Referring Change Detection in Remote Sensing Imagery

Yilmaz Korkmaz[1]   Jay N. Paranjape[1]   Celso M. de Melo[2]   Vishal M. Patel[1]

[1]Johns Hopkins University, Baltimore, USA

[2]DEVCOM U.S. Army Research Laboratory, Adelphi, USA

{ykorkma1, jparanj2, vpatel36}@jhu.edu   celso.m.demelo.civ@army.mil

Abstract

*Change detection in remote sensing imagery is essential for applications such as urban planning, environmental monitoring, and disaster management. Traditional change detection methods typically identify all changes between two temporal images without distinguishing the types of transitions, which can lead to results that may not align with specific user needs. Although semantic change detection methods have attempted to address this by categorizing changes into predefined classes, these methods rely on rigid class definitions and fixed model architectures, making it difficult to mix datasets with different label sets or reuse models across tasks, as the output channels are tightly coupled with the number and type of semantic classes. To overcome these limitations, we introduce Referring Change Detection (RCD), which leverages natural language prompts to detect specific classes of changes in remote sensing images. By integrating language understanding with visual analysis, our approach allows users to specify the exact type of change they are interested in. However, training models for RCD is challenging due to the limited availability of annotated data and severe class imbalance in existing datasets. To address this, we propose a two-stage framework consisting of (I) **RCDNet**, a cross-modal fusion network designed for referring change detection, and (II) **RCDGen**, a diffusion-based synthetic data generation pipeline that produces realistic post-change images and change maps for a specified category using only pre-change image, without relying on semantic segmentation masks and thereby significantly lowering the barrier to scalable data creation. Experiments across multiple datasets show that our framework enables scalable and targeted change detection. Project website is [here](#).*

1. Introduction

Change detection (CD) task in remote sensing refers to identifying significant changes to an area between two time intervals. CD is an essential task for urban planning, envi-
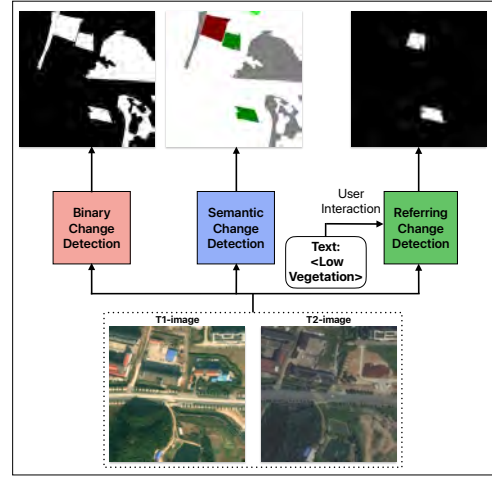

Figure 1. Comparison of different approaches in change detection. T1- and T2-images represent distinct time points, corresponding to pre-change and post-change states. Binary change detection methods produce a single map showing all changes that occurred over this time frame, while semantic change detection methods attempt to categorize all changes, both of which limit user interaction. In contrast, referring change detection enables user input to specify particular changes of interest, displaying only those changes that match the user-defined criteria.

ronmental monitoring and disaster management [9, 20–22]. However, it possesses a significant challenge because remote sensing images taken at different times frequently display data inconsistencies, including illumination changes, noise, and variations in resolution or alignment [26, 32, 46]. In recent years deep learning methods have gained significant attraction for CD utilizing convolutional neural networks [4, 18, 33], transformers [1, 7], diffusion [2] and more recently state space models [53].

Change detection has traditionally been approached as a binary classification problem on undifferentiated change areas, providing no information about changes across distinct semantic classes. While binary change detection (BCD) offer a general overview of altered areas, many applications require identifying the specific semantic classes as-

sociated with these changes. For instance, in urban planning, distinguishing between changes, especially buildings, roads, and vegetation is crucial. Similarly, in environmental monitoring, understanding whether changes involve forest cover, water bodies, or agricultural land provides valuable insights. Several semantic change detection (SCD) methods have been proposed in the literature to provide this detailed information [59]. However, SCD methods, which aim to categorize changes into predefined classes, often encounter significant limitations due to inconsistencies in class definitions across different datasets. For instance, in one of the SCD datasets, "impervious surface" and "bare ground" are treated as separate classes [58] but they are combined into a single class named "non-vegetated ground surface" in another dataset [50]. These discrepancies make it challenging for SCD models to generalize across datasets, as the same physical features may be classified differently depending on the dataset's labeling schema. The scarcity of SCD datasets intensifies these discrepancies and this lack of standardization leads to problems in accurately detecting and categorizing changes. Because the number of output channels in models corresponds directly to the number of classes, models trained on one dataset struggle to perform effectively on another with differing class definitions. Additionally, training with mixed datasets becomes impractical if class definitions do not align precisely.

SCD methods face additional challenges due to the limited availability of datasets, which often exhibit significant class imbalances. For example, in the SECOND dataset [51], the non-vegetated ground surface class covers 43% of the total area of change and appears in 2,689 images, while the "playground" class represents only 0.38% and is included in just 129 images. Similarly, in the CNAM-CD dataset [58], "water bodies" are present in 394 images, whereas "bare ground" appears in 1,453 images. This class imbalance hinders models from learning and detecting changes in underrepresented classes effectively, limiting the generalization capacity and robustness of SCD methods across diverse real-world applications.

