Treepedia 2.0: Applying Deep Learning for Large-scale Quantification of Urban Tree Cover

Bill Yang Cai*, Xiaojiang Li†, Ian Seiferling‡, and Carlo Ratti§
Senseable City Lab
Massachusetts Institute of Technology
Cambridge, MA, USA
Email: *billcai@mit.edu, †xiaojian@mit.edu, ‡ianseifs@mit.edu, §ratti@mit.edu*

Abstract—Recent advances in deep learning have made it possible to quantify urban metrics at fine resolution, and over large extents using street-level images. Here, we focus on measuring urban tree cover using Google Street View (GSV) images. First, we provide a small-scale labelled validation dataset and propose standard metrics to compare the performance of automated estimations of street tree cover using GSV. We apply state-of-the-art deep learning models, and compare their performance to a previously established benchmark of an unsupervised method. Our training procedure for deep learning models is novel; we utilize the abundance of openly available and similarly labelled street-level image datasets to pre-train our model. We then perform additional training on a small training dataset consisting of GSV images. We find that deep learning models significantly outperform the unsupervised benchmark method. Our semantic segmentation model increased mean intersection-over-union (IoU) from 44.10% to 60.42% relative to the unsupervised method and our end-to-end model decreased Mean Absolute Error from 10.04% to 4.67%. We also employ a recently developed method called gradient-weighted class activation map (Grad-CAM) to interpret the features learned by the end-to-end model. This technique confirms that the end-to-end model has accurately learned to identify tree cover area as key features for predicting percentage tree cover. Our paper provides an example of applying advanced deep learning techniques on a large-scale, geo-tagged and image-based dataset to efficiently estimate important urban metrics. The results demonstrate that deep learning models are highly accurate, can be interpretable, and can also be efficient in terms of data-labelling effort and computational resources.

Keywords-urban greenery; deep learning; computer vision;

I. INTRODUCTION

Recent advances in deep learning, particularly in deep convolutional neural networks (DCNN), provide researchers with a powerful tool to learn complex representations of unstructured data such as images. Since DCNNs beat previous benchmarks from shallow networks in competitions such as ImageNet, computer vision benchmark competitions have been dominated by researchers using deep learning methods [1]. Along with the development of deep learning, the provision of large-scale datasets has greatly catalyzed deep learning research by providing the data they require to learn complex data representations, and establishing benchmark test datasets and standard evaluation metrics to compare performance across different methods [2].

A. Traditional approaches to quantify urban tree cover

Urban greenery, and in particular trees, attract substantial interest in both academic research and urban planning since they provide wide-ranging services in cities such as carbon sequestration and oxygen production [3], and heat island effect mitigation [4]. Furthermore, the perception of urban greenery has a significant influence on the visual appeal of streets [5], such that urban greenery programs have generally received support from local residents [6].
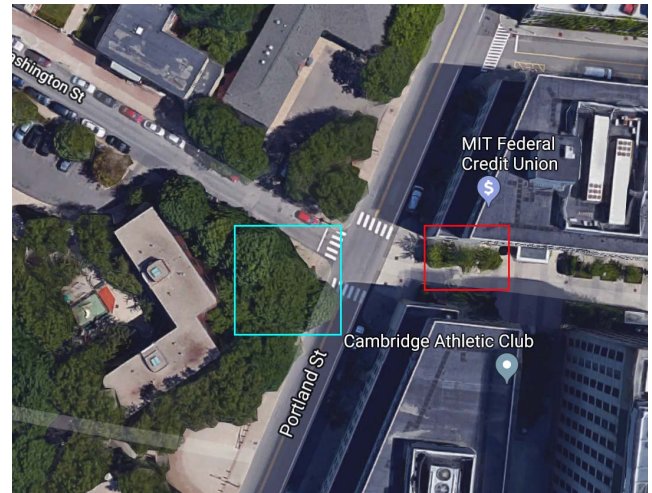

Figure 1. Satellite image of junction between Portland and Washington St in Cambridge, MA. Trees highlighted in red have less dense canopies than tress highlighted in blue.

