Change3D: Revisiting Change Detection and Captioning from A Video Modeling Perspective

Supplementary Material

1. More Experimental Details

1.1. Dataset Description

Binary Change Detection Datasets: (1) The LEVIR-CD [2] comprises 637 bitemporal image pairs sourced from Google Earth, each with a high resolution of 0.5 m/pixel. Spanning images captured from 2002 to 2018 in various locations, this dataset includes annotations for 31333 individual building changes. (2) The WHU-CD [10] dataset focuses on building change detection and contains high-resolution (0.2 m/pixel) bi-temporal aerial images, totaling 32507×15354 pixels. It primarily encompasses areas affected by earthquakes and subsequent reconstruction, mainly involving building renovations. (3) The CLCD [16] dataset consists of cropland change samples, including buildings, roads, lakes, *etc*. The bi-temporal images in CLCD were collected by Gaofen-2 in Guangdong Province, China, in 2017 and 2019, respectively, with spatial resolutions ranging from 0.5 to 2 m. Following the standard procedure detailed in [17, 21], each image of the three datasets is segmented into 256×256 patches. Consequently, the LEVIR-CD dataset is divided into 7120 pairs for training, 1024 pairs for validation, and 2048 pairs for testing. The WHU-CD dataset is partitioned into 5947 training pairs, 744 validation pairs, and 744 test pairs. The CLCD dataset is divided into 1440, 480, and 480 pairs for training, validation, and testing, respectively.

Semantic Change Detection Datasets: (1) The HRSCD [4] dataset contains a total of 291 image pairs of 10000×10000 pixels, each with a resolution of 0.5 m/pixel. The images cover a range of urban and countryside areas in Rennes and Caen, France, including five classes of land cover, *i.e*., artificial surface, agricultural areas, forest, wetland, and water. (2) The SECOND [20] dataset consists of 4662 pairs of aerial images collected from several platforms and sensors, comprising of six land-cover categories, *i.e*., non-vegetated ground surface, tree, low-vegetation, water, buildings, and playgrounds, which are frequently involved in natural and man-made geographical changes. These pairs of images are distributed over various cities, including Hangzhou, Chengdu, and Shanghai. Considering that most of the labeled areas don't change in HRSCD, *e.g*., artificial surfaces and agricultural lands only account for 0.6% of this dataset [22], we discard image pairs with less than 10% of the pixels changed. Each image of the two datasets is cropped into 256×256 non-overlap patches. Consequently, the HRSCD dataset is split into 6525, 932, and 1865 pairs for training, validation, and testing, and the SECOND dataset is divided into 11872 training pairs and 6776 testing pairs, respectively.

Building Damage Assessment Dataset: The xBD [8] is a large-scale building damage assessment dataset that provides high-resolution (0.8 m/pixel) satellite imagery with building localization and damage level labels, collected from 19 disaster events such as floods and earthquakes with an image size of 1024×1024 pixels. The dataset uses polygons to represent building instances and provides four damage categories, *i.e*., non-damage, minor damage, major damage, and destroyed for each building. The minor damage pixels represent visible roof cracks or partially burnt structures while the major damage represents a partial wall, roof collapse, or structure surrounded by water. The destroyed label means that the building structure has completely collapsed, scorched or is no longer present. All the images are cropped into 256×256 non-overlap patches, yielding 44785, 14928, and 14928 image pairs for training, holdout, and testing, respectively.

Change Captioning Datasets: (1) The LEVIR-CC [14] dataset is derived primarily from the LEVIR-CD [2], with each image having a spatial resolution of 1024×1024 pixels and a resolution of 0.5 m/pixel. These bi-temporal images are sourced from 20 regions in Texas, USA, with a time span of 5 to 15 years. Each image pair is annotated with five sentences provided by five distinct annotators to describe the differences between the images. (2) The DUBAI-CC [9] dataset focuses on urbanization changes in Dubai between 2000 and 2010. It contains 500 image tiles, each 50×50 pixels, to analyze urbanization, extracted from bi-temporal images in the visible and infrared bands. It identifies six broad categories of change: roads, houses, buildings, green areas, lakes, and islands. Each bi-temporal image is annotated by five annotators, with each description containing at least three words addressing the spatial distribution and attributes of the changes. All images are cropped or resized into 256×256 patches. The LEVIR-CC dataset is divided into 6815, 1333, and 1929 pairs for training, validation, and testing, respectively, and the DUBAI-CC dataset is split into 300 training pairs, 50 validation pairs, and 150 testing pairs.

1.2. Attention Visualization Details

As illustrated in Fig. 4 (see in our paper), to understand the feature distribution learned in each component within the models, we select three representative bi-temporal image-based methods for comparison, *i.e*., GASNet [21], AMT-

Table 1. Study the effectiveness of the proposed method with different 3D architectures on three binary change detection datasets, respectively.

Method	LEVIR-CD			WHU-CD			CLCD		
	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
I3D [1]	91.21	83.84	99.11	94.18	89.01	99.55	78.67	64.84	96.92
Slow-R50 [6]	91.39	84.14	99.13	94.37	89.34	99.56	**78.82**	**65.05**	**96.93**
UniFormer-XS [12]	91.76	84.77	99.16	94.23	89.08	99.55	78.10	64.07	96.89
X3D-L [5]	**91.82**	**84.87**	**99.17**	**94.56**	**89.69**	**99.57**	78.03	63.97	96.87

Table 2. Study the effectiveness of the proposed method with different 3D architectures on two semantic change detection datasets.

