TextSCD: Leveraging Text-based Semantic Guidance for Remote Sensing Image Semantic Change Detection

Haiyan Huang[a], Qimin Cheng[b,d,*], Duowang Zhu[a], Xiao Huang[c], Qunshan Zhao[d]

[a]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
[b]School of Electronics Information and Communications, Huazhong University of Science and Technology, Wuhan, China
[c]Department of Geosciences, University of Arkansas, Fayetteville, USA
[d]Urban Big Data Centre, School of Social and Political Sciences, University of Glasgow, Glasgow, UK
Email: chengqm@hust.edu.cn

Keywords: Semantic change detection, Vision-language representation learning, Multi-task learning, Remote sensing.

Abstract

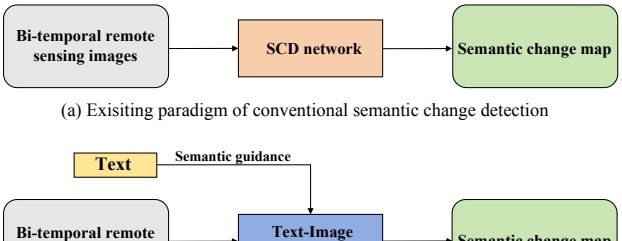
Semantic change detection (SCD) in remote sensing image aims to identify semantic alterations between bi-temporal images captured at the same geographic location. SCD is extensively applied in fields such as environmental monitoring and disaster assessment. Despite significant advancements in deep learning leading to numerous successful approaches, most existing methods primarily rely on visual representation learning, thereby overlooking the potential benefits of multimodal data. Recently, vision-language models have demonstrated outstanding performance across various downstream tasks. In this paper, we propose a novel framework named TextSCD that leverages text-based semantic information to guide the generation of semantic change maps. Our approach integrates Gemini to generate change descriptions between bi-temporal images and employs a multi-level semantic extraction method to capture features from both images and their corresponding captions. Furthermore, we introduce a semantic text-guided interaction module that facilitates the effective integration of visual and textual features, enhancing multimodal knowledge transfer and the extraction of discriminative features. This design effectively reduces false detections and omissions. We validate the effectiveness of our model on the SECOND dataset, achieving notable improvements in overall accuracy for semantic change detection.

1. Introduction

Our planet is undergoing substantial transformations due to natural phenomena and persistent human activities (Gueguen and Hamid, 2016; Kennedy et al., 2009). The rapid advancements in Earth observation technology have increased access to high-resolution image, thereby expanding the application of semantic change detection (SCD) to areas such as urban planning, disaster monitoring, and natural resource management (Ochtyra et al., 2020; Liu et al., 2021; Zheng et al., 2021). Consequently, accurate semantic change detection is essential for a comprehensive understanding of urban development processes and effective disaster monitoring.

Traditional methods for SCD predominantly rely on either pixel-based or object-based approaches. Pixel-based techniques (Yan et al., 2019; Wu et al., 2017) detect changes by classifying individual pixels and comparing these classifications over time, which aids in identifying region-specific changes and types of alterations. Object-based approaches (Lv et al., 2020; Desclée et al., 2006) aim to reduce pixel-level noise and focus on identifying target objects; however, they face challenges in complex scenes where segmentation quality directly affects change detection accuracy.

With the advent of deep learning, image encoders for SCD, such as convolutional neural networks (Xia et al., 2022) and transformers (Zheng et al., 2022; Niu et al., 2023), have surpassed traditional methods. An end-to-end SCD network paradigm is illustrated in Figure 1. These methods can be further categorized into single-branch, dual-branch, and multi-task approaches



(a) Exisiting paradigm of conventional semantic change detection

(b) Proposed paradigm for leveraging text-based semantic guidance

Figure 1. (a) Conventional approaches for semantic change detection predominantly rely on single-modal image as input to generate semantic change maps. (b) Our proposed text-guided semantic change detection framework incorporates textual information through multimodal semantic interaction mechanisms, enhancing the precision of change identification by leveraging cross-modal contextual understanding.