Traditional approaches to map vegetation cover mapping have relied on either overhead view remote sensing images or photographs, and surveyed data obtained via on-ground fieldwork. Yang et al [7] argues that methods based on overhead view images are limited in their ability to represent the profile view of street vegetation that people experience on the streets. We provide a comparison between overhead and street view images (Fig. 1 and Fig. 2 respectively). The

IEEE computer society

overhead view image does not show substantial differentiation between the spread and density of tree canopy along the pedestrian walkway on Washington St (red box in Fig. 1) and the spread and density of tree canopy at the opposite end of the road junction (blue box in Fig. 1). The street level images clearly portray differences between the tree canopy density of these two groups of trees (red and blue boxes, respectively, in Fig. 2). On the other hand, fieldwork images are limited in their coverage and are costly to acquire at citywide scales.
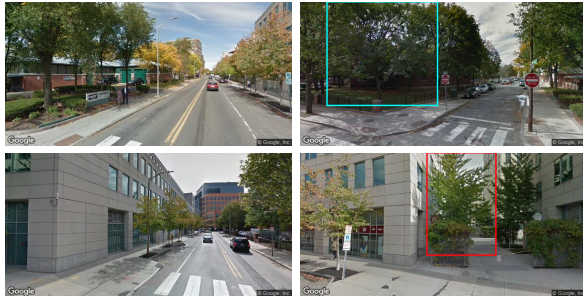

Figure 2. Google Street View images of identical Portland and Washington St junction taken at different headings. Trees highlighted in red have less dense canopies than tress highlighted in blue.

The advent of large-scale and open-source street-level image datasets such as Google Street View (GSV) and Baidu Maps street view service provide researchers a new tool for assessing urban vegetation cover [8]. In addition to their extensive coverage across most of North America, Europe and East Asia excluding China, GSV images are collected and processed to obtain high-resolution, precisely standardized and located images of streetscapes [9]. Furthermore, the recent explosion of research in autonomous vehicle technology has resulted in an abundance of large-scale labelled street-level image datasets such as the Cityscapes dataset [10] and Mapillary Vistas. Thus, GSV represents a comprehensive image dataset of the urban landscape that meets the training requirements of deep learning models.

To quantify urban tree cover, Yang et al [7] and Li. et al [8] utilize the Green View Index (GVI) metric. The GVI metric measures street-level urban tree cover cover by averaging the quantity of identified tree canopy in street-level images taken at different headings and pitches from the same geographical location. To identify tree canopy cover in these images, Li. et al [8] uses an unsupervised mean shift image segmentation algorithm [11] to measure the number of green pixels as a percentage of total pixels in a GSV image. While the benchmark unsupervised method does not require training data, segmentation results from this method are influenced by the presence of shadows and the illumination condition in the GSV images. Green features other than vegetation are also misidentified as urban tree canopies with this method. Seiferling et al. [12] first applied supervised

learning to quantify vegetation cover in GSV images, but their method involved pre-computing image features. To address these shortcomings, we make use of recent advances in supervised DCNN methods by training and testing two models. We train a DCNN semantic segmentation model to produce a densely segmented image. and we also train a DCNN end-to-end model that directly estimates GVI in GSV images.

To validate accuracy in estimating GVI, Li. et al [8] used 33 labelled GSV images and calculated the Pearson's correlation coefficient between GVI from labelled images and model estimates. In this paper, we use a larger test set of 100 GSV images randomly selected across cities in our study area. We propose two additional metrics to validate prediction accuracy of our models: (1) Mean Intersection-over-Union (IoU) is a standard metric used in the computer vision field and captures the accuracy of the location of assigned "green vegetation" pixel labels. (2) Mean Absolute Error calculates the mean of absolute differences in labelled GVI and estimated GVI across images in the test GSV dataset. These two metrics provide a direct measure of accuracy in locating urban tree cover and measuring GVI respectively, while Pearson's correlation coefficient only provides information on the strength of co-movement between labelled GVI and estimated GVI.

