SAM-CD: Change Detection in Remote Sensing Using Segment Anything Model

Faroq AL-Tam*
ELM Company
Riyadh - Saudi Arabia
faltam@elm.sa

Thariq Khalid
ELM Company
Riyadh - Saudi Arabia
tkadavil@elm.sa

Athul Mathew
ELM Company
Riyadh - Saudi Arabia
amathew@elm.sa

Andrew Carnell
ELM Company
London - UK
acarnell@elm.sa

Riad Souissi
ELM Company
Riyadh - Saudi Arabia
rsouissi@elm.sa

Abstract

In remote sensing, Change Detection (CD) refers to locating surface changes in the same area over time. Changes can occur due to man-made or natural activities, and CD is important for analyzing climate changes. The recent advancements in satellite imagery and deep learning allow the development of affordable and powerful CD solutions. The breakthroughs in computer vision Foundation Models (FMs) bring new opportunities for better and more flexible remote sensing solutions. However, solving CD using FMs has not been explored before and this work presents the first FM-based deep learning model, SAM-CD. We propose a novel model that adapts the Segment Anything Model (SAM) for solving CD. The experimental results show that the proposed approach achieves the state of the art when evaluated on two challenging benchmark public datasets LEVIR-CD and DSIFN-CD.

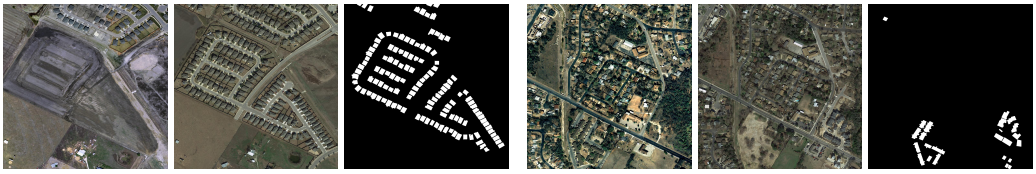
1 Introduction

Change detection (CD) refers to the detection of relevant changes while ignoring irrelevant differences (*e.g.* shadow and imagery impairments) of the same area at different periods of time. With the high rate of city development and deterioration of the natural environment, the importance of CD is higher than ever before [15]. CD plays an important role in environmental and surface change monitoring [8], urban planning [7], disaster evaluation [17], agriculture [7], among many other applications.

In bi-temporal change detection, given two input images, the pre-event image $I^{t_1} \in \mathbb{R}^{N^2 \times C}$ and the post-event image $I^{t_2} \in \mathbb{R}^{N^2 \times C}$, of the same place taken at two different times t_1 and t_2 where N and C are the spatial and spectral dimensions of the images. The binary CD refers to detecting the accumulated change, $\delta^{\tilde{t}} \in \{0,1\}^{N^2}$, during the period $\tilde{t} = t_2 - t_1$. Some examples of CD image pairs, from the LEVIR-CD dataset [5], are shown in Figure 1.

The complexity of satellite imagery data makes CD a challenging task. However, the recent advances in foundation models (FMs) [10, 16, 6] bring new opportunities. Recently prompt learning in computer vision [14] is gaining the attention of the researchers as an alternative to transfer learning and fine-tuning. It becomes possible due to the emergence of efficient self-supervised [10], and model-in-the-loop training procedures [16].

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.



(a) Changes from I^{t_2} only. (b) Changes from both I^{t_1} and I^{t_2}.

Figure 1: Binary change detection. Left to right: I^{t_1}, I^{t_2} and $\delta^{\tilde{t}}$. The pairs are from LEVIR-CD dataset.

The idea of prompt learning is to freeze the parameters of the foundation model and prompt it with new learnable features. Another powerful approach is adaptation [6], where a new module (adapter) is attached to the frozen foundation model and training is performed on the adapter parameters only.

Among the recently developed FMs is the Segment Anything Model (SAM) [16], which has demonstrated a great ability in zero-shot image segmentation of different modalities. Benefiting from the powerful capabilities of SAM, in order to solve the CD problem, has not been explored before and this article presents the first SAM-based CD model.

2 Related Work

The main scheme of solving CD follows the typical techniques used in dense prediction (*e.g.* semantic segmentation) where an encoder, with a convolutional neural network (CNN) [19], *e.g.* a ResNet [11] or Transformer [2] backbone, is used to extract features from the pre and post-event images. The features, from both images, are then fused to form a change latent space. This space is then ingested by a decoder module to predict the change masks. Regarding training, transfer learning is the most common training procedure for CD models, where the backbone layers are trained (or fine-tuned) alongside the change fusion and decoder layers.

