

## ViewDelta: Scaling Scene Change Detection through Text-Conditioning

Subin Varghese  
 University of Houston  
 4226 MLK Blvd Houston TX  
 srvargh2@cougarnet.uh.edu

Joshua Gao  
 University of Houston  
 4226 MLK Blvd Houston TX  
 jkgao@cougarnet.uh.edu

Vedhus Hoskere  
 University of Houston  
 4226 MLK Blvd Houston TX  
 vhoskere@central.uh.edu

### Abstract

We introduce a generalized framework for Scene Change Detection (SCD) that addresses the core ambiguity of distinguishing “relevant” from “nuisance” changes, enabling effective joint training of a single model across diverse domains and applications. Existing methods struggle to generalize due to differences in dataset labeling, where changes such as vegetation growth or lane marking alterations may be labeled as relevant in one dataset and irrelevant in another. To resolve this ambiguity, we propose ViewDelta, a text conditioned change detection framework that uses natural language prompts to define relevant changes precisely, such as a single attribute, a specific set of classes, or all observable differences. To facilitate training in this paradigm, we release the Conditional Change Segmentation dataset (Cseg), the first large-scale synthetic dataset for text conditioned SCD, consisting of over 500,000 image pairs with more than 300,000 unique textual prompts describing relevant changes. Experiments demonstrate that a single ViewDelta model trained jointly on Cseg, SYSU-CD, PSCD, VL-CMU-CD, and their unaligned variants achieves performance competitive with or superior to dataset specific models, highlighting text conditioning as a powerful approach for generalizable SCD. Our code and dataset are available at [github.io/viewdelta/](https://github.io/viewdelta/).

### 1. Introduction

Detecting and interpreting scene changes is a long-standing task in computer vision [3, 11, 22, 45], with applications in situational awareness [15, 53], disaster assessment [41, 44, 46, 50], and environmental monitoring [20, 31]. Although many public benchmarks exist, each encodes its own notion of what they label as a *relevant* change. SYSU-CD assumes land-use changes in satellite imagery [43], PSCD provides eight semantic street classes [42], and VL-CMU-CD focuses on changes such as cars, people, and litter [1].

Models trained on any single benchmark therefore learn a narrow, implicit definition of change and degrade sharply

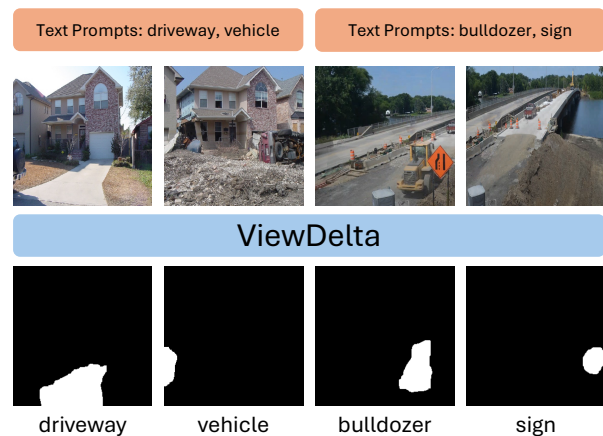


Figure 1. Given a pair of (un)aligned images with text to contextualize what change may be of relevance, ViewDelta highlights relevant changes or all changes.

when deployed elsewhere [24, 28]. Additionally, jointly training on different change detection datasets can be an ambiguous task for a model, as context is needed to understand what change is considered relevant. For example, if a model is given an image pair from a dataset that labels vegetation changes and another that ignores them, it receives contradictory signals about what features define a valid change. Resolving the ambiguity of what change is relevant is key to unlocking joint training across different sources, enabling the creation of a single model that can learn from a greater diversity of data, including from sources that have unaligned views between image pairs.

To address this ambiguity and investigate the performance of a jointly trained change detection model, we introduce the ViewDelta framework, Fig. 1, that incorporates text prompts to *contextualize relevant change* for unaligned image pairs. This approach makes it feasible to unambiguously train a single model across multiple change detection datasets that define relevant change differently, thereby producing a more generalizable model.

In order to evaluate the ability of ViewDelta in distin-

guishing relevant and irrelevant change with varying text, we introduce a new dataset, CSeg (Conditional change Segmentation), consisting of over 500,000 images and 300,000 accompanying unique prompts. CSeg uses inpainting to create synthetic changes, similar in method to COCO-Inpainted [39, 40] and Changen2 [56], to simulate realistic change scenarios. **CSeg intentionally incorporates nuisance changes into images as well as irrelevant text in prompts**, thus requiring comprehension of the accompanying prompt to contextualize what is a relevant change.

ViewDelta is jointly trained on CSeg, PSCD [42], SYSU-CD [43], VL-CMU-CD [1], and the unaligned variants of PSCD and VL-CMU-CD [28]. We compare against fully finetuned models on each dataset and achieve competitive performance relative to domain-specific models across the datasets. Additionally, we evaluated what performance gain could be achieved by finetuning the generalized ViewDelta model for each dataset, showing a consistent improvement in performance. Our contributions include:

- Introducing a novel text prompt conditioned change detection task that outputs binary segmentations based on user-specified textual descriptions of relevant changes.
- Proposing ViewDelta, a framework which enables joint training of change detection datasets that leverages prompt conditioning to disambiguate what change is relevant.
- Creating and publicly releasing the CSeg dataset, consisting of aligned and unaligned image pairs with corresponding textual prompts and annotated segmentation masks.
- Evaluation of the first generalized change detection model capable of performing scene change detection across varying views with user definable definition of relevant change or all changes at runtime across domains.

## 2. Related Work

**Scene Change Detection.** Change detection (CD) for images is the umbrella task of identifying variations between two images of the same place taken at different times. CD methods have most commonly been applied in applications such as surveillance and satellite imagery [2, 17, 19, 43, 57], where alignment between images is a reasonable assumption. CD has found widespread use in satellite imagery for both binary [6, 43] and multi-class [18, 52] change detection.