B. Deep learning in quantifying urban canopy cover

The use of DCNN in urban ecology and geographic information science (GIS) is in its early stages and relies mostly on overhead, often satellite, imagery. A recent survey by Cheng et al.'s of the use of DCNN in recent remote sensing literature [13] concluded that datasets of non-overhead images, especially of geo-tagged and profile-perspective social media photos provide a larger-scale and more visually diverse collection of images for deep learning algorithms as compared to imagery collected from satellites and airborne surveys.

Wegner et al [14] is an example of recent work that employed DCNN models to street-level images in the urban ecology field. The authors developed a sophisticated method that uses DCNN models on street-level and overhead images to accurately locate and classify tree species. While the information obtained about urban trees from Wegner et al's method is rich in terms of mapping individual trees [14], our DCNN method of estimating the GVI metric produces a much faster aggregate measure of tree cover without extensive training data. Therefore, our method is more easily applied at the city-scale with limited computational resources. Moreover, aggregate measure of tree cover along streets, in and of itself, provides rich and useful information to a variety of stakeholders.

Early applications of DCNN in urban ecology and geospatial sciences have mainly focused on the accuracy of DCNN models. To date, there have been little to no DCNN and com-

50

puter vision applications in urban ecology or GIS that have sought to understand the features learned and identified by DCNN models. As an active and important area of research amongst computer vision scientists, tools to interpret and "peer into" black-box DCNN models are important to visually validate trained models, ascertain that trained models do not contain significant biases, and convince researchers from fields outside of computer vision about the decision process of black-box DCNN models. Therefore, for DCNN models that we employ, we use a visually interpretable method called gradient-weighted class activation map (Grad-CAM) [15] to understand the learned features extracted by our trained DCNN models.

II. METHODOLOGY

A. Obtaining GSV training-validation-test dataset

We choose Cambridge (Massachusetts, USA), Johannesburg (South Africa), Oslo (Norway), São Paulo (Brazil), and Singapore (Singapore) as cities included in our training and test sets. The 5 cities chosen to for this paper are located in differing climatic regions, have various tree species assemblages and forms, and have varying levels of green cover according to the GVI scores estimated by the benchmark *Treepedia* unsupervised segmentation method. This diversity amongst cities should imply model generalization if strong predictive results are found. We use OpenStreetMap street network extracts of the 5 sample cities, and sample points 1 km apart on road networks. We then query the Google Maps API to determine the availability of GSV taken at sampled points. From each of the 5 cities, we randomly select 100 available GSV images each to form a training-validation-test set. We then divide the 500 image dataset into a 100 image test set, 320 image training set and a 80 image validation set. We produce manual labels by carefully tracing all vertical vegetation in the images for all 500 images (Fig. 3).

Figure 3. Top: GSV image in Singapore and associated vegetation labels. Bottom: GSV image in São Paulo and associated vegetation labels.

B. Augmenting training data

We augment our model training by first using the high-quality, pixel-labelled Cityscapes dataset to initially train our DCNN model. The Cityscapes dataset is currently one of

the largest, most extensive and most richly annotated image dataset of city streetscapes to date [10]. The finely labelled subset of the Cityscapes dataset contains over 5000 images taken from vehicle-mounted cameras placed in vehicles and sampled across numerous German cities. We also utilize the class labels for the Cityscapes dataset, which contain a class label for vertical vegetation. We convert the Cityscapes dataset by collapsing the original multi-class labels into binary labels for vegetation and non-vegetation. By first training our models on the larger Cityscapes dataset, we increase our training dataset with the aim of increasing our model performance.