(Tian et al., 2022; Cui and Jiang, 2023). Single-branch methods label samples based on change types and feed image difference information into a feature extraction network, directly extracting features from altered target objects (Bovolo and Bruzzone, 2006). However, these methods suffer from category imbalance due to the squared labeling of semantic categories in image segmentation. Dual-branch methods classify images at different temporal instances and compute change detection results by comparing these classifications, but they are limited by their dependency on classification results and error accumulation over time, especially at object boundaries in complex scenes (Xia et al., 2022). Multi-task frameworks, widely adop-

* Email: chengqm@hust.edu.cn

ted in recent works, combine semantic segmentation and binary change detection, achieving efficient semantic change detection through feature sharing between different tasks. This approach enhances feature extraction and change detection accuracy by jointly learning related tasks (Ding et al., 2022).

Recently, vision-language contrastive pre-training has garnered extensive attention in the field of computer vision, demonstrating superior semantic understanding by deeply learning the associations between images and text. This methodology has been successful in cross-modal retrieval (Xia et al., 2023), video captioning (Wu et al., 2023), and question answering (Parelli et al., 2023). In the domain of remote sensing, multimodal data can provide additional information to unimodal data, further enhancing deep neural networks' image understanding capabilities (Wang et al., 2024; Duan et al., 2024; Huang et al., 2023). For instance, Rahhal et al.(2022) proposed a textual description approach to improve remote sensing image retrieval. Li et al. (2021) constructed class-level semantic representations using domain knowledge to address inconsistencies between the visual image space and the semantic space, thereby improving image scene classification performance. Lu et al (2023) integrated textual information with a target detection framework to supplement missing class information in limited images, enhancing new class detection performance. Despite the remarkable zero-shot performance demonstrated by CLIP, its potential for semantic change detection in remote sensing images remains largely unexplored.

To address these challenges, we propose TextSCD, a novel and effective framework that leverages text-based semantic guidance for fine-grained semantic change detection in remote sensing images. TextSCD incorporates text modality information, capturing rich semantic content from remote sensing data. In the encoder, bi-temporal images are processed through a Siamese neural network, while text information is encoded using a text encoder. We then design a text-guided semantic fusion module to effectively utilize multimodal information. Finally, a lightweight decoder efficiently outputs the detected semantic changes.

The principal contributions of this paper are as follows:

1. We develop a novel semantic change detection framework, TextSCD. To the best of our knowledge, this is a pioneering study that incorporates multimodal vision-language information into the semantic change detection task.

2. We introduce Gemini to generate change captions from bi-temporal images. Additionally, we design a semantic text-guided interaction module to leverage multimodal information and enhance the capture of semantic changes.

3. We conduct experiments on the SECOND dataset. The experimental results demonstrate TextSCD's effectiveness and superiority in accurately detecting semantic changes. Effectively extracting and utilizing multimodal information from images provides a new perspective for the remote sensing community.

The remainder of this paper is organized as follows: Section 2 introduces the related work; Section 3 presents the architecture of TextSCD; Section 4 discusses the experimental results on the SECOND dataset; and Section 5 concludes the paper.

2. Related Work

This section reviews the pertinent literature, organized into two primary domains: deep learning-based semantic change detection and vision-language pre-training for remote sensing image analysis.

2.1 Deep Learning-Based Semantic Change Detection

SCD in remote sensing images aims to identify changes in semantic categories between bi-temporal images, providing not only the locations of these changes but also their nature (i.e., the transition from one class to another). Given its wide range of applications across various fields, SCD has attracted significant research interest. Daudt et al. (2019) made substantial contributions to the field by employing the canonical Fully Convolutional Network (FCN) architecture, achieving commendable results in both semantic segmentation and change detection tasks. Their approach laid the groundwork for subsequent advancements in SCD methodologies. Building on this foundation, Yang et al. (2021) addressed the challenge of classification ambiguity, particularly prevalent in scenarios involving asymmetric changes. They introduced the SECOND dataset—a benchmark for semantic change detection—developed using a modular approach with diverse structures. This dataset has since become instrumental in refining and evaluating change detection models. Advancing the state of the art, Cui et al. (2023) proposed MTSCD-Net, a framework designed to extract multi-scale features from bi-temporal images. Their exploration of an encoder based on Swin Transformer aggregation, coupled with the use of spatial attention for decoding with prior information, demonstrated the effectiveness of this approach in capturing fine-grained changes. To tackle the issue of limited change samples in semantic change detection, Ding et al. (2024) introduced ScanNet, a model that synergizes the strengths of Convolutional Neural Networks (CNNs) and Transformer architectures. This integration enables joint spatial and temporal modeling, establishing prior constraints that balance accuracy and efficiency in detection tasks. Collectively, these studies have significantly contributed to the evolving landscape of semantic change detection. By introducing innovative solutions to address the multifaceted challenges in this domain, they have enhanced the modeling of the intricate relationship between semantics and changes, paving the way for more accurate and efficient detection methods.

