FaceLift: A transparent deep learning framework recreating the urban spaces people intuitively love

Sagar Joglekar
King's College London, Department of Informatics
sagar.joglekar@kcl.ac.uk

Daniele Quercia
Nokia Bell Labs, Cambridge, UK
quercia@cantab.net

Miriam Redi
Nokia Bell Labs, Cambridge, UK
miriam.redi@gmail.com

Luca Aiello
Nokia Bell Labs, Cambridge, UK
lajello@gmail.com

Tobias Kauer
Nokia Bell Labs, Cambridge, UK
tobias.kauer@fh-potsdam.de

Nishanth Sastry
King's College London, Department of Informatics
nishanth.sastry@kcl.ac.uk

ABSTRACT

In computer vision, deep learning techniques have recently been used to predict whether urban scenes are likely to be considered beautiful, and it turns out that these techniques do so quite accurately. To support urban interventions, however, one needs to go beyond *predicting* beauty and tackle the challenge of *recreating* beauty. Unfortunately, deep learning techniques have not been designed with that challenge in mind. Given their "black-box nature", they cannot even explain why a scene has been predicted to be beautiful. To partly fix that, we propose a deep learning framework (which we name FaceLift) that is able to both *beautify* existing Google Street views and *explain* which urban elements make those transformed scenes beautiful. To quantitatively evaluate our framework, we cannot resort to any existing metric (as the research problem at hand has never been faced before) and need to formulate new ones. These new metrics should ideally capture the presence (or absence) of elements that make urban spaces great. Upon a review of the urban planning literature, we identify four main metrics: walkability, green, openness, and visual complexity. For all the four metrics, the beautified scenes meet the expectations set by the literature on what great spaces tend to be made of. These results suggest that, in the future, as our framework's components are further researched and become better and more sophisticated, it is not hard to imagine technologies that will be able to accurately and efficiently support architects and planners in the design of the spaces we intuitively love.

ACM Reference format:
Sagar Joglekar, Daniele Quercia, Miriam Redi, Luca Aiello, Tobias Kauer, and Nishanth Sastry. 2017. FaceLift: A transparent deep learning framework recreating the urban spaces people intuitively love. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 8 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

Conference'17, Washington, DC, USA

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Whether a street is considered beautiful is subjective, yet research has shown that there are specific urban elements that are universally considered beautiful: from greenery, to small streets, to memorable spaces [1, 14, 15]. These elements are those that contribute to the creation of what the urban sociologist Jane Jacobs called 'urban vitality' [8].

Given that, it comes as no surprise that computer vision techniques can automatically analyze pictures of urban scenes and accurately determine the extent to which these scenes are, *on average*, considered beautiful. Deep learning has greatly contributed to increase these techniques' accuracy [5].

However, urban planners and architects are interested in urban interventions and, as such, they wish to go beyond technologies that are only able to predict beauty scores. They often called for technologies that would make easier to recreate beauty in urban design [3]. Deep learning is not fit for purpose. It is not meant to recreate beautiful scenes, not least because it cannot provide provide any explanation on why a scene is beautiful.

To partly fix that, we propose a deep learning framework (which we name FaceLift) that is able to both *generate* a beautiful scene (or, better, *beautify* an existing scene) and *explain* why that scene is beautiful. In so doing, we make two main contributions:

- We propose a deep learning framework that is able to learn whether Google Street views are beautiful or not and that, based on that training, is able to both *beautify* existing views and *explain* which urban elements make these views beautiful (Section 3).
- We quantitatively evaluate whether the framework is able to actually produce beautified scenes (Section 4). We do so by proposing a family of four urban design metrics that we have formulated based on a through review of the literature in urban planning. For all these four metrics, the framework passes with flying colors: with minimal interventions, beautified scenes are twice as walkable as the original scenes, for example. Also, after building an interactive tool with "FaceLifted" scenes in Boston and presenting it to twenty experts in architecture, we found that the majority of them agreed on three main areas of our work's impact: decision making, participatory urbanism, and promotion of restorative spaces among the general public.

We conclude by pointing out some limitations that might well guide future work (Section 5).

2 RELATED WORK

Previous work has focused on collecting ground truth data about how people perceive urban spaces, on predicting urban qualities from visual data, and on generating synthetic images that enhance a given quality (e.g., beauty).

