GLANCE, FOCUS AND REFINEMENT NETWORK FOR REMOTE SENSING CHANGE DETECTION

Hao Zhang[1,2], *Zixuan Sun*[2], *Yuhui Zheng*[2], *Kaihua Zhang*[1,2⋆], *Gang Dong*[1], *Lingyan Liang*[1], *Yaqian Zhao*[1]

[1]Inspur Electronic Information Industry Co., Ltd., Beijing, China
[2]Nanjing University of Information Science and Technology, Nanjing, China

ABSTRACT

Existing change detection (CD) methods often directly fuse the multi-level features from bi-temporal remote sensing images without discriminatively considering each pixel's importance. Despite the demonstrated success, unselectively mixing the features degrades the model's performance to effectively capture the change targets due to the imbalance ratio between the change regions and the whole scene. To this end, this paper presents a glance, focus, and refinement network (GFRNet), which formulates CD as a continuous, step-by-step focusing process to mimic the human visual system. Specifically, the GFRNet first employs a transformer encoder to extract the global features from the bi-temporal images, where each feature takes a glance at the whole scene. Then, the GFRNet gradually pays attention to a cascade of salient regions, and ultimately progressively refines its focus on the desired areas of change. Comprehensive evaluations on two extensively utilized benchmark datasets, including LEVIR-CD and WHU-CD, demonstrate the superiority of our GFRNet to a variety of state-of-the-art methods.

Index Terms— Change detection, remote sensing, mask transformer, salient.

1. INTRODUCTION

Change detection (CD) [1] aims at discerning alterations that transpire between multiple images captured within the same geographical expanse but at distinct times. By virtue of its useful potential, CD technology has garnered widespread utilization across a spectrum of domains, encompassing applications such as damage assessment, urban planning, and forestry monitoring, among others [2, 3, 4].

With the fast development of deep learning (DL) in computer vision [9, 10, 11], DL has become the predominant approach for tackling CD [5, 6, 7, 8]. The fundamental architecture underlying DL-based CD methods can be categorized into two categories [5]: 1) one-stream framework, and 2) dual-stream framework. In the one-stream framework, the bi-temporal images are first added or subtracted, and then different semantic segmentation models, such as U-Net and FCN-PP [5, 12], are directly used to predict the CD map. In contrast, the dual-stream framework utilizes a Siamese network to extract bi-temporal features separately. Then, the features are fused via different fusion mechanisms, such as transformer and 3D convolution [7, 8].

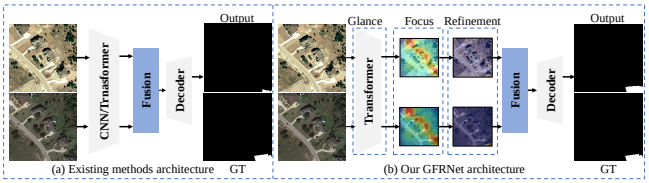

Figure 1. Structure Comparison of existing state-of-the-art (SOTA) methods [5, 6, 7, 8] and our GFRNet. Existing SOTA methods directly feed the features into the fusion module. This simply treating each pixel equally limits the model's ability to capture the desired change targets. In contrast, our GFRNet enables the model to gradually focus on the desired change areas through a coarse-to-fine feature learning process, thereby yielding a favorable performance.

Despite the demonstrated success, the DL-based methods suffer from some issues, making them remain some room for improvement. Specifically, as illustrated by Figure 1, when dealing with the remote sensing images, the relatively small size of the target results in only a small fraction of pixels being associated with the change region. Current SOTA approaches [5, 6, 7, 8] often directly concatenate the features and then fuse them for CD map prediction. However, this simple mixing strategy treats the whole image features equally, thereby failing to effectively discern the importance of distinct change regions. Consequently, these models fail to concentrate their attention on the desired change regions, leading to unsatisfying detection results.

To overcome the above challenges, inspired by the human visual system that conceptualizes CD as a coherent, progressively focused process, we propose a glance, focus, and refinement network (GFRNet) for CD. The GFRNet is composed of three novel subnet designs tailored to glance, focus, and refinement respectively. Firstly, the glance subnet

⋆ Corresponding author. Email: zhkhua@gmail.com. This work is supported in part by National Key Research and Development Program of China under Grant No. 2018AAA0100400, in part by NSFC under No. 62276141.

