We thank the editor and reviewers for their detailed and constructive comments and suggestions. We would love to thank you for allowing us to resubmit a revised copy of the manuscript and we highly appreciate your time and consideration. We have incorporated the review comments and revised the manuscript thoroughly. The review comments and revision have made the study more accurate and complete. We hope the revised manuscript is acceptable for publication. While the changes made can be seen in the revised manuscript, we also present here our detailed responses to the review comments.

****

Response to Editor:

We kindly ask you to provide a detailed point-by-point response to all referee comments and specify all changes in the revised manuscript. The response to the Referees shall be structured in a clear and easy to follow sequence: (1) comments from Referees, (2) author's response, (3) author's changes in manuscript. In addition, please provide a marked-up manuscript version showing the changes made (using track changes in Word or latexdiff in LaTeX). This version should be combined with your response file so that the Topical Editor can clearly identify what changes have been made.

Response:

Many thanks for your comments and suggestions. According to your comments and comments from the other reviewers, we have done a lot of works to revise and improve the paper.

The main revisions of the paper are:

1. The content of the Methods have been supplemented to improve understandability, including the structure of the ECA (in 3.1.1), the rationale of Point Head (in 3.1.2), and some details of discriminator (in 3.1.3);

2. The analysis of the result have been elaborated to upgrade the readability and approval of the results, from both quantitative and qualitative perspectives. In the quantitative analysis, more cities have been involved for accuracy evaluation (in 4.1); and in the qualitative analysis, the influence of external factors to UGS results have been discussed.

3. Ablation study on the discriminator have been added in the Discussion (in 5.2), in order to verify the effectiveness of the proposed framework;

4. Improvements were made to address shortcomings mentioned by the reviewers, including insufficient references, details in the figures, network configurations, etc.

****

Response to the Referee #1:

Referee: #1

Main ideas:

This paper provides a high-resolution urban green space (UGS) maps for 34 megacities in China, as well as a UGS dataset for the deep learning models. The paper is complete in structure and the results the results have shown the effectiveness of the proposed deep learning framework. I think the paper can contribute to UGS research in terms of both algorithm, dataset and products. However, some revisions still should be made to the manuscript before considering publication:

Response:

We feel great thanks for your affirmation and careful review work on our article. Your professional suggestions are very helpful to make our study clearer and more comprehensive. According to your nice suggestions, we have made extensive corrections to our previous draft, the detailed corrections are listed below. The detailed point-by-point responses are listed below.

****

1. The caption of Figure 1 should be expanded in combination with its content to improve understanding.

Response:

Thanks for pointing this out. We have expanded the caption of Figure 1 to improve readability. The caption of Figure 1 has been changed into:

"Figure 1. Diagram of the deep learning framework to generate UGS-1m. a) Pre-train the proposed UGSNet on the UGSet dataset; b) Optimize the generator (initialized by UGSNet) to different target cities with a discriminator through adversarial training; c) Apply each optimized generator to corresponding target city for large-scale mapping." (Line 92, Page 4).

****

2. The descriptions of the reminders of the paper (Line 100) should be revised, while the Conclusions are arranged in Section 7.

Response:

Thank you for your reminder. We have changed the corresponding descriptions accordingly:

"… The access to the code and data is provided in Sect. 6. Finally, conclusions will be made in

Sect. 7.” (Line 102, Page 5).

****

3. In the deep learning framework, the author use “four enhanced Coordinate attention (ECA) modules (Hou et al., 2021) to enhance feature representations” (Line 164). The descriptions of the ECA is simple with only a Figure 7. Is there any difference between the ECA and that in the reference (Hou et al., 2021)? Please elaborate.

Response:

Thanks for your valuable comments. The difference between the ECA and CA in *(Hou et al., 2021)* have been elaborated. Corresponding statements can be found at:

“… The four residual blocks are connected by four enhanced coordinate attention (ECA) modules to enhance feature representations.