Ground truth of urban perceptions. So far the most detailed studies of perceptions of urban environments and their visual appearance have relied on personal interviews and observation of city streets: for example, some researchers relied on annotations of video recordings by experts [16], while others have used participant ratings of simulated (rather than existing) street scenes [10]. The web has recently been used to survey a large number of individuals. Place Pulse is a website that asks a series of binary perception questions (such as 'Which place looks safer [between the two]?') across a large number of geo-tagged images [15]. In a similar way, Quercia *et al.* collected pairwise judgments about the extent to which urban scenes are considered quiet, beautiful and happy [14]. They were then able to analyze the scenes together with their ratings using image-processing tools, and found that the amount of greenery in any given scene was associated with all three attributes and that cars and fortress-like buildings were associated with sadness. Taken all together, their results pointed in the same direction: urban elements that hinder social interactions were undesirable, while elements that increase interactions were the ones that should be integrated by urban planners to retrofit cities for greater happiness.

Deep learning and the city. Computer vision techniques have increasingly become more sophisticated. Deep learning techniques, in particular, have been recently used to accurately predict urban beauty [5, 17], urban change [11], and even crime [4].

Generative models. Deep learning has recently been used not only to analyze existing images but also to generate new ones. Ngyuen *et al.* [12] used generative networks to create a natural-looking image that maximizes a specific neuron. In theory, the resulting image is the one that "best activates" the neuron under consideration (e.g., that associated with urban beauty). In practice, it is still a synthetic template that needs further processing to look realistic.

To sum up, a lot of work has gone into collecting ground truth data about how people tend to perceive urban spaces, and into building accurate predictions models of urban qualities. However, little work has gone into models that generate realistic urban scenes and that offer human-interpretable explanations of what they generate.

3 FACELIFT FRAMEWORK

The goal of FaceLift is to take as input a geo-located urban scene and give as output its transformed (beautified) version. To that end, it performs five steps: 1) curating urban scenes; 2) training a beauty classifier; 3) generating a synthetic beautified scene; 4) returning a realistic beautified scene; and 5) identifying the urban elements characterizing the beautified scene.

Symbol	Meaning
I_i	Original urban scene
Y	Set of annotation classes for urban scenes (e.g., beautiful, ugly)
y_i	Annotation class in Y (e.g., beautiful)
$\hat{I}_j$	Template scene (synthetic image)
I'	Target Image
C	Beauty Classifier

Table 1: Notations

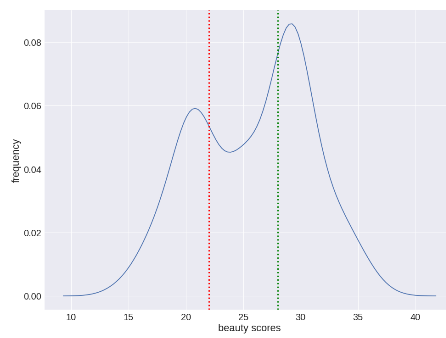


Figure 1: Frequency distribution of beauty scores. The red and green lines represent the thresholds below and above which images are considered ugly and beautiful. Conservatively, images in between are discarded.

Step 1 Curating Urban Scenes

To begin with, we need highly curated training data with labels reflecting urban beauty. We start with the Place Pulse dataset that contains 100k Google Street Views across 56 cities around the world [5]. These scenes are labeled in terms of whether the corresponding places are likely to be perceived beautiful, depressing, rich, and safe. We focus only on those scenes that are labeled in terms of beauty and that have at least three judgments. This leave us with roughly 20,000 scenes. To transform judgments into beauty scores, we use the TrueSkill algorithm [7], which gives us a way of partitioning the scenes into two sets (Figure 1): one containing beautiful scenes, and the other containing ugly scenes. The resulting set of scenes is too small for training any deep learning module without avoiding over-fitting though. As such, we need to augment such a set. We do so in two ways. First, we feed each scene's location into the Google Streetview API to obtain the snapshots of the same location at different camera angles (i.e., at $\theta \in -30°, -15°, 15°, 30°$). However, the resulting dataset is still too small for robust training. Therefore, again, we feed each scene's location into the Google Streetview API, but now we do so to obtain other scenes at distance $d \in \{10, 20, 40, 60\}$ meters. This will greatly expand our set of scenes, but it might do so at the price of introducing scenes whose beauty scores have little to do with the original scene's. To fix that, we take only the scenes that are *similar* to the original one (we call this way of augmenting "conservative translation"). To compute the similarity between a pair of scenes, we represent the two scenes with visual features derived from the FC7 layer of PlacesNet and compute the similarity between the two corresponding feature vectors [21].

