Response to the Anonymous Referee #2

General comment:

The authors used the STSR-Seg method to develop a novel BRA dataset covering 2016-2021 with a temporal interval one-year, it has been demonstrated to have better performance than Zhang et al. (2022) results and covers the whole China (rural and urban areas), with an overall accuracy of 82.85%. The manuscript has been well-written and clearly organized. However, I still have several concerns about the current manuscript as follows

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.

Comment 1: The generating training samples should be greatly improved. For example, are the sample size of 200 and standard deviation of 150 reasonable? The land-cover label of each training coordinate came from which dataset (the spatial resolution of the coordinate represent 10 m spatial resolution?). As for the reference year, why you randomly assigned during 2016-2021? And you stated '*we further rebalance the gathering samples to make sure that the built type is the majority by thresholding*', how to achieve the goal?

Response: Thank you for the reviewer's comment.

Before responding your comments. We would like to restate our motivations of this generating process to better make a response.

Our method (STSR-Seg) needs two types of reference data (i.e., supervised information) in the training process. One is the accurate building rooftop data (i.e., vector data collected from Tiandi-Map). However, it only covers 52 cities and the supervised information is not sufficient to support our mapping needs for building rooftops for multiple years across the whole country. Therefore, we collected another data for weak supervision information, i.e., coarse resolution (10m), but full coverage, as well as multi-year built-up area data. This data was collected from the Dynamic World product (Dynamicworld, 2022).

When generating training samples, we need to consider two issues.

(1) Firstly, we need to make it as inclusiveness of built-area as possible as we could. Therefore, we sampled in the vicinity of the administrative point in China, and used a Gaussian distribution for sampling. In our experiments, we found the standard deviation of 150 and sampling size of 200 are the best for covering the possible regions (Fig. A1 in the manuscript of page 33).

(2) Second, we need to consider the engineering efficiency and resources for local storage. Considering that the amount of training data used in the current state-of-the-art remote sensing large

models in deep learning field is one million (Wang et al., 2022), we want to make our training sample size approach this order of magnitude. There are 2, 844 basic points used to determine the sampling range. So, the amount of the training data size (N) is:

$$N = 2844 \times 200 \times t,$$

where t is the number of the sampling year. In this paper, to approach the magnitude of one million, we used 3 for t, i.e., "randomly assign three reference years over 2016-2021." Although more sampling years may improve the accuracy (due to more temporal samples), the resulting computational effort is also larger. Based on the current amount of data, it takes us about 1 month to finish the model training only once using our local server. We sometimes reset the hyper-parameters to re-train the model, and the actual time is longer.

Finally, not all samples should be used, because some of them have few pixels in the built-up area and result in a long-tail distribution problem (as shown in Fig. A2 in the manuscript, some samples are even in the sea area). So, we need to perform a second filtering, i.e., filtering by the built-up area pixel proportion. Specifically, those with built-up area pixel percentage less than 10% are discarded when computing the loss.

In all, (1) the sample size and the standard deviation are set to best cover the entire built-up area of China (Fig. A2). (2) The land-cover label of each training coordinate came from the Dynamic World product (built-up area). (3) The number of randomly sampled years is used to control the total number of training samples, and 3 is the optimal option based on the previous work and the upper limit of our computational resources. (4) Further filtering is done by a 10% threshold to eliminate samples that do not contain built-up areas and potentially erroneous samples.

We integrate the relating explanations into the revised manuscript, from line 239 to 263 on page 11 and 12:

…… As described in Sect. 3, the reference data for training consists of both the high-resolution building rooftop in 47 cities from Tiandi-Map and the low-resolution land cover data from the Dynamic World product. For the rooftop data, we can easily pair them with the Sentinel-2 imagery obtained in the same location and time…….

The number of sampling coordinates and standard deviation employed in the heuristic sampling strategy is based on off-line experiments, which thoroughly cover the potential urban areas of China (Fig. A2). It is important to note that only three years are randomly sampled from 2016-2021 to avoid increasing the dataset and imposing an unmanageable computational burden. Through this approach, the land cover training samples, covering both urban and rural scenes and ranging for various years, are easily and automatically gathered.

