Response to the reviews for Royal Society Open Science submission "FaceLift: A transparent deep learning framework to beautify urban scenes"

Sagar Joglekar, Daniele Quercia, Miriam Redi, Luca Aiello, Tobias Kauer, Nishanth Sastry

We would like to express our sincere thanks to the Editors for supporting this process and the reviewers for their very detailed and constructive comments. We have worked to address all their concerns in the revised version of the manuscript.

Summary of reviewer requests

Reviewers made certain important points which we try to address in this revision:

1. *Reviewer 1 expressed concern about the validity of the "Walkability" metric actually quantifying walkability of places*
2. *Reviewer 1 expressed concern about the validity of the "Openness" metric actually quantifying open nature of places. They also suggested trying out Scene Understanding (SUN) attributes from places 205 to address this.*
3. *Reviewer 2 recommended adding definition for urban informatics along with a reference.*
4. *Reviewer 2 recommended discussion the PlacePulse dataset in more detail. " how it was created, and who participated in the curation and assignment of labels. What biases are present in this dataset? Has this been analysed?"*
5. *Reviewer 2: Looking at the representative examples in Table 4, the "mechanism" or trained logic of the algorithm seems quite straightforward, as acknowledged by the authors, e.g. "adding greenery, narrow roads, and pavements." However, I also note that images in rows 3 and 4 shift from a winter scene (trees with no leaves) to a summer scene, or a grey sky to a blue sky. Further, some suggested beautifications shift entire structures such as buildings. These observations could be discussed, and then used to talk about two points: (a) limitations of the approach, and; (b) usefulness of the results (beyond what is covered in Q4).*
6. *Reviewer 2: Section Q4 did not convince me. The Likert scale sought to evaluate how well FaceLift supports decision making. It is not clear to me what is meant by decision making here. Similarly, I do not see how the substitution of an act of human creativity through a deep learning algorithm can be rated as "participatory urbanism" when there is nobody participating other than the machine.*

7. Reviewer 2 recommended adding critical reflections about the framework and have kindly given pointers to do so.

We have thoroughly revised the paper to address all the comments from the reviewers. Below, we provide detailed answers to each of their comments individually.

Requests from Reviewer 1

First of all, is there a full list of all the keywords used for each metric? This would be a useful addition to a Supplementary Information section. For "Walkability", I am concerned that the categories chosen (e.g. plaza, courtyard, park) tend to be places that people go to rest rather than walk. So I am undertrain(sic) that these categories are measuring something akin to walkability of a scene they might be measuring something else entirely. This might still be a good urban design metric, but I am just not convinced it is measuring walkability.

To classify the PlacesNet labels into the Walkable category, we rely on the definitions of Walkable areas from previous work [10], which uses 8 properties of streets to score their walkability: *Road safety, Easy to cross, Sidewalks, Hilliness, Navigation, Safety from crime, Smart and beautiful, Fun and relaxing*. We use these 8 properties as a guidance to classify the PlacesNet labels into the Walkable category. We summarize this categorization in a new Table (see the bottom of this letter and Table 8 in the paper). The elements in the Walkable category serve different functions but they are all either walkable spaces or they are most often situated in walkable areas. We do agree with the Reviewer that walkability acts often as an enabler for other desirable properties of places—like their restorative potential—and therefore it is confounded with them. In the new version of the manuscript, in Section "Q2 Are beautified scenes great urban spaces?," we acknowledge that our walkability measure could correlate to the notion of beauty because it might also act as a proxy for higher-order properties that are enabled by walkability.