Scene Change Detection (SCD), the primary focus of this work, narrows CD to ground-level scenes where additional variations not usually accounted for in satellite images are faced. Example variations include in viewpoint, scale, and illumination [1, 39, 39, 42]. Additionally, acquiring and labeling data for SCD is extremely laborious and introduces additional consideration to what constitutes as *relevant* change [1, 42, 49]. For instance, whether shift-

ing shadows, growing foliage, trash, structural degradation, or snow should be considered as a *relevant* or *nuisance* change. The designation of what is a relevant change depends entirely on the application. SCD models therefore learn what change is *relevant* implicitly through the labels used for a dataset. The PSCD [42] dataset introduced the first semantic SCD dataset with one of eight street-scene categories *structure*, *lane marking*, *object (traffic)*, *object (other)*, *human*, *vehicle*, and *barrier*. Semantic SCD allows for explicit definition of *relevant* change. Our work seeks to generalize SCD further by enabling a user to explicitly define *relevant* change through text at runtime.

### Limitations of Foundational Segmentation Models.

General purpose semantic segmentation approaches have had tremendous success with models such as the Segment Anything Model (SAM) [26, 36]. Combined with image registration approaches, general semantic segmentation can provide valuable information for change detection. However, as shown in prior works in SCD [42, 49], even a theoretically perfect open vocabulary semantic segmentation model struggles fundamentally with change detection in two common scenarios if (i) the object type remains the same (e.g., brick building) but changes still occur within an object (e.g., missing brick or new graffiti) and (ii) an object is replaced by another object of the same class (e.g., red car to blue car).

**Generalization in Change Detection.** Recent works have increasingly focused on improving the generalization of SCD models to operate across SCD datasets [24], as well as operate on larger viewpoint variation in real images [28]. GeSCF [24] introduced a framework for generalizable SCD, that utilizes SAM [26], showing both a strong performance across datasets and nearly matches the performance of tailored SCD models like C-3PO [49]. Impressively, GeSCF [24] results also indicate better generalizability than finetuned SAM [26] based CD methods in terms of generalizability. However, GeSCF does not natively provide semantics for the changes found and requires coarse alignment between images.

Towards creating a generalizable model robust to view change and generalize across datasets, Lin et al. [28] proposed using the Dinov2 [32] model with cross-attention, which we shall refer to as the Dinov2 RSCD model. To evaluate the performance of Dinov2 RSCD, Lin et al. [28] created two variants of both PSCD [42] (Diff-1 PSCD and Diff-2 PSCD)[32] and VL-CMU-CD [1] (VL-CMU-CD Diff-1 and VL-CMU-CD Diff-2 [32]). These variants were created by perturbing the coarsely aligned image sequences to the first or second-nearest neighbor, thus allowing for true parallax and occlusion effects. Evaluation of Dinov2 RSCD showed state-of-the-art performance on both VL-CMU-CD, PSCD, and their variants. Our work closely aligns with Dinov2 RSCD [28] in that we also generalize across SCD

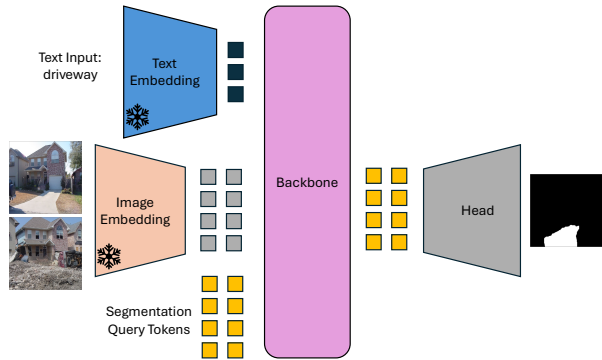


Figure 2. Overview diagram of the ViewDelta change detection framework, consisting of two primary branches to process images and text. The framework is agnostic to choice of embeddings, backbone, and segmentation head. We utilize SigLip text embeddings, Dinov2 image embeddings, a ViT backbone, and a fully convolutional mask decoder. The focus of this framework is to enable the evaluation of text prompts to enable joint training across SCD datasets.

datasets with view variations as well as leverage Dinov2 features in ViewDelta. However, we deviate in training methodology and further generalize the SCD task with text conditioning.

Large vision-language models now offer strong multi-modal reasoning. Gemini 2.5 Pro can accept two images with a natural-language prompt and returns dense masks or bounding boxes. This interface lets us query it for text-conditioned change detection, aligning exactly with ViewDelta’s output format. We therefore include Gemini 2.5 Pro as a baseline and evaluate it on CSeg.

**Synthetic Dataset Generation.** Creating SCD datasets of real images is an extremely laborious and expensive task [1, 42]. To address this, several works have explored synthetic data generation. For instance, the COCO-Inpainted dataset [39] simulates change by inpainting and applying affine transformations to an image. In a follow-up study, KC-3D generates purely synthetic 3D scenes by randomly placing objects and progressively removing them across frames with varying camera viewpoints to create change pairs. More recently, Changen2 [56], leverages diffusion transformer models to simulate temporal changes by editing, removing, or generating object masks.

### 3. ViewDelta

In this section, we introduce **ViewDelta**, our proposed framework for text-prompted change detection, as illustrated in Fig. 2. ViewDelta processes two input images,  $I_a$  and  $I_b$ , captured before and after a change event, respectively, along with a text prompt  $T$  that specifies the type of change relevant to the user. The goal is to predict a binary segmentation map  $M \in \{0, 1\}^{H \times W}$ , where  $M_{i,j} = 1$  indi-

cates the presence of the specified change at pixel location  $(i, j)$  and  $M_{i,j} = 0$  otherwise.