To address these limitations, we introduce **Referring Change Detection (RCD)** and propose a language-guided change detection network, **RCDNet**, that effectively integrates visual and textual inputs to generate accurate targeted change maps. Our RCD task inherently overcomes the constraints of traditional SCD tasks—which rely on predefined classes and a fixed, limited number of output channels—by introducing a flexible, language-guided detection process that allows users to specify semantic changes without predefined classes (see Fig. 1). Our approach facilitates the integration of heterogeneous datasets, thereby improving the model's generalizability and versatility. Furthermore, to support effective training for RCD, we propose a synthetic data generation pipeline, **RCDGen**, that leverages a latent diffusion model to simultaneously generate realistic post-change images and change maps from a given pre-change image. This allows targeting specific classes in the original image and replacing them with the desired post-change category, thus allowing us to produce a much wider spectrum of post-change images than what is available in existing datasets. Our contributions can be summarized as:

- We define Referring Change Detection (RCD) task, enabling user interaction in change detection through natural language inputs.
- We propose a synthetic data generation pipeline that simultaneously outputs post-change image and change map given a pre-change image and desired change category.
- We introduce a lightweight and efficient RCD model designed to capture only user-intended changes that can be also be adapted to SCD and BCD tasks.

2. Related Work

Traditional change detection methods: Change Detection has largely been studied as a binary segmentation problem, where the objective is to compute one change map for a given image pair. Usually, the change map of interest depends on the object of interest defined by the dataset. For instance, LEVIR-CD [5] only provides change maps for buildings. Classical CD methods involve using algebraic algorithms to compute change maps [10, 37, 43], transformative approaches which generate transforms on the pre- and post-images followed by a difference [12, 27, 38, 39]. With the advent of Deep Learning (DL), change detection has also moved from classical methods to DL-based methods. Convolutional Neural Networks (CNNs) have been widely used in CD, including FC-EF [4], FC-Siam [4], SNUNet [18], IFNet [52], and DT-SCN [33].

An alternative to CNN-based methods is the transformer-based methods, which include BiT [7], STANet [6], and ChangeFormer [1]. These utilize the Vision Transformer (ViT) [15], due to its ability to capture long-range dependencies better than CNNs. On the other hand, DDPM-CD [2] uses generative modeling to learn powerful encoders. Very recently, selective state space models (S4), i.e., Mamba, have been used instead of CNNs or transformers for CD following its computational efficiency [19]. These methods include RSMamba [55], ChangeMamba [8], CD-Mamba [53] and M-CD [41]. Our approach leverages the Mamba architecture for its computational efficiency, as in prior work, and improves it with cross-attention transformers to enable multimodality via language guidance, resulting in a Mamba–Transformer hybrid architecture.

Semantic change detection methods: The changes related to different objects of interest can vary significantly depending on the application. Therefore, generating semantic change maps with multiple classes, rather than a simple binary map, is more valuable. As a result, sev-

eral studies have aimed to address the task of Semantic Change Detection (SCD). Gousseau et al. [11] are among the first to publish a dataset for semantic change detection (SCD) and propose a fully convolutional network for the task. Zhang et al. [51] also release a dataset, named SEC-OND, and introduce asymmetrical siamese networks as a solution. In [13], the authors present an innovative CNN architecture for SCD called SSCD-1, where semantic temporal features are integrated within a deep change detection unit; they later propose a more efficient model named Bi-SRNet, which incorporates a novel semantic consistency loss. In [14], the authors introduce a CNN-based Triple Encoder-Decoder network (TED) for extracting temporal features and change representations, followed by a Semantic Change Transformer (SCanFormer) designed to explicitly capture semantic transitions in token space. They then present SCanNet, a hybrid 'CNN-Transformer' method that uses TED for feature extraction and SCanFormer to capture correlations within the embedded semantic space.

LLM-based Methods in Remote Sensing: Various concurrent works have utilized foundation models for change detection and other geo-sensing-based tasks. One of these works called SeFi-CD [54] introduces a framework where text prompts are used to generate visual prompts for frozen foundation models which further generate the change map. This method comes closest to our approach. However, a major difference in our method is that we utilize the foundation model CLIP [42] for generating the text embedding, but train a Mamba-based encoder and decoder instead of using a foundation model for encoding the image. The reason behind that is the observation that the available foundation models are not trained heavily with remote-sensing images and might hallucinate. Since implementation is not public at the time of submission we are not able to include this method in comparison. Another concurrent work called CDChat [40] tries to generate a text description of the change between two temporal images using Vicuna [56]. In addition, it can also perform visual question answering based on the input pre- and post-change images. However, this method does not generate visual change maps. A recent work, BAN [30], repurposes CLIP's ViT for bi-temporal change detection through a Bi-Temporal Adapter Network (BAN) with bridging modules that inject general features into a task-specific CD branch; however, it is trained in either the BCD or SCD setting, which keeps our approach distinct.

Synthetic Datasets in Remote Sensing: Song et al. [44] present a Stable Diffusion and 3D modeling based synthetic image generation pipeline that uses GPT-4 for prompt creation, allowing images to be generated entirely from scratch. However, their change detection dataset exclusively captures building changes, resulting in the absence of other semantic classes. Meanwhile, Tang et al. [16] propose

a Stable Diffusion based generation pipeline that depends on semantic segmentation masks during inference, thus limited with existing segmentation datasets. In a similar vein, Zheng et al. [17] introduce a GAN based synthetic image generator that incrementally adds or removes objects from pre-event images, also relying on annotated segmentation data. While these methods can yield highly realistic simulations, they differ from our approach in that we do not require semantic segmentation masks during inference. Building on [17], Zheng et al. [57] propose a self-supervised technique and latent diffusion based model, removing the need for pre-event annotations by using object contours as diffusion conditions. While this innovation reduces annotation requirements, it relies on off-the-shelf foundation models (e.g., SAM [29]) to generate the contours at inference time, potentially limiting performance based on these models' capabilities.