Method	HRSCD				SECOND			
	F_{scd}	mIoU	OA	SeK	F_{scd}	mIoU	OA	SeK
I3D [1]	70.99	66.41	81.10	23.31	61.78	71.95	87.09	20.99
Slow-R50 [6]	71.20	66.93	81.91	23.55	61.93	72.11	87.41	21.22
UniFormer-XS [12]	**73.69**	**69.30**	**83.07**	**27.31**	62.00	71.79	87.03	21.22
X3D-L [5]	73.29	68.67	82.57	26.85	**62.83**	**72.95**	**87.42**	**22.98**

Table 3. Study the effectiveness of the proposed method with different 3D architectures on the LEVIR-CC and DUBAI-CC datasets. Abbreviations B, M, R, and C refer to BLEU, METEOR, ROUGE, and CIDEr, respectively.

Method	LEVIR-CC							DUBAI-CC						
	B-1	B-2	B-3	B-4	M	R	C	B-1	B-2	B-3	B-4	M	R	C
I3D [1]	86.09	78.06	71.16	65.34	40.18	75.30	138.29	**73.11**	**60.80**	**50.42**	**40.69**	**28.68**	**60.01**	**91.18**
Slow-R50 [6]	85.56	77.65	70.46	64.52	39.94	75.01	137.52	73.05	60.14	48.43	37.83	27.22	57.56	88.81
UniFormer-XS [12]	**86.75**	**78.84**	**71.68**	**65.58**	**40.86**	**75.98**	**140.15**	70.27	57.11	45.70	35.26	25.96	54.03	81.66
X3D-L [5]	85.81	77.81	70.57	64.38	40.03	75.12	138.29	72.25	58.68	47.13	36.80	27.06	56.04	86.19

Table 4. Quantitative evaluation of attention maps.

Method	MSE
GASNet [21]	0.35
AMTNet [17]	0.16
EADTer [18]	0.25
Change3D	**0.07**

Net [17], and EATDer [18]. The outputs F_1 and F_2 of these methods represent the final layer's output from the shared-weight image encoder at time T_1 and T_2, respectively. The differential features, represented as F_C, are extracted by the change extractor from the final layer to depict alterations. In our method, F_1 and F_2 correspond to the final layer's output during video feature encoding at time T_1 and T_2, respectively, while F_C represents the perception features. For better visualization of the feature maps, we employ max and average pooling operations across the channel dimension to compress the features, then combine them via element-wise addition, and subsequently normalize them within the range of 0 to 1.

In Tab. 4, we normalize the values in the attention map to [0, 1], apply a threshold of 0.5 to create binary maps, and calculate the MSE against the ground truth. Results show that our method achieves the lowest MSE and thus is more effective in focusing on changed regions.

Table 5. Investigation on the impact of different initialization methods for perception frames across three binary change detection datasets.

Initialization	LEVIR-CD			WHU-CD			CLCD		
	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
Zeros	91.64	84.56	99.16	94.19	89.02	99.55	77.15	62.79	96.67
Ones	91.67	84.61	99.15	94.27	89.17	99.55	77.05	62.67	96.71
Uniform	91.75	84.77	99.16	94.43	89.45	99.56	77.76	63.61	96.80
Random	**91.82**	**84.87**	**99.17**	**94.56**	**89.69**	**99.57**	**78.03**	**63.97**	**96.87**

Table 6. Investigation on the impact of different initialization methods for perception frames on the xBD dataset.

Initialization	F_1^{loc}	F_1^{cls}	$F_1^{overall}$	Damage F_1 Per-class			
				Non	Minor	Major	Destroy
Zeros	85.73	75.29	78.42	94.95	57.35	73.40	86.50
Ones	85.75	75.51	78.58	95.00	57.12	74.38	86.36
Uniform	85.96	75.67	78.76	94.94	57.05	75.00	86.43
Random	**85.74**	**76.71**	**79.42**	**95.08**	**58.70**	**76.50**	**86.76**

2. More Diagnostic Experiments

Effectiveness with different architectures. The effectiveness of the proposed Change3D with different 3D architectures on BCD, SCD, and CC tasks is presented in Tab. 1-3. Notably, all video models exhibit competitive performance, underscoring the efficacy of the proposed method with various video modeling architectures.

Impact of initialization on perception frames. We explore the effects of various initialization methods on perception frames by employing four different approaches: initializing with zeros, ones, uniform values between 0 and 1, and random initialization (*i.e.*, with a mean of 0 and a standard deviation of 1). Analysis of Tab. 5-6 reveals that fixed-value initialization is less effective compared to the other methods. Random initialization yields the most favorable outcomes, which is reasonable as it enables a more robust generation of perception features.

Impact of different similarity losses. We conduct extensive ablation experiments to explore the impact of similarity losses, including L1, L2, Contrastive (with a margin of 0.5), Angular, and Cosine. Tab. 7 presents several key observations: (1) Without the similarity loss, Change3D exhibits inferior performance compared to others, highlighting the effectiveness and necessity of the similarity loss for semantic change detection task. (2) Cosine and Angular losses outperform others on both HRSCD and SECOND datasets, as they better handle changes in content rather than

Table 7. Investigation on the impact of different similarity loss functions across two semantic change detection datasets.

Similarity Loss	HRSCD				SECOND			
	F_{scd}	mIoU	OA	SeK	F_{scd}	mIoU	OA	SeK
-	72.28	67.74	82.01	25.02	61.49	71.97	86.86	21.08
L1	73.13	68.50	82.58	26.32	61.67	72.14	87.09	21.28
L2	72.61	68.17	82.11	25.84	61.93	72.16	87.08	21.39
Contrastive	72.97	68.54	82.63	26.16	62.03	72.21	87.16	21.55
Angular	73.28	**68.73**	82.66	26.65	**62.64**	72.65	87.26	22.58
Cosine	**73.29**	68.59	**82.74**	**26.73**	62.61	**72.84**	**87.40**	**22.86**

Table 8. Performance comparison of several representative methods with random initialization across three binary change detection datasets.