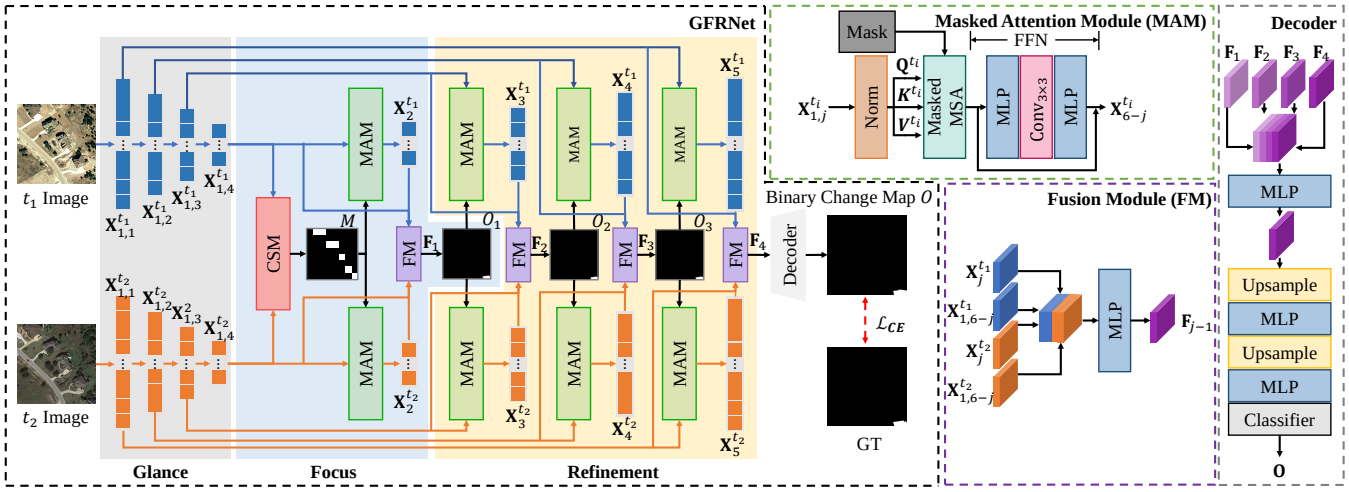

Figure 2. Architecture of the proposed GFRNet.

employs a Siamese transformer encoder to extract the global features that can capture long-range dependency in the whole images, facilitating taking a glance at the whole scene. Subsequently, the focus subnet devises a collaborative salient module (CSM) to generate a salient region mask from the bi-temporal image features [13], and then computes the masked attention (MA) [14] within these regions, attending its emphasis on the co-salient areas that contain all change regions. Following this, the refinement subnet gradually refines the features from the focused regions using the preliminary prediction as a mask for MA computation, thereby making the network focus on the desired change regions. Finally, the refined features are fed into the decoder to predict the CD map. Extensive evaluations on LEVIR-CD [15] and WHU-CD [16] demonstrate the favorable performance of our GFRNet compared to a variety of SOTA methods.

The main contributions of this work can be summarized as follows:

(1) We propose the GFRNet for CD, which is able to achieve a satisfying performance via a coarse-to-fine feature learning process.

(2) We design the focus subnet that leverages the CSM to make the network pay more attention to the salient areas that contain all the changed regions.

(3) We devise the refinement subnet that leverages the multi-scale MA to further refine the salient regions, resulting in satisfying CD results.