2.2 Vision-Language Pre-Training

Classical semantic change detection tasks typically follow a pre-training+fine-tuning paradigm, wherein models are pre-trained on datasets like ImageNet. However, such pre-training may not fully adapt to the specific requirements of remote sensing, as remote sensing images differ significantly from natural images. In contrast, multimodal learning leverages large-scale image-text datasets for pre-training, providing richer prior knowledge. Large-scale pre-trained vision-language models, such as CLIP, excel at handling diverse modalities. CLIP uses natural language as a supervisory signal to learn visual features via contrastive learning on web-scale image-text data. This semantic-level language supervision enables the visual network to acquire high-quality, semantically rich visual features, significantly enhancing performance in cross-modal and fine-grained visual tasks. Given the strong performance of large-scale vision-language pre-training, multimodal learning has garnered considerable attention for downstream tasks, proving effective across various computer vision domains, including

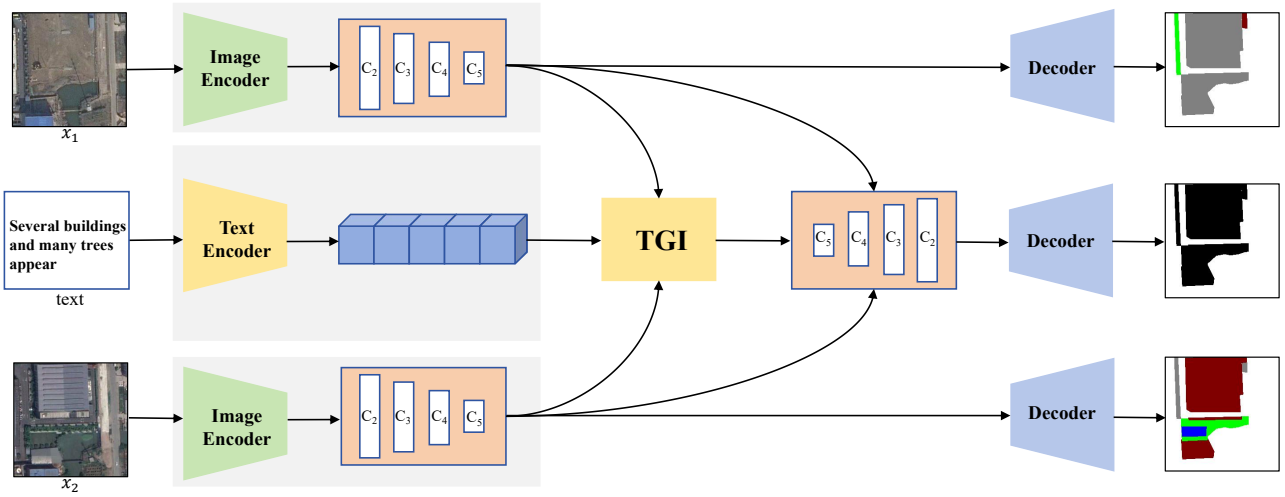


Figure 2. Overview of the proposed TextSCD framework. During training, the model is jointly fed the original images and corresponding captions, with the semantic text-guided interaction module enhancing the change detection process.

super-resolution (Zhang et al., 2024), object detection (Wei et al., 2024), and image segmentation (Zhou et al., 2023). Motivated by this progress, we aim to develop an effective framework to adapt the CLIP model for the semantic change detection task.

3. Methodology

Inspired by the success of large-scale pre-trained vision-language models, such as CLIP (Radford et al., 2021)) in downstream tasks, we construct a multimodal encoder that integrates CLIP text encoder with the ResNet34 image encoder. Subsequently, the text-guided module is proposed to facilitate the fusion of image and text features. Finally, we choose a simple yet effective decoder to restore features to the original resolution. The proposed framework TextSCD is shown in 2.

