Response to the Anonymous Referee #1

General comment:

The manuscript presents the China Building Rooftop Area (CBRA) dataset, which provides national-scale pixel-level information on individual building rooftop distribution and multi-annual dynamics from 2016 to 2021. The authors proposed an interesting and novel method for extracting high-resolution production of building rooftop from Sentinel images that could potentially reduce data acquisition costs. The study is well-structured, and the results demonstrate characteristics and superiority over previous production. The paper could be a valuable contribution to the society of urban remote sensing in terms of both the novel methodology and production. However, some revisions are necessary before accepting it for publication.

Response:

We express our gratitude to the reviewer for acknowledging our research and offering valuable suggestions. In the following section, we present a detailed response to each comment provided. The reviewer's comments are presented in black font, while our responses are presented in blue font. The revised manuscript is presented in red color, and our modifications are highlighted in yellow.

Major comments 1:

The one of the key contributions in this manuscript is the usage of a spatial generalization (SG) loss to increase the generalization capacity of a larger geographical region. However, as shown in figure 9, 12, 13, there exist some block-like result. It seems the SG loss could not perform well in these regions. Are they all caused by a resolution of less than 2.5 meters so the model can't distinguish them? The authors should provide further justification of the importance of the SG loss, including its superiority compared to directly utilizing cross-entropy.

Response:

Thank you for your valuable comments.

The block-like results are mainly caused by the buffer-like effect of a few pixels at the building edges caused by our approach, which is common in the building extraction practice (Ding et al., 2021; Zorzi et al., 2021; Guo et al., 2022; Liu et al., 2022). We have included examples of this effect in Fig. C1, where we present an example of boundary localization results using the Unet++ method (Zhou et al., 2019). Even in very high-resolution images (0.3 m), the building boundary extraction results exhibit an uncertain offset of several pixels.

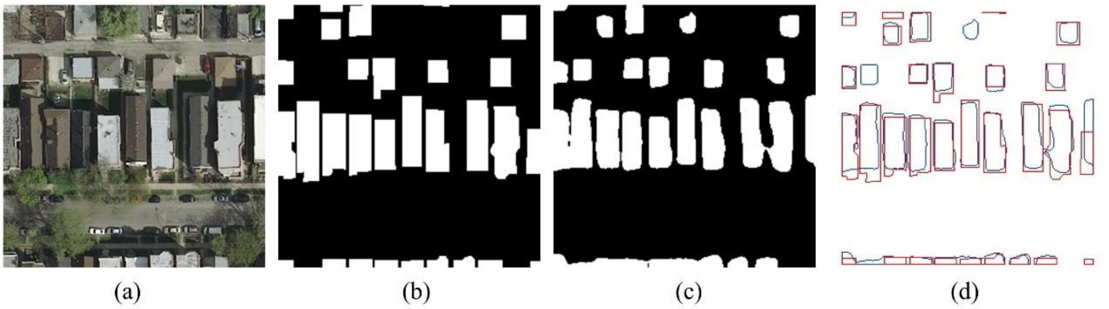

Figure C1: An example of the ambiguity of building edges from Inria building dataset (0.3 m) by the Unet++ method (Zhou et al., 2019). (a) The input image (0.3 m). (b) The ground truth. (c) The extraction result. (d) The building boundary, where red is ground truth and blue is prediction, respectively. It can be from (c) that the building boundary extraction results still shows with an offset of several pixels even from the very high-resolution images (0.3 m).

In densely residential areas, the aforementioned ambiguity is further compounded, as demonstrated in Fig. C2. When buildings are close (e.g., less than 6-7 m, i.e., 2-3 pixels at 2.5 m resolution image), our method may not be able to effectively distinguish between them. It is a common phenomenon in building extraction task, but the previous work mainly uses very high-resolution images (less than 1 m), making their results seem better than ours.