Finally, the generated samples may still exhibit redundant information, necessitating their further filtration. Specifically, those samples containing fewer built-up area pixels (i.e., below 10%) are discarded. In all, we obtain two sets for model training……

References:
Dynamicworld: https://dynamicworld.app/, last access: 24 November 2022.
Wang, D., Zhang, Q., Xu, Y., Zhang, J., Du, B., Tao, D., and Zhang, L.: Advancing plain vision transformer towards remote sensing foundation model, IEEE Transactions on Geoscience and Remote Sensing, 2022.

Comment 2: The authors used the spatiotemporal learning method to achieve the aim of downscaling and independently generate the CBRA map in each year, but the block effect is still obvious in your results. To further demonstrate the feasibility of the proposed method, can you use the Landsat image (30 m) as an example to generate the CBRA at 2.5? I believe it would make more sense (only a suggestion).

Response: Thank you for the reviewer's suggestion.

We feel that using Landsat data does not necessarily lead to better results. This is due to our methodology's reliance on a framework that integrates super-resolution and semantic segmentation techniques. We believe that the accuracy of the model's output results is contingent on the alignment between the original image resolution and the resolution of the resulting data (Shermeyer & Van Etten, 2019). As such, the greater the similarity between the two resolutions, the higher the model's accuracy. In our work, we established an output resolution of 2.5 meters. The Sentinel-2 data, with its coverage of the entire Chinese territory over many years and a resolution of up to 10 meters, represents the optimal data source for our study. In contrast, the Landsat image, with a resolution of 30 meters, may not enhance the model's accuracy due to the significant discrepancy between its resolution and the specified output resolution of 2.5 meters.

References:
Shermeyer, J., & Van Etten, A. (2019). The effects of super-resolution on object detection performance in satellite imagery. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 0.

Comment 3: In section 4.4, the authors used a lot of thresholds (0.5 and 0.2) according to their prior knowledges, however, as for a national-scale mapping method, the authors should give the analysis of why these thresholds are reasonable.

Response: Thank you for the reviewer's comment.

There are two thresholds involved. One is the threshold used to binarize the probability map (value range between 0 and 1) of the output from our STSR-Seg model. The threshold is 0.5 here since it is used to discriminate the two classes, i.e., building or not, during model training. This is a common setting based on the previous research on semantic segmentation (Liu and Tang, 2023). We followed this setup directly.

Another threshold is about thresholding the Dynamic World product. The Dynamic World product provides the probability estimation (value range between 0 and 1) of the built-up area. In Section 4.4, we want to use this product to make an intersection with our result to remove the potential false predictions. Therefore, we need to binarize the Dynamic World by a threshold. This threshold was determined though the following way:

(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. C1. 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.

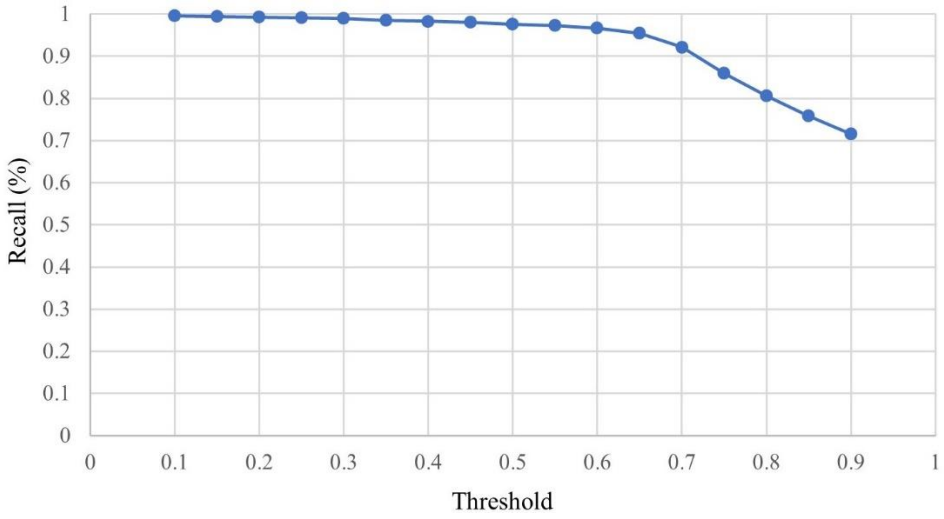


Figure C1: The recall curve for the threshold selection. The threshold interval of 0.05 was used to binarize the Dynamic World product, and it was compared with the building distribution data we collected from 52 cities in China to calculate recall respectively. When the threshold is 0.3, recall reaches 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, and finally we consider 0.2 is the best option.

