We appreciate all the reviewer comments. We have incorporated most of the comments in the new draft. Listed below are our responses to each of the comments.
First, a brief part of their comment is highlighted (as RC), followed by our response.

Response to reviewer 1:

Comment: Stability of WiFi signature. In Section 3 and Figure 1, the authors showed that WiFi RSSI is spatially distinct and can be used as location signatures. However, WiFi also suffers from the issue of temporal stability and its signals may change over time or under environment dynamics (say, people are moving around). In Section 4, the authors suggest pausing for 10 seconds for a stable Wi-Fi signal collection. Although this way can alleviate short-term signal dynamics, it cannot eradicate the issue of long-term instability. For example, in a large site with significant human dynamics (e.g., a shopping mall), the WiFi signature or pattern can hardly be similar when a robot revisits the same location after 15 mins (it is possible as the authors pause the robot a lot for every 4 meters and robots often move slowly in the shopping mall).
The authors should discuss this point or limit this as a limitation.

Response: Wi-Fi RSSI profiles may be affected over time due to environmental dynamics. However, for the purpose of online SLAM, it may be justified to use it, because the time span of the operation is not very large most of the times. Further, we empirically observe that our mechanism of using RSSI vectors through pausing and averaging across a number of measurements is relatively robust to the types of dynamics that is seen in a building in a university.
We would also like to note that we have several knobs built in to compensate for this dynamics. We could adjust the frequency of pauses for Wi-Fi RSSI collection or the time length of the pauses for averaging the data. We could adjust the Wi-Fi similarity threshold for computing similar clusters and employ lower values for a higher degree of dynamics for compensating fluctuations.
But, in general, a higher degree of dynamics cause more RSSI fluctuations which may limit the applicability of the approach.
We have added some discussion in section 7/page 17, the paragraph of Environmental Dynamics.

Comment: Clarity of Visual Edge. The term of visual edge in the paper is very misleading. At first glance, I thought it is an object edge in a picture. Later I realized this is more likely to the edge between nodes in the graph SLAM. The authors should make it clear. Furthermore, more details are also needed in section 4 to tell the reader how

such edges are constructed. High-level descriptions of RANSAC and matching is not enough, as visual and WiFi signals are overarching in their proposed approach while reliable `visual edge` detection certainly affects the overall augmentation performance.

Response: We added the definition of a visual edge in section 4/page 4, and also added a discussion paragraph in section 7/page 14, talking about the relation between our proposed approach and visual edge construction.

Comment: Related Work. This paper falls into the category of RF-augmented positioning systems. In this sense, a subsection that discusses other emerging RF-assisted positioning systems is needed. In particular, the authors should at least discuss the following work and highlight their differences between them.

Response: We appreciate the comment and have added RF related references mentioned by the reviewer in section 2/page 2, RF-assisted positioning systems paragraph.

Response to reviewer 2:

Comment: Section I - Wi-Fi Localization/SLAM:
This paragraph could be extended to add more details about the approach used in Wi-Fi SLAM and how the other approaches compare with the current approach.

Response: In this literature, RSSI or CSI is employed either for localization of the robot or for SLAM where the landmarks are Wi-Fi Access Points (APs). We also note that most of them require a training phase for the collection of Wi-Fi data. In our work, we improve visual 3D maps rather than mapping just the APs and our method does not require a training phase.
We added explanation to section 2/page 2, Wi-Fi Localization/SLAM paragraph.

Comment: Wi-Fi and Visual Association: The approach requires that the robot stands still for 10 seconds to aggregate a Wi-Fi signature, can the approach still be usable while moving at the cost of reducing the loop closure detection performance? or the signature created won't be really usable to create good clusters?

Response: Our solution to changes in RSSI due to environmental characteristics as well as background communication is to collect signal strength for several seconds and average it. This will get affected if we perform a continuous movement. Depending on

the environmental dynamics and the number of wireless clients in a given area, this might or might not affect the wifi clustering.
We have added the explanation to section 7/page 17, the paragraph of Environmental Dynamics.

Comment: What does the cluster look like in general? A figure could help with the description. A loop closure may not be found if the current frame is close the edges of the cluster radius. When close to edges of some clusters, could the approach bound loop closure search to more than one cluster (like 3 most similar and/or closest clusters)? Is there a maximum of clusters?

Response: The loop closure search is not bounded to one cluster and all similar clusters are considered for finding the right candidate. There is no bound on the number of similar clusters and any cluster with a Wi-Fi similarity higher than a defined threshold are counted as a similar cluster. So, when the current frame is close to the edge, with a high probability the nearby clusters would be identified as similar clusters and we would take into account their visual frames for loop detection.
We added Figure 3 for cluster representation.

Comment: In Section 6.1, I think "false negative" is referring more to "true negative" instead, as false negative only tells that a wrong loop closure has been missed (which is correct). A true negative would be that there was a real loop closure but the SLAM approach missed it. If it is the case, look to make the changes across the paper.

Response: We believe that we are using the term correctly in this context. False negative here means a correct loop closure is not detected.

Comment: In most tables, there could be less numbers after the points, maybe keep one or two instead of 4 not significative numbers.

Response: We appreciate the comment, we decreased the numbers after the points to 2.

Comment: The A Hall dataset is very large and there are not many blocking walls along the trajectory which causes less RSSI attenuation between different places. This leads to less number of Wi-Fi clusters." Is there a parameter that can be tuned to create more clusters in such environment?

Response: Yes, increasing the Wi-Fi similarity threshold would lead to creating more number of clusters. But it would also decrease the number of similar clusters for loop closure detection which may lead to losing some correct loop closures.
We can also define two thresholds, one threshold for creating clusters, one threshold for specifying similar clusters, but we believe this will not affect the performance of A Hall. Because although the number of clusters has increased, but more clusters would be counted as similar clusters.
We added explanation in section 6.4/page 11, mentioning the effect of similarity threshold on number of clusters.

Comment: Could the approach be used to distinguish between floors of a building? It could be interesting to see results on a highly symmetric multi-floors building (like in many offices).

Response: Our empirical observation is that RSSI vectors we use are fairly distinct at various locations, and could be useful to clearly determine the floor that the robot is on. However, this needs further study and we have added it to section 8, conclusion and future work.

Comment: As the Wi-Fi signatures are recorded in the map at specific positions, could it be possible to triangulate the APs (online or offline) in order to use that kind of information to predict better in which cluster the robot would be in next localization/mapping sessions?

Response: There is a wealth of literature on wireless localization, some of which use techniques described by the reviewer. However, the overall result of most of these papers is localization with a resolution of ~10m.
Our goal is to more deeply integrate wireless sensing into visual SLAM. Therefore, we did not go the route of using an existing wireless localization algorithm and use the location result in our SLAM. We believe that our solution is a more elegant use of wireless sensing to particularly solve the perceptual aliasing problem we have identified without doing the extra work of performing localization, which Wi-Fi is ill-suited for.

Comment: In section 6.6, "CDF" is not defined.

Response: We appreciate the comment and have added the definition in section 6.6. CDF stands for Cumulative Distribution Function of distances between estimated position and ground truth position.

Response to reviewer 3:

Comment: outline of how they deal with Wi-Fi access points that are available in one region in the environment, but not in another region.

Response: If some Wi-Fi access points are not available in some regions, we set their corresponding values to zero.
We added the explanation in section 3.2/page 4.