PERFORMANCE EVALUATION OF FUSION TECHNIQUES FOR CROSS-DOMAIN BUILDING ROOFTOP SEGMENTATION

Haolin Li [1,2] *, Jiaojiao Tian [1], Yuxing Xie [1], Chong Li [2], Peter Reinartz [1]

[1] German Aerospace Center (DLR), Remote Sensing Technology Institute, D-82234 Wessling, Germany – (jiaojiao.tian, yuxing xie, peter.reinartz)@dlr.de

[2] Sichuan Surveying and Mapping Product Quality Test & Control Center, MNR, Chengdu 610041, China – lihaolin0516@gmail.com

Commission III, WG III/6

KEY WORDS: Building roof segmentation, Neural network, Self-training, HRNet, OCRNet, Swin Transfomer

ABSTRACT:

Convolutional Neural Networks have been widely introduced to building rooftop segmentation using satellite and aerial imagery. Preparing efficient training data is still among the critical issues on this topic. Therefore, adopting available annotated cross-domain multisource dataset is needed. This paper evaluates the performance of fusing the state-of-art deep learning neural network architectures for cross-domain building rooftop segmentation. We have selected three semantic image segmentation neural networks, including Swin transformer, OCRNet and HRNet. The predictions from these three neural networks are combined with majority voting, max value and union fusion techniques, a refined building rooftop segmentation mask is therefore delivered. The experiments on two benchmark datasets show that the proposed fusion techniques outperform single models and other state-of-art cross-domain segmentation approaches.

1. INTRODUCTION

Building rooftop segmentation is one of the fundamental tasks in photogrammetry and remote sensing. In particular, an up-to-date building rooftop map is required for many applications, including urban mapping, city planning, and land use analysis. Many edge-driven and region-driven approaches are proposed in the last two decades (Cui et a., 2011, Tian and Reinartz, 2013, Qin et al., 2016; Hossain et al., 2019).

The development of machine learning and deep learning (DL) algorithms has further boosted the region-driven building extraction approaches. Especially, semantic segmentation methods based on Convolutional Neural Network (CNN) have achieved great success in extracting building rooftop segmentations (Ji et al., 2018, Yuan et al., 2021b). However, training an efficient semantic segmentation model requires large amounts of manually annotated pixel-level building masks, which requires a lot of manual work and is therefore very expensive and time-consuming (Farahani et al., 2021, Chen et al., 2021). Complex and diverse scenes are also increasing the difficulty of data labelling. Luckily nowadays more research institutes and universities are willing to publish their annotated building rooftop segmentation as benchmark datasets (Chen et al., 2020). However, introducing the available benchmark datasets for other building roof segmentation applications is not an easy task. Due to the differences in building types and distributions, which is explained as a domain gap in computer vision, the training data annotated for one city cannot be easily adapted to a different test region. Therefore, cross-domain learning is a critical research topic for building roof segmentation, in particular for the applications when diverse test regions are involved (Peng et al., 2021).

In this paper, we evaluate the performance of fusing three advanced deep neural network models for cross-domain building rooftop segmentation, including Swin Transformer (Liu et al., 2021), OCRNet (Yuan et al., 2020), and HRNet (Wang et al., 2019). Three fusion techniques are tested, which are majority voting, max value and union fusion. Moreover, in order to minimize the appearance discrepancy between the source domain and target domain images, we adopt the LAB-based image translation method in the pre-processing step. We assess the proposed fusion methods on two building extraction benchmark dataset WHU Building dataset (Liu and Ji, 2020) and Potsdam Building dataset (Rottensteiner et al. 2012). Through comparing to other cross-domain semantic segmentation approaches, the evaluation results prove that fused predictions from three state-of-art semantic segmentation models retain a more robust performance.

2. RELATED WORK

2.1 Building rooftop segmentation with deep neural networks

Building rooftop segmentation is a binary classification task, which aims to label pixels of original images as two classes: buildings and non-buildings. Driven by the trend in semantic segmentation tasks, in recent years most building rooftop segmentation works use deep neural networks, achieving state-of-the-art results in benchmark datasets. Among those works, well-known fully convolutional networks (FCNs) (Long et al., 2015) are widely employed for image semantic segmentation, such as U-Net (Ronneberger et al., 2015), SegNet (Yang et al. ,2018), DenseNet (Li et al., 2018), and HRNet (Wang et al., 2019). In detail, Ji et al. (2018) and Kang et al. (2019) adopt U-Net for building extraction from optical images. Xu et al. (2018)

* Corresponding author

and Yi et al. (2019) introduce residual blocks to U-Net in order to facilitate training. Yang et al. (2018) combines signed-distance labels with SegNet to achieve instance-level building extraction. To better utilize high-order structural features for accurate building extraction, Li et al. (2018) adopt DenseNet with an adversarial module. Inspired by HRNetv2 (Sun et al., 2019), Zhu et al. (2020) propose MAP-Net, which introduces a channel-wise attention module to adaptively squeeze multiscale features extracted from the multipath network.