We now integrate the relating explanations into the revised manuscript (Fig. A4 is the Fig. C1 above in the revised manuscript), from line 379 to 386 on page 16:

……removing the zero padding. (5) A threshold value of 0.5 is used to differentiate between candidate foreground pixels (i.e., building rooftop) and background pixels, following common practice (Liu and Tang, 2023).

…… 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).

References:

Liu, Z. and Tang, H.: Learning Sparse Geometric Features for Building Segmentation from Low-Resolution Remote-Sensing Images, Remote Sens (Basel), 15, 1741, 2023.

Comment 4: In term of temporal-optimization, authors used the temporal checking method (proposed by Li et al. (2014)) for optimize the multitemporal CBAR results. Did the authors consider whether classification errors would be introduced into the optimized results?

Response: Thank you for the reviewer's comment.

We have considered the building's characteristics when utilizing the temporal checking methodology. We assumed that the building would not undergo continuous changes in its state within a 3-year period, such as construction-demolition-construction or demolition-construction-demolition. Therefore, we limited the sliding window to 3 and did not perform iterative filtering by expanding window as in the approach proposed by Li et al. (2015). The filtering method we used was the most conservative, and any errors it introduces should be minimal compared to the accuracy improvement it provides.

We now add our consideration into the revised manuscript, from line 389 to 395 on page 17:

Due to the possible random bias of our method in locating the boundary of building rooftops (outlined in Sect. 6.2), inconsistencies in identification results over time for the same building may occur. To address this issue, a temporal homogeneity check approach has been implemented. Specifically, it is assumed that a building's state does not undergo continuous change over three successive years. Building upon the method proposed by Li et al, (2015), a 3×3 sliding window is utilized to determine the final pixel value through majority voting, as illustrated in Fig. 8. This procedure ensures that the results for adjacent years are comparable. However, for the edge year like 2016 and 2021, they are not checked due to the lack of temporal information.

References:

Li, X., Gong, P., and Liang, L.: A 30-year (1984–2013) record of annual urban dynamics of Beijing City derived from Landsat data, Remote Sens Environ, 166, 78–90, 2015.

Comment 5: As for the validation, I think the comparisons with impervious surfaces products (CLCD, GAIA) might not make sense. Instead, they should pay more attention on the comparison with CBA dataset and add more BAR dataset (such as regional dataset) as comparison dataset.

Response: Thank you for the reviewer's comment.

In our study, we have employed a rigorous approach to validate the proposed dataset CBRA

from two perspectives. Firstly, we have compared CBRA with a previously published dataset, namely 90-cities-BRA (Zhang et al., 2022), which is the only comparable and available datasetup to now. Both datasets were generated using remote sensing imagery through automated procedures, and we have reported the comparison results in Table 4 of the manuscript.

Secondly, since multi-year building rooftop data for China were not previously available, we have employed other relatively coarse resolution impervious surface or built-up area datasets, namely CLCD, GAIA, and Dynamic World, to establish trends and validate our results. Specifically, we have calculated the correlation coefficients between different datasets for the percentage of foreground pixels in a 0.10° x 0.10° grid, as presented in Fig. 15 of the manuscript. This validation approach has been utilized in other studies as well (Yang and Huang, 2021) and is considered a reasonable method to establish the validity of our dataset.

Unfortunately, we could not find any multi-annual regional datasets that could be used for fair comparison with our proposed dataset. While there are some building datasets available in the public domain, most of these datasets are created through human-labelling for competitions or method development, and do not contain crucial geo-information such as the coordinate system (Li et al., 2022). Therefore, in China, the only comparable dataset available is the 90-cities-BRA dataset.