Method	LEVIR-CD			WHU-CD			CLCD		
	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
GASNet [21]	89.38	82.59	98.86	90.85	82.36	99.06	60.83	42.35	92.50
AMTNet [17]	88.94	80.08	98.89	90.23	79.27	98.78	69.32	52.17	95.29
EATDer [18]	89.35	82.31	98.86	88.79	80.01	99.07	69.46	53.35	94.68
Change3D	**90.80**	**83.16**	**99.08**	**92.40**	**85.88**	**99.41**	**71.55**	**55.71**	**96.11**

Table 9. Performance comparison of several representative methods with random initialization on the xBD dataset.

Method	F_1^{loc}	F_1^{cls}	$F_1^{overall}$	Damage F_1 Per-class			
				Non	Minor	Major	Destroy
ChangeOS-R101 [23]	80.31	67.74	72.11	88.80	46.86	67.30	75.09
DamFormer [3]	80.70	69.48	72.19	88.23	49.53	68.02	78.34
PCDASNet [19]	80.05	68.79	72.36	90.05	48.76	71.69	78.84
Change3D	**81.00**	**71.54**	**74.38**	**94.11**	**51.99**	**71.78**	**82.53**

intensity or scale. L1, L2, and Contrastive losses are more sensitive to outliers, potentially skewing results, with Contrastive loss also being highly sensitive to the margin.

Impact of pre-training *vs*. performance. (1) Since most 2D-model-based methods are typically initialized with ImageNet pre-training, our method is pre-trained using video data, such as K400 [11], and SSv2 [7], *etc*. To eliminate the influence of pre-training, we compare Change3D with several 2D-model-based methods using random initialization, as depicted in Tab. 8-9. The table shows that **under the identical initialization setting**, Change3D consistently outperforms other approaches, highlighting the superiority of the proposed method. (2) Most current state-of-the-art methods use pre-trained weights to initialize visual encoders, *e.g*., ImageNet1K (1.2M images) for AMTNet [17], DEFO [13], and SEN [24], and CLIP-400M (400M image-text pairs) for PromptCC [15]. Our video encoders use pre-trained datasets like K400 (1.9M images), AVA (1.6M images), and SSv2 (1.34M images), which are comparable to ImageNet1K but much less than CLIP-400M. (3) Fig. 1 shows that performance improves with more pre-training data from K400 [11], saturating after 75%. (4) Tab. 10 for the BCD task illustrates that pre-trained weights can improve model performance. Besides, pre-training on the K400 dataset still yields the best results, which is consistent with the findings from the BDA task of Tab. 6 (see in our paper).

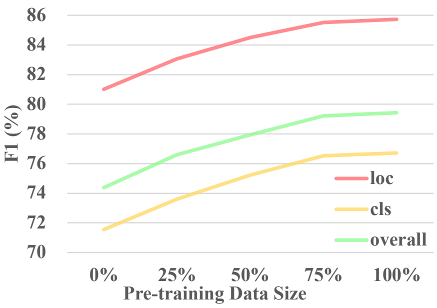


Figure 1. Pre-training data size *vs*. performance on the xBD dataset.

Table 10. Investigation on the impact of different pre-trained weights across three binary change detection datasets.

Pre-trained	LEVIR-CD			WHU-CD			CLCD		
	F1	IoU	OA	F1	IoU	OA	F1	IoU	OA
Random Init	90.80	83.16	99.08	92.40	85.88	9.41	71.55	55.71	96.11
AVA	91.27	83.93	99.12	94.26	89.14	99.55	78.61	64.76	97.01
Charades	91.23	83.87	99.11	94.05	88.77	99.54	77.92	63.83	96.89
SSv2	91.16	83.75	99.11	94.26	89.14	99.55	77.89	63.78	96.86
K400	**91.39**	**84.14**	**99.13**	**94.37**	**89.34**	**99.56**	**78.82**	**65.05**	**96.93**

Necessity of multiple perception frames. Using multiple perception frames improves the model's capacity to learn individual characteristics for each sub-task. Results in Tab. 11 demonstrate that using multiple perception frames leads to superior results, highlighting their effectiveness.

Table 11. Single *vs*. multiple perception frames on two semantic change detection and one damage assessment datasets.

Perception Frame	HRSCD				SECOND				xBD		
	F1	mIoU	OA	SeK	F1	mIoU	OA	SeK	F_1^{loc}	F_1^{cls}	$F_1^{overall}$
Single	72.61	67.80	82.14	25.61	61.09	72.00	86.94	21.34	85.03	75.29	79.00
Multiple	**73.29**	**68.67**	**82.57**	**26.85**	**62.83**	**72.95**	**87.42**	**22.98**	**85.74**	**76.71**	**79.42**

3. Detailed Architecture

As illustrated in Fig. 2, we provide a detailed architecture of Change3D, which is designed to address multiple tasks, including binary change detection, semantic change detection, building damage assessment, and change captioning. Each task involves a video encoder and task-specific decoders. Specifically, the input consists of bi-temporal images I_1 and I_2, along with K perception frames that enhance temporal modeling by enriching inter-frame interactions. These inputs are stacked into a video frame sequence and processed through a multi-layer video encoder. The video encoder, equipped with spatiotemporal modeling ca-