(a) (b)

Figure 2: Two types of augmentation: (a) rotation of the Street Views camera (based on rotation); and (b) exploration of scenes at increasing distances (based on translation).

Augmentation	Accuracy (Percentage)
None	63
Rotation	68
Rotation + Translation	64
Rotation + Conservative Translation	73.5

Table 2: Percentage accuracy for our beauty classifier trained on differently augmented sets of urban scenes.

For all scenes at increasing distance $d \in \{10, 20, 40, 60\}$ meters, we take only those whose similarity scores with the original scene is above a threshold. In a conservative fashion, we choose that threshold to be the median similarity between rotated and original scenes (those of the first augmentation step). To make sure this additional augmentation has not introduced any unwanted noise, we consider two sets of scenes: one containing those that have been taken during this last step (*taken-set*), and the other containing those that have been filtered away (*filtered-set*). Each scene is labeled with PlacesNet [21] and is represented with the five most confident scene labels. The labels are aggregated at set level by computing each label's frequency on the *taken-set* minus that on the *filtered-set* and by then characterizing each label's propensity to be correctly augmented as: $\text{prone}(label) = fr(label, taken\text{-}set) - fr(label, filtered\text{-}set)$. This reflects the extent to which a scene with a given label is prone to be augmented or not. From Figure 3, we find that, as one would expect, scenes that contain highways, fields and bridges can be augmented at increasing distances while still showing resemblances to the original scene; by contrast, scenes that contain gardens, residential neighborhoods , plazas, and skyscrapers cannot be easily augmented, as they are often found in high density parts of the city.

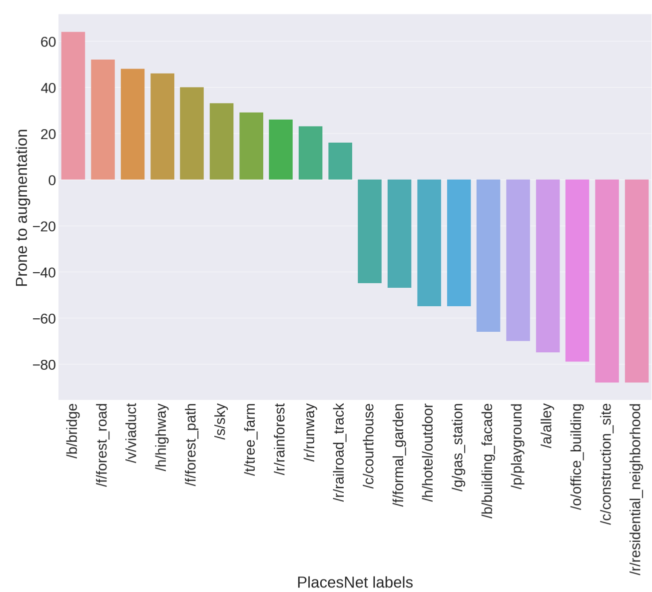


Figure 3: The types of scene that have greater propensity to be correctly augmented with similar scenes at increasing distances.

Step 2 Training a beauty classifier

Having this highly curated set of labeled urban scenes, we are now ready to train classifier C with labels reflecting our beauty assessments. More specifically, we train CaffeNet, a modified version of AlexNet[9, 19]. The training is done on a 70% split of the data, and the testing on the remaining 30%. All this is done on increasingly augmented sets of data. We start from our 20k images and progressively augment them with the snapshots obtained with the 5-angle camera rotations, and then with the exploration of scenes at increasing distance $d \in \{10, 20, 40, 60\}$ meters. The idea behind data augmentation is that accuracy would increase with it. Indeed it does (Table 2): it goes from 63% on the set of original scenes to as much as 73.5% on the set of fully augmented scenes, which is a notable increase in accuracy for such classes of classification tasks.