Recently, transformer networks such as the Swin transformer attract researchers' attention, benefit to its high efficiency and effectiveness with a shifted window based self-attention module. Yuan et al. (2021a) and Chen et al. (2022) directly apply Swin transformer with multiscale features in the building roof segmentation task.

2.2 Cross-domain learning for building roof segmentation

Due to the domain gap of images captured from distinct cities or with different shooting conditions, the performance of FCNs drops significantly on unseen datasets, which usually causes poor generalization (Peng et al., 2021). To efficiently process large-scale data with relatively low costs, effective cross-domain strategies are desired. In building rooftop segmentation works, few pioneer studies are available. For instance, Peng et al. (2021) introduces full-level domain adaptation methods including the mean-teacher model (Tarvainen et al., 2017), adversarial learning, and self-training. Although classic domain adaptation methods can achieve notable progress in reducing domain shifts between different datasets, but they are not data-friendly for complex scenarios, as the target (test) data are required in the training phase. It means that the network has to be repeatedly trained if multiple test data sets are planned.

Benefiting from advanced neural network structures and learning abilities, a single DL segmentation model can already cope with some domain shift problems. As each neural network structure can learn unique features from the image, and provides its own predictions, in this paper, we propose a data-friendly framework combining the predictions from three advanced semantic segmentation networks, thus to further improve the robustness of their performance in cross-domain building rooftop segmentation

3. METHODS

In this section, three fusion approaches are described and used to combine the prediction results from Swin Transformer (Liu et al., 2021), OCRNet (Yuan et al., 2020) and HRNet (Wang et al., 2019) models. In the pre-processing step we have adopted the LAB color translation to reduce the appearance discrepancy between the source domain and target domain images.

3.1 Image translation

Basically, two categories of approaches are available for image translation, including color transform and generative adversarial networks (GANs). It has been proven that it is difficult to train an efficient GAN model for the image translation using the current techniques (Peng et al., 2021). Therefore, we selected the CIELAB (LAB) based color translation (He et al., 2021) to reduce the domain discrepancy.

Instead of randomly select one image from the target dataset, we take 10 images and translate them to the LAB color space (LAB) (Jain, 1989). In LAB (*l*a*b*) color space, l represents for the perceptual lightness, a is relative to the green-red opponent color, while b represents the blue-yellow opponent. After that we calculate the mean and standard deviation of these ten images, which are noted as μ_T and σ_T, respectively. We project all source domain image to LAB space I_S^{LAB}, and then then shift the distribution of pixels values of each channel to the target domain as Equation. 1.

$$\hat{I}_S^{LAB} = \frac{(I_S^{LAB} - \mu_S)}{\sigma_S} * \sigma_T + \mu_T \quad (1)$$

In the end the LAB images $\hat{I}_S^{LAB}$ are translated back to RGB color space, which are used as input for the building rooftop segmentation task.

3.2 CNN based building rooftop segmentation

CNN based segmentation approaches have received increasingly interest as they are able to deliver more accurate result and robust to noises containing in the training datasets (Alzubaidi et al., 2021). In this paper we have selected three state-of-art semantic image segmentation deep neural network architectures for building rooftop extraction, including Swin transformer (Liu et al., 2021), OCRNet (Yuan et al., 2020), and HRNet (Wang et al., 2019).

Swin Transformer is one respective vision transformer proposed by Microsoft (Liu et al., 2021). The main highlight of Swin Transformer is hierarchical feature representation and its linear computational complexity with respect to input image size. Using the proposed shifted window approach to compute self-attention can significantly enhance the modelling power, thus to further improve the efficiency and effectiveness for vision tasks. Up to now, Swin Transformer achieves the state-of-the-art performance on many semantic segmentation tasks, including building extraction (Xu et al., 2021, Chen et al., 2022)

OCRNet: As its name states, Object-Contextual Representations (OCR) addresses the semantic segmentation problem with a focus on the context aggregation strategy (Yuan et al., 2021). It presents a simple yet effective approach for object-contextual representations, which characterizes each single pixel with its corresponding object representation, thus to improve the learning ability and decrease the influences of unnecessary details in images. Object region learning and object region representation computation are presented as parallel modules, and are integrated as the cross-attention module in the decoder. It has been tested on various object extraction and segmentation applications (Jin et al., 2021, Huang et al., 2021)

HRNet is an earlier semantic image segmentation network structure from Microsoft research (Wang et al., 2020). It enables the high-resolution representations through the interaction of the high-to-low resolution convolution streams in parallel. In particular, it can repeatedly exchange information across high- and low-level presentations. The benefit is that the resulting representation is semantically richer and spatially more precise, until now it has been used in a wide range of applications, including human pose estimation, semantic segmentation, and object detection. It has also a good performance in building extraction (Seong et al., 2021, Cheng et al., 2020).