The proposed framework requires the adaptation of a multimodal model to take two images as well as a text prompt as input and output a binary segmentation mask for SCD. This deviates from existing SCD models as well as multi-modal segmentation models, as two images and text are used as input. From the proven effectiveness of transformer-based models in multi-modal applications [34, 47, 54], we also utilize the Vision Transformer (ViT) [14], however we incorporate additional modifications to effectively integrate text prompts and generalizability with joint training for SCD. These modifications include integrating a frozen text encoder to transform the text prompt  $T$  into a sequence of tokens, using an image encoder instead of a traditional patch embedding network for processing the input images  $I_a$  and  $I_b$ , and introducing learnable segmentation query tokens. We employ a conventional fully convolutional segmentation head to produce the final binary segmentation map  $M$  from the learnable segmentation query tokens, rather than directly from the image tokens. These design choices, as shown in our ablation, are required to achieve strong generalizability in SCD.

#### 3.1. Text Embeddings

We leverage only the text encoder of the SigLip [54] model due to its improved performance against models such as CLIP [34], OpenCLIP [16], and EVA-CLIP [47]. We leave the SigLip model frozen in order to keep the text generalization ability of SigLip. We evaluate the generalizability of the model on the CSeg dataset, which contains unique prompts in the test set that are not seen in the training set.

#### 3.2. Image Embeddings

Patch embeddings have been widely adopted by state-of-the-art methods in both change detection and semantic segmentation. However, in our joint training configuration, we observed that optimizing patch embeddings was challenging due to the slow convergence rate on the available training data. We investigate directly using image features from a frozen Dinov2 [32] model as embeddings. Our ablation study demonstrates that Dinov2 embeddings yielded faster convergence rate with less data. We attribute the difficulty of optimizing learned patch embeddings primarily to dataset size limitations rather than inherent issues with patch embeddings themselves.

#### 3.3. Backbone

We utilize the ViT [14] transformer architecture in its base configuration as the backbone, consisting of 12 transformer encoder layers, each characterized by a hidden dimension of 768, an intermediate MLP dimension of 3072, and 12 attention heads. While all image and text tokens are fed into

the backbone, we do not employ feature aggregation strategies on the output tokens. Instead, we only pass the segmentation query tokens to the segmentation head. This differs from common aggregation strategies, such as addition, subtraction or concatenation, commonly employed in models like FC-EF [13], STANet [10], [8], Dinov2 RSCD[28], and DSAMNet [43]. As will be shown in our ablation, aggregating features during joint training showed suboptimal performance for our network configuration.

### 3.4. Segmentation Query Tokens

Drawing inspiration from DETR [5], we leverage dedicated segmentation query tokens (SQT) to enhance model flexibility. These tokens enable the model to create streamlined representations for segmentation-related features, circumventing the inherent complexities encountered if the model had to directly map and align combined image and textual tokens to a feature space suitable for change detection. This methodological choice effectively addresses issues related to adding padding tokens or deciding which tokens from the sequence of image and text tokens to use as input to the segmentation head. As will be shown in our ablation section, removal of these segmentation query tokens causes a significant performance decrease.

### 3.5. Segmentation Head

Existing state-of-the-art change detection architectures, including SwinSUNet [55], ChangeFormerV4 [4], and MambaBCD-Base [8], predominantly assume spatial alignment between input image pairs, which introduces a bias that may limit performance in scenarios involving unaligned or varying viewpoints. To remove this bias, our segmentation head operates only on segmentation query tokens, as shown in Algorithm 1. This design choice removes the necessity of explicit feature combination from spatially mismatched input images and facilitates effective fusion of visual and textual embeddings. The segmentation head itself is minimal in design consisting of a simple five-layer network following a sequence of a 2D convolution, a 2D transposed convolution, a second 2D convolution, bilinear upsampling, and a final 2D convolution, with the convolutional layers using ReLU activations.

### 3.6. Training

**Joint Dataset.** ViewDelta is trained jointly on CSeg, SYSU-CD, PSCD (both its binary and multi-class forms), VL-CMU-CD, plus the Diff-1 and Diff-2 variants of PSCD and VL-CMU-CD. For PSCD, each semantic label is turned into a separate binary mask, and the class name becomes the text prompt. Image pairs that contain no changes for the given prompt are kept in both training and testing to ensure fair comparison with existing multi-class SCD methods.

---

#### Algorithm 1 Forward prediction of proposed model

---

**Require:** Images  $I_a, I_b$  and text embeddings  $T_e$

- 1:  $I_a \leftarrow \text{ImageEmbedder}(I_a)$
- 2:  $I_b \leftarrow \text{ImageEmbedder}(I_b)$
- 3:  $T \leftarrow \text{TextEmbedder}(T_e)$
- 4:  $X \leftarrow \text{Concat}(I_a, I_b, T, \text{SQT})$
- 5:  $X \leftarrow X + \text{PositionalEmbedding}$
- 6:  $X \leftarrow \text{MLP}(X)$
- 7:  $X \leftarrow \text{ViT}(X)$
- 8:  $\text{SQT} \leftarrow X[:, -N_s :, :]$
- 9:  $\text{SegLogits} \leftarrow \text{UpsampleNetwork}(\text{SQT})$
- 10: **return** SegLogits

---

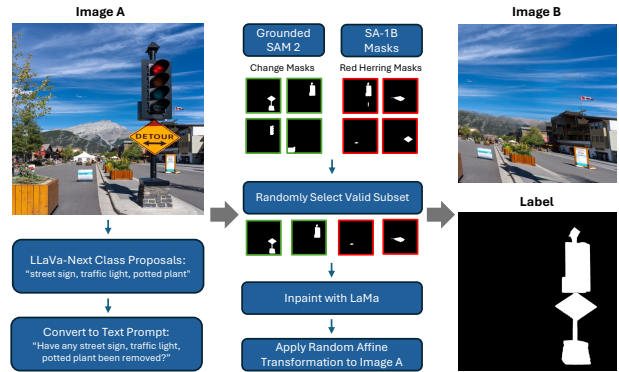


Figure 3. **Outline of the steps involved in generating CSeg.** LLaVA-Next identifies object classes in an SA-1B image, which inform a text prompt and guide Grounded SAM 2 in extracting instance masks. Some masks are then inpainted to simulate changes, while some original SA-1B masks are used as “red herrings”. This process yields a text prompt, Image A, Image B, and a change mask label.