2. PROPOSED METHOD

2.1. Architecture Overview

Figure 2 illustrates the architecture of our GFRNet. The GFRNet consists of three cascaded subnets: the glance, the focus, and the refinement subnets. Specifically, given a pair of remote sensing images $\mathbf{I}_1$ and $\mathbf{I}_2$ with size of $H \times W \times 3$ as input, the glance subnet, which is a Siamese transformer encoder with shared weights, produces the features $\{\mathbf{X}_{1,i}^{t_1}, \mathbf{X}_{1,i}^{t_2} \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}\}_{i=1}^{4}$. Afterwards, $\mathbf{X}_{1,4}^{t_1}$ and $\mathbf{X}_{1,4}^{t_2}$ are fed into the focus subnet that designs the CSM to first generate a salient region mask $M \in \mathbb{R}^{H \times W}$, and then uses M to compute the MA of $\mathbf{X}_{1,4}^{t_1}$ and $\mathbf{X}_{1,4}^{t_2}$, yielding the salient region features $\mathbf{X}_2^{t_1}$ and $\mathbf{X}_2^{t_2}$. Then, $\{\mathbf{X}_2^{t_1}, \mathbf{X}_2^{t_2}\}$ and $\{\mathbf{X}_{1,4}^{t_1}, \mathbf{X}_{1,4}^{t_2}\}$ are passed through the fusion module (FM), yielding the fused features $\mathbf{F}_1 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$ and a preliminary change map $O_1 \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32}}$. Finally, in the refinement subnet, O_1 serves as a mask to compute the MA for the features $\mathbf{X}_{1,3}^{t_1}$ and $\mathbf{X}_{1,3}^{t_2}$, yielding features $\mathbf{X}_3^{t_1}$ and $\mathbf{X}_3^{t_2}$, which are then used for feature fusion and change map prediction, yielding the fused feature $\mathbf{F}_2 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}$, and the predicted change map $O_2 \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16}}$. The whole process is repeated three times, and the four fused feature maps are upsampled and input into the decoder to yield the predicted change map $O \in \mathbb{R}^{H \times W}$.

2.2. Glance Subnet

Given the $\mathbf{I}_1$ and $\mathbf{I}_2$, our glance subnet leverages a Siamese hierarchical transformer encoder based on MiT [17], enabling efficient capture of global contextual information across the entire image:

$$[\mathbf{X}_{1,i}^{t_1}, \mathbf{X}_{1,i}^{t_2}] = \text{MiTencoder}([\mathbf{I}_1, \mathbf{I}_2]), \quad (1)$$

where $i = 1, 2, 3, 4$ denotes the i^{th} feature level. Besides, to adapt to the varying input image sizes, we use a 3×3 convolution in the feed-forward network (FFN) as an alternative to the fixed-position encoding in the MiT.

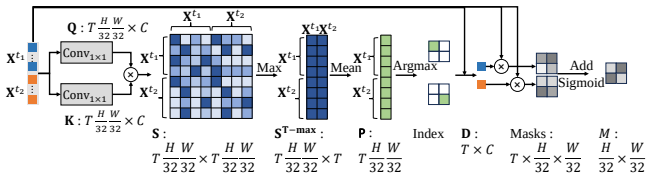

Figure 3. Architecture of the CSM.

2.3. Focus Subnet

The focus subnet includes two key modules, i.e., the CSM and the MAM, where the CSM aims to find all the change regions, while the MAM further guides the model's attention towards the desired change regions.

2.3.1. CSM

As illustrated by Figure 3, we first leverage the $\mathbf{X}_{1,4}^{t_1}, \mathbf{X}_{1,4}^{t_2}$ to generate the query $Q \in \mathbb{R}^{T\frac{H}{32}\frac{W}{32}\times C}$ and the key $K \in \mathbb{R}^{T\frac{H}{32}\frac{W}{32}\times C}$ with the number of input images $T = 2$ and the feature channel number C through a transformation module including a 1×1 convolution, a concatenation and a reshape operation. Then, we use scaled dot-product attention to calculate the pixel-wise feature similarity map:

$$S = \frac{KQ^\top}{\sqrt{d_k}}, \tag{2}$$

where $S \in \mathbb{R}^{T\frac{H}{32}\frac{W}{32}\times T\frac{H}{32}\frac{W}{32}}$, $\top$ denotes matrix transpose operator, and $\frac{1}{\sqrt{d_k}}$ denotes scaling factor.

Next, we reshape S to $\mathbf{S} \in \mathbb{R}^{T\frac{H}{32}\frac{W}{32}\times T\times\frac{H}{32}\frac{W}{32}}$, and select the value with the maximum similarity in each image, yielding T maximum similarity values for each pixel. By subsequently averaging these T maximum similarity values, we establish the co-salient probability P for each pixel:

$$P = \frac{1}{T}\sum_{t=1}^{T} S_{\max}[:, t], \tag{3}$$

where $S_{\max} = \max_{i=1\cdots HW} \mathbf{S}[:, :, i]$.

Then, P is reshaped to $\mathbf{P} \in \mathbb{R}^{T\times\frac{H}{32}\times\frac{W}{32}}$ and utilized to determine the indices corresponding to the most salient pixel within each image. These indices are then employed to extract the most salient pixels $D \in \mathbb{R}^{T\times C}$ from normalized features $||\mathbf{X}_{1,4}^{t_1}||_2$ and $||\mathbf{X}_{1,4}^{t_2}||_2$.