> " We manually select only the PlacesNet labels that are related to walkability, using a list walkability properties of streets that have been defined in previous work as a guidance [10]. These labels include, for example, abbey, plaza, courtyard, garden, picnic area, and park (Table 8 contains the exhaustive list) [...] 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. It is worth noting that walkability often acts as an enabler for other desirable properties of places, therefore our walkability measure could correlate to the notion of beauty because it might also act as a proxy for higher-order properties facilitated by walkability. "

As for "Privacy-Openness" I am not convinced by the approach used to measure openness of a scene. Where a scene with tall trees would feel cozy, a scene with tall skyscrapers is likely to feel claustrophobic rather than cozy. So I am not convinced that lower sky presence equates to coziness, as this really depends on context. It is possible to extract Scene UNderstanding (SUN) Scene Attributes from Places205. These

include elements such as far-away horizon, nohorizon, open area; perhaps this could be a helpful way of measuring the "Privacy-Openness" aspect of a scene.

The reason why we count the amount of "sky pixels" to measure openness is because we wanted to characterize this property on a continuous spectrum, which is not possible with other frameworks. However, the reviewer is right in noting that, in general, lack of openness could indicate cozyness as well as visual oppression. To test whether in our data ... we conducted an additional experiment using the Scene UNderstanding (SUN) framework [11], as the Reviewer suggested.

We first extracted scene attributes from all the beautified and uglified scenes using SUN, so that each scene gets assigned a list of attributes. Within each group, we count the number of pictures with a given SUN attribute i. We denote these counts as ϕ_i^{beauty} and ϕ_i^{ugly}. We then calculated the differences of these counts between the two groups: $\delta(phi_i) = \phi_i^{beauty} - \phi_i^{ugly}$. Figure 1 in this letter shows the distribution of these differences. Positive (negative) values indicate that the attribute is more prevalent in the beautified (uglified) group. Scene attributes like *trees, foliage, open area, vegetation, grass* are found in more numbers in beautified scenes. On the contrary, attributes like *man-made, driving, no horizon , far-away horizon, enclosed-area* are found in more numbers in the uglified scenes. This validates the hypothesis that the lack of horizon or enclosed spaces (lack of openness in general) are not conducive to the feeling of beauty.

Also note some small changes:

- *- How is similarity calculated on page 4? Is this cosine similarity?*
- *- It would be useful to add citations to support p.14 "previous literature"*

We used Euclidean distance and we made sure to mention it in the revised version in page 4 (step 4). In page 14, we added 5 citations that cover the "previous literature" we refer to.

Requests from Reviewer 2

The paper specifies and explains key terms such as urban design, deep learning, and generative models, but not urban informatics. Perhaps add a definition and reference.

As suggested by the reviewer, in the introduction we added a definition for "urban informatics" together with a couple of supporting references:

> "Our work contributes to the field of urban informatics, an interdisciplinary area of research that studies practices and experiences across urban contexts and creates new digital tools to improve those experiences [4, 6]. "

It would be useful for the reader to better understand the Place Pulse dataset, how it was created, and who participated in the curation and assignment of labels. What biases are present in this dataset? Has this been analysed? ...