3.1 Problem Formulation

The conventional paradigm of semantic change detection tasks is framed as the process of ingesting a pair of multi-temporal remote sensing images (e.g., x_1, x_2) and employing a semantic change detection network (e.g., θ_{scd}) to produce a feature map indicative of semantic changes. This network is engineered to learn a predefined change detection function $\mathcal{F}_{scd}$ corresponding to the change detection task. It can be articulated as:

$$F^m = \mathcal{F}_{scd}(x_1, x_2; \theta_{scd}). \tag{1}$$

Traditional approaches predominantly leverage datasets pre-trained on ImageNet, which are devoid of remote sensing specific prior knowledge. Our research delves into harnessing textual information to surmount the semantic comprehension deficiencies inherent in conventional image tasks, augmenting the model's semantic understanding of images, and pioneering a novel paradigm for semantic change detection in remote sensing image. Consequently, the semantic change detection task is redefined as:

$$F^m = \mathcal{F}_{scd}(x_1, x_2, T_{text}; \theta_{scd}). \tag{2}$$

3.2 The Architecture of TextSCD

3.2.1 Multimodal Encoder The image encoder in the semantic change detection task takes the multi-temporal remote sensing images as input. To extract both spatial and deep information effectively, we employ ResNet34 as the base feature extractor. The encoding process can be represented as follows:

$$F^1_{img} = \mathcal{F}^I_v(x_1), \tag{3}$$

$$F^2_{img} = \mathcal{F}^I_v(x_2), \tag{4}$$

where $x_1 \in \mathbb{R}^{H\times W\times C}$ and $x_2 \in \mathbb{R}^{H\times W\times C}$ represent the multi-temporal remote sensing images, with H and W denoting the height and width of the images, and C representing the number of channels. $\mathcal{F}^I_v$ is the image encoder.

We employ a text semantic encoder that transforms the given text T_{text} into an embedding space, which captures the semantic nuances essential for the change detection process. Leveraging the pre-trained CLIP model, renowned for its efficacy in text feature extraction, we freeze the weights of the text encoder to preserve linguistic consistency. This transformation is represented as:

$$F_{\text{text}} = \mathcal{F}^I_t(T_{text}), \tag{5}$$

where $F_{\text{text}} \in \mathbb{R}^{L\times D_t}$ represents the text semantic feature, L is the length of the text sequence, and D_t is the dimension of the text features. $\mathcal{F}^I_t$ is the text encoder. The features extracted from different yet semantically analogous texts must be in close proximity within the reduced Euclidean space.

3.2.2 Semantic text-guided interaction module To ensure the text feature generated by the text semantic encoder can guide the change process of visual perception, we propose a text-guided interaction module (TGI) to fuse visual and text features, generating context-rich feature representations. Specifically, we model the bi-temporal images with an MLP,

$$F^i_{Diff} = MLP([x^i_1; x^i_2; x^i_{diff}]), \tag{6}$$

where $[;]$ denotes the concatenation along the channel dimension. The concatenated features pass through the MLP module

to extract spatiotemporal correlations, generating bitemporal aggregated features.

The input of TGI is the visual feature $F_{Diff} \in \mathbb{R}^{C\times H\times W}$ and text feature $F_{text} \in \mathbb{R}^{L\times D_t}$. The total length of text L is 50 for the expression-based setting. We apply a linear transformation to the text features to align their dimensions with those of the visual features. To utilize multimodal information and capture semantic information changes, we employ a cross-attention mechanism guided by the text features, as shown in Figure 3. This enables the visual features to attend to the most relevant textual descriptions, enhancing the discrimination of semantic changes. The fused features are then incorporated back into the original visual feature representation using a residual connection, preserving essential spatial details while integrating textual guidance. The process is as follows:

$$F_o^i = F_{Diff}^i + CrossAtten(LN(F_{Diff}^i), LN(F_{text})), \tag{7}$$

where $CrossAtten(\cdot)$ means cross attention, $i \in (1, 2, 3)$ denotes the index of feature layers, $LN(\cdot)$ denotes Layer Normalization (Lei Ba et al., 2016). The attention weights are multiplied by the value vector (F_{text}) to generate the output features. We also use a residual connection to add the fused features to the original visual features.