Previous researches have proved that attention mechanism can bring gain effects to deep neural networks (Vaswani et al., 2017; Woo et al., 2018). Recently, a novel "coordinate attention" (CA) (Hou et al., 2021) was proposed, which improved the weakness of traditional attention mechanisms in obtaining long-range dependence by embedding location information efficiently. Specifically, in order to capture the spatial coordinate information in the feature maps, the CA uses two 1-Dimensinal (1D) global pooling layers to encode input features along the vertical and horizontal directions, respectively, into two direction-aware feature maps. However, this approach ignores the synergistic effect of features in two spatial directions. Therefore, we propose the enhanced coordinated attention (ECA). In addition to the original two parallel 1D branches encoding long-distance correlation along the vertical and horizontal direction, respectively, ECA also introduces a 2D feature encoding branch to capture the collaborative interaction of feature maps in the entire coordinate space, so as to obtain a more comprehensive coordinate-aware attention maps for feature enhancement. The structure of the ECA module is shown in Figure 7.” (Line 165-179, Page 9-10)

****

4. The Point Head in 3.1.2 sounds interesting. However, some details are not clarified. For example, how does the N sampling points obtain? And what is the rationale behind the Point Head to improve accuracies? Besides, the value of N is missed in the paper.

Response:

Thanks for your valuable comments. According to your suggestions, we have supplemented necessary details in Sect. 3.1.2.

The description of how to obtain the N sampling points is as follows:

“In the first step, how to adaptively select sample points is the key to improve the segmentation results in an efficient and effective way, so different sampling strategies are adopted in the training and inference process. At the training stage, different points are expected to be taken into account. Therefore, at first k × N points will be randomly generated from the coarse segmentation results as candidates; then, β × N (β ∈ [0,1]) points with highest uncertainty will be selected from the k × N ones; after that, the other (1 − β) × N points will be randomly selected from the remaining candidates to supplement. In the inference process, the N sampling points are directly selected from the candidate points with highest uncertainty to consider more hard points.” (Line 188-194, Page 11).

The rationale behind the Point Head have been supplemented:

“Many semantic segmentation networks directly sample high-dimensional features to obtain segmentation results of original image size, which will lead to rough results, especially near the boundary. Therefore, the point head is introduced into UGSNet, which uses the point rending strategy (Kirillov et al., 2020) to get fine-grained UGS results efficiently. Specifically, given the coarse UGS results from the backbone, the specific process in the point head includes the following three steps: 1) firstly, collect N sampling points with lower certainty; 2) then, construct point-wise features of the selected N points based on the coarse UGS results and fine-grained features from the backbone; 3) finally, reclassify the results of the selected N points through a simple multilayer perceptron (MLP). Detailed information of each step will be elaborated in the following. ” (Line 181-187, Page 10-11).

The value of N is provided:

“In our experiments, N=1024 sampling points will be collected, and the value of k and β are 3 and 0.75, respectively.” (Line 198, Page 11).

****

5. What does “D“ denote in (4) and (5)?

Response:

Thanks for your valuable comments. The “D” in the initial (4) and (5) denotes the discriminator. Similarly, we used “G” to represent the generator for simplicity. For clarity, we have added supplementary descriptions to these sentences:

"(1) Taking the pre-trained UGSNet as the start training point of the generator, the Is and It are forward to the generator *G* to get their prediction result Ps and Pt, which can be denoted as

$$P_s, P_t = G(I_s), G(I_t) \qquad (4)$$

" (Line 230, Page 13)

"(2) Input Ps and Pt into the discriminator *D* in turn to distinguish the source of the inputs;" (Line 232, Page 13)

"(3) According to the judgement result, the discriminator *D* will be optimized first, which can be denoted as …" (Line 233, Page 13)

****

6. Some training details and the network configuration are missed. For example, the details of the data augmentation strategies.

Response:

Thank you for underlining this deficiency. Important training details and network configurations have been added. Corresponding statements can be found at:

"… Data augmentation were applied during model training, including randomly clipping, rotation, and flipping. After training, all selected models were compared on the test set." (Line 250, Page 13).

****

7. The Section 4 should be elaborated to upgrade the readability and approval of the results, which is brief relatively at present.

Response:

Thanks for your valuable comments. According to your suggestions, we have elaborated Section 4 (Page 14) to improve the readability and approval of the results. The main changes are as follows:

(1) The results are analyzed from both quantitative and qualitative perspectives;

(2) In the quantitative analysis, in addition to Guangzhou, we have added four cities for accuracy evaluation (including Changchun, Beijing, Wuhan, and Lhasa);

(3) In the qualitative analysis, we further analyzed the performance of UGS extraction as well as its relationship with external factors, such as geographical location, UGS types, phenological phase, for etc.