After that, we calculate the correlation between D and each pixel of $||\mathbf{X}_{1,4}^{t_1}||_2$ and $||\mathbf{X}_{1,4}^{t_2}||_2$. Subsequent to this, we sum all the correlation maps and apply a sigmoid function to get the mask $M \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}}$ for each image.

2.3.2. MAM

After obtaining M, the MAM computes attention within the masked regions followed by an FFN, which is formulated as:

$$\begin{aligned}\mathbf{X}_2^{t_i} &= \mathrm{FFN}(\mathrm{MA}(M, Q^{t_i}, K^{t_i}, V^{t_i})) \\ &= \mathrm{FFN}(\mathrm{softmax}(\frac{M \odot Q^{t_i}{K^{t_i}}^\top}{\sqrt{d_k}})V^{t_i}),\end{aligned} \tag{4}$$

where Q^{t_i}, K^{t_i}, and V^{t_i} denotes query, key, and value, separately, which are obtained from $\mathbf{X}_{1,4}^{t_i}$ by linear projection.

Finally, we feed the obtained features $\mathbf{X}_2^{t_i}$ with $\mathbf{X}_{1,4}^{t_i}$ into the FM that is composed of concatenation and MLP, yielding the predicted preliminary change map O_1.

2.4. Refinement Subnet

After achieving the mask O_1, we first upsample O_1 and then use it to calculate the MA for $\mathbf{X}_{1,3}^{t_1}$ and $\mathbf{X}_{1,3}^{t_2}$, obtaining the features $\mathbf{X}_3^{t_1}$ and $\mathbf{X}_3^{t_2}$. Subsequently, we integrate the features before and after applying MA, enabling the model to focus on the desired change areas. To endow the features with both deep-level semantic information and shallow-level detailed information, we undertake a fusion of multi-level features:

$$\mathbf{F}_2 = \mathrm{Concat}(\mathbf{X}_{1,3}^{t_1}, \mathbf{X}_{1,3}^{t_2}, \mathbf{X}_3^{t_1}, \mathbf{X}_3^{t_2}) + \mathrm{Upsample}(\mathbf{F}_1). \tag{5}$$

Then, we use the fused feature $\mathbf{F}_2$ to produce a preliminary prediction O_2. We perform the whole process three times. Finally, we fuse the multi-scale features $\mathbf{F}_1, \ldots, \mathbf{F}_4$ to predict the final change map O.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

We conduct experimental evaluations on two publicly available CD datasets, including LEVIR-CD [15] and WHU-CD [16]. LEVIR-CD contains 637 RS image pairs with the size of 1024×1024 and WHU-CD includes one image pair with the size of $32,507 \times 15,354$. Following [7], we cut samples of size 256×256 with non-overlapping and split them into three parts, with 7,120/1,024/2,048 for LEVIR, and 5,947/743/744 for WHU, which are used to make train/validation/test datasets, respectively. We use five quantitative metrics for evaluation, including precision, recall, F1 score, intersection over union (IoU) score, and overall accuracy (OA).

3.2. Experimental Settings

The proposed model is implemented under the PyTorch1.10 framework with a GeForce GTX 2080Ti GPU. We use a mini-batch of 8 to train our model. During the training process, we follow [6] to perform data augmentation, adopt the Cross-Entropy (CE) loss ($\mathcal{L}_{CE}$) [23] and the AdamW algorithm [24] to optimize the network, with a weight decay

Table 1. Experimental results on the LEVIR-CD and WHU-CD datasets (%). The red and blue indicate the best performance and the second performance respectively.

