Response to Reviewer 2

Comment 1:
First of all, this paper needs a better justification for why a big-data driven analysis is more human-oriented. How are the GPS-based LBS data more “human-oriented” than conventional residential density or employment density? How are ANN-based scores more “human-oriented” than visual preference surveys?

Response:
LBS are locations that are run and visited by people. Their spatial distribution at street link level is spatially much closer to the daily routine and experience of people while residential and population density are area based summary that abstract such daily pattern to an arbitrary area (MAUP). The LBS map shows subtle differentiation in pattern from street to street within their street configuration. This is not captured by area-based statistics. ANN based score is systematisation of visual surveys – encoder and decoder – all link streets, every 20 m, bi-directional are decoded for an overall set of features that are semantically segmented from each image. In themselves, there are not expressing human preference yet in combination with other dimensions by:

1. association with LBS, street layout accessibility, proximity to transit etc. – people voting with their feet – although it is possible that well-used places have poor visual qualities because these places are mandatory activities – yet given the broad weekday and weekend day sampling it is very likely that they place have also an appeal to the population frequenting them

2. validation from visual guideline established according to visual surveys and corroborated by experts.

The revised manuscript addressed these issues by summarizing Literature review and knowledge gap in Section 2 (page 4-5, line 191-208), clarifying how ANN applied (section 3.3.2, page 8-9, line 278-283, line 293 to 301), and adding more details of input data (section 3 & section 4.1).

Comment 2:
As one of the goals of this paper is to develop street quality measurement with new data, I believe that the process and software used need to be described in more details. For example, how do the real-time LBS data from Tencent look like? How did the author use SegNet (some capture images to describe the process would be helpful)? How are PoIs classified into different urban functions? Especially, when computing diversity, the authors identified “the total PoIs within a walkable distance of a street segment” (p.9). I wonder what (and how many) types of facilities are exactly included. What are the “different urban functions” (p.9) used in the entropy calculation? Readers would want to know more details.

Response:
Thanks for the suggestion! Some details of the variables have been added in the revised manuscript, especially in section 3.3.2 (page 8-9, line 278-283, line 293 to 301) and section 3.3.3 (page 11, line 330 -342) to answer reviewer’s question. Figure 3 (Sample of LBS data from Tencent), Figure 5 (Distribution of SVIs collecting points), Figure 6 (a schematic architecture of SegNet), Figure 10 (Distribution on POIs) were also added for better understanding.

Comment 3:
Even though the cluster analysis results are validated qualitatively, some quantitative measures need to be also explored. The authors should describe which clustering algorithm was used for what
reasons (e.g., single linkage, average linkage, Ward). Also, because a cluster analysis is unsupervised and any number of clusters is possible, some statistics exist to determine the optimal number of clusters, such as Calinski and Harabasz index and Silhouette index. If the authors’ goal was to classify the street segments into three groups—from poor to high quality, they could have used k-means clustering (or a similar one) instead of hierarchical clustering (plus a subsequent qualitative classification from eight to three).

Response:

*Thank you very much for raising this concern! The specific clustering algorithm has been added into the manuscript.*

Although a fixed number of clusters is possible to be determined by using the statistics suggested by the reviewer or simply using an “elbow” method. However, the ‘best’ number of clusters is not always appropriate especially in a study with a large sample size (N=1,231, the ‘best’ number of clusters would be 20-40 in this study). Since the determination of this number is not an aim of this study, we chose to use hierarchy cluster, the number of groups need not be known a priori.

We agree that there is a confusion in the part of classifying 8 groups into 3 types based only on radar charts. Therefore, in the revised version, we re-classify the 8 clusters by taking both the tree structure and radar charts into account (section 4.2, page 14, line 418 -433). The new three types are more consistent with clusters, and also show a better agreement with experts’ evaluation.

Comment 4:

Lack of reliability test for the image classification for Design factor is another limitation of this research. An interrater reliability measure (e.g., kappa statistics) may help readers believe that the subjective “design” measures from experts are reliable.

Response:

*Thank you for the suggestion! To achieve a more subjective reliability evaluation, we adopted kappa statistics (using STATA) in revised section 4.3 (page 16, line 466 - 471) to measure the inter-rater reliability.*

Comment 5:

There is more literature to be reviewed regarding “how streetscape features affect the quality of life, and the following effects on physical activities and public health.” (p.1)

Response:

Thank you for offering the references! The revised paper has expanded literature review. (References [5-12], [59-63]). Knowledge gap and main contributions of this study were also summarized at the end of Section 2 (page 4-5, line 191-208).

We highly appreciate the reviewer taking the time to offer us comments and insights related to the paper. We would be glad to respond to any further questions and comments that you may have.