Step 3 Generating a synthetic beautified scene

Having this trained classifier at hand, we can then build a generator of synthetic beautified scenes. This is a model that, given the two classes ugly y_i and beautiful y_j, transforms any original scene I_i of class y_i (e.g., ugly scene) into template scene $\hat{I_j}$ that maximizes class y_j (e.g., beautified template scene).

More specifically, given an input image I_i known to be of class y_i (e.g., ugly), our technique outputs $\hat{I_j}$, which is a more beautiful version of it (e.g., I_i is morphed towards the average representation of a beautiful scene) while preserving I_i's details. The technique does so using the "Deep Generator Network for Activation Maximization" (*DGN-AM*) [12]. Given an input image I_i, *DGN-AM* iteratively re-calculates the color of I_i's pixels in a way the output image $\hat{I_j}$ both maximizes the activation of neuron y_j (e.g., the "beauty neuron") and looks "photo realistic", which is done by conditioning the maximization to an "image prior". This is equivalent to finding the feature vector f that maximizes the following expression:

$$\hat{I_j} = G(f) : \arg\max_f (C_j(G(f)) - \lambda||f||), \quad (1)$$

where:

- $G(f)$ is the image synthetically generated from the candidate feature vector f;
- $C_j(G(f))$ is the activation value of neuron y_j in the scene classifier C (the value to be maximized);
- λ is a L_2 regularization term.

Here the initialization of f is key. If f were to be initialized with random noise, then the resulting $G(f)$ would be the average representation of category y_j (of, e.g., beauty). Instead, since f is initialized with I_i, then the resulting $G(f)$ is I_i's version "morphed to become more beautiful".

Step 4 Returning a realistic beautified scene

We now have template scene $\hat{I_j}$ (which is a synthetic beautified version of original scene I_i) and need to retrieve a realistic looking version of it. We do so by: *i)* representing each of our original scenes in Step 1 (including $\hat{I_j}$) as a 4096 dimensional feature vector derived from the FC7 layer of a pre-trained deep network [21]; *ii)* computing the similarity (as L_2 Norm) between $\hat{I_j}$'s feature vector and each of the original scene's feature vector; and *iii)* selecting the original scene most similar to $\hat{I_j}$. This results into the selection of the beautified scene I_j.

Step 5 Identifying characterizing urban elements

Since original scene I_i and beautified scene I_j are real scenes and we make sure that they maintain the same structural characteristics (e.g., point of view, layout), we can easily compare them in terms of presence or absence of SegNet's and PlacesNet's labels. That is, we can determine how the original scene and its beautified version differ in terms of urban design elements.

4 EVALUATION

The goal of FaceLift is to transform existing urban scenes into versions that: *i)* people perceive more beautiful; *ii)* contain urban elements typical of great urban spaces; *iii)* are easy to interpret; and *iv)* architects and urban planners find useful. To ascertain whether FaceLift meets that composite goal, we answer the following questions next:

Q1 Do individuals perceive"FaceLifted" scenes to be beautiful?

Q2 Does our framework produce scenes that possess urban elements typical of great spaces?

Q3 Which urban elements are mostly associated with beautiful scenes?

Q4 Do architects and urban planners find FaceLift useful?

Q1 People's perceptions of beautified scenes

To ascertain whether FaceLifted scenes are perceived by individuals as they are supposed to, we run a crowd-sourcing experiment on Amazon Mechanical Turk. We randomly select 200 scenes, 100 beautiful and 100 ugly (taken at the bottom 10 and top 10 percentiles of the Trueskill's score distribution of Figure 1). Our framework then transforms each ugly scene into its beautified version, and each beautiful scene into its corresponding 'uglified'. These scenes are arranged into pairs, each of which contains a beautiful scene and an ugly one. On Mechanical Turk, we only select verified masters for our crowd-sourcing workers (those with an approval rate above 90% during the past 30 days), pay them \$0.1 per task, and ask each of them to choose the beautiful scene for given pairs. We make sure to have at least 3 votes for each scene pair. Overall, our workers end up selecting the scenes that are actually beautiful 77.5% of the times, suggesting that FaceLifted scenes are correctly perceived most of the times.

Q2 Are beautified scenes great urban spaces?

