We appreciate all the reviewer comments. First, we summarize our contributions. We use a novel $2D^2CCA$ loss to assure similar semantics are captured from the RGB and depth domains; we design a new model structure for sparse depth completion, which exploits the relationship between RGB and depth data; our method achieves the state of the art (SOTA) on several indoor/outdoor datasets.

Response to common concerns: First, we will release codes including the data processing tools. Second, we fixed the symbol/annotation inconsistency and added symbols on Figure 2. We will carefully check grammar in the whole paper.

Response for reviewer #1:

1. "The difference between using the complementary RGB image and the dense RGB image is quite small." - The difference between known and predicted depth is small (see ablation study, line 229-233 of paper). However, Table 1 validates that using only sparse and complementary information for depth completion is sufficient and can yield better performance than using dense RGB information.

2. "For Table 3 a version of CFCNet with 0 points is missing (only RGB information)" - The core elements of our CFCNet are $2D^2CCA$ and a range transformer. With only RGB information, the network structure would be just an image encoder-decoder, and reduced to a normal FCN. This contradicts our purpose of multi-modal learning. Table 1 with only image information is for the ablation study. It shows that our core elements can improve performance.

3. Need to explain the result reported in the abstract. - The numbers in the abstract are compared with CNN-SLAM. The "13.03" is the number for that sequence. The "58.89" is a typo - the correct number is "64.34" in Table 5. We agree that using the result of a single sequence for comparison could be biased and unfair. Hence, we instead calculate the average performance on SLAM RGBD datasets. We attain a +194% improvement on the datasets.

4. "Which dataset was used for the results in Table 1? Were the networks trained for the specific input configuration in Table 1?" - We used the NYUv2 dataset, as described in the supplementary material. NYUv2 dataset was pre-processed by a 100-point stereo sparsifier and then used as the training data. We will add the information to the body of the paper.

5. "Do you backpropagate to the parameters of the sparse depth branch from the 2D CCA loss?" - Yes, we use Eq. (4) in the paper to backpropagate gradients through both the image branch and sparse depth branch. The whole network is end-to-end trainable.

6. For the K->K experiment the number of sparse depth points is missing. - We state 100 depth points in the title of the table. We will add it to the table body as well.

Response for reviewer #2:

We would like to clarify that our proposed loss is CCA-based but not SSA-based.

1. "False color visualizations, eg Fig 6, are only very qualitative and do allow to really compare performances of the methods." - Most of the related works also use false color to visualize results. We follow this practice and use color codes in the same way. Visualizing depth maps with false color is indeed qualitative. Hence we also provide numerical results in paper to compare with others. We will add the numeric data of each example shown in Fig. 6 in the paper.

2. "Many tables (eg 1,2,3,4,5) have no units." - Meters for MAE and RMSE. Percentage for all δ. We will add this info in our final version.

3. "Model seems to work but is just another model." & "The model is a combination of existing elements but reasonably engineered." - We respectfully disagree. Our method is not a simple combination of existing elements. Rather, it is a multi-modal learning framework effectively uses $2D^2CCA$ to exploit the relationship between different types of data. We want to demonstrate to other researchers a meaningful direction, in which the correlation between multi-modal features for depth completion is fully utilized. Since our novel loss is independent from existing elements such as the VGG-16 architecture, future researchers may use our loss along with other models to boost performance.

4. "The paper misses an important previous work: [P1] which should be reviewed and included into the comparison." - We thought it might be inappropriate to compare our results with theirs because the study objectives are different. Our study aims to complete depth when observable measurements are highly limited, while [P1] aims to complete missing data with significantly more observable depth measurements. The authors note that their performance would be negatively impacted if the observed depth samples are fewer than 2000. Our experiments use samples typically fewer than 500.

Response for reviewer #3:

1. Meaning of many “directional”s. - "Directional" is synonymous to "dimensional" here, used in some statistical communities. Indeed this may be confusing. We changed “directional” to “dimensional” to avoid confusion.

2. “The authors already demonstrate results when sampling 100 points from LiDAR data, but I am still interested in what will happen when using all LiDAR measurements.” - Thank you for your interest. We show numerical results for training and evaluation on KITTI train/val dataset with an additional visual result here (left to right: image, depth completion using full LiDAR points (MAE=0.2822), depth completion using 500 LiDAR points (MAE=0.7871)).

Sample#	MAE	RMSE	δ_1	δ_2	δ_3
500 (from paper,Table 2)	1.197	2.964	94.0	98.0	99.3
full points	0.596	1.568	97.5	99.3	99.8