****

8. The introductions of Figure 12 (Line 260) should also be extended.

Response:

Thanks for your valuable comments. We have tried to extended the analysis of Figure 14 (the Figure 12 in the old version). Corresponding statements can be found at:

"Figure 14 further demonstrates the performance of different methods on different kinds of UGS. It can be seen that after fully training on UGSet dataset, each model can identify the approximate region of various green spaces, including SegNet, which has shown poor performance in quantitative comparisons. Therefore, the superiority of green space identification results is mainly reflected in two aspects. One aspect is the ability to extract UGS of great inter-class similarity. As shown in the first row of Figure 14, the UGSNet can accurately identify the yellow box area, in which the green space of park has confused most comparative methods. Another aspect is the capability to grasp fine-grained edges, especially for small-scale UGS, such as the attached UGS in the last row of Figure 14." (Line 321-327, Page 18)

****

9. There are several typos and statements without any references or evidence. For example, the references of the "Guidance of the General Office of the State Council on Scientific Greening" (Line 108) should be provided.

Response:

Thanks for your valuable comments. According to your suggestion, we have revised corresponding statements and added necessary citations, which can be found at

"In recent years, in order to satisfy the concept of ecological civilization and sustainable development, scientific urban green space planning and management have been paid more and more attention in China (General Office of the State Council, PRC, 2021). Therefore, how to improve the rationality of UGS classification system and layout distribution to build a healthy and livable city has been the focus of government and scholars in recent years (Ministry of Housing and Urban-Rural Development, PRC., 2019; Chen et al., 2022). To this end, this paper selects 34 major cities/areas in China as study area, aiming to construct a comprehensive UGS dataset for deep learning model training under the official classification system, and generate high-resolution green space mapping for each city/area." (Line 105-111, Page 5).

References**:**

- Chen, B., Tu, Y., Wu, S., Song, Y., Jin, Y., Webster, C., Xu, B., and Gong, P.: Beyond green environments: multi-scale difference in human exposure to greenspace in China, Environment International, 166, 107 348, 2022.
- Ministry of Housing and Urban-Rural Development, PRC: Urban Green Space Planning Standard (GB/T51346-2019), https://www.mohurd.gov.cn/gongkai/fdzdgknr/tzgg/201910/20191012_242194.html (accessed April 9, 2019), 2019.
- General Office of the State Council, PRC: Guidelines on scientific greening, [Online], https://www.mee.gov.cn/zcwj/gwywj/202106/ t20210603_836084.shtml (accessed June 3, 2021), 2021.

****

10. For writing, the manuscript needs proofreading.

Response:

We are sorry for the deficiency. And we have made full-text inspection to rectify the spelling and grammar errors in the article.

****

****

Response to the Referee #2:

The paper makes a clear contribution to the field of urban green space mapping by providing a new annotated dataset of urban green spaces, a new deep learning framework and a detailed green space map of 34 cities in China. However, I still have several question/comments that need to be addressed before considering publication.

Response:

We really appreciate your professional review work on our article. As you are concerned, there are several problems that need to be addressed. According to your useful suggestions, we have made extensive corrections to our draft as soon as possible. The detailed point-by-point responses are listed below.

****

General remarks

- One of the key contributions of the paper is the use of adversarial training is the use of adversarial training to increase the generalisation capacity of the deep learning model. Yet the actual added value of the adversarial part of the model is only addressed to a limited extent. If I understand correctly (this should also be clarified in the paper), the other semantic segmentation models were not finetuned using adversarial training. This means that one could deduce the effectiveness of the adversarial training approach from the difference in accuracy with the other models. But this will always leave the question: is the improvement related to the model structure of the generator or due to the inclusion of a discriminator? In short, I would have liked to see the difference in performance with/without the discriminator.

Response:

Thanks for your valuable comments. We are really sorry that the problems you considered were not clearly clarified in the previous version, and we have made corresponding supplements in the new version. According to your question, we have elaborated the following two points:

(1) The improvement is related to the model structure of the generator, since the comparative experiments with other deep learning models (in Sect. 5.1) does not include a discriminator.