To answer that question, we need to understand what makes a space great. After a careful review of the urban planning literature, we identify four factors [1, 6] (summarized in Table 3): great places mainly tend to be walkable, offer greenery, feel cozy, and be visually rich.

To automatically extract visual cues related to these four factors, we select 500 ugly scenes and 500 beautiful ones at random, transform them into their opposite aesthetic qualities (i.e., ugly ones are beautified, and beautiful ones are 'uglified'), and compare which urban elements related to the four factors distinguish uglified scenes from beautified ones.

We extract labels from each of our 1,000 scenes using two computer vision algorithms. First, using PlacesNet [21], we label each of our scenes according to a classification containing 205 labels (reflecting, for example, landmarks, natural elements), and retain the five labels with highest confidence scores for the scene. Second, using Segnet [2], we label each of our scenes according to a classification containing 12 labels. Segnet is trained on dash-cam images, and the resulting labels are twelve: road, sky, trees, buildings, poles,

Figure 4: Example of "FaceLifting".

Metric	Description
Walkability	Walkable streets increase the social capital of a place, and they appeal to the exploring nature of the human psyche [6, 13, 18].
Green Spaces	The presence of greenery has repeatedly been found to impact people's well being [1]. Under certain conditions, it could also promote social interactions [14]. This suggest that not all greenery has to be considered in the same way though: dense forests or unkempt greens might well have a negative impact [8].
Privacy-Openness	A sense of privacy (as opposed to a sense of openness) impacts a place's perception [6].
Visual Complexity	Visual complexity is a measure of how diverse a urban scene is in terms of design materials, textures, and objects [6].

Table 3: Urban Design Metrics

signage, pedestrians, vehicles, bicycles, pavement, fences, and road markings.

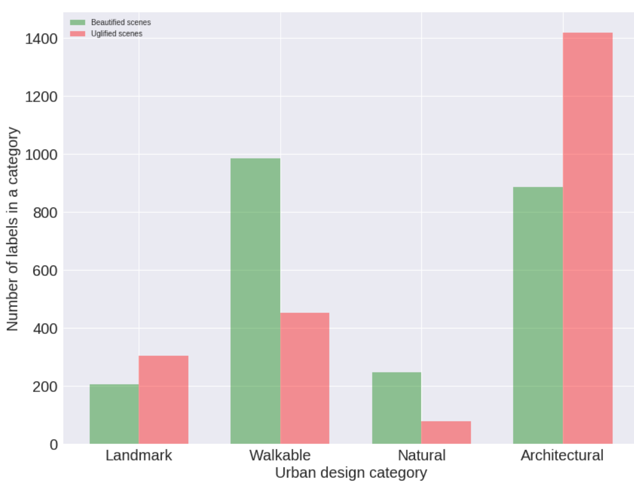


Figure 5: Number of labels in specific urban design categories (on the x-axis) found in beautified scenes as opposed to those found in uglified scenes.

Having these two ways of segmenting scenes, we can now test whether the expectations set by the literature describing metrics of great urban spaces (Table 3) are met in the FaceLifted scenes.

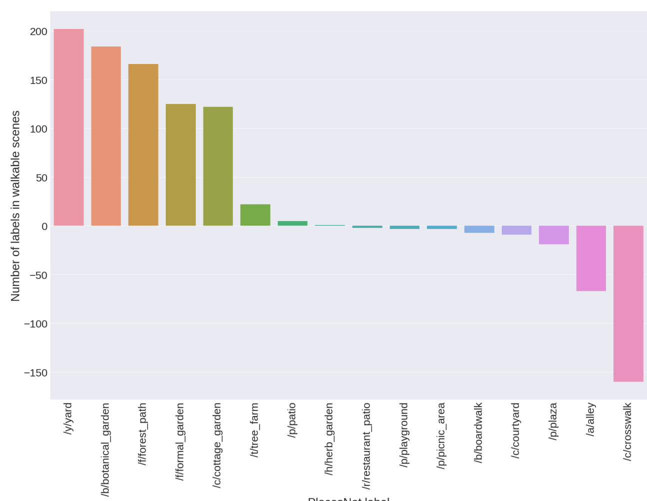


Figure 6: Count of specific walkability-related labels (on the x-axis) found in beautified scenes minus the count of the same labels found in uglified scenes.