Figure 4. Two sample images from Cityscapes dataset and their associated vegetation labels

C. Evaluation metrics

We propose two evaluation metrics to compare tree cover estimation: mean IoU for measuring the accuracy of the location of labelled vegetation labels, and Mean Absolute Error for measuring the accuracy in estimating overall GVI. The definitions of the two metrics are provided:

1) Mean IoU:

n = number of images in test set

TP_i = true positive predicted vegetation labels for image i

FP_i = false positive predicted vegetation labels for image i

FN_i = false negative predicted vegetation labels for image i

$$IoU_i = \frac{TP_i}{TP_i + FP_i + FN_i}, \quad \bar{IoU} = \frac{1}{n} \cdot \sum_{i=1}^{n} IoU_i$$

2) Mean Absolute Error:

n = number of images in test set

M = number of pixels in a single image

V_i = binary variable for pixel i, 1 indicates actual vegetation, 0 indicates actual non-vegetation

$\hat{V}_i$ = binary variable for pixel i, 1 indicates predicted vegetation, 0 indicates predicted non-vegetation

$$\delta_j = \frac{1}{M} \sum_{i}^{M} I(\hat{V}_i == 1) - I(V_i == 1), \quad \bar{\delta} = \frac{1}{n} \cdot \sum_{j=1}^{n} |\delta_j|$$

Model	Method	Training and Calibration	Prediction Output
Benchmark *Treepedia* unsupervised segmentation	Mean shift unsupervised segmentation [11]	No training, calibrated by Li. et al [8] on other GSV images	Fully pixel-segmented GSV image
DCNN semantic segmentation	Pyramid scene parsing network (PSPNet) [16] with 65,818,363 parameters	Pre-trained on full Cityscapes dataset, then trained on modified Cityscapes dataset, and finally on our labelled GSV dataset	Fully pixel-segmented GSV image
DCNN end-to-end	Deep residual network (ResNet) [17] with 28,138,601 parameters	Pre-trained on ImageNet dataset, then trained on modified Cityscapes dataset, and finally on our labelled GSV dataset	Single real-valued GVI number between 0 and 1

Table I

SUMMARY OF METHODS, TRAINING AND CALIBRATION TECHNIQUES AND PREDICTION OUTPUT OF MODELS

D. Models

1) Benchmark Treepedia unsupervised segmentation: We use Li. et al's original method as a benchmark metric to compare our performance against. Li. et al uses a mean shift segmentation algorithm that identifies "patches" in images with similar color characteristics [11]. Following the same calibrated threshold that the original *Treepedia* project used, the benchmark unsupervised segmentation method identifies "green" patches and labels them as vegetation pixels. Using this method we can see that green but non-tree objects, such as grass or green-colored objects like green shipping containers, can be misclassified (top image of Fig. 5).

2) DCNN semantic segmentation: Semantic segmentation is a supervised learning problem. Therefore, unlike Li. et al's unsupervised approach, each pixel or patch of pixels is assigned a label of semantic meaning learned by the DCNN from prior training. We adopt Zhao. et al's [16] Pyramid Scene Parsing Network (PSPNet)'s architecture for this purpose. PSPNet won the 2016 ImageNet Scene Parsing Challenge and is currently regarded as state-of-the-art for semantic segmentation. We first use pre-trained weights from the original PSPNet trained on the original Cityscapes datasets with its 19 class labels. We then remove the 19 sigmoid units in the last layer and replaced them with 2 sigmoid units. We then pre-train the network again on the aforementioned transformed Cityscapes dataset with only binary labels for vegetation a and non-vegetation. Finally, we train the network on the small labelled GSV dataset itself. Qualitatively, we see that the DCNN semantic segmentation model avoids misclassifying objects as vegetation (top image of Fig. 6).

3) DCNN end-to-end learning: As an alternative approach to the pixel-wise segmentation of vegetation as described above to ultimately compute GVI, we directly estimate GVI with a DCNN model. To conduct end-to-end direct learning of a single GVI value for each image, we adapt He et al's [17] deep residual network (ResNet) architecture. Since He et al [17], other network architectures including PSPNet have widely adopted residual connections to allow for effective feature learning through a high number

Actual GVI: 21.9%, Estimated GVI: 40.6%
Absolute Error: 18.7%, IoU: 32.4%

Actual GVI: 25.0%, Estimated GVI: 3.07%
Absolute Error: 21.9%, IoU: 7.36%

Figure 5. Example results from segmentation algorithm from benchmark unsupervised segmentation method. Highlighted areas are patches labelled as vegetation with this method.