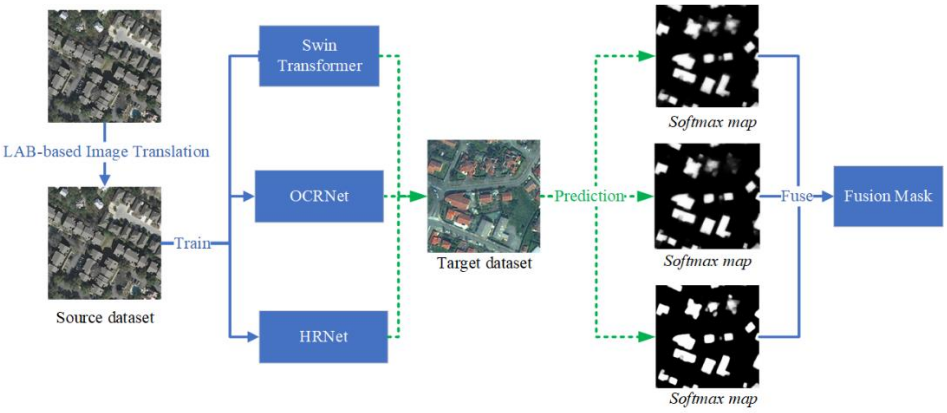


Figure 1. Flowchart of the proposed method

3.3 Fusion methods

Each CNN neural network model output can be presented as a softmax probability maps z_i, which approximately indicate the certainty that each pixel belongs to the building rooftop class. We explore three fusion approaches to generate a final segmentation mask.

3.3.1 Majority Voting: Majority voting is widely used in image processing and classification tasks (Jimenez et al., 1999; Hajdu et al., 2013). Under this fusion scheme, each segmentation model can provide a separate decision after giving a predefined threshold value (T). Thus, three labelling results are provided for each pixel. If at least two segmentation models classify one pixel into building rooftop class, the majority voting recognizes it to belong to building rooftop. They are defined as Equation (2) and (3).

$$p_i = \begin{cases} 1, & z_i \geq T \\ 0, & z_i < T \end{cases} \quad (2)$$

$$\mathrm{y} = \begin{cases} 1, & \sum_{i=1}^{n} p_i \geq \frac{n}{2} \\ 0, & \sum_{i=1}^{n} p_i < \frac{n}{2} \end{cases} \quad (3)$$

where y, z_i, n and p_i denote the category label, softmax probability map value, number of models and the segmentation results, and 1 is building rooftop, 0 means background.

3.3.2 Max Value: In the max value fusion, we firstly generate a fused softmax probability map by taking the maximum value of each pixel among three probability prediction, which are generated by Swin Transformer, OCRNet and HRNet model, respectively. Then, we generate a building mask by setting a threshold value on this fused softmax probability map. It's defined as Equation (4).

$$\mathrm{y} = \begin{cases} 1, & \max\limits_{i=1,\ldots,n} z_i \geq t_m \\ 0, & \max\limits_{i=1,\ldots,n} z_i < t_m \end{cases} \quad (4)$$

where t_m denotes threshold value, which is related with maximum value method.

3.3.3 Union: In union fusion, we sum up the probability maps that are generated by Swin Transformer, OCRNet and HRNet model. The category label is predicted by comparing the summed probability value to a given threshold. It is defined as Equation (5).

$$\mathrm{y} = \begin{cases} 1, & \sum_{i=1}^{n} z_i \geq t_u \\ 0, & \sum_{i=1}^{n} z_i < t_u \end{cases} \quad (5)$$

where t_u denotes threshold value, which is related with union fusion method.

4. EXPERIMENTS

4.1 Descriptions of Datasets

To verify the effectiveness and efficiency of the proposed method, WHU Building dataset (Ji et al., 2019) and Potsdam Building dataset (Rottensteiner et al. 2012) are employed in the experiment. We use WHU and Potsdam Building dataset alternately as source and target domain datasets.