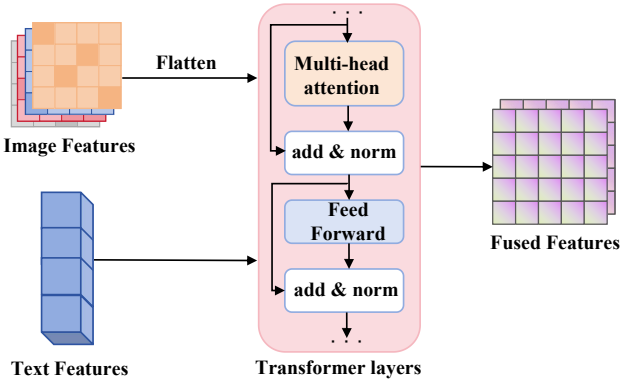


Figure 3. Illustration of the semantic text-guided interaction module.

3.2.3 Decoder Different complex designs of existing methods, we employ a simple decoder, which shows the effectiveness of TextSCD. Specifically, we perform upsampling and reconstruction of feature maps for semantic segmentation tasks. Upon initialization, it sets up a series of sequential layers, including both convolutional and transposed convolutional layers, to progressively upscale feature maps from the encoder. The upsampling process for each layer can be represented as:

$$p_i^j \leftarrow Deconv_{4\times4}(Conv_{1\times1}(p_{i+1}^j)) + p_i^j, \tag{8}$$

where i represents the layer index. After upsampling to the original resolution, a 2D convolution with a kernel size of 3×3 is applied as the classifier to generate the probability map.

3.2.4 Loss Function The loss functions employed to supervise the learning of TextSCD include dice loss (Milletari et al., 2016), cross-entropy loss, and cos similarity loss (Ding et al., 2022).

$$L_{dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2}, \tag{9}$$

where N is the total number of pixels in the image. p_i and g_i take binary values (0 or 1).

$$L_{bce}(Y, P) = -\frac{1}{N}\sum_{i=1}^{N}[Y_i \log_2 P_i + (1 - Y_i)\log_2(1 - P_i)], \tag{10}$$

where Y and P denote the ground truth and predicted change map, respectively, and N is the number of total pixels of a change map.

$$L_{sim} = \begin{cases} 1 - \cos\left(P_1^{i,j}, P_2^{i,j}\right), & \text{if } GT^{i,j} = 0 \\ \cos\left(P_1^{i,j}, P_2^{i,j}\right), & \text{if } GT^{i,j} = 1 \end{cases} \tag{11}$$

where $P_1^{i,j}, P_2^{i,j}$ denote the predicted maps in (i, j), $GT^{i,j}$ represents the ground truth, cos denotes the function of cosine similarity.

$$L = L_{dice} + L_{bce} + L_{sim}. \tag{12}$$

4. Experiments

4.1 Experimental Settings

Implementation details. We resize the image to 512×512. The training batch size is set to 8, with an initial learning rate of 0.1. The Stochastic Gradient Descent (SGD) optimizer is utilized with a momentum of 0.9 for faster convergence and a weight decay of 5×10^{-4} to prevent overfitting. For the text encoder, we use the pre-trained weights on CLIP (ViT-B-16). All experiments are executed on two NVIDIA 3090 GPUs within the pytorch framework.

Dataset. To validate the effectiveness of our model, we use the commonly used SECOND dataset (Yang et al., 2021). The resolutions range from 0.3m to 0.5m. This dataset focuses on detecting changes across six distinct land cover types: non-vegetated ground, trees, low vegetation, water, buildings, and playgrounds. The ratio of training and testing is 4:1.

Evaluation metrics. We use four common evaluation metrics to assess the performance of the semantic change detection algorithm, including OA, mIoU, SeK and F1. OA represents the numeric ratio between the correctly classified pixels and the total image pixels, while mIoU and SeK are calculated based on the discrimination of change/no-change classes. F1 specifically focuses on the accuracy obtained in the change areas.

Text rephrasing. For the change captioning of the bi-temporal remote sensing images, we adopt Gemini to generate change descriptions using this instruction: Change detection task: Use a sentence to summarize the differences in buildings, roads, vegetation, etc.

4.2 Results and analysis

We adopted an encoder-decoder framework that utilizes text guidance to generate semantic change detection maps. The quantitative experimental results on the SECOND dataset are

Figure 4. Qualitative semantic change detection results of various comparative methods on the SECOND dataset.

presented in Table 1, it reports the comparative experimental results of TextSCD on the SECOND dataset against other classical and the latest semantic change detection methods, including HRSCD-S4 (Daudt et al., 2019), SCDNet (Peng et al., 2021), MTSCD (Cui and Jiang, 2023), HGINet (Long et al., 2024).