Siamese neural networks are currently the most successful architectures [3, 18], and diverse feature fusion methods have been developed. Multi-scale fusion methods are developed in [5, 18] to enrich the change latent space. In [4] relation and scale-aware modules are developed to capture interactive information of the change in both images. Auxiliary losses can guide feature fusion and improve the quality of the detected masks [19]. The wide field of view, offered by Transformer backbones, is shown to improve the context of the change features [2], making it easier to capture long-range dependencies of the change features within the same image (using self-attention) and between the pre and post-event images (with cross-attention).

3 SAM-CD

SAM was created to segment an input image using a prompt [16]. It can take multiple sparse (points and bounding boxes) and/or dense prompts (binary masks). It was trained with model-in-loop with 1 billion masks. Please see [16] and Appendix 6.1 for further details.

The main challenge in solving CD using SAM is *how to model and prompt the change?* SAM takes a single input image and produces segmentation masks based on prompts (*e.g.* points) provided by the user. On the other hand, using SAM for CD requires decoding a change latent space (from both images) and learning how to prompt the change to the SAM's decoder.

To that end, two new components are developed: the change modeler and prompter (Figure 2a). The former receives both the pre and post-event images and creates a change latent space. This space is then used by the change prompter to instruct the decoder how to decode the change.

3.1 Change Modeler

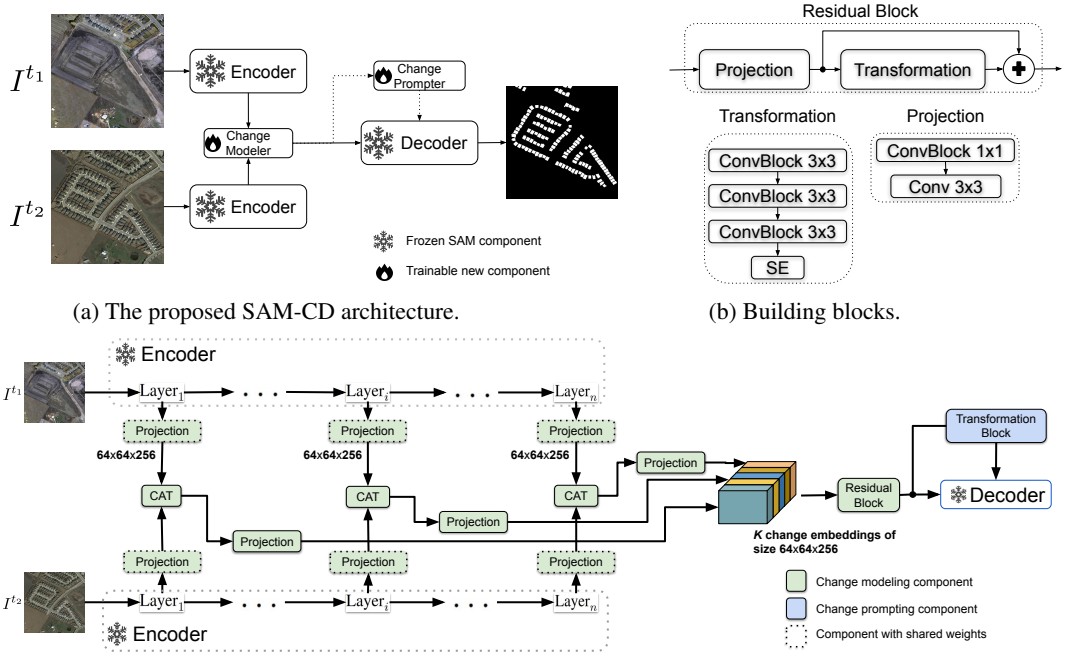
As shown in Figure 2c, both images are fed to the encoder. A subset, of length K, of feature maps are extracted from uniformly sampled layers of the encoder and transformed as follows: each feature map is projected from $64 \times 64 \times 678$ to an embedding of size of $64 \times 64 \times 256$, using a projection

block (Figure 2b). The resulting embeddings from both images are concatenated, on the channel dimension, and projected from $64 \times 64 \times 512$ to $64 \times 64 \times 256$ to form K change embeddings of size $64 \times 64 \times 256$. These are then concatenated and passed through a residual block to obtain a global change embedding of size $64 \times 64 \times 256$, ready to be used by the decoder.