of convolutional layers. Using a 50 layered ResNet as the base architecture, we add 3 more layers of dense connections at the end, with a final layer consisting of a single sigmoid unit. We first initialize the network with weights that have been pretrained on the ImageNet dataset. We then pre-train the modified ResNet with the transformed Cityscapes dataset and associated true GVI labels, before training on the small labelled GSV dataset. It is worth noting that our DCNN end-to-end model has 28,138,601 parameters, as compared to our DCNN semantic segmentation model, which has 65,818,363 parameters. At testing time, this difference implies that

52

Actual GVI: 21.9%, Estimated GVI: 22.1%,
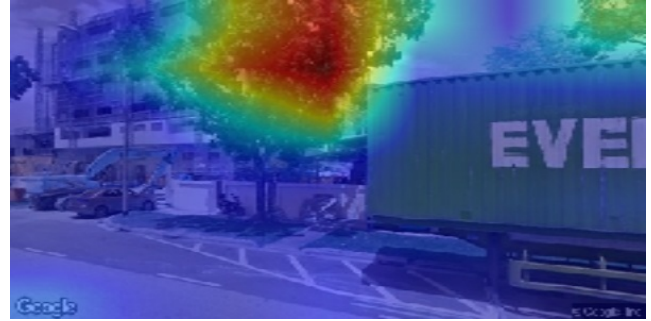Absolute Error: 0.20%, IoU: 89.8%

Actual GVI: 21.9%, Estimated GVI: 27.5%,
Absolute Error: 5.55%

Actual GVI: 25.0%, Estimated GVI: 5.22%,
Absolute Error: 19.78%, IoU: 20.4%

Actual GVI: 25.0%, Estimated GVI: 21.63%,
Absolute Error: 3.37%

Figure 6. Example results from DCNN semantic segmentation algorithm adapted from PSPNet. Highlighted areas are patches labelled as vegetation with this method.
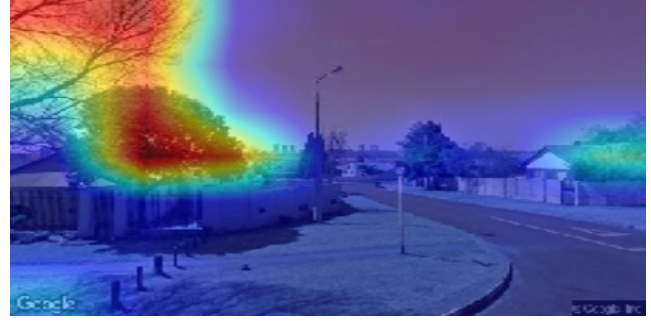
Figure 7. Example results from applying Grad-CAM on our trained DCNN end-to-end models. Areas closer to red have a more positive contribution to a higher prediction of GVI than the contribution of areas closer to blue.

DCNN end-to-end model would outperform DCNN semantic segmentation as the reduced model complexity translates to an increased number of GSV images per evaluation batch.

E. Using Grad-CAM to visualize the DCNN end-to-end model

The lack of an intermediate image segmentation mask makes it difficult to interpret the learning process of the DCNN end-to-end model. It becomes particularly difficult to confirm or communicate the features that the DCNN end-to-end model has learned in order to estimate GVI. Selvaraju et al [15] developed Grad-CAM in order to produce visual explanations for features learned in particular convolutional layers. Through partial linearization and taking the rectified linear function of linear combination of feature maps and neural importance weights, we can visualize areas of the original input image that contribute positively to the prediction of a particular class. Selvaraju et al found that DCNN models may learn biases that may not be completely generalizable and may introduce unwanted prejudice, thus visualizing the active features can help identify those biases. Hence, we apply Grad-CAM to our DCNN end-to-end model to understand whether our model has learned generalizable features.

