Response to the reviews of ACM-TOMM submission “FaceLift: A transparent deep learning framework recreating the urban spaces people intuitively love”

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. Below, we first provide a summary of our major changes, followed by a detailed response to each comment of every reviewer.

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. *. 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. *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 recommending adding critical reflections about the framework and have kindly given pointers to do so.*

We have thoroughly revised the paper in order to follow the guidance provided. A summary response to the points above:

1. We discussed the Reviewer 1's concern about Walkability as a part of limitations. Indeed we believe that the method of quantifying walkability could actually quantify some other abstract quality akin to, but not exactly, "walkability". To further help with the reflection of walkability, we add the table of labels curated using the literature on design measurement in the paper.
2. We try out the experiments with SUN labels for quantifying openness. Sagar: Time permitting, otherwise we can add this as a future work.
3.
4. We grounded our choice of the User survey scale into the previous literature that excluded a neutral option to encourage experts to avoid a de-facto position and take a stance based on their knowledge. **[++could not read the comments to this one++]**
5. We revised the whole text of the paper to fix spelling and grammar mistakes.

A detailed breakdown of the actions taken can be found next.

Requests from Reviewer 1

Reviewer 1 had no additional comments to be addressed.

Requests from Reviewer 2

> *Comment: Related work section is not detailed enough, and I doubt the number of references is enough for a comprehensive literature review.*

To address this, we did an extra round of literature survey to contextualize this paper. We added an additional subsection to ground our work with urban design metrics in the literature. We have ended up adding more than 30% to the literature review.

> *Comment: The description of the framework is not clear enough, and some details are missing, such as the network structure of the deep learning model.*

Following the reviewer's suggestion, we expanded the description of both deep learning models: the one for classification of image beauty, and the other for reproducing urban scenes to a high degree of fidelity. We also clarified the functioning of the generative model by adding examples of how generative frameworks work. We then added a schema in Figure 5 to describe how the combination of the generative model and the classifier is used to maximize beauty in urban images. We added additional text describing the architecture of the classifier in Section **"Training a beauty classifier"**. We further cite the Generator architecture we have used, and clarify how we train this generator in Section **"Generating a synthetic beautified scene"**.

Comment: Given an input image which is beautiful, does the framework return a more beautiful image or an ugly image? If it is the former, how to measure whether it really becomes more beautiful.

The beautification process happens through activation maximization of the beauty classifier's output neuron. If the input image is inherently beautiful, the output neuron corresponding to the concept of "beauty" should be in a maximal activation state. In such a setup, the activation maximization would not be useful—the image is already beautiful and there is little room for improvement. For this reason, we test the pipeline only to turn the ugly pictures (low Trueskill scores) into beautiful ones and viceversa (as mentioned in the caption of Figure 2). In the revised version, we have added the following text in Section **"Generating a synthetic beautified scene"**:

Maximizing beauty of an already beautiful image would yield a saturated template $\hat{I}_j$. For this reason, to generate an image that maximizes the beauty neuron in the classifier C, the network needs to be supplied with an image that lies in the class y_i. The constraint is reverse for maximizing for class y_j.

Requests from Reviewer 3

Comment: There are two main contributions at this paper. First is generating better urban scenes and second is able to explain deep learning framework with detail. But I think these two aspects are different, and address both of them will not highlight the main topic or innovation well. Actually, I cannot find, or the authors do not explicitly point out how to make "black-box" of CNN more apparently with detail description. So I suggest the authors only highlight one of them.

This is a valid concern. We have now de-emphasised the 'black-box' claim **[++could not read the rest of the text here++]**

Comment: About "Q4 Do architects and urban planners find it useful?", as the results shown in Table 6, I find that the rating from "'definitely not" to "definitely" without a "neutral" option, and this will lead to bias.

The question of providing a neutral option has been debated for decades, with convincing arguments on both sides. Previous work [1, 3] showed that neutral choices end up commonly used to express *lack of knowledge or indifference*. All our participants were experts in their respective fields of urban design, data visualization and architecture. Hence we consciously wanted them to express an opinion about the utility of such a tool in their practice. We clarified this aspect with the following text:

Since our respondents were experts in different areas, we wanted them to express an opinion about the utility of such a technology in their practices, and only give non-neutral responses. In accordance with this constraint, we designed the survey based on a non-neutral response Likert scale, as critical studies suggest [1, 3]

Comment: In Fig.8 there are five urban design metrics, but in the section of abstract and Table 4, there are only four metrics: walkability, green, openness, and visual complexity.

We thank the reviewer for pointing this out. We have added back the fifth metric of 'landmarks' into the paper. The metric is described in Table 4 and referred to when needed. Past studies have shown the role of landmarks in shaping the perception of 'goodness' and 'memorability' **[++could not make the new term here++]** of a city, at the macroscopic level of neighborhoods [4, 2].

Some of the positive comments...

Reviewers noted many positive aspects about this paper. Common across all three reviews were an appreciation of the usefulness of the work for the community as well as the innovation in this piece of work. We are grateful to Reviewer 1 for nominating this work for best paper and commending the 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] A. Baka, L. Figgou, and V. Triga. 'neither agree, nor disagree': a critical analysis of the middle answer category in voting advice applications. *International Journal of Electronic Governance*, 5(3-4):244–263, 2012.

[2] K. Lynch. *The image of the city*, volume 11. 1960.

[3] G. Moors. Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity*, 42(6):779–794, 2008.

[4] 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.