3. Methodology

The RCD task can be defined similar to the BCD task, albeit with an additional user-input selecting the change of interest. Thus, the goal of RCD is to learn a function f that satisfies $M = f(I_{pre}, I_{post}, C)$ where I_{pre} and I_{post} denote the pre- and post-change images respectively, and M denotes the change map. The user intent is specified by C. In our method, we use CLIP-based text embeddings of the change category to model the user intent. Next, we describe our RCDGen pipeline and RCDNet architecture.

3.1. RCDGen

Semantic change detection datasets are often small in scale and contain limited semantically annotated examples, often accompanied by severe class imbalance, making it challenging to train robust models and increasing the risk of overfitting. To address these limitations for RCD, we propose RCDGen, a generative pipeline that synthesizes post-change images conditioned on existing pre-change images and target classes, allowing us to produce any number of post-change images with the desired class, ensuring a well-balanced and diverse dataset. We illustrate the RCDGen's training and inference procedures in Fig. 2. We modify the InstructPix2Pix pipeline [3] and utilize existing SCD datasets to learn to model the joint distribution of post-change images and change maps conditioned on pre-change images. More specifically, we first use a Variational Auto Encoder (VAE) to generate latents for the pre- and post-change images. Then, we randomly select a change class and generate a change map for this class by masking other pixels in the annotated change map, followed by downsampling to match the shape of the latents. Both the latents and the change map are concatenated and passed through the forward diffusion process, where random noise is added to only the post-image latent and the change map. During the
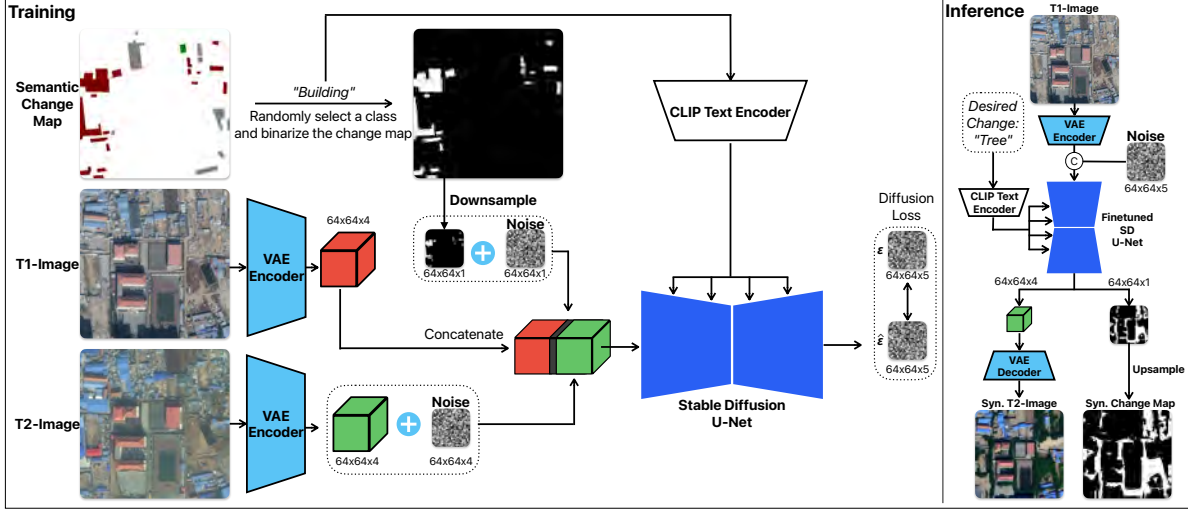
Figure 2. The training and inference scheme of RCDGen is illustrated. In the training phase, the pre-change latent is shown in red, and the post-change latent in green. The change maps are downsampled to match the resolution of the latents (64×64) to enable concatenation. During inference, pre-change latent is concatenated with noise of shape $64 \times 64 \times 5$. After denoising, the generated latent is split back to match the channels used during training. The generated downsampled mask is upsampled via interpolation, and the generated latent is decoded via the VAE Decoder.
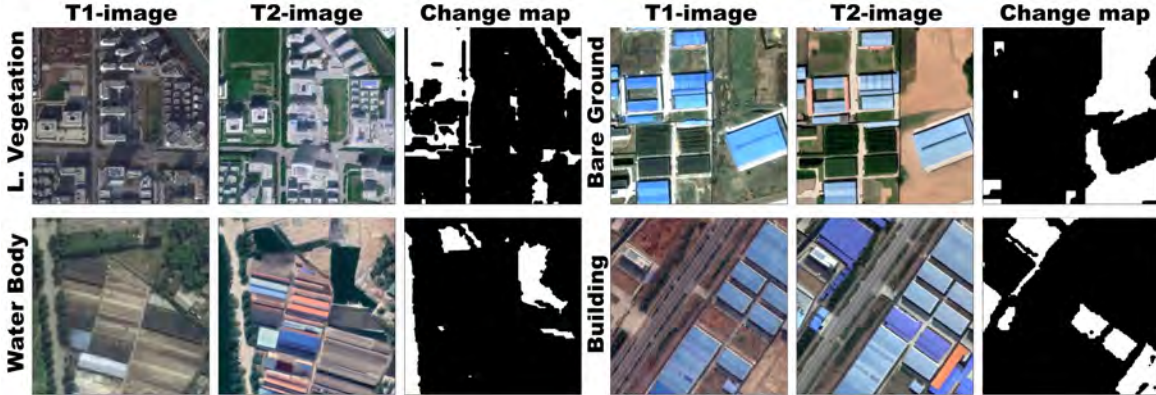

Figure 3. Synthetic samples generated by our pipeline: T1-images (pre-change) are sourced from real datasets, while T2-images (post-change) and corresponding change maps are generated. Labels on the left side of each image pair indicate the specific change category.

denoising process, we utilize the CLIP text embeddings of the class name to guide the UNet to predict the noise added during the forward process. This setup is trained using the canonical diffusion loss [24] denoted as follows:

$$L = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1)} \left[\left\| \epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(T1), c_T) \right\|_2^2 \right], \quad (1)$$

where $\mathcal{E}$ denotes VAE encoder, ϵ_θ denotes UNet, c_T represents the text embeddings, t is a random timestep and z_t is defined as:

$$z_t = \sqrt{\alpha_t} \times [\mathcal{E}(T2) \oplus \tilde{m}] + \sqrt{1 - \alpha_t} \times \epsilon, \quad (2)$$

where $\oplus$ denotes concatenation in channel dimension, α_t corresponds to the noise schedule term, ϵ denotes the added noise, and $\tilde{m}$ is the downsampled change map.