WHU Building Dataset. The dataset consists both aerial and satellite imagery over Christchurch, New Zealand (Ji et al., 2019). In our experiment we take only the aerial dataset, which covers an area of 450 km² with an original resolution of 0.075 *m*. Over 220 000 independent buildings various types and locations were manually digitalized and corrected from the New Zealand government published building footprint vectors (https://data.linz.govt.nz/).

Potsdam Building Dataset. This dataset is generated from The International Society for Photogrammetry and Remote Sensing (ISPRS) 2-D Semantic Labeling Contest's dataset, where binary building masks were extracted based on the semantic label maps. It covers an area of 3.42 km² and consists of 38 VHR aerial images tiles with a size of 6000×6000 pixels with the GSD of 0.05 m. The dataset was collected over the city of Potsdam, Germany, which is a typical historical European city with large building blocks and dense settlement structures (Rottensteiner et al. 2012).

WHU Potsdam

Figure 2. Example images of the WHU and Potsdam benchmark datasets.

4.2 Experiment setup and training details

To reduce the influences of image resolution differences and the domain gap between the Potsdam and WHU dataset, both datasets are down sampled into a GSD of 0.3 m, which is also a generally recommended resolution for building segmentation.

After that all images are cropped in the 512 ×512 pixels patches, which results in a total of 8189 tiles for WHU Building dataset and 152 tiles for Potsdam dataset. Meanwhile, we use the officially recommended method to divide the dataset into training, testing and validation set, and we calculate the result accuracy using the officially provided testing datasets. The proposed method is implemented under the MMSegmentation framework (Chen et al., 2019a), and all the experiments were conducted on 4 GeForce RTX 2080Ti GPUs.

4.3 Evaluation Method

Two parameters F1-score (F1) and the Intersection-Over Union (IoU) of the building rooftop segments are calculated to evaluate the accuracy of the extracted building rooftop segments. They are defined as Equation (6) and (7).

$$\mathrm{F1} = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}) \quad (6)$$
$$\mathrm{IoU} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP} + \mathrm{FN}) \quad (7)$$

where TP, FP, and FN denote the pixel numbers of True Positives, False Positives, and False Negatives, respectively. Note that higher F1-score and IoU denote better overall performance.

4.4 Experimental Results

The aim of this section is to evaluate the fused cross-domain building segmentation approach by comparing them to single segmentation models and other state-of-art cross-domain segmentation approaches.

4.4.1 Compare to single training models: Table I summarizes the F1 and IoU metrics yielded by single DL models, including Swin Transformer (Swin), OCRNet, HRNet, as well as different fusion approaches. For the majority voting a commonly used thresholding value $T = 0.5$ is selected. However, for the max value and union method, we have tested 4 threshold values t_m for Max Value method [0.60, 0.65, 0.70, 0.75] and 6 threshold values t_u for the Union fusion method [1.1, 1.2, 1.3, 1.4, 1.5, 1.6]. We analyse the results for each case study.

Case 1: Potsdam→WHU: Firstly, comparing to OCRNet and HRNet, Swin transformer has a generally better performance on cross-domain datasets. However, fusing the predictions by OCRNet and HRNet can still further improve the accuracy. As Table I shows, the proposed Union-1.6 approach archives the increase of IoU and F1 by 1.11% and 0.74%, respectively. It also outperforms Majority Voting method with an IoU gain of 3.47% and 2.34%. The Majority voting and Max value fusion approaches cannot overstep the results from Swin transformer.

For the visual comparison, we have selected five image patches and presented in Figure. 3. The buildings derived by Swin, OCRNet, HRNet and the best results from each fusion approach method are presented together with the ground truth. As presented in the first row, the building rooftop segments obtained by Union-1.5 method are almost identical to the ground truth, and it has more precise edge than the segments predicted by HRNet, Swin and OCRNet model. In the third row of Figure.3, Swin transformer has shown more miss-detections than the other two models. After the fusion steps, more building rooftops are correctly detected. The last two row of Fig.3 clearly demonstrates that the Union-1.5 method is capable of identifying small sized buildings, and it can help to correct some recognition errors.