The proposed framework is composed of a generator and a discriminator, in which the adversarial training was adopted to help model transfer learning, as you have mentioned above. Before the start of adversarial training, the generator will be initialized by a pre-trained UGSNet on UGSet, which means that the generator actually has the same model structure as the UGSNet.

Therefore, in the discussion section (Sect. 5.1.3) of the previous version, we compared the performance of UGSNet and other deep learning models on the UGSet. This stage is actually a fair comparison of candidate generators in the pre-training stage, and no discriminator is involved. Therefore, the precision improvement of UGSNet in Table 3 over other models can be fully attributed to its model architecture, which also proved the superiority of the UGSNet as the generator.

To clarify this point more clearly, we modified the sub-title of Sect. 5.1 into “5.1 Comparative experiments at pre-training stage”, and added necessary statements at the beginning of Sect. 5.1:

“As we have mentioned above, before the start of adversarial training stage, the generator will be initialized by a pre-trained UGSNet on UGSet. Therefore, in order to fully verify the advancement of UGSNet and its qualification to initialize the generator, this section introduces several state-of-the-art (SOTA) deep learning models as candidate generators for comparison. Noted that the comparative experiment is completely conducted on UGSet, and no discriminator is introduced. After all the models have been fully trained on training set of the UGSet, the best-trained model of each model will be evaluated on the testing set of the UGSet. The comparative results are provided in Table 3." (Line 309-314, Page 17-18).

(2) The difference in performance with/without the discriminator have been added in the "Sect 5.2 Ablation study on the discriminator" of the revised version.

Thank you for pointing out this problem. In the revised version, we have supplemented a new experiment on with/without the discriminator in the newly added Sect 5.2. Corresponding statements can be found at:

"As we have mentioned above, the proposed framework is composed of a generator and a discriminator, which adopts the adversarial training to help model transfer learning. In order to test the effectiveness of the proposed framework, this section further conducted ablation experiments on the with and without the discriminator, which respectively correspond to:

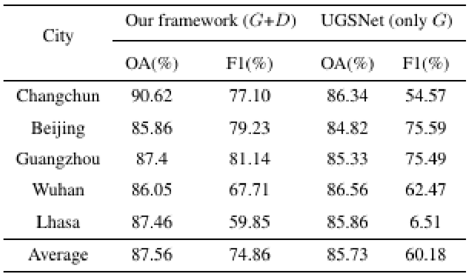
(1) Our framework (G+D): contains a generator and a discriminator, in which the generator is initialized by the UGSNet pre-trained on UGSet, and the discriminator is employed at the adversarial training stage to overcome domain shifts and obtain a refined UGSNet for each target city/area, before generating the UGS map for it;

(2) UGSNet (only D): no discriminator is involved, simply applying the pre-trained UGSNet to each target city/area and generate their UGS maps, regardless of the domain shifts between the UGSet and images from different target cities/areas.

The result of "Our framework (G+D)" comes from quantitative results of UGS-1m in Sect. 4. In order to test the effect of "UGSNet (only G)", the pre-trained UGSNet is applied to the same sample areas for accuracy evaluation. The final ablation results are shown in Table 4. It can be seen from the results that when the discriminator is not used, the OA of almost all cities decreases to a certain extent. Generally speaking, the average OA decreases from 87.56% to 85.73%. The F1 score shows a sharp decline, with the average F1 score dropping from 74.86% to 60.18%.

Specifically, the decline of F1 score in Guangzhou and Beijing is relatively small, which indicates that the difference between the images of these two cities and the UGSet images is not that significant. Therefore, the pre-trained model can capture some UGS. It is worth noting that the use of discriminator can significantly improve the results in Changchun, according to the great growth of F1 score of 22.53%. Moreover, the results in Lhasa, only have an F1 score of 6.51% without D, which can reach 59.85% when using D. The ablation experiment fully proves the effectiveness and potential of the proposed framework for large-scale green space mapping. ”(Line 329-346, Page 20-21).

Table 4. Ablation study on the discriminator(D).



