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. *The description of the "Facelift" pipeline was not clear enough and missed on the details of the deep learning model*
2. *The literature review was not at par.*
3. *Some clear adjustment in the description or over arching vision of the work to be either a Urban scene generator, or a framework to make the "black-box" more accessible*
4. *Clarity around the scale of the User survey*
5. *Over all proof-reading to sort out some spelling mistakes*

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

1. We improved the description of the machine learning pipeline and of the individual blocks that compose it. We added a supportive diagram of the pipeline to clarify how the two deep learning models are combined. We also clarified the requirements and constraints of the beautification process.
2. We increased the literature review by more than 30% by adding references to frame our work in a broader context.
3. We clarified what we mean by "explaining black-box inferences". Specifically, we rephrased the manuscript to best reflect the actual contribution of the work: instead of claiming that we are trying

to completely unravel the deep learning black-box framework, we now make clearer that one of our objective was to simply make deep learning methods' outputs interpretable by practitioners.

4. Questions were asked with a non-neutral response Likert scale. That is because previous work [1, 3] has shown that a non-neutral scale: *(i)* pushes respondents to "take a stance", given the absence of a neutral response; and *(ii)* works best if respondents are experts in the subject matter of the survey as responses of the "I don't know" type tend to be rare (as it is has been the case for our survey).

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 expand the presentation of our generator, and clarify how we train it 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"**:

> *It makes little sense to beautify an already beautiful image, not least because such beautification process would result in a saturated template $\hat{I}_j$ in our framework. For this reason, to generate an image that maximizes the beauty neuron in the classifier C, we restrict the corresponding input image to be in class y_i (i.e., ugly scenes as per the divisions in Figure 2). We do the opposite when maximizing the ugly neuron.*

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-emphasized the 'black-box' claim. In the manuscript, we now make clear that the main goal of FaceLift is to automatically beautify an existing urban scene. Then, as a secondary goal, to make FaceLift usable for practitioners, it is also able to explain which urban elements have been added/removed in the beautification process.

> *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:

> *Questions were asked with a non-neutral response Likert scale . . . That is because previous work [1, 3] has shown that such a scale:* (i) *pushes respondents to "take a stance", given the absence of a neutral response; and* (ii) *works best if respondents are experts in the subject matter of the survey as responses of the "I don't know" type tend to be rare (as it is has been the case for our survey).*

> *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' in Table 4. Past studies have indeed shown how the presence of landmarks positively impacts a city's 'memorability' and 'navigability' [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 their positive comments, 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.