In practice, setting $K = 5$ layers is found sufficient. The `ConvBlock` in Figure 2b is a sequence of: a 2D convolution layer (`Conv2D`), a layer normalization [1] (`LayerNorm`), and a GELU activation [12] (`GELU`). `SE` in the Transformation Block (Figure 2b) is a squeeze-and-excitation layer [13] used to recalibrate the transformed embedding before feeding it to SAM. `CAT` in Figure 2c is the concatenation operator applied on the channel dimension.

3.2 Change Prompter

The prompter is a simple transformation block (Figure 2b), which transforms the change embedding (obtained by the change modeler) into a prompt embedding of size $64 \times 64 \times 256$. Both embeddings from the modeler and prompter are fed to the decoder to obtain the change masks, as shown in Figures 2a and 2c.



(a) The proposed SAM-CD architecture.

(b) Building blocks.

(c) Change modeling and prompting.

Figure 2: The proposed SAM-CD architecture and its main building blocks.

The aim of keeping the design of SAM-CD simple, *i.e.* using simple building blocks and operators without complex feature fusion modules, is to minimize the complexity of SAM-CD in order to validate the effectiveness of using SAM in solving CD.

4 Results

To evaluate the proposed work, it is compared to the state of art methods using the same Key Performance Indicators (KPIs) found in the literature: Precision (Pre), Recall (Rec), F1-score (F1), and the Intersection over Union (IoU) between the predicted and Ground Truth (GT) masks. The evaluation protocol and implementation procedure are detailed in Appendix 6.2.

4.1 Datasets

4.1.1 LEVIR-CD

This is a very high-resolution imagery dataset for building change detection [5]. Each image is an RGB tile with a size of 1024×1024 pixels, with binary labels. LEVIR-CD is collected from Google Earth and contains 637 image pairs of which 448 are used for training, 64 for validation, and 129 for testing. The bi-temporal difference ranges from 5 to 14 years in some pairs.

4.1.2 DSIFN-CD

This is a high-resolution imagery dataset collected from six different cities in China using Google Earth [20]. It contains 3940 pairs (of size 512×512 cropped from six large tiles). The training and validation sets are selected from five cities with 3600 and 340 pairs for training and validation, respectively. The testing set contains 48 pairs from the sixth city only, and this dataset contains various classes of changes: roads, buildings, croplands, and water bodies with bi-temporal difference ranges from 5 to 17 years. Making it a challenging dataset due to change complexity and the intra-city diversity between the training and testing sets.

4.2 Discussion

The results of SAM-CD and other models for both datasets are shown in Table 1. Regarding LEVIR-CD, the proposed model achieved the best results and is ahead of the second best model USSFC-NET by $\approx +2.46\%, -2.51\%, 0.02\%$ and 0.48% in precision, recall, IoU, and F1-score, respectively. When considering the results of DSIFN-CD, which is more challenging, the proposed model achieves the state of art in all metrics except the recall. Outperforming the second best model USSFC-NET with $\approx +0.77\%, -2.07\%, +2.44\%$, and $+1.55\%$ in in precision, recall, IoU, and F1-score, respectively.

It deserves noting that, the proposed model is consistent in both datasets, which is reflected by the high F1-score in both datasets. In addition, visual inspection results are reported in Appendix 6.3.

Model	**Year**	**LEVIR-CD**				**DSIFN-CD**			
		Pre	**Rec**	**IoU**	**F1**	**Pre**	**Rec**	**IoU**	**F1**
FC-Siam-Conc [3]	2018	91.99	76.77	71.96	83.69	59.08	62.80	43.76	60.88
STANet [9]	2020	83.81	91.00	77.40	87.26	51.48	36.40	27.11	42.65
BIT [2]	2022	89.24	89.37	80.68	89.31	56.36	62.79	42.25	59.40
USSFC-NET [18]	2023	89.70	**93.42**	84.36	91.04	63.73	**76.32**	53.20	69.47
SAM-CD (proposed)		**92.16**	90.91	**84.38**	**91.52**	**68.42**	74.25	**55.64**	**71.02**

Table 1: The change detection results of the LEVIR-CD and DSIFN-CD *test* sets.

5 Conclusion

This work presented SAM-CD, a change detection model developed based on the Segment Anything Model (SAM). Two main components were added to the SAM architecture. The change modeler and prompter. During training, these two new components were the only trainable parts of SAM-CD architecture, while the parameters of SAM were kept frozen. SAM-CD not only achieved state of the art when evaluated on two challenging CD datasets, but also produced consistent results.

We believe that employing foundation models for CD will help to accelerate the analysis of climate change problems and this work will unlock new CD solutions that leverage prompt learning and foundation models adaptation. In the future, this work will be extended with new prompting designs and advanced feature fusion methods, and more CD use-cases will be evaluated.