For training, We utilize images from the training sets of the SECOND [51], CNAM-CD [58], LEVIR-CD [5] and LEVIR-CD+ [5] datasets. For inference, InstructPix2Pix [3] extends the classifier-free guidance [23] sampling strategy by incorporating dual conditioning inputs—an image and an instruction—each applied with separate coefficients to control their respective strengths. We adapt this strategy to balance the desired change with consistency to the pre-image. The modified noise estimation is given below:

$$\tilde{e}_\theta\left(z_t, \mathcal{E}(T1), c_T\right) = e_\theta\left(z_t, \varnothing, \varnothing\right)$$
$$+ s_I \cdot \left(e_\theta\left(z_t, \mathcal{E}(T1), \varnothing\right) - e_\theta\left(z_t, \varnothing, \varnothing\right)\right) \quad (3)$$
$$+ s_T \cdot \left(e_\theta\left(z_t, \mathcal{E}(T1), c_T\right) - e_\theta\left(z_t, \mathcal{E}(T1), \varnothing\right)\right),$$

where $\varnothing$ represents the null condition, $\tilde{e}_\theta$ denotes the modified estimate, s_I controls the strength of the image condi-

4

tion, and s_T controls the strength of the text condition.

To create a more comprehensive dataset, we additionally incorporate a relatively small number of inpainted samples within our synthetic dataset, generated by adapting stable diffusion for inpainting. These inpainted samples introduce subtle, realistic changes that complement the larger transformations in the primary synthetic data, achieving a more balanced representation of change types. For this inpainting process, we leverage an existing remote sensing semantic segmentation dataset [48], where specific classes are randomly masked using semantic segmentation masks and modified based on prompts, following a similar pipeline but without the InstructPix2Pix sampling scheme. This combination of subtle and significant changes allows our synthetic dataset to better capture a diverse range of transformations. Our final synthetic dataset includes 29,500 images synthesized from SECOND, 17,500 from CNAM-CD, and 8,300 inpainted samples, creating a robust and varied resource for referring change detection. We utilize this synthetic dataset for pretraining of the RCDNet. We show sample synthetic images generated by our pipeline in Fig. 3. A more comprehensive set of samples are provided in the supplementary document. We believe this dataset and generation pipeline would be highly beneficial for the community.

3.2. RCDNet

We illustrate the architecture of our proposed model in Fig. 4. Our design is primarily inspired by the state-of-the-art Mamba-based semantic segmentation model [47], which has demonstrated strong performance in binary change detection (BCD) tasks [41]. Specifically, we use a VMamba-based [34] Siamese Encoder Block (SEB) to process both the pre-change (T1-image) and post-change (T2-image) images in parallel, extracting meaningful features from each. These features are then passed to a Fusion Module (FM), which performs a comparative analysis and produces fused latent representations. Finally, these fused latents are input to a series of lightweight Mask Decoder Blocks (MDB), which generate the change map conditioned on user intent, provided as text embeddings corresponding to the desired change class. We detail the key components of RCDNet in the following sections.

Siamaese Encoder Block (SEB): Siamese Encoder Block is built with Visual State Space Blocks (VSSB) [34], as shown with black frame in Fig. 4. A series of consecutive SEBs forms an encoder network featuring two identical weighted branches (Siamese architecture) that process both T1- and T2-images, representing distinct time points. In each SEB, the resolution of feature maps is halved while the number of channels is doubled, enhancing feature representation. Each encoded feature is directed to separate fusion modules, enabling the selection of differential features to be shared with the Mask Decoder Blocks.

Fusion Module (FM): The fusion module receives features from the two branches and produces a single feature vector for the mask decoder for each resolution, shown with green frame in Fig. 4. FM blocks consists of linear, depth-wise convolution and Concat & Selective Scan (CSS) modules, following [47] and [41]. In the CSS module, the two features are initially flattened and then concatenated along the sequence length dimension. Here, the concatenation of feature vectors from both branches allows selective scan module to process both feature vectors together, enhancing the cross-branch long range contextual sensitivity unlike relying solely on convolution layers for feature fusion.

Mask Decoder Block (MDB): The Mask Decoder Blocks, highlighted with blue in Fig. 4, are responsible for generating binary change maps based on text embeddings using cross-attention transformer layers. To enhance the learning of inter-channel dependencies, we incorporate both average-pooling and max-pooling operations on feature channels, following [47]. The CLIP Text Encoder generates text embeddings using the target class names as inputs. These text embeddings serve as keys and values in the cross-attention transformer, guiding the selection of visual queries.

4. Experiments and Results

To enable fair comparison against existing SCD and BCD methods, we use only the corresponding change category (class name) as the text prompt and conduct extensive experiments, spanning from classical SCD to cross-domain tests. In all tables, cases where the synthetic data is used for pretraining are indicated in parentheses. For within-domain scenarios, all models are trained zsolely on the target dataset. For the cross-domain cases, all SCD methods are trained exclusively on the SECOND dataset, as it includes buildings as a sub-class.