III. RESULTS

A. Model Performance

We find that DCNN models outperform the original *Treepedia* unsupervised segmentation model significantly (Table II). DCNN semantic segmentation achieve a higher Mean IoU and a lower Mean Absolute Error, while DCNN end-to-end achieve the lowest Mean Absolute Error out of all three models. Comparing the model results using the correlation coefficient between model predicted values and the labelled test set values, we find that the DCNN models significantly outperform the original unsupervised segmentation model (Table II). Furthermore, we report the 5%-95% of the GVI estimation error to indicate the spread and centrality of the estimation errors. For this metric, we want the spread to be centered on 0 and we want the spread to be small. In this regard, we see that DCNN models outperform the original unsupervised model significantly in prediction accuracy.

B. Interpretability of DCNN end-to-end model

The Grad-CAM method applied to our trained DCNN end-to-end model and test images suggests that the DCNN end-to-end model learned to identify vertical vegetation in GSV images, without any noticeable systematic biases (Fig. 8).

53

Model	Mean IoU (%)	Mean Absolute Error (%)	Pearson's Correlation Coefficient	5%-95% of GVI Estimation Error (%)
Benchmark *Treepedia* unsupervised segmentation	44.7	10.1	0.708	-26.6, 18.7
DCNN semantic segmentation	61.3	7.83	0.830	-20.0, 12,37
DCNN end-to-end	NA	4.67	0.939	-10.9, 7.97

Table II

ACCURACY COMPARISON BETWEEN MODELS. DCNN END-TO-END MODEL DOES NOT PROVIDE AN INTERMEDIATE IMAGE SEGMENTATION, HENCE IT IS NOT POSSIBLE TO DIRECTLY EVALUATE ITS MEAN IoU PERFORMANCE.

C. Running time and scalability

Training and testing of the DCNN semantic segmentation and DCNN end-to-end learning models were conducted on a system equipped with a single NVIDIA 1080Ti GPU with 11GB of memory. The training of both models took about 48 hours of training time each.

A critical issue that affects the quantification of urban tree cover is the scalability of our GVI algorithm. Cities with complex and large street networks such as London and New York have up to 1 million GSV images. Therefore, fast evaluation time is essential towards the feasibility of large-scale quantification of a city's tree cover.

As a test of scalability, we tested the speed of our models in processing 100, 1000 and 10000 GSV images, and report the running time in terms of wall-clock time. To speed up computation, we parallelize the Python-based code for the benchmark *Treepedia* unsupervised segmentation, and we increase the evaluation batch size to the maximum that our single GPU system can handle. The ResNet-based DCNN end-to-end model evaluates 10000 GSV images under 40 seconds, which is a 0.1% of the evaluation time needed for the benchmark *Treepedia* unsupervised segmentation (Table 2). The DCNN end-to-end model is also significantly faster than the DCNN semantic segmentation model, due to the difference in model complexity resulting in a larger evaluation batches for the DCNN end-to-end model. Using linear extrapolation, our fastest model, the DCNN end-to-end model, can calculate the GVI of street-level images in a large city like London with 1 million images in slightly over one hour of processing time.

IV. CONCLUSION

We demonstrate the potential of deep learning in achieving better performance as measured by GVI correlation, Mean IoU and Mean Absolute Error, as compared to the original unsupervised segmentation models used in the *Treepedia* [18] project and website. Notably, our DCNN end-to-end model achieves both significantly better model performance and evaluation speed.

Besides accuracy, we also introduce the use of Grad-CAM to inspect features learned by our DCNN end-to-end model. We find that our trained DCNN end-to-end model accurately learned to pick out vertical vegetation from GSV images without any noticeable biases.

By using street-level GSV images, we are able to utilize the abundance of open-source manually labelled street-level image datasets to heavily pre-train our DCNN models. After pre-training, we tune our DCNN models to GSV images by training on a small-scale dataset. This saves significant resources that would normally be associated with building a large custom training dataset. Nonetheless, we produced a carefully curated and labelled training set specific to GSV images and vertical vegetation which will be released and hopefully built upon. It provides future researchers a benchmark validation set that they could use to compare their model performance with our models.