Table 1. Comparison of Different Semantic Change Detection Methods

Method	F1	mIoU	OA	SeK
HRSCD-S4	57.04	69.94	85.97	16.72
SCDNet	59.34	71.28	86.57	19.26
MTSCD	60.50	71.52	86.70	20.26
HGINet	61.02	71.54	87.21	21.28
Baseline	60.67	72.25	87.18	20.90
+TGI (TextSCD)	61.90	72.38	87.67	21.66

From this table, we can see that HGINet significantly outform the other comparative methods, this enhancement is mainly attributed to the introduction of GCN can effectively model the interaction of bi-temporal images. Our proposed TextSCD achieves the best performance across four SCD evaluation metrics, which improved the F1 score by 0.88% and mIoU score by 0.84% compared with the state-of-the-art method. Figure 4 demonstrates qualitative semantic change detection results of various comparative methods on the SECOND dataset. The results show that the proposed TextSCD can effectively mitigate false detections and omissions. The baseline of our model uses a ResNet34 encoder to process the features of dual-time remote sensing images, rather than utilizing multimodal features as in the method presented in this paper. Unlike previous methods, the text-guided interaction module in TextSCD helps to extract richer semantic information, which plays a crucial role in mitigating false negatives and missed detections. This combination of the visual and textual branches enables TextSCD to outperform the baseline methods by effectively reducing errors in the RSICD task.

5. Conclusion

In this paper, we propose TextSCD, a novel text-guided semantic change detection framework for high-resolution remote sensing images, called TextSCD that fully leverages information from multiple modalities. In the first stage, we innovatively introduced Gemini to generate change descriptions for remote sensing images, employing a multimodal encoder to extract both textual and visual features. In the second stage, we proposed a semantic text-guided interaction module that effectively fuses image and text features under the guidance of textual information. Finally, a simple yet effective decoder was utilized to generate semantic change maps. Experimental results demonstrate that our semantic text-guided guided method provides sufficient proposals for the SCD task, achieving a high F1 score and improving the accuracy of semantic change detection compared to classical and state-of-the-art methods. In future work, we will explore the integration of multimodal information in semi-supervised or unsupervised semantic change detection tasks to overcome challenges associated with pixel-level annotations.

6. Acknowledgement

This work was supported by the National Key Research and Development Program of China (Grant No.2023YFB3906102), the China Scholarship Council (No.202406160037), Fundamental Research Fund Program of LIESMARS (4201-420100071) and the National Natural Science Foundation of China under Grant (No.42271352).

References

Al Rahhal, M. M., Bazi, Y., Alsharif, N. A., Bashmal, L., Alajlan, N., Melgani, F., 2022. Multilanguage transformer for improved text to remote sensing image retrieval. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 9115–9126.

Bovolo, F., Bruzzone, L., 2006. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1), 218–236.

Cui, F., Jiang, J., 2023. MTSCD-Net: A network based on multi-task learning for semantic change detection of bitemporal remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 118, 103294.

Daudt, R. C., Le Saux, B., Boulch, A., Gousseau, Y., 2019. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187, 102783.

Desclée, B., Bogaert, P., Defourny, P., 2006. Forest change detection by statistical object-based method. *Remote sensing of environment*, 102(1-2), 1–11.

Ding, L., Guo, H., Liu, S., Mou, L., Zhang, J., Bruzzone, L., 2022. Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14.

Ding, L., Zhang, J., Guo, H., Zhang, K., Liu, B., Bruzzone, L., 2024. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*.

Duan, C., Zheng, X., Li, R., Wu, Z., 2024. Urban flood vulnerability Knowledge-Graph based on remote sensing and textual bimodal data fusion. *Journal of Hydrology*, 633, 131010.

Gueguen, L., Hamid, R., 2016. Toward a generalizable image representation for large-scale change detection: Application to generic damage analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 54(6), 3378–3387.

Huang, H., Shao, Z., Cheng, Q., Huang, X., Wu, X., Li, G., Tan, L., 2023. MC-Net: multi-scale contextual information aggregation network for image captioning on remote sensing images. *International Journal of Digital Earth*, 16(2), 4848–4866.

Kennedy, R. E., Townsend, P. A., Gross, J. E., Cohen, W. B., Bolstad, P., Wang, Y., Adams, P., 2009. Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote sensing of environment*, 113(7), 1382–1396.