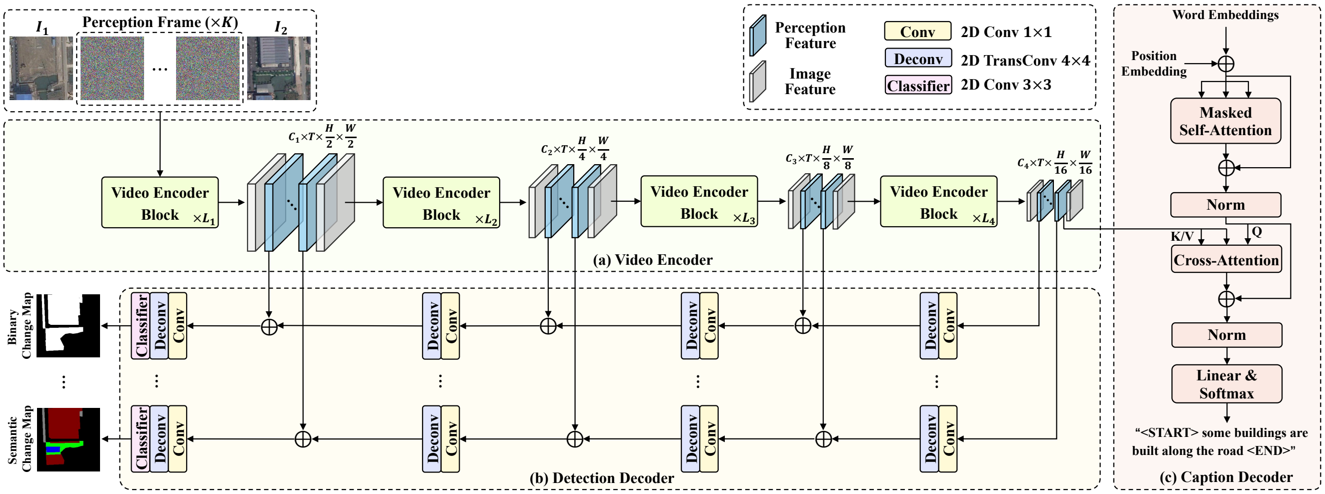

I_1
Perception Frame (×K)
...
I_2
Perception Feature
Image Feature
Conv
2D Conv 1×1
Deconv
2D TransConv 4×4
Classifier
2D Conv 3×3
Video Encoder Block ×L_1
$C_1 \times T \times \frac{H}{2} \times \frac{W}{2}$
Video Encoder Block ×L_2
$C_2 \times T \times \frac{H}{4} \times \frac{W}{4}$
Video Encoder Block ×L_3
$C_3 \times T \times \frac{H}{8} \times \frac{W}{8}$
Video Encoder Block ×L_4
$C_4 \times T \times \frac{H}{16} \times \frac{W}{16}$
(a) Video Encoder
Binary Change Map
Semantic Change Map
Classifier
Deconv
Conv
(b) Detection Decoder
Word Embeddings
Position Embedding
Masked Self-Attention
Norm
K/V
Q
Cross-Attention
Norm
Linear & Softmax
"<START> some buildings are built along the road <END>"
(c) Caption Decoder

Figure 2. Detailed architecture of Change3D. $L_{\{1-4\}}$ denote the number of blocks in each layer.

pabilities, extracts robust features by integrating spatial details with temporal relationships, effectively capturing dynamic changes. The extracted features are then forwarded to task-specific decoders. For change map prediction, the framework leverages multi-layer perception features, while the highest-level semantic features from the encoder's final layer are used for caption generation. Each decoder employs distinct parameters tailored to its respective task.

Change3D eliminates the need for complex, task-specific change extractors, providing a unified framework for diverse change detection and captioning tasks.

4. Theoretical Analysis

To establish a comprehensive theoretical foundation and illustrate the superiority of our proposed method, we present a theoretical analysis of video models applied to change detection and captioning tasks. Our approach diverges significantly from existing methods, particularly in differential feature extraction. Therefore, we provide a detailed analysis of the image encoding and change extraction in the previous paradigm, as well as the video encoding introduced in our proposed paradigm.

Our proposed video models (*i.e.*, video encoder) can be conceptualized as a conditional probabilistic model that utilizes the video encoder's inter-frame relation modeling capabilities to capture changes from the entire input frame sequence. Our method treats bi-temporal images and perception frames as sequences of video frames. Through video encoding, perception frames comprehensively capture contextual information among the images, thereby producing perception features for effective change representation.

4.1. Previous Paradigm

In the previous paradigm, bi-temporal image pairs are treated as separate inputs, each processed individually by a shared-weight image encoder to extract spatial features, followed by a dedicated change extractor. A decoder then makes predictions, as detailed below:

- **Image Encoding:** Each image, I_1 and I_2, is independently encoded to produce feature representations F_1 and F_2:
$$P(F_1 \mid I_1) \text{ and } P(F_2 \mid I_2), \tag{1}$$
where $P(F_1 \mid I_1)$ and $P(F_2 \mid I_2)$ describe the conditional probability distributions of extracting features F_1 and F_2 from images I_1 and I_2, respectively.
- **Change Extraction:** The differential features F_C are derived from a change extraction module:
$$P(F_C \mid F_1, F_2), \tag{2}$$
where this expression represents the conditional probability of obtaining the change features F_C, given the extracted features F_1 and F_2 from the two images.
- **Decoding:** The decoder transforms the differential features to change maps or captions O:
$$P(O \mid F_C), \tag{3}$$
where $P(O \mid F_C)$ denotes the conditional probability of generating output O, such as a change map or caption, based on the differential features F_C.
- **Joint Probability:** The joint probability for generating outputs given the inputs is influenced by independent image encoding and change extraction. Combined with Eqs. (1) to (3), we obtain:
$$\begin{aligned} P(O \mid I_1, I_2) = P(F_1 \mid I_1) \cdot P(F_2 \mid I_2) \cdot P(F_C \mid F_1, F_2) \\ \cdot P(O \mid F_C), \end{aligned} \tag{4}$$

which describes the process of generating the output conditioned on bi-temporal images, including the independent extraction of features and change detection.

