Response to Reviewer 1 Comments

Thank you a lot for your careful reading and helpful comments on our manuscript. In this revision, we have addressed all of these comments/suggestions. We highlighted all the revisions in red/blue color. Below are our point-to-point responses to your questions. We tried our best to improve the manuscript and made changes in the manuscript. We appreciate for Editors/Reviewers' careful review and hope for your favorable consideration regarding this manuscript.

Line 22 : Define what you mean by GPS-tagged RS image - is it an ortho photo, what is the accuracy of the GPS, what is the elevation model, etc. etc.

Response 1: GPS-tagged RS image means that the coordinates of the image are already known. Currently, we consider cross-view geo-localization as an image retrieval problem, where the accuracy of GPS coordinates and elevation models is not our main focus. Additionally, the datasets we use do not provide such information.

Line 29 : What is the elevation model, etc. etc. the success is a function of the quality of the coordinates of the street view image - this needs to be addressed

Response 2: Same as **Response 1**, this is not our main focus.

Line 33 and 40: What is the elevation model, etc. etc. the success is a function of the quality of the coordinates of the street view image - this needs to be addressed

Response 3: We offered a more understandable expression.

Line 82 , 202, 244, 249, 264, 269, 279, 333, 373, 458, 479 : Spelling error.

Response 4: We have corrected these errors.

Line 271: How is this threshhold determined?

Response 5: We introduced a method for threshold selection in Section 4.5.2. Due to the coverage area of a RS image being 144 meters × 144 meters, we choose 0 meters, 36 meters, 72 meters and 144 meters as distance thresholds for meter-level evaluation.

Line 273 : Reference for PyTorch

Response 6: We added references to PyTorch in the paper.

Line 287-290 : How did you determine these settings

Response 7: We initialize our model with off-the-shelf pre-trained weights on ImageNet-1K with 8 × 8 window size, so we we set different learning rates for Swin-Geo and other modules. The weight (α) of the loss function is set to 10, following the approach mentioned in the TransGeo model described in the paper. We consider both retrieval loss and contrastive learning loss equally important, so the weights (β, γ, δ) are set to 0.5, 0.25 and 0.25.

Line 321-322 : How can you say these are significant improvements as less than 2%?

Response 8: On one hand, our model achieves a significant accuracy improvement of nearly 5% on the CVACT dataset. On the other hand, although the accuracy improvement of our model on the CVUSA dataset is less than 2%, it's important to note that the accuracy of previous methods on this dataset already exceeds 94%, reaching a bottleneck. Achieving improvements of over 2% in performance is already challenging, as evidenced by the latest two state-of-the-art methods(TransGeo and L2LTR), which only provide accuracy improvements of 1.49% and 0.03% on CVUSA dataset, respectively, compared to previous methods. Therefore, we consider our accuracy improvement significant.

Line 333-334 : Why were these values selected?

Response 9: Same as Response 5.

Line 342-343, Line 321-333 : Are these really significant improvements?

Response 10: Same as Response 8.

Line 354-355 : Are top-1 and top-1% the same thing?

Response 11: They are not the same thing. We added explanations in Section 4.5.1.

Line 386, 388-389, 390-397 : Barely more than zero improvement.

Response 12: Same as Response 8. When performance improvement reaches a bottleneck, even minor performance gains are still noteworthy. Additionally, as shown in Figure 1 and Figure 2, while the improvement in our recall metric may be modest, the consistent performance enhancement observed throughout the training process underscores the reliability of this improvement and the effectiveness of our approach.

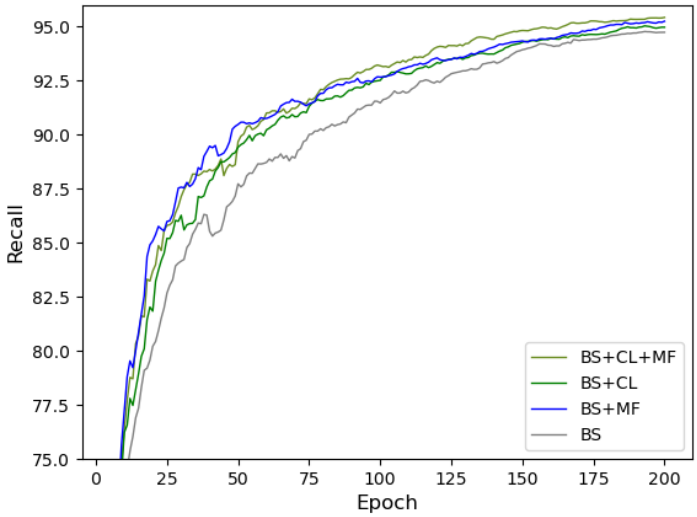


Figure 1 The Recall on CVUSA dataset

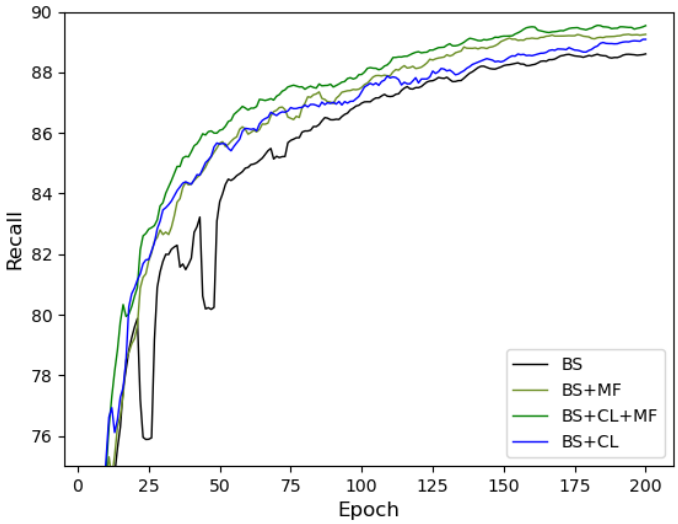


Figure 2 The Recall on CVUSA dataset

Line 409 and 446 : Values areas makes no sense

Response 13: We have added a description regarding significance of thermal values areas in paper. Due to the different fields of view and perspectives between street-view images and RS images, street-view images may contain significant occlusions. This implies that areas closer to the center of the RS image and regions that are not occluded are more likely to appear in the street-view image. Therefore, to achieve better performance, the model should allocate higher attention to areas that are present in both street-view images and RS images. This implies that these regions should have higher thermal values, while other areas should have lower thermalvalues.