Acknowledgments and Disclosure of Funding

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Wele Gedara Chaminda Bandara and Vishal M. Patel. A transformer-based siamese network for change detection. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210, 2022.

[3] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067, 2018.

[4] Chao-Peng Chen, Jun-Wei Hsieh, Ping-Yang Chen, Yi-Kuan Hsieh, and Bor-Shiun Wang. Saras-net: Scale and relation aware siamese network for change detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[5] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10), 2020.

[6] Shoufa Chen, Chongjian GE, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16664–16678. Curran Associates, Inc., 2022.

[7] Begüm Demir, Francesca Bovolo, and Lorenzo Bruzzone. Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1):300–312, 2013.

[8] Leila M.G. Fonseca, Thales S. Körting, Hugo do N. Bendini, Cesare D. Girolamo-Neto, Alana K. Neves, Anderson R. Soares, Evandro C. Taquary, and Raian V. Maretto. Pattern recognition and remote sensing techniques applied to land use and land cover mapping in the brazilian savannah. *Pattern Recognition Letters*, 148:54–60, 2021.

[9] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan. Augfpn: Improving multi-scale feature learning for object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12592–12601, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[14] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.

[15] Lazhar Khelifi and Max Mignotte. Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis. *IEEE Access*, 8:126385–126400, 2020.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[17] Maja Kucharczyk and Chris H. Hugenholtz. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sensing of Environment*, 264:112577, 2021.

[18] Tao Lei, Xinzhe Geng, Hailong Ning, Zhiyong Lv, Maoguo Gong, Yaochu Jin, and Asoke K. Nandi. Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[19] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[20] 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 Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020.

[21] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing.

6 Appendix

6.1 A: SAM-CD Formulation

SAM has three main building blocks: the image and prompt encoders, and a mask decoder. At the inference time, SAM receives an image $I \in \mathbb{R}^{N^2 \times C}$ and a prompt P and produces a set of masks.

The image encoder, E_{in}, is a masked auto-encoder (MAE) model [10], and there are two prompt encoders, E_{sparse} for sparse prompts, and E_{dense} for dense prompts. For sparse prompts, E_{sparse} is a simple layer that learns positional embeddings of points and bounding boxes. For dense prompts, E_{dense} is a CNN module. The mask decoder D_{mask} is a Transformer module that decodes the image embeddings using the prompt embeddings. Formally, we can model SAM as:

$$\begin{aligned} &\textbf{SAM} : D_{\text{mask}}(E_{\text{in}}(I)) \\ &\text{prompted by:} \quad \left\{E_{\text{dense}}(P_{\text{mask}}), \texttt{Cat}_C\left(E_{\text{sparse}}(P_{\text{points}}), \tau^{\text{out}}\right)\right\} \end{aligned} \tag{1}$$

where τ^{out} is the set of output trainable tokens, and $\texttt{Cat}_C$ is a concatenation operator applied on the channel dimension.

For a given CD dataset $\mathcal{X}$, and a pre-trained SAM model with frozen parameters θ, and the change trainable parameters ϑ, the CD objective is to find ϑ that minimizes the change prediction function f:

$$\min_{\vartheta} \mathbb{E}_{\left\{(I^{t_1}, I^{t_2}), \delta^{\tilde{t}}\right\} \sim \mathcal{X}} \mathcal{L}(f(I^{t_1}, I^{t_2}; \theta, \vartheta), \delta^{\tilde{t}}) \tag{2}$$

where f is the SAM-CD model, and $\mathcal{L}$ is the adopted loss function which is a combination of the dice and binary cross entropy losses [21]:

$$\mathcal{L}(M^{\tilde{t}}, \delta^{\tilde{t}}) = -\frac{1}{B} \sum_{b=1}^{B} \left(\frac{1}{2} \delta_b^{\tilde{t}} \log M_b^{\tilde{t}} + \frac{2\delta_b^{\tilde{t}} M_b^{\tilde{t}}}{\delta_b^{\tilde{t}} + M_b^{\tilde{t}}} \right), \tag{3}$$

where $M^{\tilde{t}}$ is the obtained logits from f and B is the batch size.