- **Entropy:** The entropy expression is defined as follows:

$$H(O, F_C, F_1, F_2)_{\text{prev}} = H(O \mid F_C) + H(F_C \mid F_1, F_2) + H(F_1 \mid I_1) + H(F_2 \mid I_2), \tag{5}$$

  where each term represents the uncertainty at different stages: the entropy of the output given the change features, the entropy of the change features given the image features, and the entropy of each image feature conditioned on the respective input image.

- **Mutual Information:** The mutual information between the input and differential features is defined as:

$$I(F_C; I_1, I_2)_{\text{prev}} = H(F_C)_{\text{prev}} + H(I_1, I_2) - H(F_C, I_1, I_2) \tag{6}$$

  which quantifies the information between the change features F_C and the input images I_1 and I_2.

4.2. Our Paradigm

Our approach redefines change detection and captioning tasks from a video modeling perspective. By incorporating learnable perception frames between the bi-temporal images, a video encoder facilitates direct interaction between the perception frame and the images to extract differences, as follows:

- **Video Encoding:** The bi-temporal images I_1, I_2 incorporated with perception frames I_P are stacked along the temporal dimension to construct a video, then a video encoder processes it to produce differential features F_C, which is formulated as follows:

$$P(F_C | I_1, I_P, I_2), \tag{7}$$

  where this expression reflects the conditional probability of obtaining the differential features F_C directly from the sequence of input images and the perception frame, effectively modeling the inter-frame relations for change extraction.

- **Decoding:** A decoder is applied to predict the change maps or captions O:

$$P(O \mid F_C), \tag{8}$$

  where as before, describes the likelihood of generating output O from the differential features F_C.

- **Joint Probability:** The joint probability benefits from holistic video encoding, which incorporates perception frames. Combined with Eqs. (7) and (8), we get:

$$P(O \mid I_1, I_P, I_2) = P(F_C | I_1, I_P, I_2) \cdot P(O \mid F_C), \tag{9}$$

  where this formulation emphasizes the integrated processing of temporal sequences through video encoding. This approach connects input frames with differential feature extraction seamlessly, eliminating the need for change extractor designs.

- **Entropy:** The entropy expression is formulated as:

$$H(O, F_C)_{\text{our}} = H(O \mid F_C) + H(F_C \mid I_1, I_P, I_2), \tag{10}$$

  which reflects the more efficient processing of information with reduced uncertainty across the change detection pipeline when using perception frames.

- **Mutual Information:** The mutual information between the input and differential features is defined as:

$$I(F_C; I_1, I_2, I_P)_{\text{our}} = H(F_C)_{\text{our}} + H(I_1, I_2, I_P) - H(F_C, I_1, I_2, I_P), \tag{11}$$

  illustrating enhanced mutual information and interdependence achieved by integrating perception frames within video encoding.

4.3. Comparsion

- **Probabilistic Model Comparison:** Our paradigm predicts output more accurately due to the inclusion of perception frames and holistic video encoding, which captures richer inter-frame information:

$$P(O \mid I_1, I_2)_{\text{prev}} < P(O \mid I_1, I_P, I_2)_{\text{our}} \tag{12}$$

- **Entropy Comparison:** Our approach shows reduced overall entropy, indicating that the feature representations are more deterministic and less uncertain, leading to more reliable predictions.

$$H(O, F_C, F_1, F_2)_{\text{prev}} > H(O, F_C)_{\text{our}} \tag{13}$$

- **Mutual Information Comparison:** Our paradigm captures a higher amount of information between the inputs and features, promoting enhanced understanding of changes as a result of direct interaction via video encoding.

$$I(F_C; I_1, I_2)_{\text{prev}} < I(F_C; I_1, I_2, I_P)_{\text{our}} \tag{14}$$

4.4. Summary

Our proposed paradigm demonstrates significant improvements in both probabilistic and information-theoretic measures. By incorporating perception frames into the video encoding process, our approach achieves:

- Lower overall entropy, reflecting more deterministic feature representations.
- Higher mutual information, indicating better capture of the complex interdependencies and information among sequences.

- Enhanced joint probability model, delivering more accurate and reliable predictions in change detection and captioning tasks.

These advantages underscore the theoretical and empirical superiority of our paradigm, making it a simple yet effective framework for change detection and captioning tasks.

5. Qualitative Results

To qualitatively compare our method with previous approaches, we present comprehensive samples randomly selected from eight datasets, as illustrated in Fig. 3-7. Several key observations can be made from these samples: (1) Fig. 3 shows that our proposed method outperforms all compared methods on the binary change detection task across various scenarios, including small, large, dense, and sparse changes. (2) Fig. 4-5 demonstrate that Change3D accurately identifies land cover types and produces clearer boundaries. (3) Fig. 6 indicates that Change3D generates more accurate and semantically consistent assessment maps reflecting damage levels. (4) Fig. 7 shows that Change3D provides more precise descriptions of the changes. These achievements are primarily attributed to the effective feature interaction between learnable perception frames and bi-temporal images in capturing differences, demonstrating the effectiveness of Change3D for multiple change detection and captioning tasks.