Lei Ba, J., Kiros, J. R., Hinton, G. E., 2016. Layer normalization. *ArXiv e-prints*, arXiv–1607.

Li, Y., Zhu, Z., Yu, J.-G., Zhang, Y., 2021. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12), 10590–10603.

Liu, T., Gong, M., Lu, D., Zhang, Q., Zheng, H., Jiang, F., Zhang, M., 2021. Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–17.

Long, J., Li, M., Wang, X., Stein, A., 2024. Semantic change detection using a hierarchical semantic graph interaction network from high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 211, 318–335.

Lu, X., Sun, X., Diao, W., Mao, Y., Li, J., Zhang, Y., Wang, P., Fu, K., 2023. Few-shot object detection in aerial imagery guided by text-modal knowledge. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–19.

Lv, Z., Liu, T., Benediktsson, J. A., 2020. Object-oriented key point vector distance for binary land cover change detection using VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(9), 6524–6533.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, Ieee, 565–571.

Niu, Y., Guo, H., Lu, J., Ding, L., Yu, D., 2023. SMNet: symmetric multi-task network for semantic change detection in remote sensing images based on CNN and transformer. *Remote Sensing*, 15(4), 949.

Ochtyra, A., Marcinkowska-Ochtyra, A., Raczko, E., 2020. Threshold-and trend-based vegetation change monitoring algorithm based on the inter-annual multi-temporal normalized difference moisture index series: A case study of the Tatra Mountains. *Remote Sensing of Environment*, 249, 112026.

Parelli, M., Delitzas, A., Hars, N., Vlassis, G., Anagnostidis, S., Bachmann, G., Hofmann, T., 2023. Clip-guided vision-language pre-training for question answering in 3d scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5606–5611.

Peng, D., Bruzzone, L., Zhang, Y., Guan, H., He, P., 2021. SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery. *International Journal of Applied Earth Observation and Geoinformation*, 103, 102465.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *International conference on machine learning*, PMLR, 8748–8763.

Tian, S., Zhong, Y., Zheng, Z., Ma, A., Tan, X., Zhang, L., 2022. Large-scale deep learning based binary and semantic change detection in ultra high resolution remote sensing imagery: From benchmark datasets to urban application. *ISPRS Journal of Photogrammetry and Remote Sensing*, 193, 164–186.

Wang, L., Dong, S., Chen, Y., Meng, X., Fang, S., Fei, S., 2024. MetaSegNet: Metadata-collaborative Vision-Language Representation Learning for Semantic Segmentation of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*.

Wei, M., Chen, L., Ji, W., Yue, X., Zimmermann, R., 2024. In Defense of Clip-based Video Relation Detection. *IEEE Transactions on Image Processing*.

Wu, C., Du, B., Cui, X., Zhang, L., 2017. A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion. *Remote Sensing of Environment*, 199, 241–255.

Wu, W., Luo, H., Fang, B., Wang, J., Ouyang, W., 2023. Cap4video: What can auxiliary captions do for text-video retrieval? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10704–10713.

Xia, H., Tian, Y., Zhang, L., Li, S., 2022. A deep Siamese postclassification fusion network for semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–16.

Xia, X., Dong, G., Li, F., Zhu, L., Ying, X., 2023. When CLIP meets cross-modal hashing retrieval: A new strong baseline. *Information Fusion*, 100, 101968.

Yan, J., Wang, L., Song, W., Chen, Y., Chen, X., Deng, Z., 2019. A time-series classification approach based on change detection for rapid land cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 158, 249–262.

Yang, K., Xia, G.-S., Liu, Z., Du, B., Yang, W., Pelillo, M., Zhang, L., 2021. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–18.

Zhang, D., Schroeder, A., Yan, H., Yang, H., Hu, J., Lee, M. Y., Cho, K. S., Susztak, K., Xu, G. X., Feldman, M. D. et al., 2024. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nature Biotechnology*, 1–6.

Zheng, Z., Zhong, Y., Tian, S., Ma, A., Zhang, L., 2022. ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 228–239.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., Zhang, L., 2021. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265, 112636.

Zhou, Z., Alabi, O., Wei, M., Vercauteren, T., Shi, M., 2023. Text promptable surgical instrument segmentation with vision-language models. *Advances in Neural Information Processing Systems*, 36, 28611–28623.