**Training Parameters.** During training, the Adam optimizer [25] with weight decay, a one-epoch warm-up, and cosine annealing learning rate decay, starting from an initial learning rate of  $2 \times 10^{-5}$  is used. Training is performed using a batch size of 4 per GPU using 4 Nvidia A100 GPUs, resulting in an effective batch size of 16. We utilize DeepSpeed ZeRO Stage 2 [35] for sharding optimizer states and gradients, activation checkpointing within attention layers and large linear layer, and 16-bit mixed-precision training.

## 4. CSeg Dataset

Inspired by the COCO-Inpainted dataset [40], we propose a procedure as shown in Fig. 3 that leverages state-of-the-art vision models to generate the CSeg dataset. CSeg is a large-scale, synthetic dataset designed to evaluate a model’s ability to distinguish relevant from nuisance changes in images based on a given text prompt to contextualize what may be relevant.

## 4.1. Generating Changes and Text Prompts

Similar to the COCO-Inpainted dataset [40], we simulate changes between two images by inpainting and applying an affine transformation for view variation. However, our approach is unique in its use of large vision-language models to identify objects to generate diverse text prompts. We first select an SA-1B [26] image and leverage LLaVa-Next [29] to identify a maximum of ten different objects in the image to be used as classes. To promote dataset balance, we select one to five of the least represented classes during generation. These classes are then used directly as text prompts or applied to a natural language prompt template. These prompt templates are a list of 45 predefined and manually validated templates generated with GPT-4o that have the intent of identifying change. To simulate change, Grounded SAM 2 [23, 26, 30, 36–38] is then used to generate instance masks across classes, and we randomly select a subset of at most 10 masks to be inpainted in Image B with LaMa [48]. We combine these inpainted masks to produce the change label. Finally, we apply a random affine transformation to either Image A or Image B to simulate a perspective change. To account for potential occlusion caused by the affine transformation, we check the masks of the inpainted objects and remove any occluded masks from the change label.

ViewDelta should be generalized to detect all changes in an image and not limited to those specified by a prompt, which eliminates the need for explicit class enumeration. To generate these “all” image pairs, we randomly inpaint five to ten SA-1B masks on Image B to simulate change. The text prompt is randomly selected from a predefined list of 96 prompts generated with GPT-4o that convey an intent for all changes.

## 4.2. Combating Inpainting Noise

As stated in [40], the inpainted regions tend to have inpainting artifacts (also observed in other studies [11, 12, 21, 51]). We follow a similar strategy employed in the COCO-Inpainted dataset [40] where we include “red herring” masks to prevent the model from learning artifacts instead of image changes. These red herring masks are a randomly selected subset of zero to ten SA-1B masks that are not classified as any of the classes in the text prompt. Since SA-1B doesn’t provide mask classifications, we use Grounded SAM 2 with a low confidence threshold to ensure that the SA-1B mask doesn’t coincide with a change class instance. These red herring masks are inpainted on image B, but not added to the change mask. During training, this forces the model to learn the change between the two images that are consistent with the text prompt, rather than labeling inpainting noise as changes. Note that adding red herring masks to “all” image pairs is not applicable, since every change needs to be accounted for, which would in-

clude all inpaints.

## 4.3. CSeg Dataset Statistics

There is a total of 501,153 change image pairs in CSeg split into 451,090 pairs in train and 50,063 in test. There are 24,298 unique classes and 339,348 unique prompts in train and 7,285 unique classes and 7,326 unique prompts in test. There are 1,408 unique classes and 35,271 unique prompts not seen in train. 53,257 and 5,947 pairs were “all” image pairs in the train and test sets respectively at about a 12% of the total for each set. CSeg’s class counts are top-heavy in that more common classes such as “clouds”, “sky”, “people”, “water”, “trees” and “roof” appear exponentially more often than less common classes like “calcite”, “aldravanda”, and “ice hammer” as depicted in Fig. 4. This is due to how the majority of images in SA-1B include these common objects. During class proposals, we select the classes that are least used in CSeg to mitigate this problem.

Although LLaVa-Next [29] and Grounded SAM 2 [23, 26, 30, 36–38] are very impressive vision models, they cannot produce flawless class proposals and instance masks for all images across all classes. This consequently affects the quality of CSeg. To quantitatively measure the quality of CSeg, we manually check randomly sampled images for change label correctness until the percent accuracy clearly converges. We validated 500 images - about 0.1% of the entire dataset - and observed an accuracy of 94.0%. This results in a margin of error of  $\pm 2.22\%$  at a 95% statistical confidence. The large majority of errors stem from Grounded SAM 2 yielding masks that are misclassified, especially with more challenging classes. We provide additional examples in Fig. 5 for further evaluation.

## 5. Experiments

We evaluate ViewDelta in two configurations: (1) a general model jointly trained across multiple datasets (CSeg, PSCD, VL-CMU-CD, SYSU-CD, and their unaligned variants), and (2) specialized models that start from the general model and are then fine-tuned on individual datasets. To assess effectiveness across diverse scenarios, we test on five benchmarks: CSeg (our proposed dataset with complex text prompts), PSCD [42] (street-view semantic changes), SYSU-CD [43] (satellite imagery), VL-CMU-CD [1] (street-view with label noise), and the Diff-1/Diff-2 unaligned variants of PSCD and VL-CMU-CD [28] (testing robustness to viewpoint changes). These datasets span indoor, outdoor, street-view, and satellite imagery domains.