Method	Pub.	Year	LEVIR-CD					WHU-CD				
			Precision↑	Recall↑	F1↑	IoU↑	OA↑	Precision↑	Recall↑	F1↑	IoU↑	OA↑
FC-EF [5]	ICIP	2018	86.91	80.17	83.40	71.53	98.39	71.63	67.25	69.37	53.11	97.61
IFNet [18]	ISPRS	2020	94.02	82.93	88.13	78.77	98.87	96.91	73.19	83.40	71.52	98.83
BIT [7]	TGRS	2021	89.24	89.37	89.31	80.68	98.92	86.64	81.48	83.98	72.39	98.75
ChangeFormer [6]	IGARSS	2022	92.05	88.80	90.40	82.48	99.04	88.22	79.86	83.83	72.16	98.89
P2V-CD [19]	TIP	2022	93.32	90.60	91.94	-	-	95.48	89.47	92.38	-	-
TransUNetCD [20]	TGRS	2022	92.43	89.82	91.11	83.67	-	93.59	89.60	93.59	84.42	-
EGRCNN [21]	TGRS	2022	88.58	91.13	89.84	81.55	98.85	90.92	89.41	90.16	82.08	99.29
ENCL-CD [22]	GRSL	2022	91.27	89.72	90.49	82.63	99.04	94.56	86.77	90.50	82.64	99.34
Ours	Submission		93.79	90.70	92.22	85.56	99.22	97.35	92.93	95.09	90.64	99.62

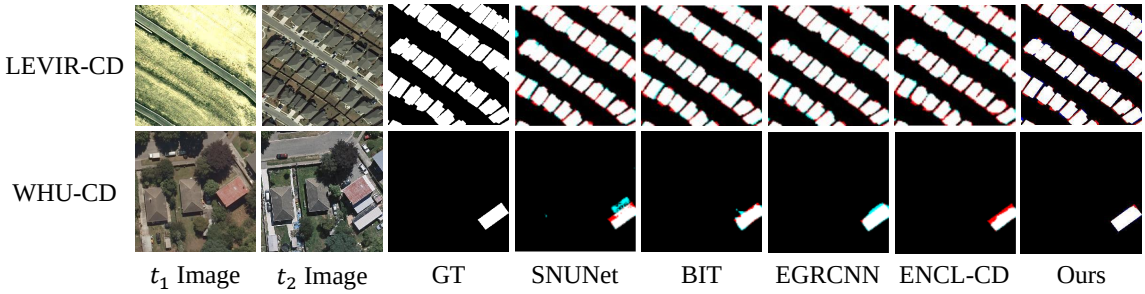

Figure 4. Visual comparisons on the LEVIR-CD and WHU-CD datasets, where the FNs and FPs are indicated in red and blue, respectively.

Table 2. Ablation experiments on LEVIR-CD (%) for the focus subnet and the refinement subnet.

ID	Subnet			LEVIR-CD				
	Glance	Focus	Refinement	Precision↑	Recall↑	F1↑	IoU↑	OA↑
1	✓			92.05	88.80	90.40	82.48	99.04
2	✓	✓		91.47	**91.95**	91.71	84.69	99.15
3	✓		✓	93.06	90.74	91.89	84.99	99.18
4	✓	✓	✓	**93.79**	90.70	**92.22**	**85.56**	**99.22**

of 0.01 and beta values of (0.9, 0.999). The initial learning rate is set to $1e-4$ and decays linearly to 0 until training 200 epochs.

3.3. Experimental Results

Table 1 lists the quantitative results of our GFRNet with other state-of-the-art approaches. Specifically, on LEVIR-CD, our model achieves the best scores of 92.22%, 85.56%, 99.22% in terms of three important metrics. Although IFNet and EGRCNN achieve relatively high precision and recall individually, they do not strike a balance between the two metrics, resulting in one of the metrics even falling below that of earlier methods. On WHU-CD, our model achieves the best scores on all metrics among all the evaluated methods. Figure 4 illustrates the qualitative results of the evaluated methods, where our method achieves the least false positives (FPs) and false negatives (FNs). These favorable test results confirm our model's efficacy across diverse application scenarios.

3.4. Ablation Study

To validate the effectiveness of the key designs of our model, we conduct ablation experiments on the LEVIR-CD. Our baseline is derived from ChangeFormer, featuring only the glance subnet. As listed in Table 2, incorporating the focus subnet into the model results in substantial improvements, notably a 1.31% increase in terms of F1 score, a 2.2% improvement in terms of IoU, and the highest performance in terms of the recall. These enhancements stem from the focus subnet's capacity to steer the model's attention towards salient regions related to the changing areas. Furthermore, the refinement subnet contributes to a 1.49% boost in terms of F1 score and a 2.51% enhancement in terms of IoU, demonstrating its efficacy in guiding the model to concentrate on change areas.