References:

Zhang, Z., Qian, Z., Zhong, T., Chen, M., Zhang, K., Yang, Y., Zhu, R., Zhang, F., Zhang, H., and Zhou, F.: Vectorized rooftop area data for 90 cities in China, Sci Data, 9, 1–12, 2022.

Yang, J. and Huang, X.: The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019, Earth Syst Sci Data, 13, 3907–3925, 2021.

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.

Comment 6: The quantitative analysis in Section 5.1.1 is interesting, can you give the reasons why the BAR dataset achieved higher accuracy in rural areas than that of the urbans.

Response: Thank you for the reviewer's comment.

We believe this comes mainly from the difference between the urban test sample and the rural test sample we use. For urban areas we used a reliable survey-based sample for testing, containing 245,458 buildings. These samples are very accurate and detailed.

However, since it is still very challenging to find reliable test samples in rural areas, we collected 33,736 building rooftops from Open Street Map and manually corrected them using high-resolution images. We labeled more obvious buildings due to limitations in image resolution. Consequently, the rural test sample was relatively less difficult for the model to identify compared to the urban sample. The report accuracy shows higher accuracy in rural areas than in urban areas. The same conclusion was drawn for GHSL in our benchmark. Its recall value in rural areas was also

higher than in urban areas.

In the original manuscript, we mainly conducted product-to-product comparisons without delving into the differences between our products in urban and rural areas. We concur it is necessary to illustrate this. Therefore, we added a description on this aspect to Section 5.1.1 in the revised manuscript, from line 448 to 451 on page 19:

……However, the CBRA is very close to it (78.94%), with a gap of only 1.95%, indicating its reliability in predicting building rooftops in rural areas. Considering the varieties of urban and rural test samples, it should be mentioned that the presented results in Table 4 intend to compare product-to-product, rather than to demonstrate performance differences between urban and rural areas.

Comment 7: The figure 10 illustrated that the CBAR dataset has obvious advantages than BRA dataset in their previous study. However, why the BRA has high geometry accuracy in top left corner of the Figure 10c and suffers obvious offset in the central areas (red circle) over such local area.

Response: Thank you for your comment.

The compared data (i.e., the 90-cities-BRA in Fig. 10c) is produced from Google Earth Satellite (GES) imagery. When acquiring larger scale GES images, stitching of images from multiple sensors is generally required. When stitching, due to the elevation change of the topology, there will be large errors at the stitching regions, especially in places with large height fluctuations (e.g., high-rise buildings). Examples are shown in the below, and the images are captured from Google Earth Pro software.

(a) (b)

Figure C2: 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.

In Fig 10 of our manuscript, the red circle designates the high-rise buildings, highlighting that the bias is more prominent than in the surrounding low-rise buildings. Given that the 90-cities-BRA manuscript by Zhang et al. (2022) did not provide additional information on the assessment of image quality, we speculate that their errors may have arisen due to the aforementioned reason.

Details regarding this phenomenon are introduced in the revised manuscript (Fig. A1 is the Fig. C2 above), line 159 on page 6:

However, the GES images are collected from various kinds of high-resolution satellites, and have two potential problems when applied to large-scale mapping: (1) inconsistent geographical offset (illustrated in Fig. A1), and (2) inconsistent acquisition time (e.g., the image is obtained from various satellite sensors with different acquisition times) which results in spatio-temporal inconsistency in the generated product.

References:

Zhang, Z., Qian, Z., Zhong, T., Chen, M., Zhang, K., Yang, Y., Zhu, R., Zhang, F., Zhang, H., & Zhou, F. (2022). Vectorized rooftop area data for 90 cities in China. Scientific Data, 9(1), 1–12.

Comment 8: The temporal analysis in Section 5.2 should be strengthen, as the increased/decreased BRA is great small than the stable BRAs, combining the stable and changed BRAs for analyzing the accuracy metrics cannot illustrate the performance of proposed method in the temporal dimension. For example, you can use the changed validation points to analyze the changed CBRAs.