H1 Beautified scenes tend to be walkable. We manually select only the labels that are related to walkability. These labels include, for example, *abbey, plaza, courtyard, garden, picnic area, and park*. To test hypothesis *H1*, we count the number of walkability-related labels

Pair of urban elements	β_1	β_2	β_3	Error Rate (Percentage)
Buildings - Trees	-0.032	0.084	0.005	12.7
Sky - Buildings	-0.08	-0.11	0.064	14.4
Roads - Vehicles	-0.015	-0.05	0.023	40.6
Sky - Trees	0.03	0.11	-0.012	12.8
Roads - Trees	0.04	0.10	-0.031	13.5
Roads - Buildings	-0.05	-0.097	0.04	20.2

Table 4: Coefficients of logistic regressions run on one pair of predictors at the time.

found in beautified scenes as opposed to those found in uglified scenes (Figure 5): the former contain twice as many walkability labels than the latter. We then determine which types of scenes are associated with beauty (Figure 6). Unsurprisingly, beautified scenes tend to show gardens, yards, and small paths. By contrast, uglified ones tend to show built environment features such as shop fronts and broad roads.

H2 Beautified scenes tend to offer green spaces. We manually select only the PlacesNet's labels that are related to greenery. These labels include, for example, *fields, pasture, forest, ocean, and beach*. Then, in our 1,000 scenes, to test hypothesis *H2*, we count the number of nature-related labels found in beautified scenes as opposed to those found in uglified scenes (Figure 5): the former contain more than twice as many nature-related labels than the latter. To test this hypothesis further, we compute the fraction of 'tree' pixels (using SegNet's label 'tree') in beautified and uglified scenes, and find that beautification adds 32% of tree pixels, while uglification removes 17% of them.

H3 Beautified scenes tend to feel private and 'cozy'. To test hypothesis *H3*, we count the fraction of pixels that Segnet labeled as 'sky' and show the results in a bin plot in Figure 7a: the x-axis has six bins (each of which represents a given range of sky fraction), and the y-axis shows the percentage of beautified *vs.* uglified scenes that fall into each bin. Beautified scenes tend to be cozier (lower sky presence) than the corresponding original scenes.

H4 Beautified scenes tend to be visually rich. To quantify to which extent scenes are visually rich, we measure their visual complexity [6] as the amount of disorder in terms of distribution of (Segnet) urban elements in the scene:

$$H(X) = -\sum p(i) \log p(i) \tag{2}$$

where i is the i^{th} Segnet's label. The total number of labels is twelve. The higher $H(X)$, the higher the scene's entropy, that is, the higher the scene's complexity. To test hypothesis *H4*, we show the percentage of scenes that fall into a complexity bin (Figure 7b): beautified scenes are of low to medium complexity, while uglified ones are of high complexity.

Q3 Urban elements of beautified scenes

To determine which urban elements are the best predictors of urban beauty and the extent to which they are so, we run a logistic regression, and, to ease interpretation, we do so on one pair of predictors at the time:

$$Pr(beautiful) = logit^{-1}(\alpha + \beta_1 * V_1 + \beta_2 * V_2 + \beta_3 * V_1.V_2) \tag{3}$$

where $V1$ is the fraction of the scene's pixels marked with one Segnet's label, say, "buildings" (over the total number of pixels), and $V2$ is the fraction of the scene's pixels marked with another label, say, "trees". The result consists of three beta coefficients: β_1 reflects $V1$'s contribution in predicting beauty, β_2 reflects $V2$'s contribution, and β_3 is the interaction effect, that is, it reflects the contribution of the dependency of $V1$ and $V2$ in predicting beauty. We run logistic regressions on the five factors that have been found to be most predictive of urban beauty [1, 6, 14], and show the results in Table 4.