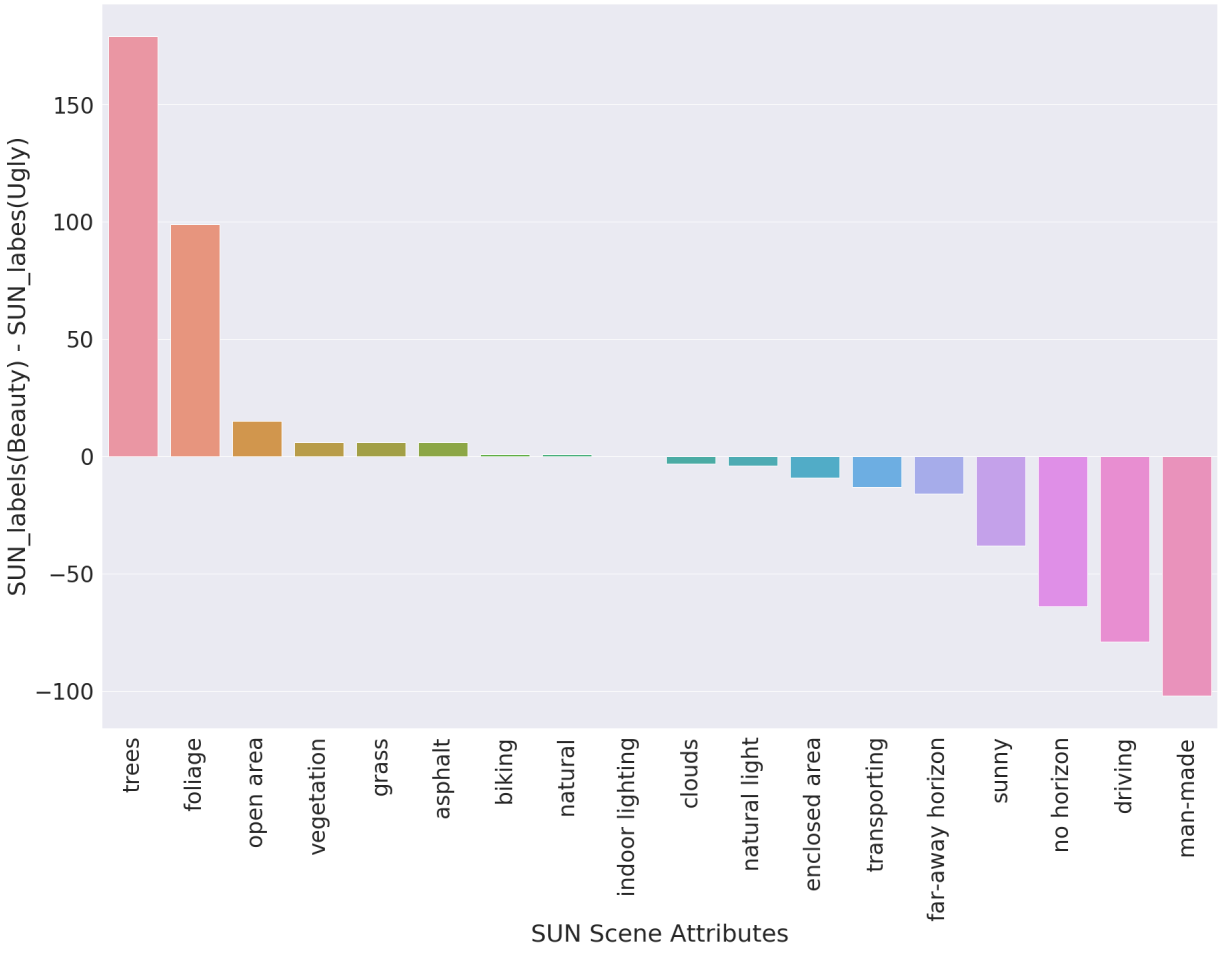

Fig. 1:

We agree with the reviewer that a more self-contained description of the Place Pulse dataset and of its potential biases is in order. We added the following paragraph in the Section "Curating Urban Scenes" (page 4):

> " To begin with, we need highly curated training data with labels reflecting urban beauty. We start with the Place Pulse dataset that contains a set of 110k Google street view images from 56 major cities across 28 countries around the world [3]. The pictures were labeled by volunteers through an ad-hoc crowdsourcing website[1]. Random pairs of images were shown on the website to online volunteers. For each pair, multiple people were asked to select which scene looked more beautiful, safe, lively, boring, wealthy, and depressing. At the time of writing, 1.2 million pairwise comparisons were generated by 82k online volunteers from 162 countries, with a

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

good mix of people residing in both developed and developing countries. To our knowledge, no systematic analysis of the biases of the Place Pulse dataset has been conducted yet. However, it is reasonable to expect that representation biases are minimized by the substantial size of the dataset, the wide variety of places represented, and the diversity of gender, racial, and cultural background of the raters. ”

Looking at the representative examples in Table 4, the "mechanism" or trained logic of the algorithm seems quite straightforward, as acknowledged by the authors, e.g. "adding greenery, narrow roads, and pavements." However, I also note that images in rows 3 and 4 shift from a winter scene (trees with no leaves) to a summer scene, or a grey sky to a blue sky. Further, some suggested beautifications shift entire structures such as buildings. These observations could be discussed, and then used to talk about two points: (a) limitations of the approach, and; (b) usefulness of the results (beyond what is covered in Q4).