City	Our framework (G+D)		UGSNet (only G)	
	OA(%)	F1(%)	OA(%)	F1(%)
Changchun	90.62	77.10	86.34	54.57
Beijing	85.86	79.23	84.82	75.59
Guangzhou	87.4	81.14	85.33	75.49
Wuhan	86.05	67.71	86.56	62.47
Lhasa	87.46	59.85	85.86	6.51
Average	87.56	74.86	85.73	60.18

****

- Related to the previous remark I would have liked a description of the spatial differences in accuracy in the discussion section. Do some cities show a lower accuracy? Could we explain this by the type of green space that is dominant? Difference in phenological phase when the imagery was taken?...

Response:

Thanks for your valuable comments. According to your suggestions, in addition to Guangzhou, we have added four cities for precision verification to test the effect of green space extraction (including Changchun, Beijing, Wuhan, and Lhasa) and further analyzed the impact of different factors on the UGS results from both quantitative and qualitative perspectives.

After analysis, we found that the extraction result has nothing to do with the geographical location. In addition, we also found that due to the influence of image taking angle and phenological phase, some cities have shown lower accuracy in quantitative analysis. The analysis and discussion are supplemented in Sect. 4 of the manuscript.

Some important statements can be found in:

“The evaluation results are summarized in Table 2, which are evaluated by OA, Pre, Rec and F1. It can be seen that in the five cities for verification, the average OA in all cities is 87.56%, while the

OA of each city is higher than 85%. Among them, the highest OA is 90.62% in Changchun, while the lowest OA also reaches 85.86% in Beijing, indicating that the UGS results in different cities is basically good. In terms of F1 score, Guangzhou has the highest F1 score of 81.14%, followed by Beijing and Changchun with the F1 of 79.23% and 77.23%, respectively. Though the F1 scores of Wuhan and Lhasa are relatively low, of 67.71% and 59.85%, respectively, the average F1 score of the final UGS results also reaches 74.86%. Moreover, the average Recall of 76.61% also denotes a relatively low missed-detection rate of the UGS extraction results, which is significantly important in applications. In general, after quantitative validation in several different cities, the availability of UGS-1m is preliminarily demonstrated. ” (Line 276-284, Page 15).

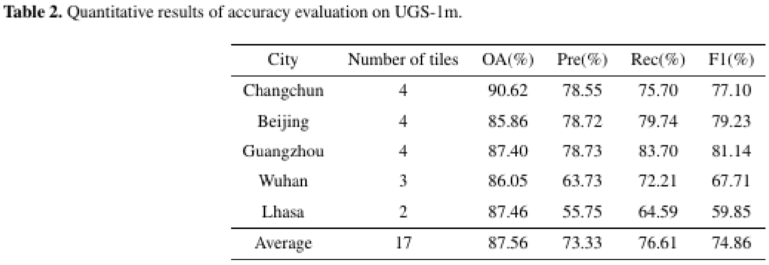


Table 2. Quantitative results of accuracy evaluation on UGS-1m.

City	Number of tiles	OA(%)	Pre(%)	Rec(%)	F1(%)
Changchun	4	90.62	78.55	75.70	77.10
Beijing	4	85.86	78.72	79.74	79.23
Guangzhou	4	87.40	78.73	83.70	81.14
Wuhan	3	86.05	63.73	72.21	67.71
Lhasa	2	87.46	55.75	64.59	59.85
Average	17	87.56	73.33	76.61	74.86

“From the overview image of Changchun and Guangzhou (Figure 11 and Figure 12), it can be seen that the extracted UGS results are in good agreement with the reference map, which is mainly reflected in the good restoration of UGS of various scales in each example image. The zoom-in area of each image further shows the details of UGS-1m for extracting different kinds of UGS, including park, square, green buffer, as well as the attached green space. Specifically, the UGS-1m performs well in the extraction of green space attached to residential buildings, although they are complex and broken in morphology compared to other UGS types. Notably, although Changchun and Guangzhou are geographically far away, distributed in the northernmost and southernmost regions of China respectively, the UGS results in these two cities are both good. This shows that the performance of the proposed USG extraction framework is unlikely to be affected by the difference of geographical location, which may attribute to the adversarial training strategy to model transferring.” (Line 289-297, Page 15).

