Taking Yangpu District in Shanghai as an example, this paper proposes a comprehensive approach to measure street quality with multiple big data. The manuscript is well-written, the conceptual framework is solid, and the methodology is sound. On the other hand, some data processing and analysis steps need more justifications and explanations to enable this research to be valid and replicable.

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?

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 Pols classified into different urban functions? Especially, when computing diversity, the authors identified “the total Pols 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.

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).

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.

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)

I recommend “Destination accessibility” instead of “Accessibility” in Figure 1.

Typo

- P.2 line 64: “Some studies have found that there is a strong, positive coefficient” -> “positive correlation”
- P.2 line 82: “studies are of a wide variety and but are limited” -> “variety but are”
- P.13. Table1: Double check the numbers. “Type C: Group 5 (112 of 1,232, 45%), poor quality” -> should be 9%. Also, 559 + 557 + 112 + 1 does not equal to 1,232 (total street segments). Lastly, I would consider group 8 as medium quality rather than high quality.