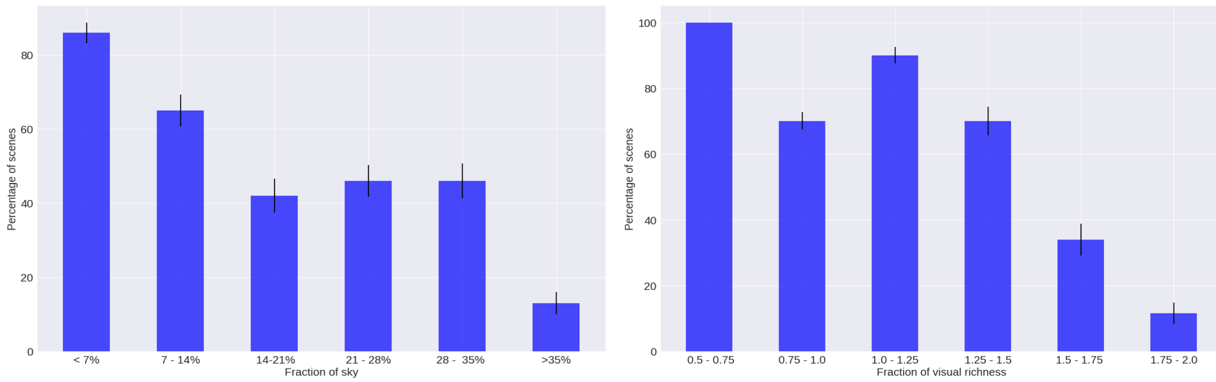
Since we are using logistic regressions, the quantitative interpretation of the beta coefficients is eased by the "divide by 4 rule" [20]: we can take β coefficients and "divide them by 4 to get an upper bound of the predictive difference corresponding to a unit difference" in beauty [20]. For example, take the results in the first row of Table 4. In the model $Pr(beautiful) = logit^{-1}(\alpha - 0.032 \cdot buildings + 0.084 \cdot trees + 0.005 \cdot buildings \cdot trees)$, we can divide - 0.032/4 to get -0.008: a difference of 1 in the fraction of pixels being buildings corresponds to no more than a 0.8% *negative* difference in the probability of the scene being beautiful. In a similar way, a difference of 1 in the fraction of pixels being trees corresponds to no more than a 0.021% *positive* difference in the probability of the scene being beautiful. By considering the remaining results in Table 4, we find that, across all pairwise comparisons, trees is the most positive element associated with beauty, while roads and buildings are the most negative ones. Since these results go in the direction one would expect, one might conclude that the scenes beautified by our framework are in line with previous literature, adding further external validity to our work.

Q4 Do architects and urban planners find it useful?

To ascertain whether practitioners find FaceLift potentially useful, we built an interactive map of the city of Boston in which, for selected points, we showed pairs of urban scenes before/after beautification (Figure 9). We then sent that map along with a survey to 20 experts in architecture, urban planning, and data visualization around the world. The experts had to complete tasks in which they rated FaceLift based on how well it supports decision making, participatory urbanism, and promotion of green spaces among the general public. The results are show in Figure 8: according to our experts, the tool can very probably supports decision making, probably support participatory urbanism, and definitely promote green spaces. These results are qualitatively supported by our experts' comments, which include: "*The maps reveal patterns that might not otherwise be apparent*", "*The tool helps focusing on parameters to identify beauty in the city while exploring it*", and "*The metrics are nice. It made me think more about beautiful places needing a combination of criteria, rather than a high score on one or two dimensions. It made me realize that these criteria are probably spatially correlated*".

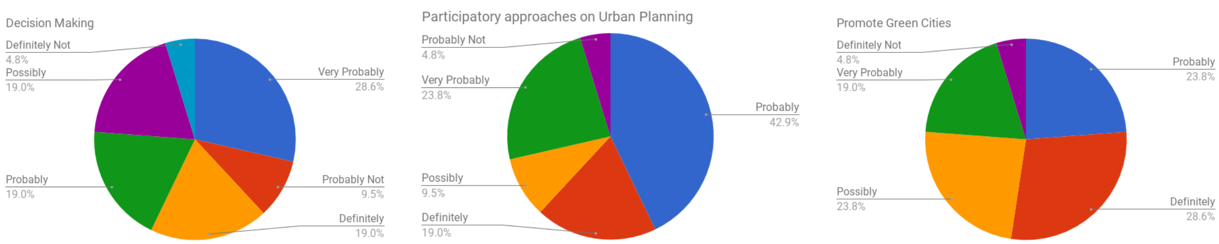
5 CONCLUSION

FaceLift is a transparent framework that beautifies existing urban scenes. This translates into two main technical advancements. First, FaceLift is able to generate realistic scenes as opposed to existing approaches based on Generative Adversarial Networks whose final transformations are quite coarse as they still take the form of abstract



(a) (b)

Figure 7: The percentage of scenes (y-axis): (a) having an increasing presence of sky (on the x-axis); and (b) having an increasing level of visual richness (on the x-axis).