Method	Potsdam→WHU		WHU→Potsdam	
	IoU [%]	F1 [%]	IoU [%]	F1 [%]
HRNet	67.96	80.92	65.30	79.01
Swin	72.54	84.08	72.30	83.93
OCRNet	68.89	81.58	67.06	80.28
Majority Voting	70.18	82.48	67.39	80.52
Max Value 0.60	70.22	82.51	73.50	84.72
Max Value 0.65	71.51	83.39	71.19	83.17
Max Value 0.70	72.55	84.10	68.36	81.21
Max Value 0.75	73.34	84.62	65.29	79.00
Union 1.1	68.91	81.60	**75.08**	**85.76**
Union 1.2	70.19	82.48	73.90	84.99
Union 1.3	71.30	83.25	72.57	84.11
Union 1.4	72.24	83.88	71.08	83.09
Union 1.5	73.02	84.41	69.42	81.95
Union 1.6	**73.65**	**84.82**	67.60	80.67

Table 1. Summary of the accuracy obtained by different methods

Trans Method	Model	Potsdam→WHU		WHU→Potsdam	
		IoU [%]	F1 [%]	IoU [%]	F1 [%]
LAB-based	HRNet	67.96	80.92	65.30	79.01
	Swin	**72.54**	**84.08**	**72.30**	**83.93**
	OCRNet	68.89	81.58	67.06	80.28
Normalization	HRNet	66.17	79.64	63.17	77.43
	Swin	**70.63**	**82.79**	**69.68**	**82.13**
	OCRNet	67.42	80.54	67.29	80.45

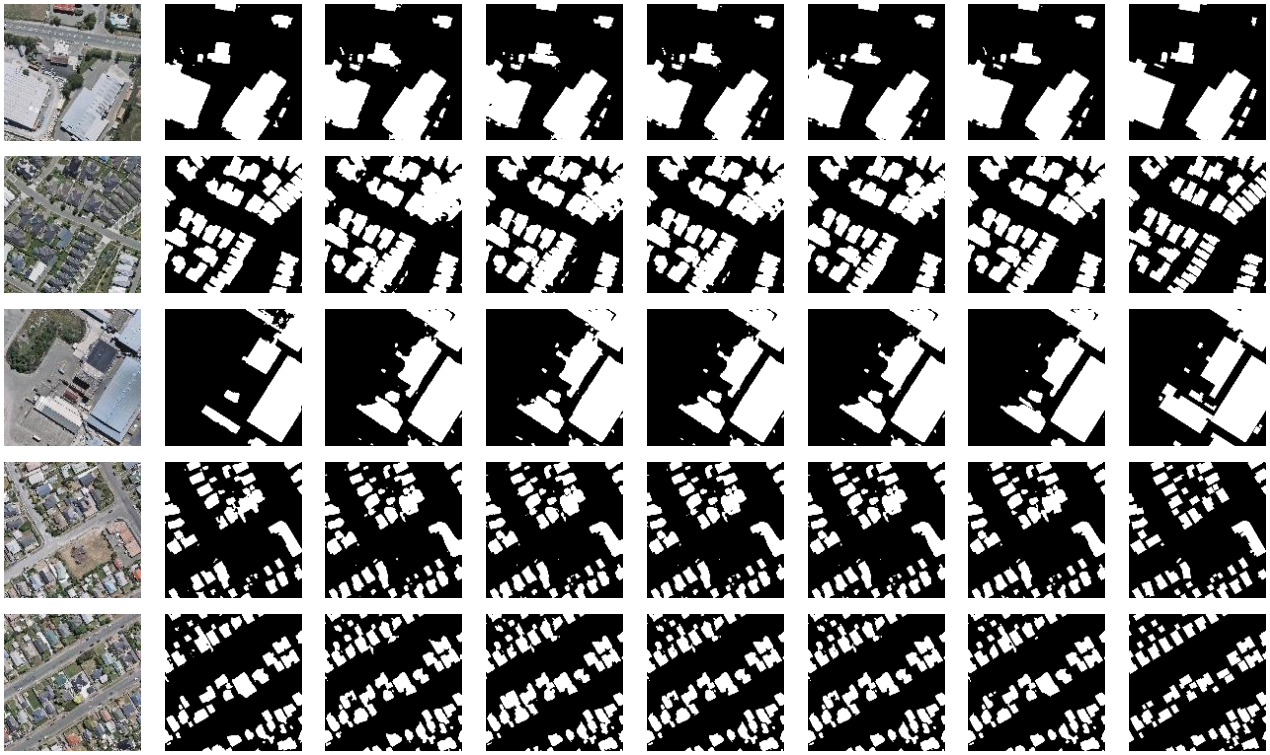
Table 2. Comparison of the LAB-based image translation method with the normalization method

Method	Potsdam→WHU		WHU→Potsdam	
	IoU [%]	F1 [%]	IoU [%]	F1 [%]
DAugNet	58.27	73.63	59.63	74.71
DATA	65.99	79.51	70.42	82.64
OSA	57.37	72.91	69.57	82.05
LTA	65.75	79.34	71.59	83.44
JPRNet	62.77	77.13	60.19	75.15
Union 1.10	68.91	81.60	**75.08**	**85.76**
Union 1.60	**73.65**	**84.82**	67.60	80.67

