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

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 explain how we have done so.

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 in order to follow the guidance provided. A summary response to the points above:

1. *We added some more references to contextualize our work*
2. *We added some more clarity around the pipeline specifics*
3. *We clarified the beautification process and the required constraints which might have caused confusion as rightly pointed out.*
4. *We clarified what we mean by "explaining black-box inferences" in the context of urban beauty models. We also reshaped our introduction to reflect the said clarifications.*
5. *We tried to addressed the concern about non-neutral postion in the likert scale, but citing some works around this scale as well as specifying our motivation for this choice in the paper.*
6. *We addressed some minor misses.*

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

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 understand if there are any more related works that we can use to contextualize this paper. The idea of using machine learning models to enrich urban design experience is relatively new. [SJ: Add some more fluff around which citations were added to the realated work]

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.

In this paper, we are training a couple of tried and tested network architectures on our data. We draw the reader's attention to the details through the relevant citations seen in the following statements in the section **Training a beauty classifier** and section **Generating a synthetic beautified scene**

> *We use the CaffeNet architecture, a modified version of AlexNet [2, 7]. 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.*

> *The technique does so using the "Deep Generator Network for Activation Maximization" (*DGN-AM*) [5]. 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".*

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 output neuron. If the input images is inherently beautiful, the output neuron corresponding to beauty could be expected to already be in a maximal activation state. In such a setup, the activation maximization would not yield any usable result. For this reason, we test the pipeline only on images which have the Trueskill scores to be in the lower or upper end of the spectrum (as mentioned in the caption for Fig.2) and transform them to the opposite end. This prevents these situations and ensures that the input images are either inherently un-aesthetic or inherently beautiful. To clarify this caveat, we have added the below extension to the section **Generating a synthetic beautified scene**

The legibility of the transformed image is highly dependent on the initial state of the neural activation that you are trying to maximize. Maximizing beauty of an already beautiful image, would yield in a saturated, illegible template $\hat{I}_j$. For this reason, to generate an image, that maximizes the beauty neuron in the classifier C , you need to supply an apriori image that most definitely 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 of the reviewer. And to our end we try to address this by explaining what we mean by addressing the "black box" problem. We try to clarify what we mean by 'explaining' the deep learning inference in the **Introduction** section of the paper. Through this paper we are rationalizing the way facelift generates beautified scenes, through the lens of urban metrics. That way, when facelift presents the user with a 'beautified' version of an input google street view, it presents them with the different variations in the 5 metrics which happen in the due course of the transformation. We also do a statistical analysis of a large sample set of images 'beautified' through this process, and test certain hypothesis, which were inspired from urban design, and urban vitality literature. These steps, when articulated in the right way, provide a more transparent picture of the Generative pipeline to the users, as noted by several participants of the expert-survey.

We hope the modifications to the introduction addresses the reviewer's question about the postion of this paper.

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.

This is a very valid concern raised by the reviewer. The question of providing a neutral option has been debated for decades, with convincing arguments on both side. We inspired our choice from previous works[4][1] around the critique of the neutral choice response. Here the argument against neutral choices conveyed that, neutral choices end up commonly used to express lack of knowledge or indifference. To that extent, 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.

To that effect we have added clarification of the choice along with the citations in **Q4** with the following text

Being experts in their respective fields, we wanted the survey takers to express a clear opinion about the utility of such a technology in their areas of practice. In accordance with this constraint, we designed the survey based on a non neutral response Likert scale, as explored in previous critical studies [1, 4]

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.

Thank you 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. The utility and link of presence of landmarks to the perception of 'goodness' and 'memorableness' of a city has been explored at a macroscopic level in previous studies [6, 3]. But our setup was not suitable to do any form of hypothesis testing with landmarks at an urban scene level. Perhaps we would like to extend the study in a follow up

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 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] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

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

[5] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, 2016.

[6] 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.

[7] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.