We report standard change detection metrics including Intersection over Union (IoU), F1 score, Recall, and Precision. For each dataset, we adopt task-appropriate prompting strategies: dataset-specific class names for PSCD, a comprehensive change description for SYSU-CD, and a description of object categories for VL-CMU-CD. Section

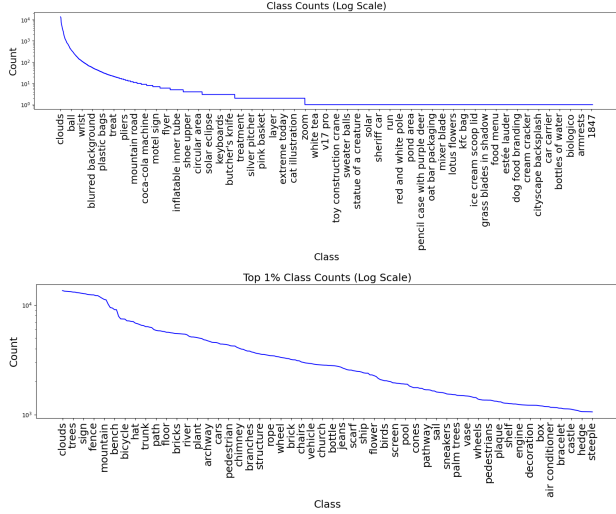


Figure 4. **Class count distributions.** We separate the top 1% most frequent classes and the complete dataset for easier evaluation. These figures don’t include the “all” class.

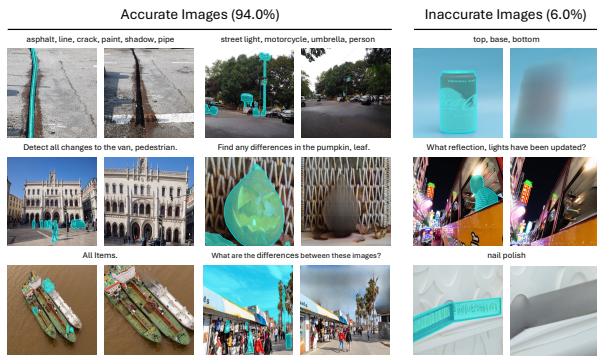


Figure 5. **Evaluation of CSeg label quality.** Change mask labels are highlighted in blue. Manual inspection of the generated images show a 94% agreement with a human reviewer.

5 presents quantitative comparisons with state-of-the-art methods, qualitative analysis of model predictions, and ablation studies examining key architectural components.

### 5.1. Quantitative Evaluation

We evaluate ViewDelta’s core capabilities: (1) interpreting natural language to identify relevant changes, (2) generalizing across diverse scene types, and (3) handling viewpoint variations. Our experiments compare the general jointly-trained model against dataset-specific fine-tuned versions and state-of-the-art baselines.

**Text-Prompted Change Detection (CSeg).** Table 1 demonstrates ViewDelta’s ability to interpret complex natural language prompts on our proposed CSeg dataset. The general ViewDelta achieves 83.80% IoU, significantly outperforming Gemini 2.5 Pro. The high precision (95.07%)

Model	IoU	F1	Rec	Prec
ViewDelta	83.80	91.19	87.61	95.07
Gemini 2.5	37.15 ± 1.48	54.18 ± 1.63	58.31 ± 1.72	50.59 ± 1.80

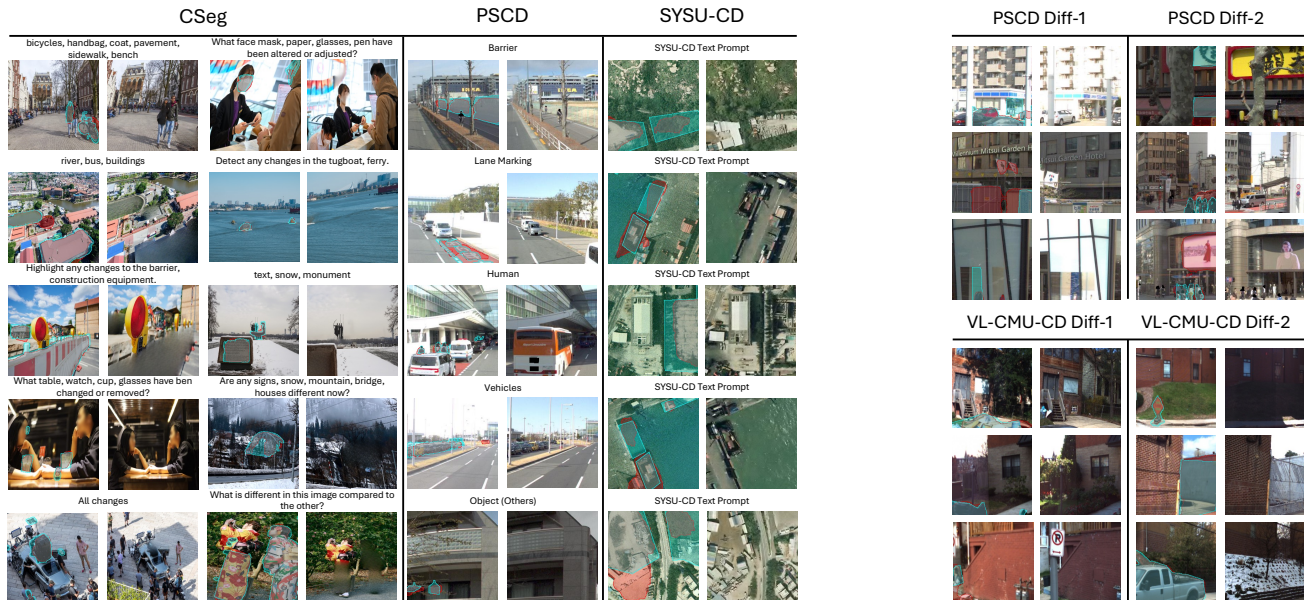
Table 1. **Quantitative Results on the CSeg Dataset.** The table compares ViewDelta, evaluated on the full test set, with a baseline from Gemini 2.5 Pro. Due to the nature of API-based model access, Gemini’s performance was assessed on a randomly selected subset of 2000 test samples. We report the standard error at a 95% statistical confidence for these results to reflect the statistical variance inherent in this sampling approach.