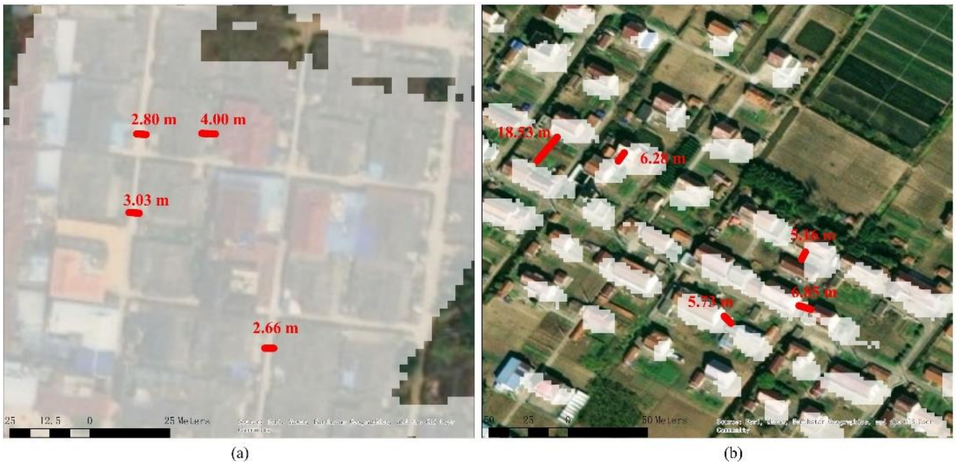

Figure C2: Two examples of blob-like areas and the measured distances between adjacent buildings. (a) Densely residential area (101.302089ºE, 21.298532ºN). (b) Relatively discrete residential area (121.634662 ºE, 31.746674 ºN). Imagery © 2023 Maxar Techonlogies.

As for the SG loss, as discussed in Sect. 6.1 “Importance of the spatio-temporal aware learning”, using SG loss as an auxiliary loss could greatly improve the performance than only using cross entropy loss (2.38% improvement in F1 score). The SG loss mainly helps us to obtain more weakly supervised information in rural areas, thus improving the generalization performance of the model. However, due to the spatial resolution limitation, some regions are more seriously adhered.

Therefore, we now strengthen the statement of limitation (Sect. 6.2) in the revised manuscript (Fig. 19 in the revised manuscript is the Fig. C2 above), from line 608 to 614 on page 30:

Although our STSR-Seg framework is scalable, allowing larger areas to be monitored (e.g., national scale). there remain some limitations to our approach. Specifically, the segmentation results for densely populated residential areas may present certain rooftops as a single block, rather than individual buildings. Our analysis suggests that this occurrence is primarily due to the resolution of the results, which is 2.5 m. Furthermore, the semantic segmentation technique utilized in the approach may introduce some uncertainty at the edges of buildings, resulting in additional pixels, up to three pixels, at the boundary. Consequently, up to 7.5 meters of buffering may occur, exacerbating the problem of building adhesion. Examples of this issue are presented in Fig. 19

References:

Ding, L., Tang, H., Liu, Y., Shi, Y., Zhu, X. X., & Bruzzone, L. (2021). Adversarial shape learning for building extraction in VHR remote sensing images. IEEE Transactions on Image Processing, 31, 678–690.

Zorzi, S., Bittner, K., & Fraundorfer, F. (2021). Machine-learned regularization and polygonization of building segmentation masks. 2020 25th International Conference on Pattern Recognition (ICPR), 3098–3105.

Guo, H., Du, B., Zhang, L., & Su, X. (2022). A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 183, 240–252.

Liu, Z., Tang, H., & Huang, W. (2022). Building Outline Delineation From VHR Remote Sensing Images Using the Convolutional Recurrent Neural Network Embedded With Line Segment Information. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–13.

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging, 39(6), 1856–1867.

Major comments 2:

While the SG loss might be reasonable for regions with missing high-resolution supervised information, the authors assign a "built" land cover type to each building rooftop reference as described in line 260, which might potentially confuse the original high-resolution information in my personal viewpoint. Therefore, a detailed description of the pipeline is required when both high-resolution and low-resolution references are available.

Response:

Thank you for your valuable comments.

As shown in Fig.6 of the manuscript, the SG loss is used to provide additional supervised information with coarse resolution (i.e., the “built-up” area in Dynamic World product) to ensure that the model can effectively generalize to a larger geographic region, especially in the absence of reliable high-resolution building references (i.e., the building rooftop from Tian-di Map). We also find that using low-resolution labels in the presence of high-resolution labels as well improves the accuracy of the model, as shown in the following table:

Table C1: The impact of collaborative training with high-resolution references and low-resolution references. When using the SG loss, there are two strategies. One involves the use of low-resolution references solely in the absence of high-resolution supervised information, while the other entails training both simultaneously (i.e., utilizing low-resolution information for supervision at the same location even in the presence of high-resolution references). Our empirical results indicate that the second approach, i.e., collaborative training, can achieve superior accuracy compared to the former.