4.1. Implementation Details

We train our model using the AdamW optimizer [36] with an initial learning rate of 6×10^{-5} and a weight decay of 0.01. The learning rate follows a polynomial decay schedule with an exponent of 0.9. Training is conducted over 200 epochs with a batch size of 4 with an NVIDIA RTX A6000 GPU with 48GB of memory. For the synthetic data pipeline, we fine-tune stable diffusion for 15,000 iterations using a batch size of 4. Fine-tuning is carried out with the Hugging Face Diffusers library [45] and leverages the pretrained InstructPix2Pix checkpoint [3]. During synthetic data generation, we set s_I to 1.5 and s_T to 7.0.

4.2. Datasets

We use the benchmark SCD datasets SECOND [51] and CNAM-CD [58] for within-domain comparisons for the SCD tasks. For cross-domain experiments and comparison
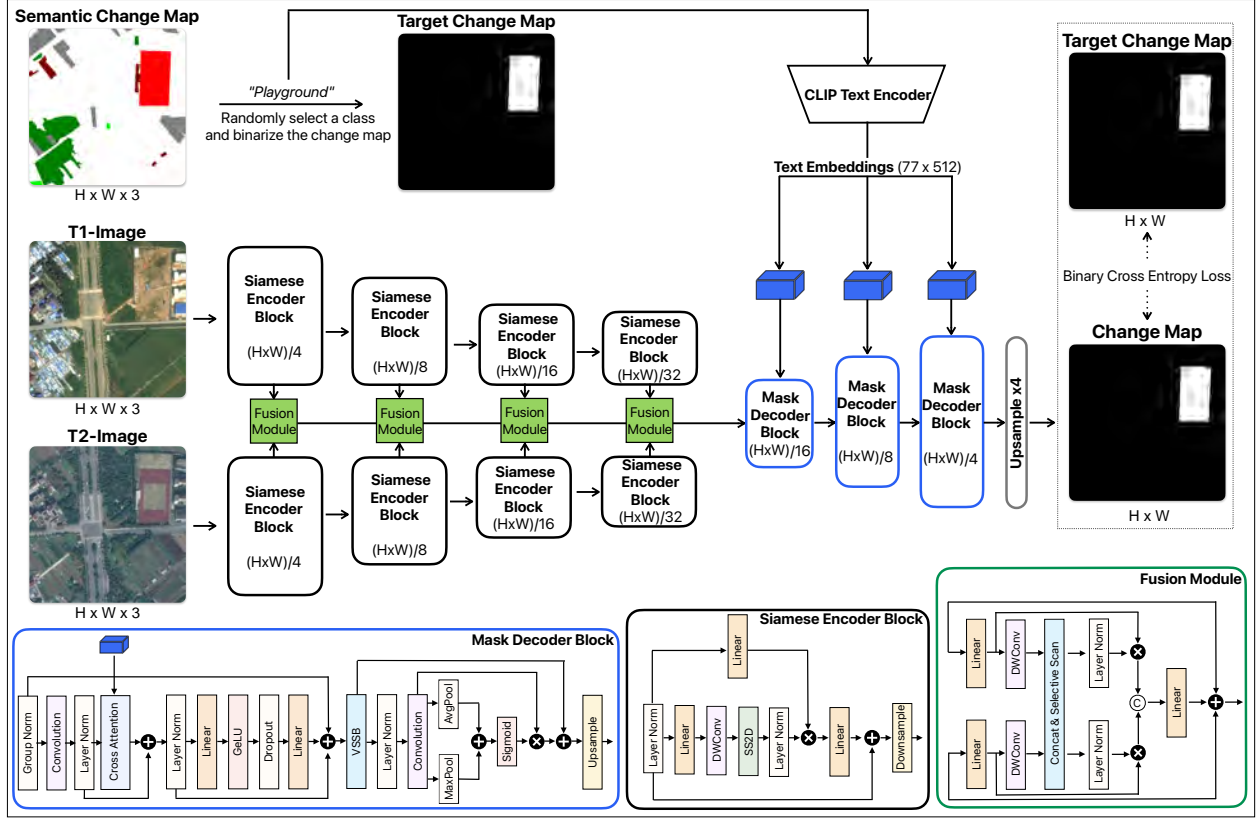
Figure 4. The RCDNet and its training scheme are illustrated. T1- and T2-images represent distinct time points, corresponding to pre-change and post-change states. The change category (class) is randomly selected from unique values in the semantic change map and guides the network through text embeddings. The change map is binarized for that specified category, enabling the use of canonical binary cross-entropy loss as training objective.

Table 1. Within-domain and cross-domain performance of SCD methods.

Method	SECOND [51]				CNAM-CD [58]				WHU-CD [28]		LEVIR-CD [5]	
	mIoU (↑)	SeK (↑)	OA (↑)	F_{scd} (↑)	mIoU (↑)	SeK (↑)	OA (↑)	F_{scd} (↑)	IoU (↑)	OA (↑)	IoU (↑)	OA (↑)
Experiment	within-domain				within-domain				cross-domain		cross-domain	
HRSCD-str.3 [11]	66.74	13.21	84.63	56.51	57.36	0.62	77.24	28.38	43.75	96.04	42.74	95.74
BAN [30]	66.55	10.87	84.59	53.15	62.51	2.63	77.25	35.76	29.78	93.23	18.71	91.30
HRSCD-str.4 [11]	72.24	20.56	87.60	63.21	67.94	3.31	78.94	34.14	43.97	95.90	48.29	96.40
SSCD-1 [13]	73.03	22.50	88.07	65.36	71.17	13.47	82.08	46.80	46.14	96.05	49.64	96.01
Bi-SRNet [13]	73.04	22.26	88.19	65.21	69.11	19.22	82.25	58.55	46.57	96.27	44.74	95.28
TED [14]	**73.53**	22.90	88.14	65.28	67.00	9.65	78.98	44.78	44.46	95.69	51.43	96.11
SCanNet [14]	73.42	23.90	88.20	66.66	71.38	22.68	84.69	61.21	38.32	94.58	47.29	95.29
RCDNet	73.04	24.57	89.00	68.29	72.83	29.49	87.35	70.41	59.05	97.73	53.98	96.78
RCDNet (synthetic)	73.47	**25.25**	**89.24**	**68.90**	75.32	**33.81**	**88.95**	**73.50**	**63.14**	**97.99**	**60.21**	**97.50**

with BCD methods, we employ the BCD datasets LEVIR-CD [5] and WHU-CD [28]. Detailed description of each dataset is provided in the supplementary document.