We agree with the reviewer that the examples in Table 4 expose some of the limitations of our approach. These can be broadly summarized by saying that generative image models are still hard to control, especially when dealing with complex scenes with several elements. This shortcoming is compounded by the restricted size of training data. We briefly mentioned this limitation in the previous version of the paper; in the new manuscript, we expand on that point:

“ The main limitation is that generative image models are still hard to control, especially when dealing with complex scenes with multiple elements. Some of the beautifications suggested by our tool modify the scene too dramatically (e.g., shifting buildings or broadening roads) to use them as a blueprint for urban interventions. This undesired effect is compounded by the restricted size and potential biases of data that we use both for training and for selecting the scene most similar to the machine-generated image—which might result for example in generating scenes set in seasons or weather conditions that differ from the input image. To address these limitations, more work has to go into offering principled ways of fine-tuning the generative process, as well as into collecting reliable ground truth data on human perceptions. This data should ideally be stratified according to the people's characteristics that impact their perceptions. ”

Even though this limitation partly restricts the capacity of our tool, we still argue for its potential to simplify and democratize the process of creating restorative spaces, as we detail in the next reply.

Section Q4 did not convince me. The Likert scale sought to evaluate how well FaceLift supports decision making. It is not clear to me what is meant by decision making here. Similarly, I do not see how the substitution of an act of human creativity through a deep learning algorithm can be rated as "participatory urbanism" when there is nobody participating other than the machine.

We thank the Reviewer for allowing us to clarify this point, as we realize we could be clearer on the purpose for which FaceLift is intended. We added the following discussion in the conclusions section, hoping that it will serve to clarify the intended use cases for our tool:

“ We conceived FaceLift not as a technology to *replace* the decision making process of planners and architects, but rather as a tool to *support* their work. Facelift could integrate the creative

process of beautification of a city by suggesting imagined versions of what urban spaces could become after applying certain sets of interventions. We do not expect machine-generated scenes to equal the quality of designs done by experts. However, unlike the work of an expert, Facelift is able to generate beautified scenes very fast (in seconds) and at scale (for an entire city), while quickly providing a numerical estimate of how much some urban elements should change to increase beauty. The user study we conducted suggests that these features make it possible to inspire the work of decision makers and to nudge then into considering alternative approaches to urban interventions that might not otherwise be apparent. We believe this source of inspiration could advantage non-experts too, for example by helping residents to imagine a possible future for their cities and motivate citizen action in the deployment micro-interventions. ”

This brings me to suggest the addition of a critical reflection and limitations section. Two examples of points that could be explored here:

(a) The mechanistic/positivist way the algorithm beautifies urban scenes risks becoming a cookie cutter as it does not take into account the full spectrum of authentic ways urban scenes can be activated and then perceived as beautiful. Similarly to how a leafless tree in winter is perceived less beautiful than a lush, leafy tree in summer, there are influences of people, urban policies, placemaking initiatives that impact on the notion of "beauty." Norberg-Schulz (1980) uses a phenomenology approach to describe the "essence" of a place, which is socio-culturally and time-specific. Brand (1997) traces the development of a street scene / building faade over time as it changes through renovations, modifications, and customisations and as a result, perceptions change. In my own work (2017), I reviewed placemaking interventions and explored participatory forms of citymaking.