Model	IoU	F1	Rec	Prec
<i>Fine-tuned models</i>				
CSCDNet				
+ SSCDNet [42]	22.3	–	–	–
CSSCDNet [42]	32.2	–	–	–
ViewDelta	<b>55.5</b>	<b>78.3</b>	69.5	<b>73.5</b>
<i>General model</i>				
ViewDelta	51.2	67.8	<b>75.2</b>	61.7

Table 2. **Quantitative Results for semantic SCD on the PSCD [42] Dataset.** Scores are averaged over all classes. To evaluate our text-conditioned model on this multi-class benchmark, we prompt ViewDelta with each class name (e.g., “human”) for every image pair. This procedure is followed regardless of which changes are present, ensuring a fair comparison against standard multi-class SCD methods.

Method	IoU	F1	Rec	Prec
<i>Fine-tuned models</i>				
FC-EF [13]	61.04	75.81	75.17	76.47
FC-Siam-diff [13]	61.01	75.79	75.30	76.28
FC-Siam-conc [13]	60.23	75.18	76.75	73.67
BiDateNet [33]	62.52	76.94	72.60	81.84
STANet [10]	63.09	77.37	<b>85.33</b>	70.76
DSAMNet [43]	64.18	78.18	81.86	74.81
ChangeFormerV4 [4]	65.03	78.81	77.90	79.74
BIT-50 [7]	66.14	79.62	77.90	81.42
TransUNetCD [27]	66.79	80.09	77.73	82.59
SwinSUNet [55]	68.89	81.58	79.75	83.50
MambaBCD [8]	<b>71.10</b>	<b>83.11</b>	<b>80.31</b>	<b>86.11</b>
ViewDelta	<b>70.09</b>	<b>82.41</b>	79.91	85.08
<i>General model</i>				
ViewDelta	67.05	80.27	73.90	<b>87.91</b>

Table 3. **Quantitative Results on SYSU-CD Dataset.** The highest values are highlighted for **First** and **Second**. ViewDelta is given a constant prompt of “urban development, suburban expansion, pre-construction groundwork, vegetation alteration, road widening, and coastal construction” to allow for a fair evaluation.



(a) **Qualitative Results:** We show predictions from the jointly trained ViewDelta model in red, along with the ground truth in blue. **The SYSU-CD Text Prompt** is “urban development, suburban expansion, pre-construction groundwork, vegetation alteration, road widening, and coastal construction”.

(b) **Qualitative Results on unaligned image pairs.** Jointly trained ViewDelta model predictions (red) and ground truth labels (blue).

Figure 6. Qualitative evaluation results showing model predictions across different scenarios.

Method	Aligned	Diff-1	Diff-2
<i>Fine-tuned models</i>			
DR-TANet[9]	19.0	16.9	12.5
C-3PO[49]	43.3	24.6	16.5
Dinov2 RSCD[28]	44.2	28.4	19.1
ViewDelta	<b>63.1</b>	<b>63.6</b>	<b>62.5</b>
<i>General model</i>			
ViewDelta	56.2	58.6	62.2

Table 4. **Quantitative Results (F1 scores) for binary change detection on the unaligned variants of PSCD [28, 42].** ViewDelta is given the prompt “all” to allow for a fair comparison between methods.

indicates ViewDelta effectively filters nuisance changes based on text specifications.

**Cross-Domain Generalization.** Tables 2 and 3 reveal ViewDelta’s generalization across street-view (PSCD) and satellite (SYSU-CD) domains. On PSCD semantic SCD, our general model achieves 51.23% IoU without any dataset-specific training, only 4.31% below the fine-tuned version (55.54%). This small performance gap demonstrates effective transfer learning from joint training. Notably, the general model achieves higher recall (75.16% vs 69.47%), suggesting it learns more inclusive change repre-

Method	Aligned	Diff-1	Diff-2
<i>Fine-tuned models</i>			
DR-TANet[9]	60.7	57.7	56.9
C-3PO[49]	79.5	72.1	69.3
Dinov2 RSCD[28]	79.5	76.0	73.9
ViewDelta	81.4	<b>79.5</b>	<b>78.2</b>
<i>General models</i>			
GeSCF[24]	75.4	-	-
ViewDelta	<b>81.8</b>	78.9	76.1

Table 5. **Quantitative Results (F1 scores) for binary change detection on the unaligned variants of VL-CMU-CD [1, 28].** Due to label noise [24, 49] in VL-CMU-CD, we use a single prompt “Bins, Signs, Traffic-signs, Vehicles, Refuse, Construction, Maintenance Work, Buildings” as these are the changes we have found to be consistent in the human labels. The fine-tuned model’s decrease relative to the general model reflects overfitting to noise.

sentations across datasets.

On SYSU-CD satellite imagery, the general ViewDelta achieves 67.05% IoU, outperforming several methods but falling short from state of the art. The fine-tuned version reaches 70.09% IoU, placing second only to MambaBCD [8]. Remarkably, the general model achieves the highest precision among all methods (87.91%)

**Robustness to Viewpoint Changes.** Tables 4 and 5 evaluate performance on unaligned image pairs with increasing viewpoint differences. On PSCD variants, ViewDelta shows minimal performance degradation. We attribute this robustness to our design choice to avoid spatial alignment assumptions in the segmentation head.