	IoU	OA	Recall	F1
STSR-Seg w/o collaborative training	45.15	82.73	74.51	62.21
STSR-Seg w/ collaborative training	45.51	82.85	74.66	62.55

We add more details about it in Section 6.1 (Table B4 refers to the Table C1 above), from line 583 to line 584 on page 29 :

……of training resources, and therefore greatly improves the accuracy of our data-driven method (+2.38% in terms of F1 score). Even when high-resolution references are available, incorporating low-resolution land-cover information in the training process through collaboration as supervised information is found to be beneficial (Table B4). In addition, ……

Major comments 3:

By comparing with the 90 cities BRA data, the authors claimed that the results of the 90 cities BRA data are spatially inconsistent than the CBRA. It is necessary to ensure that the supervised data is similar between the CBRA and the 90 cities BRA to rule out any differences that could be causing this inconsistency.

Response:

Thank you for your valuable comments.

According to the technical report of 90 cities BRA data, the author did not reveal more details about the training samples. This makes it difficult to accurately determine any differences between their training samples and ours. However, we believe the Google Earth Satellite imagery potentially have geographical offsets, particularly in regions where high-rise buildings are present and display shifts, as illustrated in Fig. C3. This error can be attributed to variations in the topography's elevation and is less apparent in the low-rise buildings. In all, it is believed that the spatial inconsistency is primarily attributed to image shifts.

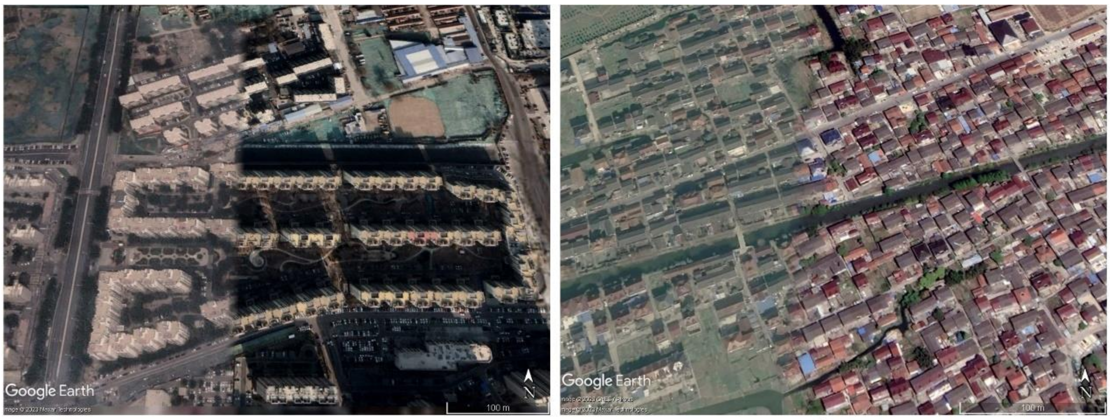

Figure C3: Two examples about the stitching part of GES imagery. (a) The high-rise buildings. (b) The low-rise buildings. It could be seen that the offset is more obvious in the high-rise buildings than that in the low-rises. Imagery © 2023 Maxar Techonlogies.

Major comments 4:

The paper highlights the limited availability of datasets covering the entire China. However, there exist some BRA datasets derived from sub-metric aerial images that cover specific regions. It would be valuable to show a comparison of the CBRA dataset with other small yet region-specific BRA datasets to gain more insights.

Response:

Thank you for your valuable comments.

We have attempted to procure regional building datasets to conduct an in-depth analysis of our product. However, it is regrettable that most of the locally-distributed building datasets in China are labeled manually and lack geographic information, such as coordinate systems, necessary for unbiased comparison (Li et al., 2022). Therefore, it is challenging to locate a regional dataset to compare our product fairly.

Nonetheless, our manuscript showcases the results of comparing our product with the 90-cities-BRA, comprising of 245, 458 independent samples. Additionally, we have conducted trend analysis using CLCD, GAIA, and Dynamic World product, illustrating the temporal consistency of the CBRA.

References:

Li, J., Huang, X., Tu, L., Zhang, T., and Wang, L.: A review of building detection from very high resolution optical remote sensing images, GIsci Remote Sens, 59, 1199–1225, 2022.

Major comments 5:

The multi-annual dynamic world product is presented as a probability map. Therefore, it is necessary to provide a more detailed technical description of the threshold, which might be used to binarize it in a meaning way.

Response:

Thank you for your valuable comments.