Response: Thank you for the reviewer's comment.

We appreciate the significance of using updated validation points in our research; however, implementing such a validation process has posed several challenges.

Firstly, identifying validation samples from publicly available Sentinel or Landsat imagery through manual labeling is feasible, with the exception of building rooftops, where the task is complicated by the unavailability of year-by-year high-resolution images from 2016-2021.

Secondly, although we can obtain some imagery from Google Earth, the inability to interact with commonly used GIS software, such as ArcGIS or QGIS, makes human labeling difficult. Additionally, Google Earth lacks images from certain years, further restricting our ability to carry out the labeling process.

Thirdly, we have been unable to locate any annual building rooftop or footprint datasets, including regional data, thereby hindering our ability to perform thorough testing of the CBRA using appropriate samples.

Therefore, in the original manuscript (Section 5.2), we assume that the buildings in the old town area of Beijing, Hong Kong and Macao are stable, and we use these samples to test our annual performance. Besides, we calculate the correlation coefficient of the CBRA and other annual products to check the trend consistency.

It is important to note that the assumption of building stability may not be entirely accurate, and we acknowledge that this may introduce some level of uncertainty in our validation process. However, given the limitations outlined earlier, we believe that utilizing the stable buildings in these areas was the most reasonable approach to validate our model's performance annually.

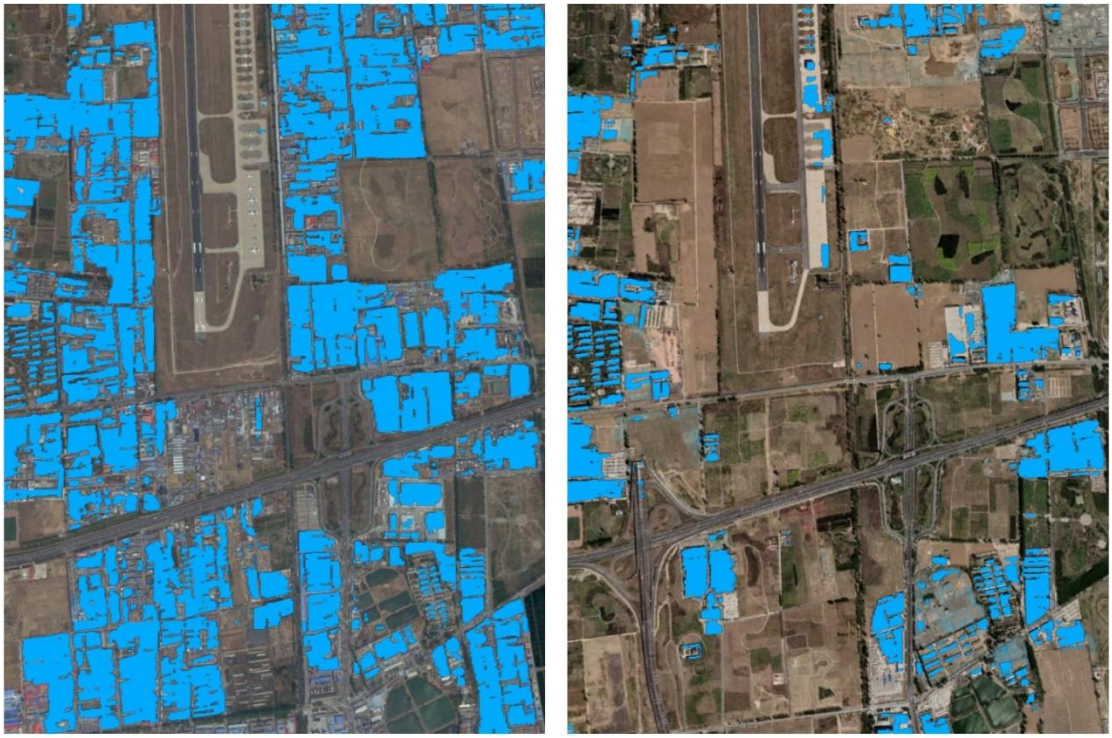
Comment 9 #1: Figure 18 is interesting, and two enlargements intuitively show the good performance of the CBAR. I suggest the authors added the high-resolution imagery over two enlargements to make the analysis more intuitive.