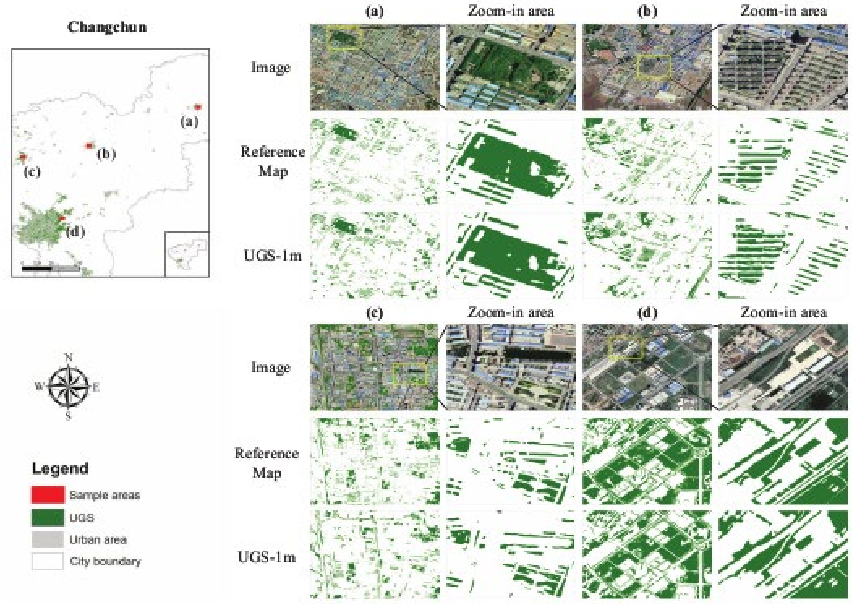


Figure 11. Qualitative analysis on UGS-1m: case study in Changchun City. (a)-(d) are four example areas collected from Changchun (Images © Google Earth 2020).

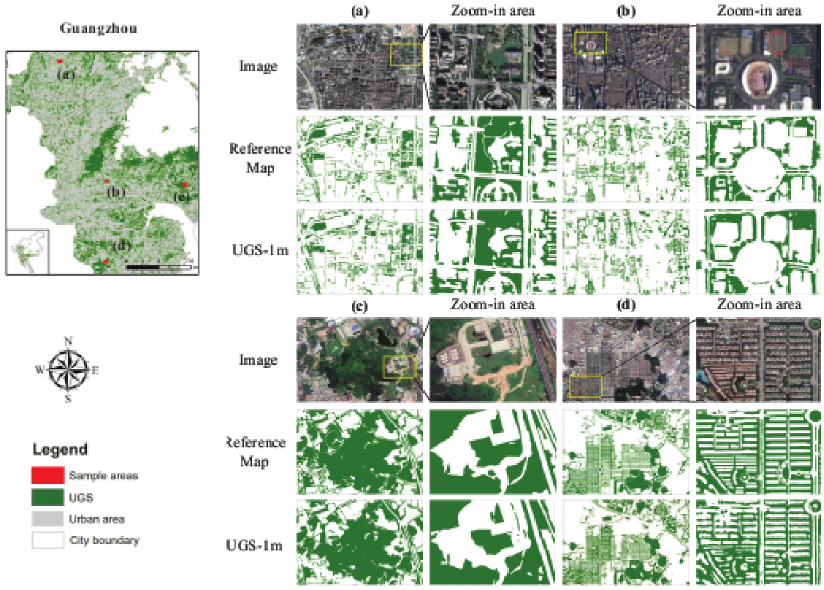


Figure 12. Qualitative analysis on UGS-1m: case study in Guangzhou City. (a)-(d) are four example areas collected from Guangzhou (Images © Google Earth 2020).

“The visualization result of Wuhan is further provided in Figure 13 for analysis. The UGS extraction results in Wuhan are mainly influenced by the shadow of buildings. On the one hand, the UGS features are sometimes blocked by building shadows, resulting in relatively poor extraction effect, such as the zoom-in area of Figure 13 - (b). On the other hand, the building shadows can easily be extracted as attached green space, according to Figure 13 - (c). This shows that the result of green space extraction is related to the image taking angle. When the angle is larger, it is more likely to have building shadows in the image and thus affecting the subsequent UGS extraction, especially the green space that attached to buildings. In addition, the results are also affected by phenological phase, as shown in Figure 13 - (a). On the whole, it can be seen that the UGS with higher and denser vegetation canopy is easier to be identified accurately, and on the contrary, the lower and sparser UGS is more easily to be misclassified due to the similar appearance with other land types, such as the bare land." (Line 298-306, Page 16-17).