(a) (b) (c)

Figure 8: Urban experts polled about the extent to which an interactive map of "FaceLifted" scenes promotes: (a) decision making; (b) citizen participation in urban planning; and (c) promotion of green cities.

Figure 9: Interactive map of FaceLifted scenes in Boston.

templates. Second, it augments the deep learning black-box with a module that offers explanations on what has been transformed, making that box more transparent.

There are still important limitations though. One is data bias. The framework is as good as its training data, and more work has to go into collecting reliable ground truth of human perceptions. This data should ideally be stratified according to the people's characteristics that impact their perceptions. The other main limitation is that generative models are hard to control, and more work has to go into offering principled ways of fine-tuning the generative process.

Despite these limitations, FaceLift has the potential to support urban interventions in scalable and replicable ways: it can be applied to the scale of an entire city, and that can be replicated in other cities. The advantage of shifting the focus of research away from predictive analytics towards urban interventions is that people could be part of discussions on works of architecture more than they are nowadays. To turn existing spaces into something more beautiful, that will still be the duty of architecture. Yet, with technologies similar to FaceLift more readily integrated in the architecture discussions, the complex job of recreating restorative spaces in an increasingly urbanized world will be greatly simplified. After all, "we delight in complexity to which genius have lent an appearance of simplicity." [3] In the context of future work, that genius is represented by the future technologies that we will contribute to build to deal with the complexity of our cities.

REFERENCES

[1] C. Alexander, S. Ishikawa, and M. Silverstein. 1977. *A Pattern Language: Towns, Buildings, Constructions*. Oxford University Press.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2015. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561* (2015).

[3] A. De Botton. 2008. *The Architecture of Happiness*. Knopf Doubleday Publishing Group.

[4] Marco De Nadai, Radu Laurentiu Vieriu, Gloria Zen, Stefan Dragicevic, Nikhil Naik, Michele Caraviello, Cesar Augusto Hidalgo, Nicu Sebe, and Bruno Lepri. 2016. Are Safer Looking Neighborhoods More Lively?: A Multimodal Investigation into Urban Life. In *Proceedings of the ACM on Multimedia Conference (MM)*.

[5] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. *arXiv preprint arXiv:1608.01769* (2016).

[6] Reid Ewing and Otto Clemente. 2013. *Measuring urban design: Metrics for livable places*. Island Press.

[7] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkillâĎć: a Bayesian skill rating system. In *Advances in neural information processing systems*. 569–576.

[8] J. Jacobs. 1961. *The Death and Life of Great American Cities*. Random House.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[10] Pall Jakob Lindal and Terry Hartig. 2012. Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology* (2012).

[11] Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. 2017. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences* 114, 29 (2017), 7571–7576.

[12] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*. 3387–3395.

[13] Daniele Quercia, Luca Maria Aiello, Rossano Schifanella, and Adam Davies. 2015. The Digital Life of Walkable Streets. In *Proceedings of the 24th ACM Conference on World Wide Web (WWW)*. 875–884.

[14] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. 2014. Aesthetic capital: what makes London look beautiful, quiet, and happy?. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 945–955.

[15] Philip Salesses, Katja Schechtner, and César A Hidalgo. 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS one* 8, 7 (2013), e68400.

[16] Robert J. Sampson and Stephen W. Raudenbush. 2004. Seeing Disorder: Neighborhood Stigma and the Social Construction of Broken Windows. *Social Psychology Quarterly* 67, 4 (2004).

[17] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. 2017. Using deep learning to quantify the beauty of outdoor places. *Royal Society open science* 4, 7 (2017), 170170.

[18] J. Speck. 2012. Walkable City: How Downtown Can Save America, One Step at a Time. In *Farrar, Straus and Giroux*.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[20] Brandon K Vaughn. 2008. Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J. *Journal of Educational Measurement* 45, 1 (2008), 94–97.

[21] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.