Response: Thank you for your comment.

Two enlargements are shown below (Fig. C3 and C4), which contain the demolition process and construction process between 2016 and 2021. Fig. C4 is a zoomed part of Fig. 18(b) in the manuscript. It shows the removal of the block consisting shack houses. Figure C5 is the zoomed part of Fig. 18(c) in the manuscript, it shows the construction of the modern apartments.

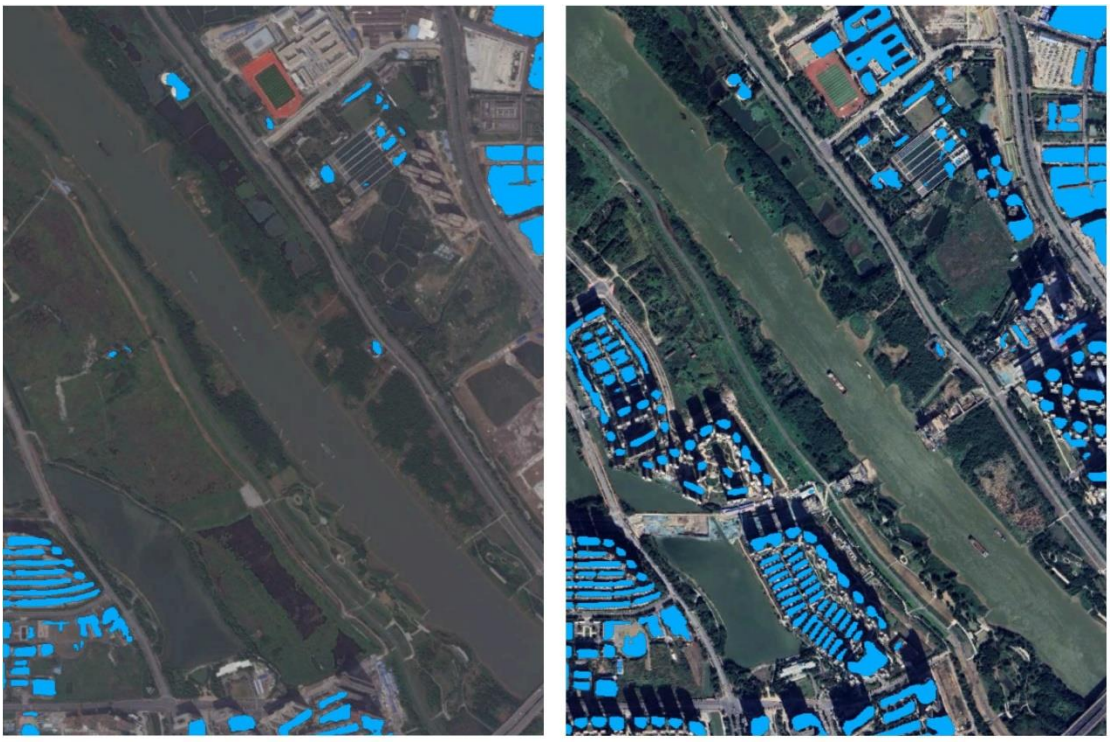
To increase the visibility, we pick the high-resolution images with high quality and large scale in the Google Earth Pro software. Owing to their magnified dimensions, they could not be accommodated in the primary body of the manuscript and were therefore included in the Appendix, and add some descriptions about them in the manuscript (Fig. A5 and A6 in the revised manuscript refer to the Fig. C3 and C4 in the author response), line 550 on page 25 :

……in the rural area, and the establishment of buildings (e.g., apartments) in the urban area, respectively. More comprehensible references about the building change can be found in Fig. A5, A6 and A7.. ……



(a) (b)

Figure C3: A zoomed result of the demolition process with high-resolution images (116.387674°E, 39.766427°N). (a) The identified building rooftop area of 2016. (b) The identified building rooftop area in 2021. Imagery © 2023 Maxar Techonlogies.