Table 3. Comparison of the proposed Union method with other methods

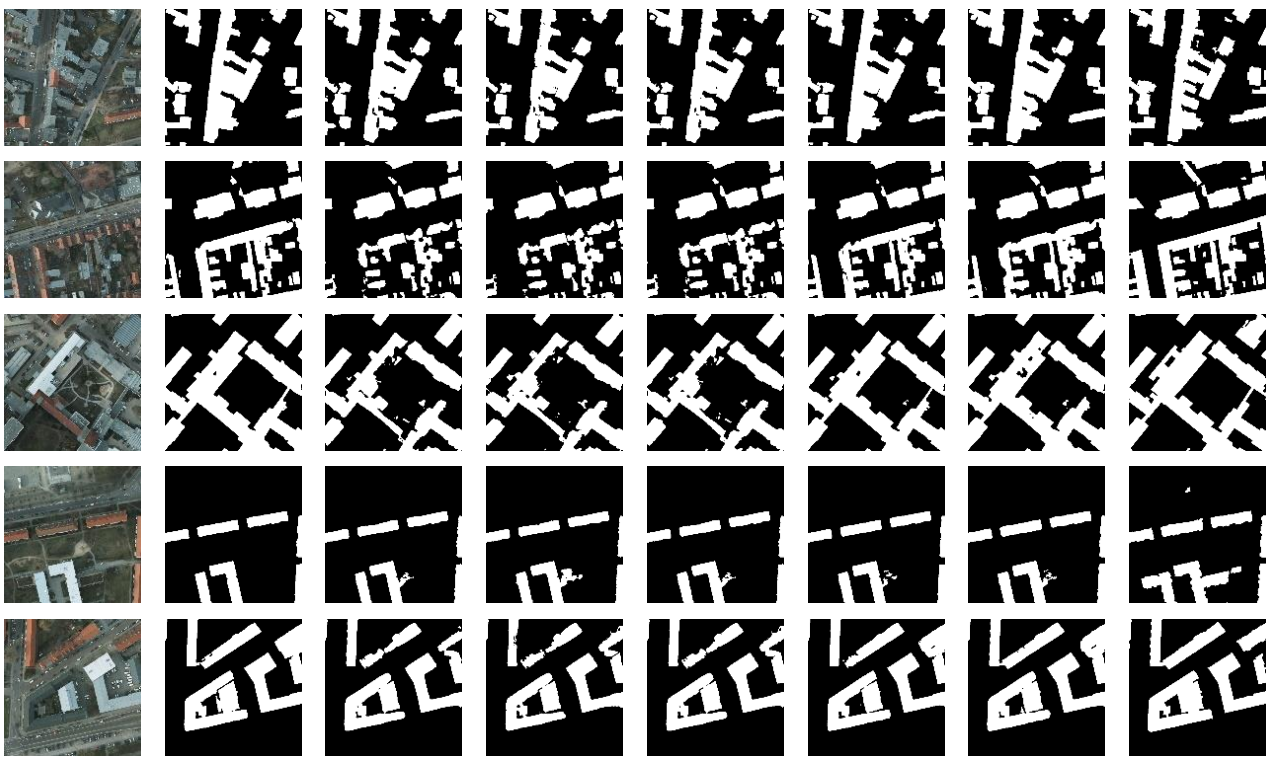
Case2: WHU → Potsdam: The same evaluation has been performed when using WHU as source domain and Potsdam as target domain dataset. WHU is a much larger dataset, it is surprisingly to see that similar accuracies have been achieved as Case1, probably due to the same domain shift. Different to case 1, in this example all three fusion techniques have achieved better IoU and F1 values than using single Swin, HRNet and OCRNet models. The predictions have generally lower value in the softmax output than case 1, thus a lower threshod for the Max value and union fusion have achieved a higher IoU and F1 values.

The visually comparison in Figure 4 shows a similar trend as Figure 3. Building rooftop segments from Swin transformer have much sharpen edges than results from other single models, especially for the first and third examples. Union 1.1 method is advantaged in identifying tiny buildings than other methods. However, at the fourth row, OCRNet and HRNet could detection the additional part the large building, which is not included in the Swin transformer prediction results



Target Images Swin OCRNet HRNet MajorityVoting MaxValue-0.75 Union-1.5 Ground Truth

Figure.3 Examples of building extraction maps obtained by different methods for the Case1 Potsdam→WHU.



Target Images Swin OCRNet HRNet Majority Voting MaxValue-0.6 Union 1.1 Ground Truth

Figure.4 Examples of building extraction maps obtained by different methods for the Case 2 WHU→Potsdam.