With an accurate, interpretable and efficient GVI estimation technique, city planners can apply our models to understand the spatial distribution of urban greenery, and prioritize public tree maintenance and planting efforts. Our technique can also be confidently used by researchers for a wide variety of research topics that relies on accurate estimations of urban greenery or tree cover including environmental inequality [19] and urban heat island models [20].

The collection of street-level images from Google, Baidu and Tencent represent one of the few geographically extensive, and informationally rich datasets available to researchers. By using DCNN models, researchers can utilize vast street-level image datasets to learn and understand complex environmental and socio-economic phenomena, in addition to urban tree cover. From setting metrics to measuring evaluation time, we demonstrate the process of rigorously applying deep learning to street-level images. Especially for researchers from fields including GIS and urban studies who lack computational resources, our paper serves as an example of an effective application of deep learning tools in a computationally efficient manner.

ACKNOWLEDGMENT

54

Model	Running Time for 100 images (seconds)	Running Time for 1000 images (seconds)	Running Time for 10000 images (seconds)
Benchmark *Treepedia* unsupervised segmentation	39.6	381	3665
DCNN semantic segmentation	20.2	207	2064
DCNN end-to-end	1.81	4.06	38.9

Table III
RUNNING TIME OF MODELS IN COMPUTING GVI

REFERENCES

[1] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27 – 48, 2016, recent Developments on Deep Big Vision.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[3] D. Nowak, R. Hoehn, and D. Crane, "Oxygen production by urban trees in the united states," *Arboriculture and Urban Forestry*, vol. 33, no. 3, pp. 220–226, 2007.

[4] R. Lafortezza, G. Carrus, G. Sanesi, and C. Davies, "Benefits and well-being perceived by people visiting green spaces in periods of heat stress," *Urban Forestry & Urban Greening*, vol. 8, no. 2, pp. 97–108, 2009.

[5] L. Bain, B. Gray, and D. Rodgers, *Living streets: Strategies for crafting public space*. John Wiley & Sons, 2012.

[6] C. Jim and W. Y. Chen, "Perception and attitude of residents toward urban green spaces in guangzhou (china)," *Environmental management*, vol. 38, no. 3, pp. 338–349, 2006.

[7] J. Yang, L. Zhao, J. Mcbride, and P. Gong, "Can you see green? assessing the visibility of urban forests in cities," *Landscape and Urban Planning*, vol. 91, no. 2, pp. 97–104, 2009.

[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] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google street view: Capturing the world at street level," *Computer*, vol. 43, no. 6, pp. 32–38, 2010.

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[11] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[12] I. Seiferling, N. Naik, C. Ratti, and R. Proulx, "Green streets-quantifying and mapping urban trees with street-level imagery and computer vision," *Landscape and Urban Planning*, vol. 165, pp. 93–101, 2017.

[13] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: benchmark and state of the art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[14] J. D. Wegner, S. Branson, D. Hall, K. Schindler, and P. Perona, "Cataloging public objects using aerial and street-level images-urban trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 6014–6023.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *arXiv preprint arXiv:1610.02391*, 2016.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] Treepedia: http://senseable.mit.edu/treepedia. Senseable City Lab.

[19] X. Li, C. Zhang, W. Li, Y. A. Kuzovkina, and D. Weiner, "Who lives in greener neighborhoods? the distribution of street greenery and its association with residents' socioeconomic conditions in hartford, connecticut, usa," *Urban Forestry & Urban Greening*, vol. 14, no. 4, pp. 751–759, 2015.