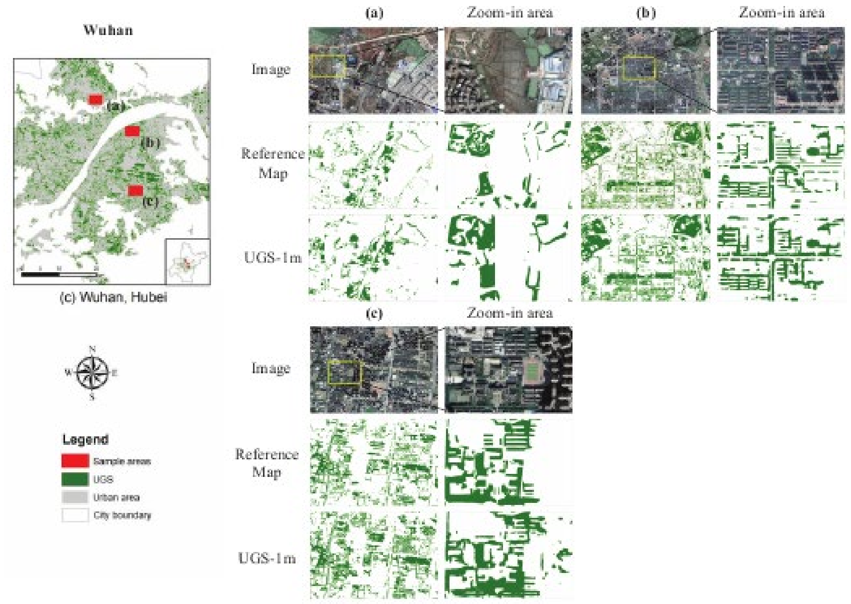


Figure 13. Qualitative analysis on UGS-1m: case study in Wuhan City. (a)-(c) are three example areas collected from Wuhan (Images © Google Earth 2020).

****

- The discussion of the type of errors in the resulting green space map is very limited. Do you mainly notice problems at the edges of green spaces? For high/low or dense/sparse vegetation?

Response:

Thank you for underlining this deficiency. We have elaborated the discussion the type of errors of the UGS results in Sect. 4, from the perspective of the UGS types, and the vegetation canopy. Specifically, under the great influence of building shadows, the green space that attached to buildings is more likely to be the main type of errors. Besides, the UGS with higher and denser vegetation canopy is easier to be identified accurately, while the lower and sparser UGS is more easily to be misclassified.

Corresponding statements can be found in:

“The UGS extraction results in Wuhan are mainly influenced by the shadow of buildings. On the one hand, the UGS features are sometimes blocked by building shadows, resulting in relatively poor extraction effect, such as the zoom-in area of Figure 13 - (b). On the other hand, the building shadows can easily be extracted as attached green space, according to Figure 13 - (c). This shows that the result of green space extraction is related to the image taking angle. When the angle is larger, it is more likely to have building shadows in the image and thus affecting the subsequent UGS extraction, especially the green space that attached to buildings.” (Line 298-304, Page 16-17).

“On the whole, it can be seen that the UGS with higher and denser vegetation canopy is easier to be identified accurately, and on the contrary, the lower and sparser UGS is more easily to be misclassified due to the similar appearance with other land types, such as the bare land.” (Line 304-306, Page 17).

****

- Are there any remarks in relation to the dataset? For example, in figure 12 image 3 (starting from above) I noticed that the residential area has been indicated as non-vegetated while there is clearly vegetation next to the buildings.