4.3. Evaluation Metrics

We evaluate the performance using four standard metrics for semantic change detection: overall pixel accuracy (OA), semantic mean intersection over union (mIoU) [51], separated kappa coefficient (SeK) [51], and semantic change

detection F1 score (F_{scd}) [13]. Detailed definitions are provided in the supplementary document.

4.4. Within-domain in SCD

We evaluate our proposed model against state-of-the-art SCD methods, conducting within-domain tests where the training and test sets originate from the same SCD dataset. Although such aggregation is unnecessary in real-world settings, to align RCD results with SCD we combine the class-specific binary change maps, obtained from multiple forward passes, into a unified semantic change map for each image. We focus on post-change maps (i.e., growth changes) for both datasets. Qualitative results, as shown in Fig. 5, illustrate the model's ability to capture precise, user-specified changes, highlighting the added detail and accuracy enabled by our approach. The left part of Tab. 1 presents within-domain performance for SECOND and CNAM-CD, where RCDNet demonstrates superior performance with or without synthetic data, outperforming the second-best methods by notable margins: on SECOND, it surpasses in SeK by 1.35, OA by 1.04, and F_{scd} by 2.24; on CNAM-CD, it achieves an additional 3.94 in mIoU, 11.13 in SeK, 4.26 in OA, and 12.29 in F_{scd}. These quantitative and qualitative improvements underscore the advantages of RCD's targeted approach, where user-defined change categories enable RCDNet to focus on specific, contextually relevant transformations. This targeted approach, combined with synthetic data pretraining, delivers more nuanced and accurate results than standard SCD methods, which broadly classify all changes without customization for user needs.

4.5. Within-domain in BCD

We evaluate performance on building change detection using the single-class BCD datasets LEVIR-CD and WHU-CD. Given the single-class focus, SCD metrics are not applicable; thus, we rely on the original BCD metrics: IoU and overall pixel accuracy (OA). Tab. 3 show within-domain performance comparison for LEVIR-CD and WHU-CD datasets. On WHU-CD, RCDNet (synthetic) outperforms the state-of-the-art best IoU (M-CD [41]) by 0.8 and matches the top OA at 99.7. On LEVIR-CD, it achieves a 0.8 improvement in IoU over M-CD [41] and shares the highest OA of 99.2. These results demonstrate RCDNet's state-of-the-art performance in the BCD task, with no performance drop despite the focus on user-specified categories, underscoring the robustness of RCD-Net's targeted approach and its adaptability to BCD along with the SCD tasks.

4.6. Cross-domain Capabilities

We evaluate the performance in a cross-domain scenario, where the target class (building) is familiar to the model through training on the SECOND dataset, but the test
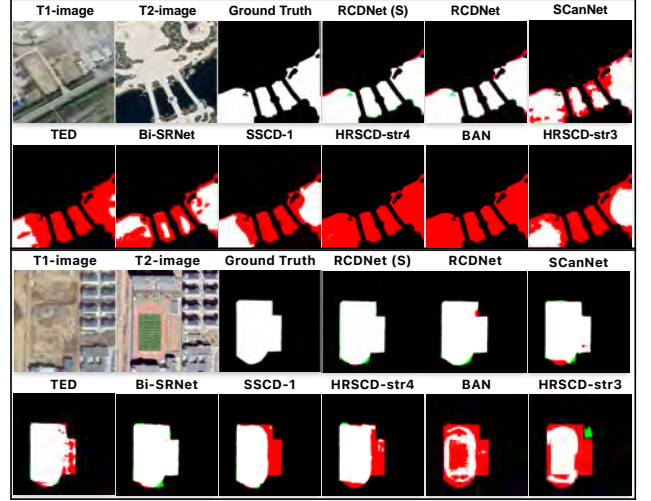

Figure 5. Qualitative results corresponding to the "Water Bodies" category in CNAM-CD (upper row) and "Playground" category in SECOND (lower row) datasets. T1-images represent pre-change (earlier time point) and T2-images represent post-change (later time point). In the results, white pixels indicate true positives, black pixels indicate true negatives, green pixels denote false positives, and red pixels denote false negatives. The notation "(S)" refers to synthetic pretraining applied to RCDNet.

Table 2. Mixed dataset performance in SCD datasets.

Metric	RCDNet (mixed)	RCDNet (distinct)
SECOND [51]		
mIoU (↑)	72.98	**73.04**
SeK (↑)	24.48	**24.57**
OA (↑)	**89.27**	89.00
F_{scd} (↑)	68.26	**68.29**
CNAM-CD [58]		
mIoU (↑)	**73.39**	72.83
SeK (↑)	**30.00**	29.49
OA (↑)	**87.51**	87.35
F_{scd} (↑)	**70.61**	70.41

images are drawn from the WHU-CD and LEVIR-CD datasets. The right part of Tab. 1 shows the cross-domain performance for the LEVIR-CD and WHU-CD datasets. Results shows that RCDNet with synthetic data significantly outperforms the other methods in cross-domain performance on the WHU-CD and LEVIR-CD datasets. Compared to the next best methods, it achieves a notable IoU improvement of 16.57 on WHU-CD and 8.78 on LEVIR-CD, along with higher OA scores. Additionally, synthetic data boosts RCDNet's performance over its non-synthetic version by 4.09 IoU on WHU-CD and 6.23 IoU on LEVIR-CD, highlighting the effectiveness of synthetic data in enhancing RCDNet's adaptability across datasets. The diversity introduced through our synthetic data generation pipeline plays a key role in enabling this benefit, as it exposes RCDNet

Table 3. Within-domain performance in BCD datasets.