By comparing the LAB-based image translation method with conventional normalization method in the Table II, we can observe that LAB-based Image Translation approach can help to obtain better and stable results. The HRNet, Swin and OCRNet model can gain 1.47%~1.79% and 1.04%~1.28% in terms of IoU and F1 for the Potsdam→WHU case. Meanwhile, the HRNet and Swin model also can gain 2.13%~2.62% and 1.58%~1.8% in terms of IoU and F1 for the WHU→Potsdam case, and the OCRNet model can achieve similar performance.

4.4.2 Compare to the other approaches: We compare our approach with other state-of-art cross-domain building rooftop segmentation approaches. By comparing the proposed Union method with other methods in the Table III, we can observe that the Union fusion method is able to obtain the best and stable results against other comparative methods, such as DAugNet (Peng et al., 2021, Tasar et al., 2020), DATA (Na et al., 2020), OSA (Tsai et al., 2018), LTA (Hoffman et al., 2016) and JPRNet (Shi et al., 2020). The proposed Union method using 1.6 as threshold can gain 7.66%~16.28% and 5.31%~11.91% in terms of IoU and F1 for the Potsdam→WHU case, and it also can gain 3.49%~15.45% and 2.32%~11.05% in terms of IoU and F1 for the WHU→Potsdam case.

5. DISCUSSION AND CONCLUSION

Adapting annotated benchmark multisource datasets to building rooftop segmentation task is a crucial issue. With the traditional matching learning approaches, the classifier trained in one dataset can hardly be used on another dataset due to the various of building types and distributions, which is now defined as domain shift or domain gap in computer vision. Benefiting from the development of deep learning techniques and neural network architectures, the learning ability of the DL based classification and segmentation approaches have been largely improved, some of which can be directly performed on cross-domain datasets. In this paper, we have compared three state-of-art DL based segmentation models and various fusion techniques for cross-domain building rooftop segmentation datasets. Our experiments on two cross-country benchmark datasets have shown that combing the predictions from more segmentation models can bring a considerable improvement to the accuracy and robustness. Benefit partly from the advanced segmentation neural network architectures, our fusion approach has also outperformed other cross-domain segmentation approaches. Union fusion approach has achieved the highest accuracy compare to other approaches when a proper threshold value is provided. The light weighted LAB based image translation can help to reduce the appearance discrepancy between the source domain and target domain images, thus improve the performance of segmentation models. It has to be noted, the selection of threshold values has a direct influence on the accuracy of the Union and Max Value fusion approaches. In the next step, we plan to use adaptive methods to solve this problem and introduce advanced fusion techniques.

REFERENCES

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1), pp.1-74.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J. and Zhang, Z., 2019a. MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European conference on computer vision (ECCV) (pp. 801-818).

Chen, Q., Wang, L., Wu, Y., Wu, G., Guo, Z. and Waslander, S.L., 2019b. TEMPORARY REMOVAL: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings. ISPRS journal of photogrammetry and remote sensing, 147, pp.42-55.

Chen, X., Yuan, Y., Zeng, G. and Wang, J., 2021. Semi-supervised semantic segmentation with cross pseudo supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2613-2622).

Chen, X., Qiu, C., Guo, W., Yu, A., Tong, X. and Schmitt, M., 2022. Multiscale feature learning by transformer for building extraction from satellite images. IEEE Geoscience and Remote Sensing Letters. 19, 2503605.

Cheng, Z. and Fu, D., 2020. Remote sensing image segmentation method based on HRNet. In IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium (pp. 6750-6753). IEEE.

Cui, S., Yan, Q., Reinartz, P. and Mansour, N., 2011. Graph search and its application in building extraction from high resolution remote sensing imagery. Search Algorithms and Applications, pp.133-150.

Farahani, A., Voghoei, S., Rasheed, K. and Arabnia, H.R., 2021. A brief review of domain adaptation. Advances in Data Science and Information Engineering, pp.877-894.

Hajdu, A., Hajdu, L., Jonas, A., Kovacs, L. and Toman, H., 2013. Generalizing the majority voting scheme to spatially constrained voting. IEEE Transactions on image processing, 22(11), pp.4182-4194.

He, J., Jia, X., Chen, S. and Liu, J., 2021. Multi-source domain adaptation with collaborative learning for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11008-11017).

Hoffman, J., Wang, D., Yu, F. and Darrell, T., 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649.

Hossain, M.D. and Chen, D., 2019. Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. ISPRS Journal of Photogrammetry and Remote Sensing, 150, pp.115-134.

Huang, S., Han, W., Chen, H., Li, G. and Tang, J., 2021. Recognizing Zucchinis Intercropped with Sunflowers in UAV Visible Images Using an Improved Method Based on OCRNet. Remote Sensing, 13(14), p.2706.