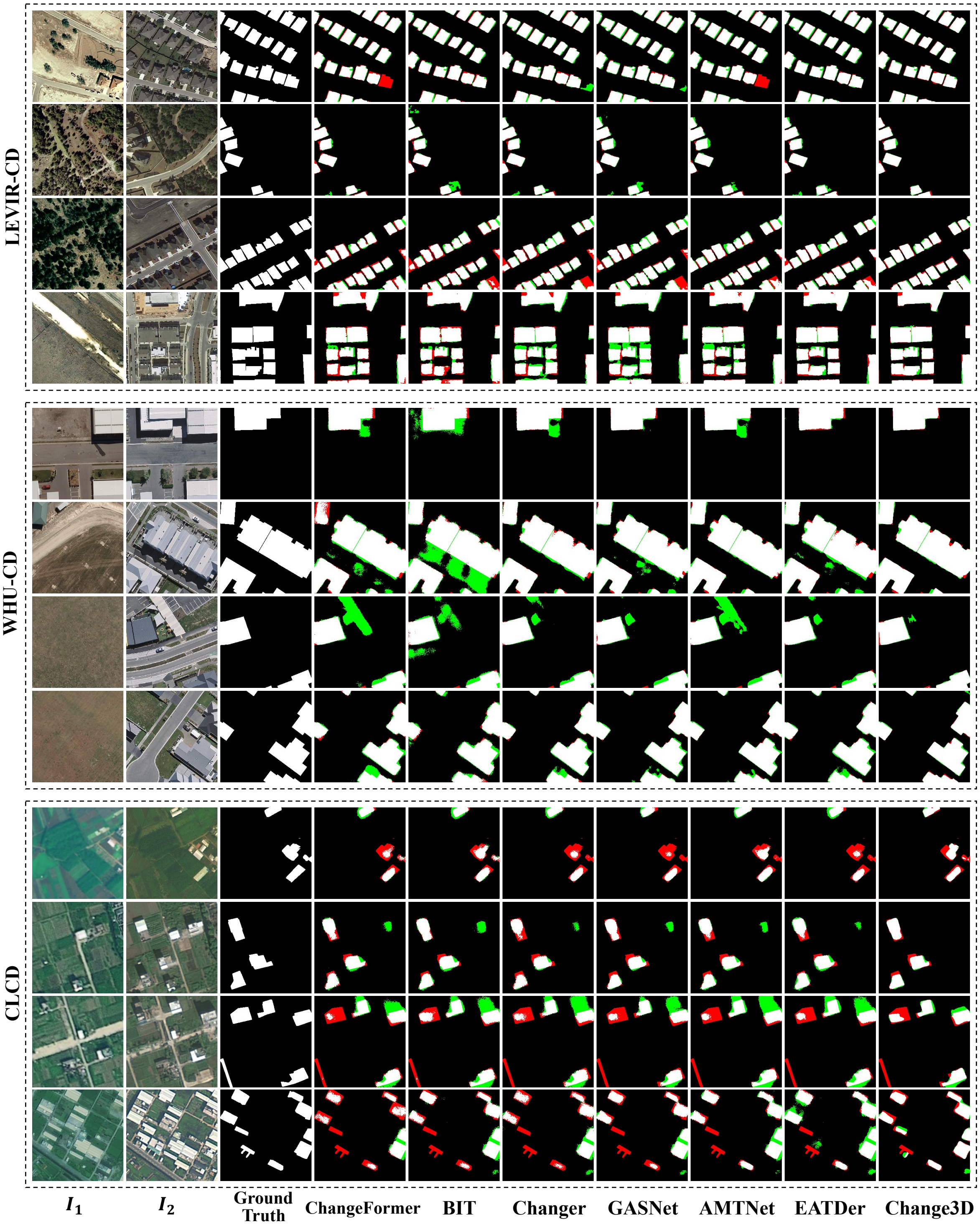


Figure 3. Qualitative comparison of several representative methods on three binary change detection datasets, *i.e*., LEVIR-CD, WHU-CD, and CLCD. White represents a true positive, black is a true negative, green indicates a false positive, and red is a false negative. Fewer green and red pixels represent better performance.

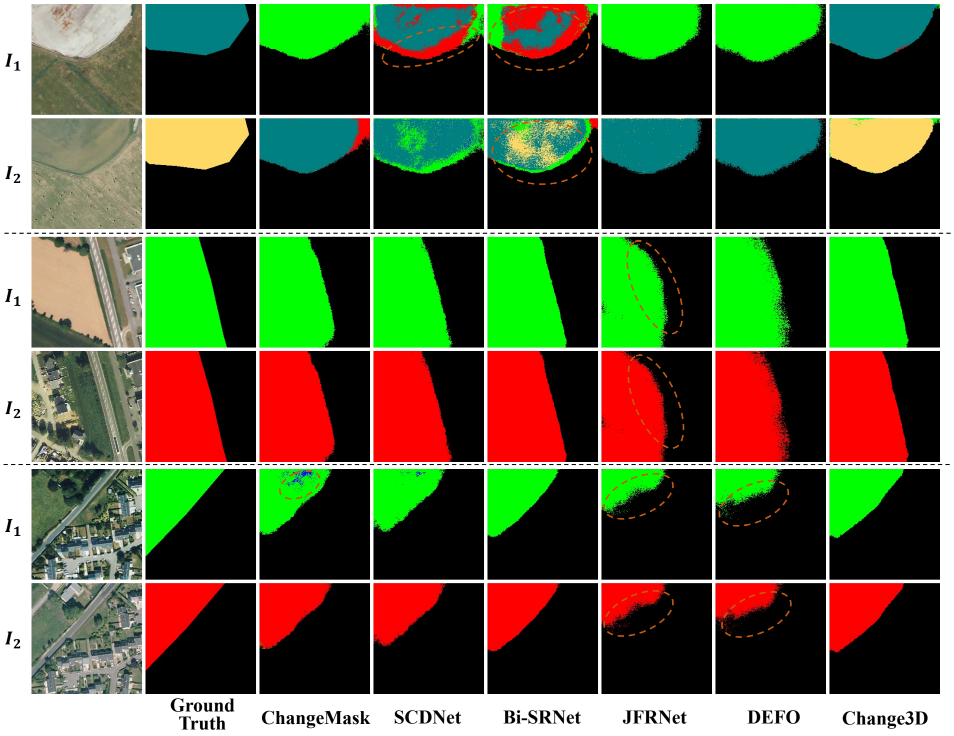


Figure 4. Qualitative comparison of several representative methods on the HRSCD dataset. Black represents non-change, red denotes artificial surfaces, green indicates agricultural areas, blue means forests, yellow represents wetlands, and teal indicates water.