## 5.2. Qualitative Evaluation

We present qualitative results of our model on example image pairs from the CSeg, PSCD, and SYSU-CD test sets in Fig. 6a. For fair comparison on the SYSU-CD test set, all images use the text prompt: “urban development, suburban expansion, pre-construction groundwork, vegetation alteration, road widening, and coastal construction.” For the PSCD test set, each image uses a prompt corresponding to the classes defined in the dataset, such as “vehicle” or “structure.” Qualitative results are also shown for the unaligned variants of PSCD and VL-CMU-CD in Fig. 6b. Additional qualitative analysis is provided in the supplementary material presenting the effect of prompt variations across datasets, such as misspellings, semantic equivalents, and interrogative formulations.

## 5.3. Ablation Study

We conduct comprehensive ablations to understand the contribution of each architectural component to ViewDelta’s performance. Our analysis reveals critical design choices for effective text-prompted change detection.

**Choice of Embeddings.** Table 6 evaluates different combinations of text and image encoders on a subset of CSeg (22,624 images, 12-hour training). The SigLip + Dinov2 combination achieves 56.44% IoU, more than doubling the performance of SigLip + Patch Embedding (26.16%). This improvement demonstrates that pretrained visual features are crucial for data-efficient learning. Interestingly, SigLip consistently outperforms CLIP regardless of image encoder, validating our choice of text encoder for its superior vision-language alignment capabilities.

**Component Analysis.** Table 7 examines the contribution of each architectural component through systematic removal.

1. *Frozen Dinov2 features improve performance across domains.* Replacing Dinov2 with trainable patch embeddings reduces performance on all datasets, with PSCD showing the largest relative decrease and CSeg the smallest.

2. *Segmentation Query Tokens (SQT) are critical for multimodal fusion.* Removing SQT substantially reduces performance, especially on PSCD.

3. *Text prompts significantly impact performance on semantic tasks.* SYSU-CD maintains reasonable performance, while performance drops substantially on CSeg and PSCD. This pattern suggests text prompts are most criti-

cal when distinguishing between multiple semantic categories, as in PSCD, versus binary change detection tasks like SYSU-CD.

Embedding Configurations	IoU
SigLip + Dinov2	<b>56.44</b>
SigLip + Patch Embedding	26.16
CLIP + Dinov2	53.46
CLIP + Patch Embedding	27.56

Table 6. Impact of varying embedding combinations on ViewDelta on a subset of CSeg.

Dataset	CSeg	PSCD [42]	SYSU-CD [43]
ViewDelta	<b>85.91</b>	<b>52.24</b>	<b>68.59</b>
w/o Dinov2	61.05	16.87	34.81
w/o SQT	62.48	8.43	53.55
w/o Prompts	77.72	10.38	67.34

Table 7. IoU of ViewDelta’s main components when jointly trained on Cseg, PSCD, and SYSU-CD. “w/o Dinov2”: Replaces the frozen Dinov2 backbone with a trainable patch embedding. “w/o SQT”: Removes the segmentation query tokens and generates the change mask from the image tokens. “w/o Prompt”: ViewDelta trained with no prompt.

## 5.4. Limitations

Although CSeg supplies a diverse set of prompts for evaluating how well a model identifies relevant changes, its variations are largely synthetic. Though we evaluate the effects of real parallax and occlusions in PSCD and VL-CMU-CD, the prompt vocabularies are restricted to eight and one classes, respectively. To the best of our knowledge, no large scale real dataset currently exists with varying views and varying prompts to contextualize relevant change.

## 6. Conclusion

ViewDelta introduces text-conditioned scene change detection, enabling joint training across datasets with different labeling conventions by using natural language prompts to define relevant changes. The architecture leverages frozen pretrained embeddings and segmentation query tokens to handle viewpoint variations without spatial alignment assumptions. Experiments on our CSeg dataset (500K+ image pairs, 300K+ unique prompts) and existing benchmarks show that a single jointly-trained model achieves performance competitive with dataset-specific methods. Text conditioning proves effective for building generalizable change detection systems that adapt to user-specified relevance at inference time.