(a) (b)

Figure C4: A zoomed result of the construction process with high-resolution images (113.037558ºE, 23.014343ºN). (a) The identified building rooftop area of 2016. (b) The identified building rooftop area in 2021. Imagery © 2023 Maxar Techonlogies.

Comment 9#2: In addition, why the North China Plain (especially in Henan province) shows such a marked increase? An enlargement in Henan should be added.

Response: Thank you for your comment.

The North China Plain mainly contains three provinces that contain Hebei, Henan, Shandong provinces and two municipality Beijing and Tianjin. In the previous research conducted by Gong et al, (2019), the top urban expansion provinces of China are Shandong, Jiangsu, Hebei, Guangdong, and Henan by 2017. The results of our analysis show the consistent trend, i.e., the roof area of buildings in the North China Plain, which is composed of Shandong, Hebei and Henan, produced a large increase from 2016-2021.

Regarding Henan province, we collected statistical data from the National Bureau of Statistics of China on the floor space of commercial and residential buildings sold between 2016 and 2021. Our analysis showed that Henan ranked fifth in the country in terms of commercial and residential building floor space sold, while Shandong ranked third. Although buildings in China are primarily tall buildings such as apartments, this indicator is related to building rooftop area and thus provides insight into the growth of building construction in Henan.

Table C1: The top 5 change of sold floor space of commercialized and residential buildings (10000 sq.m) between 2016 and 2021 (from National Bureau of Statistics of China)

Province	Change of floor space of commercialized buildings sold between 2016 and 2021	Change of Floor Space of Residential Buildings Sold between 2016 and 2021	Total change
Sichuan	4392.44	3028.05	7420.49
Jiangxi	2984.37	2540.71	5525.08
Shandong	2482.97	2033.47	4516.44
Jiangsu	2589.73	1703.84	4293.57
Henan	1970.92	2121.70	4092.62

Furthermore, the CBRA dataset we used includes various types of buildings, including residential, commercial, industrial, and agricultural. Henan is a major province for agriculture and industry, with the third population of 988 million in 2021 in China, making it more likely that building activities will be more frequent and have a greater impact on building rooftop area.

Lastly, we examined high-resolution images provided by Google satellite imagery alongside the CBRA results but did not observe any distinctive features of building change, such as a significant increase in a particular type of building. However, we noted that Henan Province is located in a plain with numerous human settlements, and each settlement experiences changes in buildings that contribute to the overall increase in the rooftop area of buildings in the province.

It is important to note that we did not conduct a more detailed analysis of the building change process as it would require additional methodologies or auxiliary data beyond the scope of this manuscript, Therefore, our focus was mainly on the overall changes in building rooftop area across China and major city clusters. The following Fig. C5 is an enlargement of the Henan province, and two examples of the new construction buildings between 2016 and 2021.

We add this figure to the revised manuscript (Fig. A7 is the Fig. C5 in the author response), line 550 on page 25:

……in the rural area, and the establishment of buildings (e.g., apartments) in the urban area, respectively. More comprehensible references about the building change can be found in Fig. A5, A6 and A7.. ……