Response:

Thank you for underlining this problem. As analyzed above, the extraction of attached green space can be more easily to be affected by external factors, like image taking angle, and the process of annotation is the same. Therefore, after manual annotation and inspection, there may still be a small number of missing labels in the dataset, especially for attached green spaces. We are really sorry for this. However, despite the possible deficiencies of the dataset, as you have noticed, the model can still learn from a large number of accurate annotations, capture the characteristics of different types of green spaces, and accurately identify the parts that should have been marked, due to the generalization of the deep learning model. Besides, the effectiveness of applying UGSet as source domain dataset for large scale green space mapping have also been proved from experiments in Sect. 5.2. Therefore, we have explained and looked forward to this problem in “5.3 Limitations". We still hope that in the following work, more attempts can be made on the problems of attached green space labeling and identification. Corresponding statements can be found at:

“Besides, even though the UGSet have proved to be practicable for UGS mapping, we still have to point out that there may be a small number of missing labels in the dataset, especially for attached green spaces. As analyzed above, the extraction of attached green space can be more easily to be affected by external factors, like image taking angle, and the process of annotation is the same. Fortunately, despite the possible deficiencies, the DL model can still learn from a large number of accurate annotations, and capture the characteristics of different types of green spaces due to the strong generalization ability. We still hope that in the following work, more attempts can be made on the problems of labeling and identification of hard UGS types." (Line 368-375, Page 22)

****

Specific remarks

- Which value did you use for N in section 3.1.2?

Response:

Thanks for your valuable comments. According to your suggestions, we have added the value of N:

"In our experiments, N=1024 sampling points will be collected, and the value of k and β are 3 and 0.75, respectively." (Line 198, Page 11).

Besides, some necessary details have also been supplemented in Sect. 3.1.2.

****

- Figure 8: I believe the final 8x8x1 image is flattened before it is fed through the softmax function?

Response:

Thank you for underlining this problem. After your kind reminder we have checked the mentioned figure and relative statements. Actually, the softmax is applied to the prediction results from the generator, and its not used in the discriminator. We sincerely apologize for our mistakes. We have revised the Figure 8 and corresponding statements, which can be found at

"Given an input of the softmax prediction map from the generator, $P \in R^{H \times W \times C}$, the discriminator will output a discriminant result of the input, $D(P) \in R^{h \times w \times 1}$. After that, the discriminator will optimize itself according to the discrimination accuracy through the cross-entropy loss in (4)." (Line 204-206, Page 12)

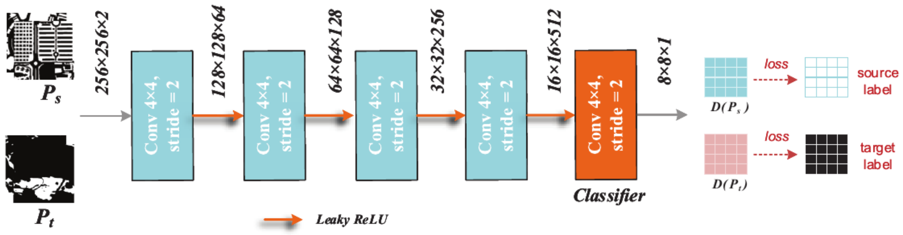


Figure 8. The structure of the discriminator.

****

- How was the parameter optimization performed?

    1. A batch size of 8 image pairs is rather small, was this decided through parameter optimization experiments or because of memory limitations?

    2. The same learning rate, batch size and number of epochs was used for all models? Was this done to facilitate the experiments?

Response:

1. Due to the GPU memory limitation, a batch size of 8 can just allow our model and some other models to run on a GeForce RTX 2080Ti. In addition, we also found that, under the GPU limitation, the mini-batch of 8 was often used in the many literatures [1-2] to obtain better training speed and training effect at the same time. Therefore, we adopt a batch size of 8 image pairs in the experiments.

    Corresponding statemenst have been revised to clarify this point:

    "A batch size of 8 sample pairs is adopted due to the limitation on GPU memory." (Line 250).

References:

[1] Zheng Z, Zhong Y, Wang J, et al. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4096-4105.

[2] Zheng Z, Ma A, Zhang L, et al. Deep multisensor learning for missing-modality all-weather mapping[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 174: 254-264.

2. All models have adopted the same learning rate, batch size and number of epochs to facilitate a fair comparison.

****

Hope these responses are clear enough for your further reviewing and we are prepared to discuss on possible issues if necessary.