In Section 4.4, we use 0.2 to binarize the Dynamic World product and use it to intersect with our result to filter the potential false predictions. To find a reliable threshold, we need to make sure the binarized map would not filter out the correct predictions. Therefore, two considerations are involved:

(1) First, with a threshold interval of 0.05, we binarize the dynamic world product and calculate its recall on the building rooftop area data we collected from 52 cities, and the results are shown in Fig. C4. It can be found that recall reaches 0.99 when the threshold value is 0.3, which shows that using a threshold value of 0.3 and below ensures that the correct results are not filtered out incorrectly in urban areas.

(2) Secondly, since the above reasonable thresholds are calculated based on urban samples, we also need to consider rural areas. Due to the lack of reliable building distribution data for rural areas, we judge the thresholds by visual observation of the images, and finally we consider 0.2 as a more robust threshold because it does not filter out our correct prediction results.

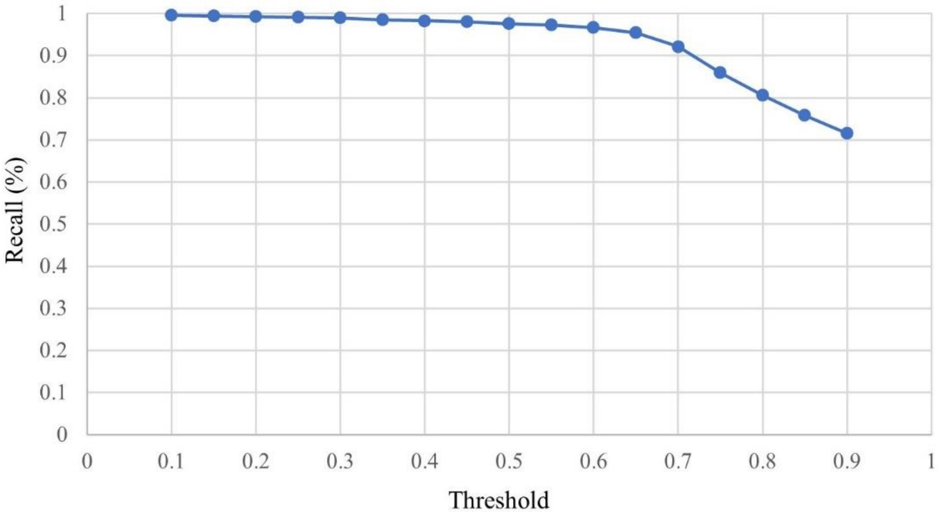

Figure C4: The recall curve for the threshold selection. The threshold interval of 0.05 was used to binarize the Dynamic World product, and it was compared with the building distribution data we collected from 52 cities in China to calculate recall respectively. When the threshold is 0.3, recall reaches 0.99, indicating a threshold value of 0.3 and below ensures that the correct results are not filtered out incorrectly in urban areas. As for the rural area, due to the lack of reliable building rooftop annotations, we judge the thresholds by visual observation of the images. Finally, we consider 0.2 is the best option.

We integrate the relating explanations into the revised manuscript (Fig. A3 is the Fig. C4 above), from line 383 to 386 on page 16.

…… The built area provided by Dynamic World is a possibility estimation ranging from 0-1. A low threshold of 0.2 is utilized to distinguish between built-up and unbuilt areas, as this threshold does not filter out correct prediction results (further discussions in Fig. A4).

Minor Comments 1:

The introduction and background sections have some repetitive descriptions. For example, the second paragraph could be shortened for there already has a detailed description in Section 2.2.

Response:

Thank you for your comment. The second paragraph in the introduction is shorten in the revised manuscript.

Minor Comments 2: line 60: please provide the source of the statistics about the urbanization rate and the population structure of China.

Response:

Thank you for your comment. The source of the statistics is from the National Bureau of Statistics of China, we now add this reference into the manuscript.

Minor Comments 3:

line 159: “apple” should be “apply”

Response:

Thank you, confirmed and corrected in the revised manuscript.

Minor Comments 4: l

line 220: please unify the description of “arcgis”

Response:

Thank you, confirmed and corrected in the revised manuscript.

Minor Comments 5:

Figure 5: what is the meaning of the four rectangle and arrow in subfigure (C)?

Response:

Thank you, it is a schematic diagram about the sliding window for obtaining the Sentinel-2 images

Minor Comments 6:

Figure 6: the direction of blue arrow and red arrow in the legend should be the same.

Response:

Thank you, the Fig. 6 is updated with a new legend in the revised manuscript.

Minor Comments 7:

line 268: correct "red backward arrow" to "red arrow"

Response:

Thank you, confirmed and corrected in the revised manuscript.