Parsons, L. G. Tassinary, R. S. Ulrich, M. R. Hebl, and M. Grossman-Alexander. The view from the road: Implications for stress recovery and immunization. *Journal of environmental psychology*, 18(2):113–140, 1998.

[13] D. Quercia, N. K. O'Hare, and H. Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 945–955. ACM, 2014.

[14] D. Quercia, R. Schifanella, and L. M. Aiello. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 116–125. ACM, 2014.

[15] N. Runge, P. Samsonov, D. Degraen, and J. Schöning. No more autobahn!: Scenic route generation using googles street view. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 147–151. ACM, 2016.

[16] C. I. Seresinhe, T. Preis, and H. S. Moat. Quantifying the impact of scenic environments on health. *Scientific reports*, 5:16899, 2015.

[17] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[19] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev. Improving image classification with location context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1008–1016, 2015.

[20] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016.

[21] R. S. Ulrich. Human responses to vegetation and landscapes. *Landscape and urban planning*, 13:29–44, 1986.

[22] O. Vinyals, M. Fortunato, and N. Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015.

[23] S. Workman, R. Souvenir, and N. Jacobs. Quantifying and predicting image scenicness. *arXiv preprint arXiv:1612.03142*, 2016.

[24] L. Xie and S. Newsam. Im2map: deriving maps from georeferenced community contributed photo collections. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 29–34. ACM, 2011.

[25] A. Zamir and M. Shah. Accurate image localization based on google maps street view. *Computer Vision–ECCV 2010*, pages 255–268, 2010.