[20] P. A. Mirzaei and F. Haghighat, "Approaches to study urban heat island–abilities and limitations," *Building and environment*, vol. 45, no. 10, pp. 2192–2201, 2010.
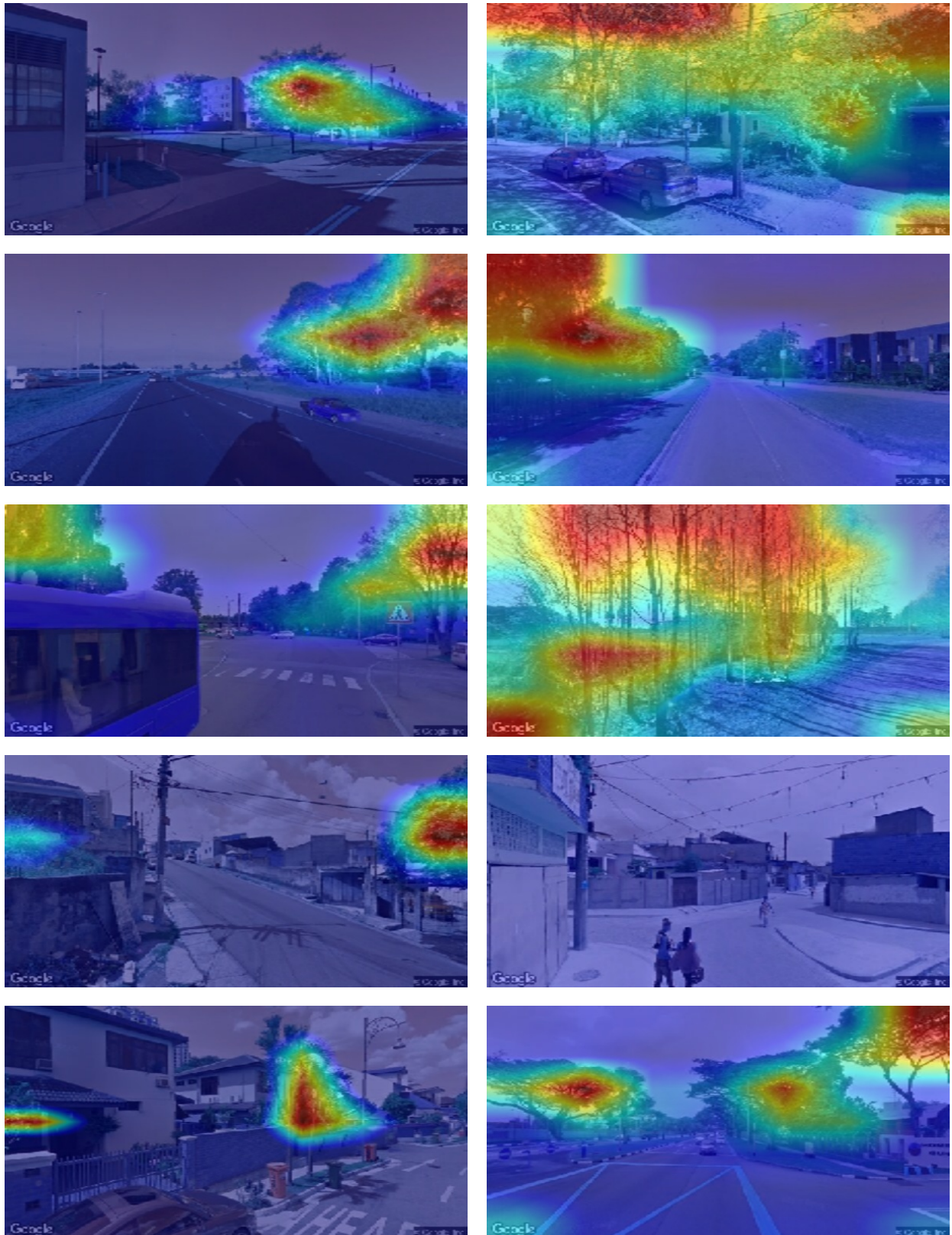
55

Figure 8.   Results from applying Grad-CAM on our trained DCNN end-to-end modele to understand features learned in the last convolutional layer. 2 images from Cambridge, Johannesburg, Oslo, Sao Paulo, and Singapore are shown in order. Areas closer to red have a more positive contribution to a higher prediction of GVI than the contribution of areas closer to blue.