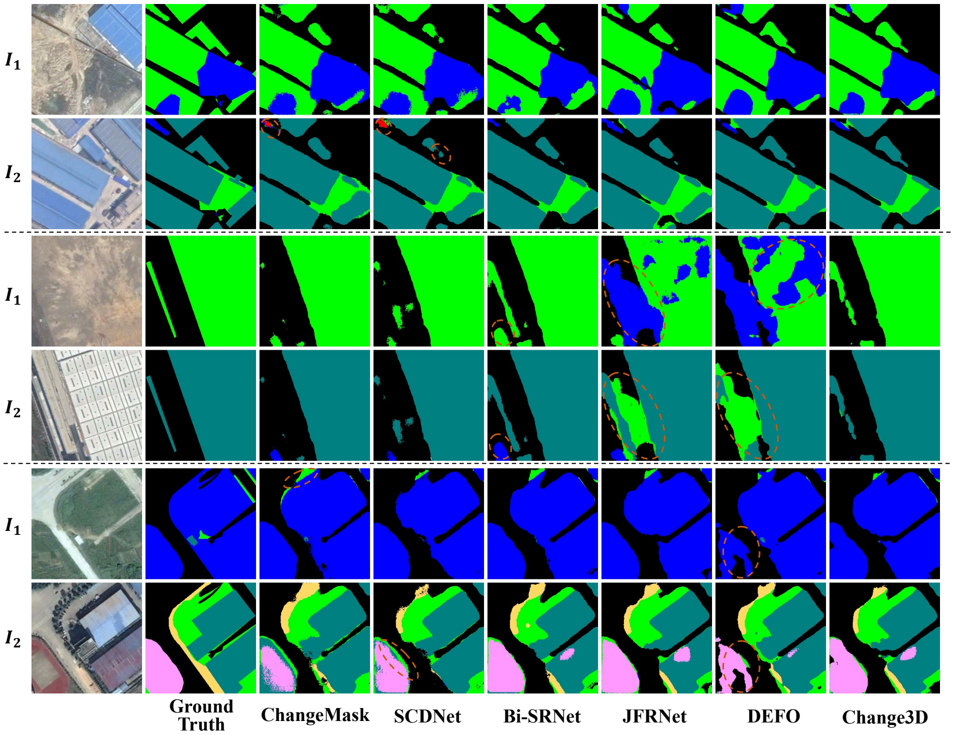


Figure 5. Qualitative comparison of several representative methods on the SECOND dataset. Black represents non-change, red denotes low-vegetation, green indicates non-vegetated ground surface, blue means trees, yellow represents water, teal indicates buildings, and violet denotes playgrounds.

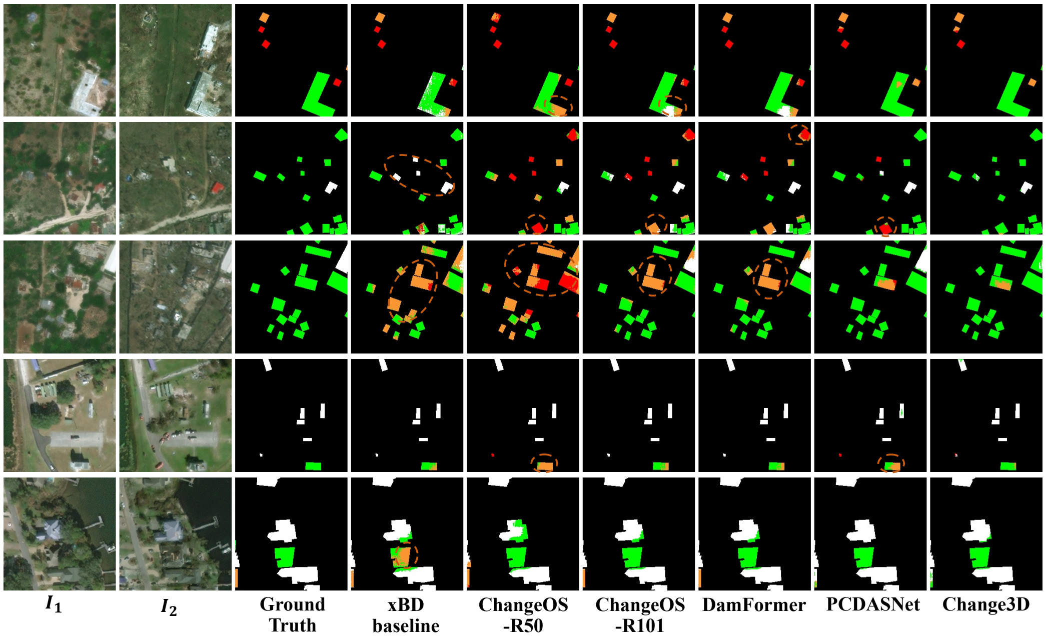
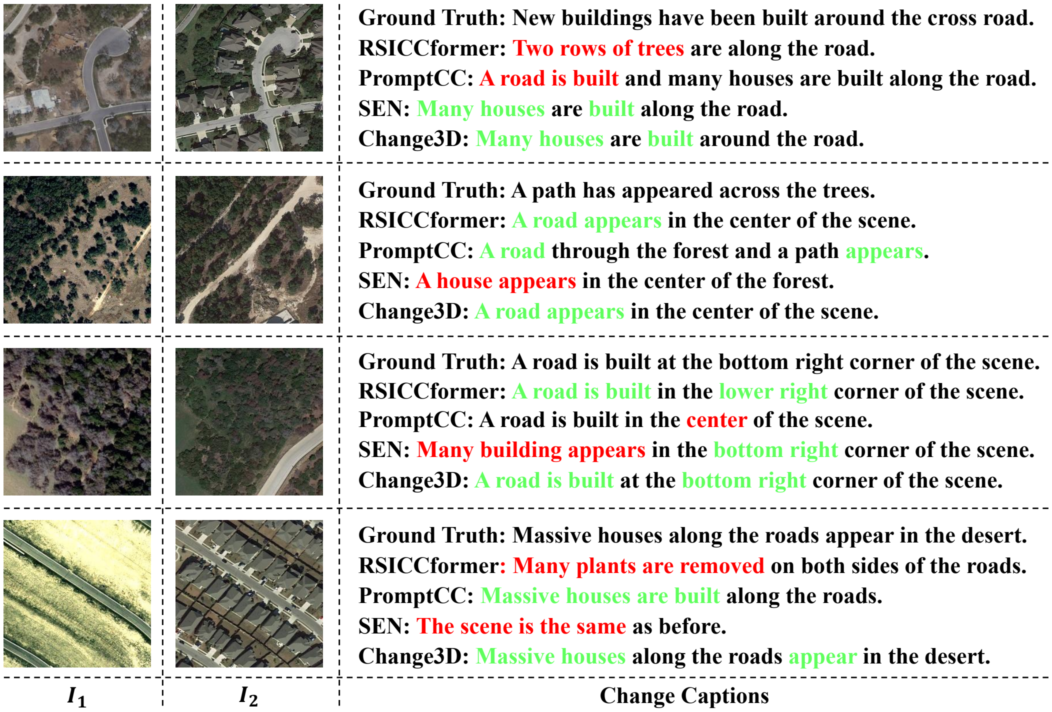