- *Norberg-Schulz, C. (1980). Genius loci: Towards a phenomenology of architecture. New York, NY: Rizzoli.*
- *Brand, S. (1997). How Buildings Learn: What Happens After Theyre Built (Rev.). London: Phoenix Illustrated.*
- *Foth, M. (2017). Lessons from Urban Guerrilla Placemaking for Smart City Commons. In Proceedings of the 8th International Conference on Communities and Technologies (C&T '17). ACM, New York, NY, USA, 32-35. DOI: https://doi.org/10.1145/3083671.3083707*

(b) The positivist paradigm of urban science has been critiqued for its technocratic worldview, and the FaceLift study would benefit from a critical reflection by the authors that picks up on some of these points, e.g.:

- *Kitchin, R. (2017). Thinking critically about and researching algorithms. Information, Communication and Society, 20(1), 1429. https://doi.org/10.1080/1369118X.2016.1154087*
- *Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. Big Data & Society, 3(2). https://doi.org/10.1177/2053951716665128*
- *Kitchin, R. (2016). The ethics of smart cities and urban science. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 374(2083). https://doi.org/10.1098/rsta.2016.0115*

We fully agree with these remarks and with the need of emphasizing such limitations and potential risks. We took the suggestion onboard and further expanded our limitations section following what the Reviewer exposed so expertly in their comment.

> “ There exists a wide spectrum of authentic ways urban scenes could be considered beautiful, because the “essence” of a place is socio-culturally and time-specific [9]. The collective perception of the urban environment evolves over time as its appearance and function change [1] as a result of shifting cultures, new urban policies, and placemaking initiatives [5]. An undiscerning, mechanistic application of machine learning tools to urban beautification is undesirable because current technology does not take into account most of these crucial aspects. Facelift is no exception, and this is why we envision its use as a way to support new forms of citymaking rather than a tool to replace traditional approaches. Nevertheless, we emphasize the need of a critical reflection on the implications of deploying such a technology, even if just in support of placemaking activities. In particular, it would be beneficial to study the impact of the transformative effect of Facelift-inspired interventions on the ecosystem of the city [2, 8] and well as the need to pair its usage with practices and principles that might reduce any potential undesired side effects [7]. ”

A suggestion for future work: The Living Building Challenge (LBC) is a performance assessment framework for the built environment that introduces non-traditional and qualitative measures such as beauty. Those buildings and architectural projects that have been assessed by the LBC could perhaps offer a complementary dataset for additional ground truthing from another perspective: https://living-future.org/lbc/beauty-petal/

We thank the Reviewer for this relevant pointer. We added a mention to LBC as a possible source of validation data orthogonal to what we considered in this work.

We want to express our sincere thanks to the Editors and to the Reviewers for all the constructive feedback as well as the positive comments above, and hope they will find the new version of the paper much improved.

References

[1] S. Brand. *How buildings learn: What happens after they're built.* Penguin, 1995.

[2] P. Dourish. Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, 3(2):2053951716665128, 2016.

[3] A. Dubey, N. Naik, D. Parikh, R. Raskar, and C. A. Hidalgo. Deep learning the city: Quantifying urban perception at a global scale. *arXiv preprint arXiv:1608.01769*, 2016.

[4] M. Foth. *Handbook of research on urban informatics: The practice and promise of the real-time city.* Information Science Reference Hershey, PA, 2009.

[5] M. Foth. Lessons from urban guerrilla placemaking for smart city commons. In *Proceedings of the 8th International Conference on Communities and Technologies*, pages 32–35. ACM, 2017.

[6] M. Foth, J. H.-j. Choi, and C. Satchell. Urban informatics. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 1–8. ACM, 2011.

[7] R. Kitchin. The ethics of smart cities and urban science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160115, 2016.

[8] R. Kitchin. Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1):14–29, 2017.

[9] C. Norberg-Schulz. *Genius Loci: Towards a Phenomenology of Architecture*. Rizzoli, 1980.

[10] D. Quercia, L. M. Aiello, R. Schifanella, and A. Davies. The digital life of walkable streets. In *Proceedings of the 24th international conference on World Wide Web*, pages 875–884. International World Wide Web Conferences Steering Committee, 2015.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016.