Method	WHU-CD [28]		LEVIR-CD [5]	
	IoU (↑)	OA (↑)	IoU (↑)	OA (↑)
FC-Siam-conc [4]	66.5	98.5	72.0	98.5
SNUNet [18]	71.7	98.7	78.8	98.8
IFNet [52]	71.5	98.8	78.8	98.9
DT-SCN [33]	84.2	99.3	78.1	98.8
STANet [6]	70.0	98.5	77.4	98.7
BIT [7]	83.4	99.3	80.7	98.9
ChangeFormer [1]	79.5	99.1	82.5	99.0
ChangeMamba [8]	86.1	99.4	82.1	99.0
CDMamba [53]	88.2	99.5	83.1	99.0
DDPM-CD [2]	86.3	99.4	83.3	99.1
M-CD [41]	91.1	99.6	85.0	**99.2**
RCDNet	91.6	**99.7**	**85.8**	**99.2**
RCDNet (synthetic)	**91.9**	**99.7**	**85.8**	**99.2**

to a wider range of visual variations and change scenarios during pretraining. By simulating diverse instances of the target class across various contexts and conditions, the synthetic data allows RCDNet to learn more generalized feature representations. This improved representation enables the model to handle previously unseen imagery more effectively, capturing complex patterns and subtle changes that might otherwise be missed. We also experiment in a completely zero-shot setting where the target change category is unknown to the model. While some results are promising, performance is not consistently reliable across different images and categories. Nevertheless, if the text embedding of an unseen category is sufficiently similar to those of seen categories, zero-shot transfer should in principle be feasible. Due to these inconsistent results, we refrain from making strong claims and leave this as a promising direction for future research.

4.7. Mixed Dataset Performance

To evaluate the impact of combining datasets on the performance of our RCDNet model, we conduct experiments using a merged dataset. This approach consolidates multiple datasets into a single training phase, eliminating the need for separate training sessions for each dataset and thus streamlining the overall process. For comparison, we combine the SECOND and CNAM-CD datasets and train our model on this mixed dataset. The performance metrics are shown in Tab. 2, indicating that any performance degradation is negligible compared to training on individual datasets and even surpasses performance on CNAM-CD. These results confirm that our approach effectively preserves model accuracy and robustness, even when trained on a diverse mix of datasets, overcoming a key limitation of traditional SCD methods, which cannot be effectively trained with combined datasets due to their strict requirement for matching class numbers and names.

4.8. Ablation Studies

We perform extensive ablation studies to refine our architecture, using a composite evaluation metric of $0.3 \times$ mIoU $+ 0.7 \times$ SeK following [58] and report this score in SECOND validation set. Key findings are summarized below:

1. **Encoder Backbone:** We evaluate several encoder backbones of different sizes, including Swin Transformer-B (medium), Swin Transformer-S (small) [35], Segformer-b2 (small), Segformer-b5 (large) [49], and the VMamba series in tiny, small, and base configurations [34] (see supplementary document for comparison graph). Based on these tests, we select VMamba-small as our encoder, which yielded the best results, confirming the observations of [41].

2. **CLIP Variants:** To assess the value of domain-specific knowledge, we compare the remote sensing-specific RemoteCLIP model [31] with OpenAI's general CLIP models [42]. Results show no advantage of RemoteCLIP over the original CLIP weights, highlighting the strong generalization capabilities of OpenAI's CLIP, which has been pretrained on a very extensive and diverse dataset. (see supplementary document for comparison).

3. **Pretraining Dataset:** Models are pretrained on synthetic data, ImageNet, or initialized from scratch. Pretraining on synthetic data provides a substantially stronger initial performance and better overall performance. See the supplementary document for a comparison over training epochs and performance advantage in SCD and BCD tasks provided by synthetic dataset pretraining.

4. **Decoder Backbone:** Both MLP-based and Mamba-based decoders are evaluated. The Mamba-based decoder outperforms the MLP-based option with comparable computational expense. More details are available in the supplementary document.

5. **CLIP Freezing:** We assess the impact of freezing versus fine-tuning the CLIP weights using LORA [25]. Supplementary document includes performance comparisons for both cases.

5. Conclusion

We introduce RCD, a flexible approach for user-directed change detection in remote sensing imagery. RCD addresses the limitations of traditional binary and semantic change detection methods by integrating pre-trained language models with a change detection network, allowing users to specify desired detection categories via natural language prompts. Additionally, our diffusion-based synthetic data generation pipeline produces realistic post-change images and change maps, mitigating data scarcity and class imbalance to improve overall and cross-domain performance of RCDNet, our proposed RCD model.