## References

- [1] Pablo Fernández Alcantarilla, Simon Stent, Germán Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42:1301–1322, 2016. 1, 2, 3, 5, 7
- [2] Abdulaziz Amer Aleissae, Amandeep Kumar, Rao Muhammad Anwer, Salman Khan, Hisham Cholakkal, Gui-Song Xia, and Fahad Shahbaz Khan. Transformers in remote sensing: A survey. *Remote Sensing*, 15(7):1860, 2023. 2
- [3] Ting Bai, Le Wang, Dameng Yin, Kaimin Sun, Yepi Chen, Wenzhuo Li, and Deren Li. Deep learning for change detection in remote sensing: a review. *Geo-spatial Information Science*, 26(3):262–288, 2023. 1
- [4] 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. IEEE, 2022. 4, 6
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 4
- [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), 2020. 2
- [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, 2021. 6
- [8] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatio-temporal state space model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 4, 6, 7
- [9] Shuo Chen, Kailun Yang, and Rainer Stiefelhagen. Dr-tanet: Dynamic receptive temporal attention network for street scene change detection. *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 502–509, 2021. 7
- [10] Gong Cheng, Ceyuan Yang, Xiwen Yao, Lei Guo, and Junwei Han. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE Transactions on Geoscience and Remote Sensing*, 56:2811–2821, 2018. 4, 6
- [11] Guangliang Cheng, Yunmeng Huang, Xiangtai Li, Shuchang Lyu, Zhaoyang Xu, Hongbo Zhao, Qi Zhao, and Shiming Xiang. Change detection methods for remote sensing in the last decade: A comprehensive review. *Remote Sensing*, 16(13):2355, 2024. 1, 5
- [12] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004. 5
- [13] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection, 2018. 4, 6
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 3
- [15] Zhihang Fu, Yaowu Chen, Hongwei Yong, Rongxin Jiang, Lei Zhang, and Xian-Sheng Hua. Foreground gating and background refining network for surveillance object detection. *IEEE Transactions on Image Processing*, PP:1–1, 2019. 1
- [16] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [17] Andrew Gonzalez, Jonathan M Chase, and Mary I O’Connor. A framework for the detection and attribution of biodiversity change. *Philosophical Transactions of the Royal Society B*, 378(1881):20220182, 2023. 2
- [18] Ritwik Gupta, Bryce Goodman, Nirav N. Patel, Richard Hosfelt, Sandra Sajeed, Eric T. Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew E. Gaston. xbd: A dataset for assessing building damage from satellite imagery. *ArXiv*, abs/1911.09296, 2019. 2
- [19] 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. 2
- [20] Ji Han, Xing Meng, Xiang Zhou, Bailu Yi, Min Liu, and Wei-Ning Xiang. A long-term analysis of urbanization process, landscape change, and carbon sources and sinks: A case study in china’s yangtze river delta region. *Journal of Cleaner Production*, 141:1040–1050, 2017. 1
- [21] Jiaya Jia and Chi-Keung Tang. Image repairing: Robust image synthesis by adaptive nd tensor voting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages I–I. IEEE, 2003. 5
- [22] Huiwei Jiang, Min Peng, Yuanjun Zhong, Haofeng Xie, Zemin Hao, Jingming Lin, Xiaoli Ma, and Xiangyun Hu. A survey on deep learning-based change detection from high-resolution remote sensing images. *Remote Sensing*, 14(7):1552, 2022. 1
- [23] Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. T-rer2: Towards generic object detection via text-visual prompt synergy, 2024. 5
- [24] Jaewoo Kim and Uehwan Kim. Towards generalizable scene change detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 2, 7
- [25] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4
- [26] 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. 2, 5

- [27] Qingyang Li, Ruofei Zhong, Xin Du, and Yu Du. Transunetcd: A hybrid transformer network for change detection in optical remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 6
- [28] Chun-Jung Lin, Sourav Garg, Tat-Jun Chin, and Feras Dayoub. Robust scene change detection using visual foundation models and cross-attention mechanisms. *ArXiv*, abs/2409.16850, 2024. 1, 2, 4, 5, 7
- [29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5
- [30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 5
- [31] Tim Newbold, Lawrence N Hudson, Samantha LL Hill, Sara Contu, Igor Lysenko, Rebecca A Senior, Luca Börger, Dominic J Bennett, Argyrios Choimes, Ben Collen, et al. Global effects of land use on local terrestrial biodiversity. *Nature*, 520(7545):45–50, 2015. 1
- [32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 2, 3
- [33] Maria Papadomanolaki, Sagar Verma, Maria Vakalopoulou, Siddharth Gupta, and Konstantinos Karantzalos. Detecting urban changes with recurrent neural networks from multi-temporal sentinel-2 data, 2019. 6
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [35] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 4
- [36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 2, 5
- [37] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding dino 1.5: Advance the "edge" of open-set object detection, 2024.
- [38] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 5
- [39] Ragav Sachdeva and Andrew Zisserman. The change you want to see. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3982–3991, 2022. 2, 3
- [40] Ragav Sachdeva and Andrew Zisserman. The change you want to see (now in 3d). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2060–2069, 2023. 2, 4, 5
- [41] Keiko Saito, Robin JS Spence, Christopher Going, and Michael Markus. Using high-resolution satellite images for post-earthquake building damage assessment: a study following the 26 january 2001 gujarat earthquake. *Earthquake spectra*, 20(1):145–169, 2004. 1
- [42] Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection, 2022. 1, 2, 3, 5, 6, 7, 8
- [43] Qian Shi, Mengxi Liu, Shengchen Li, Xiaoping Liu, Fei Wang, and Liangpei Zhang. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–16, 2021. 1, 2, 4, 5, 6, 8
- [44] Wenzhong Shi, Min Zhang, Rui Zhang, Shanxiong Chen, and Zhao Zhan. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, 12(10):1688, 2020. 1
- [45] Ashbindu Singh. Review article digital change detection techniques using remotely-sensed data. *International journal of remote sensing*, 10(6):989–1003, 1989. 1
- [46] Jérémie Sublime and Ekaterina Kalinicheva. Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the tohoku tsunami. *Remote Sensing*, 11(9):1123, 2019. 1
- [47] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3
- [48] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 5
- [49] G. Wang, Bin-Bin Gao, and Chengjie Wang. How to reduce change detection to semantic segmentation. *ArXiv*, abs/2206.07557, 2022. 2, 7
- [50] Chuyi Wu, Feng Zhang, Junshi Xia, Yichen Xu, Guoqing Li, Jibo Xie, Zhenhong Du, and Renyi Liu. Building damage detection using u-net with attention mechanism from pre-and post-disaster remote sensing datasets. *Remote Sensing*, 13(5):905, 2021. 1
- [51] Haiwei Wu, Jiantao Zhou, Jinyu Tian, and Jun Liu. Robust image forgery detection over online social network shared

- images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13430–13439, 2022. [5](#)
- [52] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Semantic change detection with asymmetric siamese networks, 2021. [2](#)
- [53] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance. *Trans. Img. Proc.*, 25(9):4354–4368, 2016. [1](#)
- [54] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. [3](#)
- [55] Cui Zhang, Liejun Wang, Shuli Cheng, and Yongming Li. Swinsunet: Pure transformer network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. [4](#), [6](#)
- [56] 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*, 47:725–741, 2024. [2](#), [3](#)
- [57] 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](#)