4. CONCLUSION

This paper has presented a GFRNet for CD, which treats CD as a coarse-to-fine feature learning process, allowing the model to gradually focus on the desired change areas. Among it, we have designed the CSM to make the model focus on salient regions first. Then, guided by several preliminary predictions, the refinement subnet pays more attention to the desired changing regions. Extensive experiments on two widely used benchmark datasets have demonstrated the superiority of our proposed GFRNet to the SOTA methods.

5. REFERENCES

[1] Huiwei Jiang, Min Peng, Yuanjun Zhong, Haofeng Xie, Zemin Hao, Jingming Lin, Xiaoli Ma, and Xiangyun Hu, "A survey on deep learning-based change detection from high-resolution remote sensing images," *RS*, vol. 14, no. 7, 2022.

[2] Joseph Z Xu, Wenhan Lu, Zebo Li, Pranav Khaitan, and Valeriya Zaytseva, "Building damage detection in satellite imagery using convolutional neural networks," *arXiv preprint arXiv:1910.06444*, 2019.

[3] Baudouin Desclée, Patrick Bogaert, and Pierre Defourny, "Forest change detection by statistical object-based method," *RSE*, vol. 102, no. 1-2, pp. 1–11, 2006.

[4] Hui Luo, Chong Liu, Chen Wu, and Xian Guo, "Urban change detection based on dempster–shafer theory for multitemporal very high-resolution imagery," *RS*, vol. 10, no. 7, pp. 980, 2018.

[5] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch, "Fully convolutional siamese networks for change detection," in *ICIP*. IEEE, 2018, pp. 4063–4067.

[6] Wele Gedara Chaminda Bandara and Vishal M Patel, "A transformer-based siamese network for change detection," in *IGARSS*. IEEE, 2022, pp. 207–210.

[7] Hao Chen, Zipeng Qi, and Zhenwei Shi, "Remote sensing image change detection with transformers," *TGRS*, vol. 60, pp. 1–14, 2021.

[8] Guanghui Wang, Bin Li, Tao Zhang, and Shubi Zhang, "A network combining a transformer and a convolutional neural network for remote sensing image change detection," *RS*, vol. 14, no. 9, pp. 2228, 2022.

[9] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] Yang Wu, Hao Zhang, Lingyan Liang, Yaqian Zhao, and Kaihua Zhang, "Group-wise co-salient object detection with siamese transformers via brownian distance covariance matching," in *ICASSP*. IEEE, 2023, pp. 1–5.

[11] Yang Wu, Huihui Song, Bo Liu, Kaihua Zhang, and Dong Liu, "Co-salient object detection with uncertainty-aware group exchange-masking," in *CVPR*, June 2023, pp. 19639–19648.

[12] Tao Lei, Yuxiao Zhang, Zhiyong Lv, Shuying Li, Shigang Liu, and Asoke K Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *GRSL*, vol. 16, no. 6, pp. 982–986, 2019.

[13] Siyue Yu, Jimin Xiao, Bingfeng Zhang, and Eng Gee Lim, "Democracy does matter: Comprehensive feature mining for co-salient object detection," in *CVPR*, 2022, pp. 979–988.

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022, pp. 1290–1299.

[15] Hao Chen and Zhenwei Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *RS*, vol. 12, no. 10, pp. 1662, 2020.

[16] Shunping Ji, Shiqing Wei, and Meng Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *TGRS*, vol. 57, no. 1, pp. 574–586, 2019.

[17] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *NeurIPS*, vol. 34, pp. 12077–12090, 2021.

[18] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS*, vol. 166, pp. 183–200, 2020.

[19] Manhui Lin, Guangyi Yang, and Hongyan Zhang, "Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images," *TIP*, vol. 32, pp. 57–71, 2022.

[20] Qingyang Li, Ruofei Zhong, Xin Du, and Yu Du, "Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images," *GRSL*, vol. 60, pp. 1–19, 2022.

[21] Beifang Bai, Wei Fu, Ting Lu, and Shutao Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *TGRS*, vol. 60, pp. 1–13, 2021.

[22] Mingwei Zhang, Qiang Li, Yuan Yuan, and Qi Wang, "Edge neighborhood contrastive learning for building change detection," *GRSL*, 2022.

[23] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z Li, "Multi-label cnn based pedestrian attribute learning for soft biometrics," in *ICB*. IEEE, 2015, pp. 535–540.

[24] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.