6. Acknowledgement

References

[1] 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. 1, 2, 8

[2] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M. Patel. Ddpm-cd: Denoising diffusion probabilistic models as feature extractors for change detection, 2024. 1, 2, 8

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3, 4, 5

[4] 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. 1, 2, 8

[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. 2, 4, 6, 8

[6] 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):1662, 2020. 2, 8

[7] Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 1, 2, 8

[8] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatio-temporal state space model, 2024. 2, 8

[9] Nicholas C Coops, Piotr Tompalski, Tristan RH Goodbody, Alexis Achim, and Christopher Mulverhill. Framework for near real-time forest inventory using multi source remote sensing data. *Forestry*, 96(1):1–19, 2023. 1

[10] E. Brondízio D. Lu Corresponding author, P. Mausel and E. Moran. Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401, 2004. 2

[11] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019. 3, 6

[12] M. Dharani and G. Sreenivasulu. Land use and land cover change detection by using principal component analysis and morphological operations in remote sensing applications. *International Journal of Computers and Applications*, 43(5): 462–471, 2021. 2

[13] Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo Bruzzone. Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14, 2022. 3, 6, 7

[14] Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3, 6

[15] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[16] Tang et al. Changeanywhere: Sample generation for remote sensing change detection via semantic latent diffusion model. *arXiv preprint arXiv:2404.08892*, 2024. 3

[17] Zheng et al. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21818–21827, 2023. 3

[18] Sheng Fang, Kaiyu Li, Jinyuan Shao, and Zhe Li. Snunet-cd: A densely connected siamese network for change detection of vhr images. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1, 2, 8

[19] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[20] Ebrahim Hamidi, Brad G Peter, David F Muñoz, Hamed Moftakhari, and Hamid Moradkhani. Fast flood extent monitoring with sar change detection using google earth engine. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–19, 2023. 1

[21] Md Ariful Haque, Sharmin Shishir, Anannya Mazumder, and Mehedi Iqbal. Change detection of jamuna river and its impact on the local settlements. *Physical Geography*, 44(2): 186–206, 2023.

[22] Xiaoning He, Shuangcheng Zhang, Bowei Xue, Tong Zhao, and Tong Wu. Cross-modal change detection flood extraction based on convolutional neural network. *International Journal of Applied Earth Observation and Geoinformation*, 117:103197, 2023. 1

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 8

[26] Jordi Inglada, Vincent Muron, Damien Pichard, and Thomas Feuvrier. Analysis of artifacts in subpixel remote sensing image registration. *IEEE transactions on Geoscience and Remote Sensing*, 45(1):254–264, 2006. 1

[27] Y. H. Deng J. S. Deng, K. Wang and G. J. Qi. Pca-based land-use change detection and analysis using multitemporal and multisensor satellite data. *International Journal of Remote Sensing*, 29(16):4823–4838, 2008. 2

[28] Shunping Ji, Shiqing Wei, and Meng Lu. 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):574–586, 2018. 6, 8

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[30] Kaiyu Li, Xiangyong Cao, and Deyu Meng. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024. 3, 6

[31] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 8

[32] Jun Liu, Xing Wang, Min Chen, Shuguang Liu, Zhenfeng Shao, Xiran Zhou, and Ping Liu. Illumination and contrast balancing for remote sensing images. *Remote Sensing*, 6(2):1102–1123, 2014. 1

[33] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18(5):811–815, 2021. 1, 2, 8

[34] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 5, 8

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 8

[36] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[37] Luigi Tommaso Luppino, Filippo Maria Bianchi, Gabriele Moser, and Stian Normann Anfinsen. Unsupervised image regression for heterogeneous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9960–9975, 2019. 2

[38] Allan Aasbjerg Nielsen. The regularized iteratively reweighted mad method for change detection in multi- and hyperspectral data. *IEEE Transactions on Image Processing*, 16(2):463–478, 2007. 2

[39] Allan A. Nielsen, Knut Conradsen, and James J. Simpson. Multivariate alteration detection (mad) and maf postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64(1):1–19, 1998. 2

[40] Mubashir Noman, Noor Ahsan, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Cdchat: A large multimodal model for remote sensing change description, 2024. 3

[41] Jay N. Paranjape, Celso de Melo, and Vishal M. Patel. A mamba-based siamese network for remote sensing change detection, 2024. 2, 5, 7, 8

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3, 8

[43] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693, 2019. 2

[44] Jian Song, Hongruixuan Chen, and Naoto Yokoya. Syntheworld: A large-scale synthetic dataset for land cover mapping and building change detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8287–8296, 2024. 3

[45] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 5

[46] Xue Wan, Jianguo Liu, Shengyang Li, John Dawson, and Hongshi Yan. An illumination-invariant change detection method based on disparity saliency map for multitemporal optical remotely sensed images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1311–1324, 2018. 1

[47] Zifu Wan, Pingping Zhang, Yuhao Wang, Silong Yong, Simon Stepputtis, Katia Sycara, and Yaqi Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. *arXiv preprint arXiv:2404.04256*, 2024. 5

[48] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 5

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 8

[50] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Semantic change detection with asymmetric siamese networks. *arXiv preprint arXiv:2010.05687*, 2020. 2

[51] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2022. 2, 3, 4, 5, 6, 7

[52] Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change de-

tection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020. 2, 8

[53] Haotian Zhang, Keyan Chen, Chenyang Liu, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Cdmamba: Remote sensing image change detection with mamba, 2024. 1, 2, 8

[54] Ling Zhao, Zhenyang Huang, Dongsheng Kuang, Chengli Peng, Jun Gan, and Haifeng Li. Sefi-cd: A semantic first change detection paradigm that can detect any change you want, 2024. 3

[55] Sijie Zhao, Hao Chen, Xueliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. Rs-mamba for large remote sensing image dense prediction, 2024. 2

[56] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 3

[57] Zhuo Zheng, Stefano Ermon, Dongjun Kim, Liangpei Zhang, and Yanfei Zhong. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[58] Yanpeng Zhou, Jinjie Wang, Jianli Ding, Bohua Liu, Nan Weng, and Hongzhi Xiao. Signet: A siamese graph convolutional network for multi-class urban change detection. *Remote Sensing*, 15(9):2464, 2023. 2, 4, 5, 6, 7, 8

[59] Qiqi Zhu, Xi Guo, Ziqi Li, and Deren Li. A review of multi-class change detection for satellite remote sensing imagery. *Geo-spatial Information Science*, 27(1):1–15, 2024. 2