Supplementary Material for Fully Transformer Network

Tianyu Yan, Zifu Wan, and Pingping Zhang[0000−0003−1206−1444]

School of Artificial Intelligence, Dalian University of Technology, China
{tianyuyan2001,wanzifu2000}@gmail.com;zhpp@dlut.edu.cn

1 Novelties of Our Framework

We agree that Transformers (including Fully Transformer Networks) are not new in current computer vision fields. However, there are some key differences between our work and previous methods: 1) As far as we know, our work is the earliest Transformer-based one, which **explicitly handles incomplete regions and irregular boundaries** for remote sensing image CD. 2) In our framework, we utilize a Siamese structure to process dual-phase remote sensing images. Besides, we introduce a pyramid structure to aggregate multi-level visual features from Transformers for feature enhancement. **These designs are totally different from existing works, especially in [1] and [2]** which mainly use a simple encoder-decoder+U-Net structure for single image feature extraction. 3) We utilize the deeply-supervised learning with **multiple boundary-aware loss functions to better train the framework**. These losses are very helpful for more accurate CD.

2 Deep Supervision

To optimize our framework, we adopt the deeply-supervised learning [3–5] for each feature level. To this goal, we first take the features of the PCP, *i.e.*, $\mathbf{F}_P^k(k = 1, 2, ..., 5)$, and use a deconvolutional layer for the corresponding prediction $\mathbf{P}^s$ as side-outputs. Note that the $\mathbf{F}_P^4$ and $\mathbf{F}_P^5$ have the same resolution. Thus, the corresponding prediction can be represented as:

$$\mathbf{P}^s = Deconv(\mathbf{F}_P^k), k = 1, 2, 3, 4, 5. \tag{1}$$

Finally, we concatenate them for the final fusion prediction,

$$\mathbf{P}^f = \mathrm{Conv}[\mathbf{P}^1, ..., \mathbf{P}^5], \tag{2}$$

All side-outputs and the final fusion prediction are supervised by the proposed hybrid losses. The deeply-supervised learning can improve the performance and optimize the framework easily.

3 More Experimental Results

3.1 Visual effects of different losses

In this work, we introduce multiple loss functions to improve the CD results. Tab. 1 shows the quantitative effects of these losses. In this supplementary material, we display some typical examples for the visual effects, as shown in Fig. 1. From the results, one can see that using the WBCE loss can help the model focus on the most change regions. With the SSIM loss, the framework can improve the structural information of the change regions. Using the SIoU loss can ensure the global completeness. As a result, combining all of them can achieve the best results, which prove the effectiveness of all loss terms. This fact is consistent with the quantitative results in Tab. 1.

Table 1. Performance comparisons with different losses on LEVIR-CD.

Losses	Pre.	Rec.	F1	IoU	OA
(a) BCE	90.68	86.91	88.75	79.78	98.88
(b) WBCE	91.65	88.42	90.01	81.83	99.00
(c) WBCE+SSIM	91.71	88.57	90.11	82.27	99.01
(d) WBCE+SSIM+SIoU	92.71	89.37	91.01	83.51	99.06

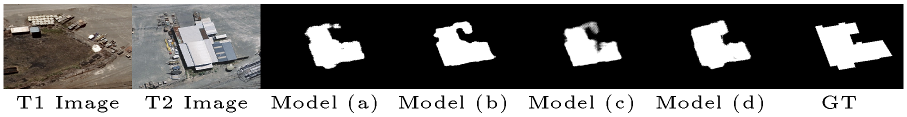

Fig. 1. Visual comparisons of predicted change maps with different models.

3.2 Scaling to higher resolutions

Since our work processes high-resolution remote sensing images, the scaling concern is very valuable. In fact, most of compared methods utilize cropping for generating low-resolution input images. In our main paper, we follow them and adopt a low resolution (256×256), mainly considering the fairness of comparisons. However, our work indeed can process a higher resolution with SwinT-Base/Small/Tiny. Tab. 2 shows the performance analysis with different resolutions and computation on WHU-CD. One can see that our method can naturally scale to higher resolutions and show slightly better results.

Table 2. Performance analysis with different input resolutions on WHU-CD.

Input resolution	Pre.	Rec.	F1	IoU	OA	Flops(G)
256×256	93.09	91.24	92.16	85.45	99.37	45
384×384	93.83	90.58	92.17	85.48	99.37	134
512×512	94.21	90.25	92.19	85.51	99.38	198

References

1. He, X., Tan, E.L., Bi, H., Zhang, X., Zhao, S., Lei, B.: Fully transformer network for skin lesion analysis. Medical Image Analysis **77**, 102357 (2022)
2. Wu, S., Wu, T., Lin, F., Tian, S., Guo, G.: Fully transformer networks for semantic image segmentation. arXiv:2106.04108 (2021)
3. Zhang, P., Liu, W., Wang, D., Lei, Y., Wang, H., Lu, H.: Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps. Pattern Recognition **100**, 107130 (2020)
4. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: IEEE International Conference on Computer Vision. pp. 202–211 (2017)
5. Zhang, P., Wang, L., Wang, D., Lu, H., Shen, C.: Agile amulet: Real-time salient object detection with contextual attention. arXiv:1802.06960 (2018)