Figure 6. Qualitative comparison of several representative methods on the xBD dataset. Black represents non-change, white denotes non-damage, green indicates minor damage, orange represents major damage, red indicates destroyed.



Ground Truth: New buildings have been built around the cross road.
RSICCformer: Two rows of trees are along the road.
PromptCC: A road is built and many houses are built along the road.
SEN: Many houses are built along the road.
Change3D: Many houses are built around the road.

Ground Truth: A path has appeared across the trees.
RSICCformer: A road appears in the center of the scene.
PromptCC: A road through the forest and a path appears.
SEN: A house appears in the center of the forest.
Change3D: A road appears in the center of the scene.

Ground Truth: A road is built at the bottom right corner of the scene.
RSICCformer: A road is built in the lower right corner of the scene.
PromptCC: A road is built in the center of the scene.
SEN: Many building appears in the bottom right corner of the scene.
Change3D: A road is built at the bottom right corner of the scene.

Ground Truth: Massive houses along the roads appear in the desert.
RSICCformer: Many plants are removed on both sides of the roads.
PromptCC: Massive houses are built along the roads.
SEN: The scene is the same as before.
Change3D: Massive houses along the roads appear in the desert.

Figure 7. Qualitative comparison of several representative methods on the LEVIR-CC dataset. Green indicates correct captions, while red indicates incorrect predictions.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[2] 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. 1

[3] Hongruixuan Chen, Edoardo Nemni, Sofia Vallecorsa, Xi Li, Chen Wu, and Lars Bromley. Dual-tasks siamese transformer framework for building damage assessment. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 1600–1603. IEEE, 2022. 3

[4] 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. 1

[5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 2

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2

[7] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3

[8] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 10–17, 2019. 1

[9] Genc Hoxha, Seloua Chouaf, Farid Melgani, and Youcef Smara. Change captioning: A new paradigm for multitemporal remote sensing image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 1

[10] 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. 1

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[12] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *International Conference on Learning Representations*, 2022. 2

[13] Zhe Li, Xiaoxin Wang, Sheng Fang, Jianli Zhao, Shuqi Yang, and Wen Li. A decoder-focused multi-task network for semantic change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[14] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 1

[15] Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 3

[16] Mengxi Liu, Zhuoqun Chai, Haojun Deng, and Rong Liu. A cnn-transformer network with multiscale context aggregation for fine-grained cropland change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4297–4306, 2022. 1

[17] Wei Liu, Yiyuan Lin, Weijia Liu, Yongtao Yu, and Jonathan Li. An attention-based multiscale transformer network for remote sensing image change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:599–609, 2023. 1, 2, 3

[18] Jingjing Ma, Junyi Duan, Xu Tang, Xiangrong Zhang, and Licheng Jiao. Eatder: Edge-assisted adaptive transformer detector for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 3

[19] Jiaqi Wang, Haonan Guo, Xin Su, Li Zheng, and Qiangqiang Yuan. Pcdasnet: Position-constrained differential attention siamese network for building damage assessment. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3

[20] 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, 2021. 1

[21] Ruiqian Zhang, Hanchao Zhang, Xiaogang Ning, Xiao Huang, Jiaming Wang, and Wei Cui. Global-aware siamese network for change detection on remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199:61–72, 2023. 1, 2, 3

[22] Manqi Zhao, Zifei Zhao, Shuai Gong, Yunfei Liu, Jian Yang, Xiong Xiong, and Shengyang Li. Spatially and semantically enhanced siamese network for semantic change detection in high-resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2563–2573, 2022. 1

[23] Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. 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, 2021. 3

[24] Qing Zhou, Junyu Gao, Yuan Yuan, and Qi Wang. Single-stream extractor network with contrastive pre-training for remote sensing change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3