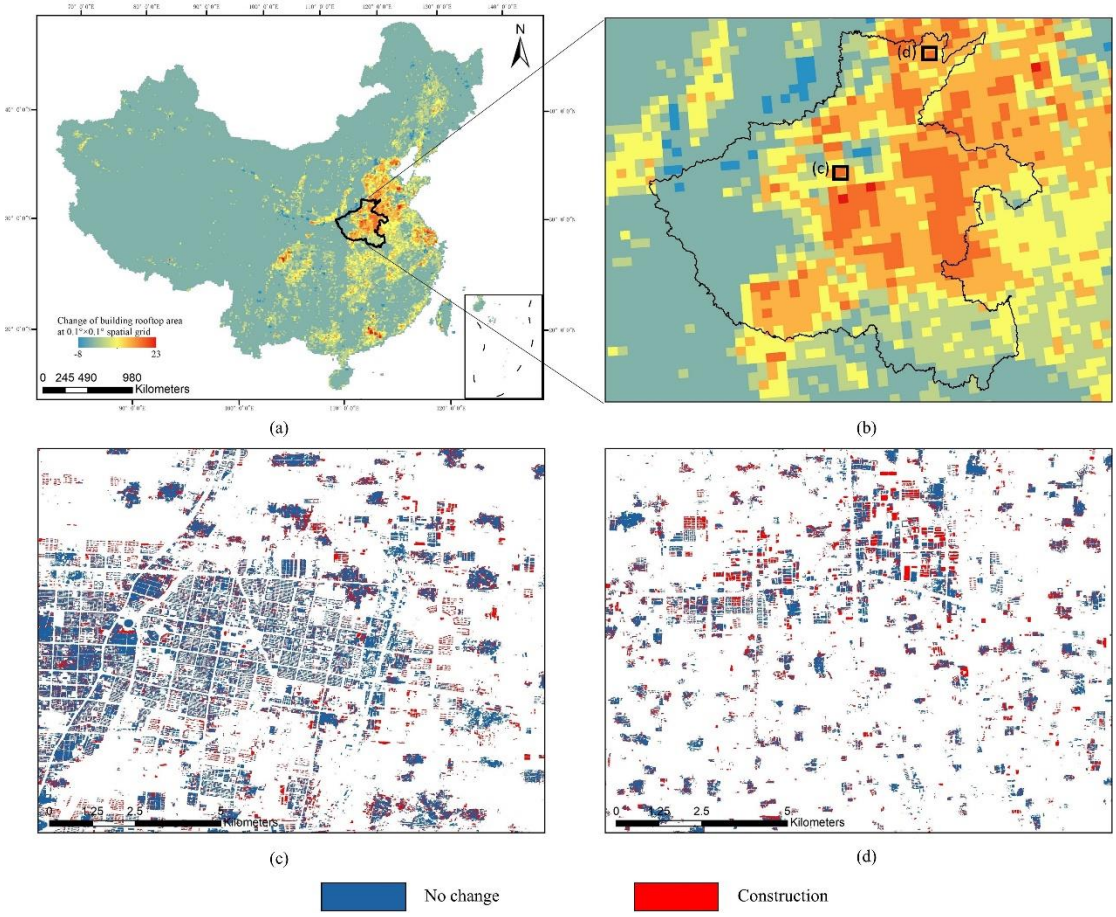


Figure C5: Enlargement of the Henan province of China. (a) The spatial distribution of the annual change of building rooftop area over the period of 2016-2021. The area fraction is aggregated within the $0.10° \times 0.10°$ spatial grid (base map © OpenStreetMap contributors 2023. Distributed under the Open Data Commons Open Database License (ODbL) v1.0). (b) The enlargement of the Henan province. (c) and (d) are two examples of the new construction of buildings from 2016 to 2021, located at "115.087461ºE, 35.769057ºN" and "113.686257ºE, 34.517389ºN", respectively.

References:

Gong, P., Li, X., & Zhang, W. (2019). 40-Year (1978–2017) human settlement changes in China reflected by impervious surfaces from satellite remote sensing. Science Bulletin, 64(11), 756–763.

Comment 10: In Line 383, the Radoux et al. (2014) mainly emphasized the spatial heterogeneity, while this part focuses on the temporal consistency, so the reference might be incorrect.

Response:

Thank you for your careful inspection. We agree that the reference to Radoux et al. (2014) is not appropriate for the point we were trying to make. To address this issue, we have revised the text and removed the reference to Radoux et al. (2014) in that sentence, please refer to line 389 to 395 on page 17:

Due to the possible random bias of our method in locating the boundary of building rooftops

(outlined in Sect. 6.2), inconsistencies in identification results over time for the same building may occur. To address this issue, a temporal homogeneity check approach has been implemented. Specifically, it is assumed that a building's state does not undergo continuous change over three successive years. Building upon the method proposed by Li et al, (2015), a 3×3 sliding window is utilized to determine the final pixel value through majority voting, as illustrated in Fig. 8. This procedure ensures that the results for adjacent years are comparable. However, for the edge year like 2016 and 2021, they are not checked due to the lack of temporal information.