Ji, S., Wei, S. and Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on Geoscience and Remote Sensing, 57(1), pp.574-586.

Kang, W., Xiang, Y., Wang, F. and You, H., 2019. EU-net: An efficient fully convolutional network for building extraction from optical remote sensing images. Remote Sensing, 11(23), p.2813.

Li, X., Yao, X. and Fang, Y., 2018. Building-a-nets: Robust building extraction from high-resolution remote sensing images with adversarial networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(10), pp.3680-3687.

Liu, J. and Ji, S., 2020. A novel recurrent encoder-decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6050-6059).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

Jain, A.K., 1989. Fundamentals of digital image processing. Prentice-Hall, Inc..

Jin, Z., Gong, T., Yu, D., Chu, Q., Wang, J., Wang, C. and Shao, J., 2021. Mining contextual information beyond image for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7231-7241).

Jimenez, L.O., Morales-Morell, A. and Creus, A., 1999. Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting, and neural networks. IEEE Transactions on Geoscience and Remote Sensing, 37(3), pp.1360-1366.

Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P., 2017, July. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 3226-3229). IEEE.

Na, Y., Kim, J.H., Lee, K., Park, J., Hwang, J.Y. and Choi, J.P., 2020. Domain adaptive transfer attack-based segmentation networks for building extraction from aerial images. IEEE Transactions on Geoscience and Remote Sensing, 59(6), pp.5171-5182.

Peng, D., Guan, H., Zang, Y. and Bruzzone, L., 2021. Full-Level Domain Adaptation for Building Extraction in Very-High-Resolution Optical Remote-Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 60, pp.1-17.

Qin, R., Tian, J. and Reinartz, P., 2016. Spatiotemporal inferences for use in building detection using series of very-high-resolution space-borne stereo images. International Journal of Remote Sensing, 37(15), pp.3455-3476.

Ronneberger, O., Fischer, P. and Brox, T., 2015, U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pp. 234-241.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S. and Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-3 (2012), Nr. 1, 1(1), pp.293-298.

Seong, S. and Choi, J., 2021. Semantic segmentation of urban buildings using a high-resolution network (HRNet) with channel and spatial attention gates. Remote Sensing, 13(16), p.3087.

Shi, L., Wang, Z., Pan, B. and Shi, Z., 2020. An end-to-end network for remote sensing imagery semantic segmentation via joint pixel-and representation-level domain adaptation. IEEE Geoscience and Remote Sensing Letters, 18(11), pp.1896-1900.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W. and Wang, J., 2019. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514.

Tarvainen, A. and Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30.

Tasar, O., Giros, A., Tarabalka, Y., Alliez, P. and Clerc, S., 2020. Daugnet: Unsupervised, multisource, multitarget, and life-long domain adaptation for semantic segmentation of satellite images. IEEE Transactions on Geoscience and Remote Sensing, 59(2), pp.1067-1081.

Tian, J. and Reinartz, P., 2013, April. Fusion of multi-spectral bands and DSM from Worldview-2 Stereo imagery for building extraction. In Joint Urban Remote Sensing Event 2013 (pp. 135-138). IEEE.

Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H. and Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7472-7481).

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X. and Liu, W., 2020. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 43(10), pp.3349-3364.

Xu, Y., Wu, L., Xie, Z. and Chen, Z., 2018. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. Remote Sensing, 10(1), p.144.

Xu, Z., Zhang, W., Zhang, T., Yang, Z. and Li, J., 2021. Efficient transformer for remote sensing image segmentation. Remote Sensing, 13(18), p.3585.

Yang, H.L., Yuan, J., Lunga, D., Laverdiere, M., Rose, A. and Bhaduri, B., 2018. Building extraction at scale using convolutional neural network: Mapping of the united states. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(8), pp.2600-2614.

Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W. and Zhao, T., 2019. Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. Remote sensing, 11(15), p.1774.

Yuan, W. and Xu, W., 2021a. MSST-Net: A Multi-Scale Adaptive Network for Building Extraction from Remote Sensing Images Based on Swin Transformer. Remote Sensing, 13(23), p.4743.

Yuan, X., Shi, J. and Gu, L., 2021b. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications, 169, p.114417.

Yuan, Y., Chen, X. and Wang, J., 2020, August. Object-contextual representations for semantic segmentation. In European conference on computer vision (pp. 173-190). Springer, Cham.

Zhu, Q., Liao, C., Hu, H., Mei, X. and Li, H., 2020. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery. IEEE Transactions on Geoscience and Remote Sensing, 59(7), pp.6169-6181.