Following the presented formulation for SAM in (1), SAM-CD can be formulated likewise as:

$$\begin{aligned} &\textbf{SAM-CD} : D_{\text{mask}}(S^{\tilde{t}}) \\ &\text{prompted by:} \quad \left\{P^{\tilde{t}}, \texttt{Cat}_C(\phi, \tau^{\text{out}})\right\} \end{aligned} \tag{4}$$

where $S^{\tilde{t}}$ is the change embedding obtained by the change modeler, and $P^{\tilde{t}}$ is the change prompt embedding obtained from the change prompter. $\texttt{Cat}_C$ is the concatenation operator applied on the channel dimension, and ϕ is the SAM's pre-trained `no-prompt` token.

6.2 B: Evaluation protocol and implementation details

The KPIs used in the evaluation are the same ones used in the literature:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{Intersection over Union} = \frac{TP}{TP + FN + FP} \tag{7}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

The proposed architecture is implemented in PyTorch, and the adopted SAM model is the `sam-vit-h` [16], where its weights are frozen while the change modeler and prompter blocks are initialized randomly. The number of layers sampled from SAM's encoder to model the change is $K = 5$. The optimizer used to train the SAM-CD is the `AdamW`, with the adopted hyper-parameters as shown in Table 2.

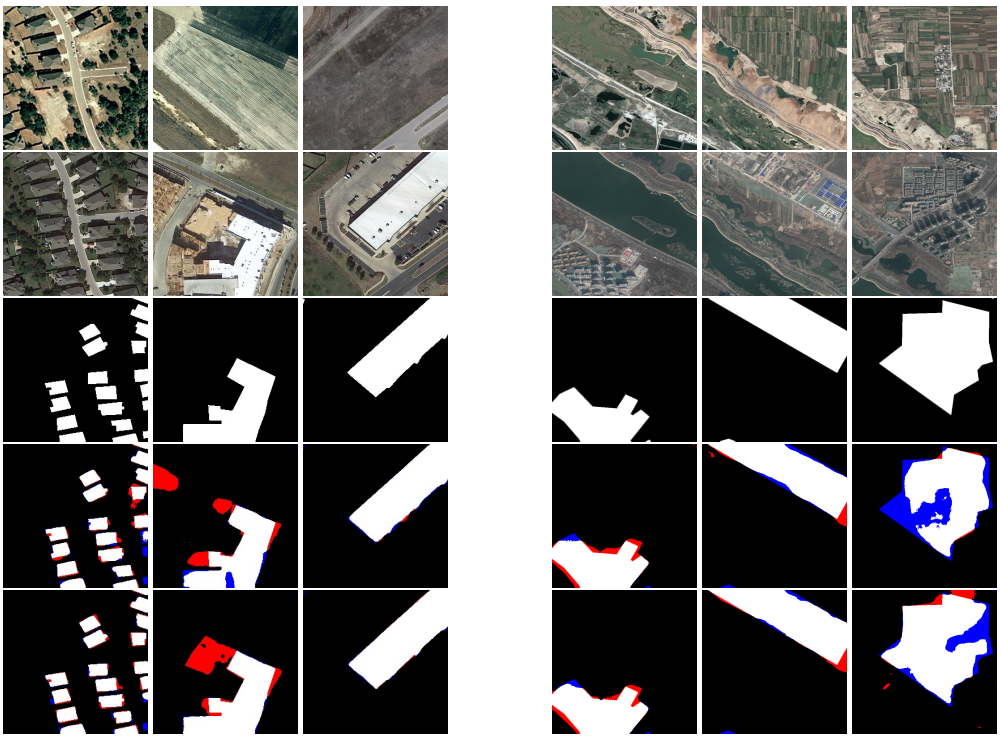
Parameter	**Value**
Initial RL	0.001
Momentum	0.9
Weight decay	0.00001
Training Schedules	400 epochs (for LEVIR-CD) and 100 epochs (for DSIFN-CD)
Scheduler	CosineSchedualer with minimum RL = 0
Hardware	V100 GPU
Batch Size	8

Table 2: The parameters being adopted for training.

During training, flip, rotate, scale, and clip augmentations were used and the input is scaled to 1024×1024 for both datasets, and no augmentation was applied during the inference time.

6.3 C: Visual Results

The visual inspection results are shown in Figures 3a and 3b. Both models are able to obtain good masks. As can be seen, SAM-CD predicts less false positive/negative areas compared to USSFC-NET. This is more clear in the DSIFN-CD test set, where it is more challenging due to the complexity of the background and diversity of the change classes.



(a) Results from the LEVIR-CD test set. (b) Results from the DSIFN-CD test set.

Figure 3: The results from both LEVIR-CD and DSIFN-CD test sets. Blue: false negative, and red: false positive. Rows top to down: pre-event image, post-event image, ground truth, USSFC-NET [18] predictions, and SAM-CD predictions.