The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-4/W16-2025
9th International Conference on Smart Data and Smart Cities (SDSC), 2–5 September 2025, Kashiwa, Japan

VLM-Based Building Change Detection with CNN-Transformer

Zeinab Gharibbafghi[1], Peter Reinartz[2]

[1] University of Osnabrueck, Germany - zgharibbafgh@uni-osnabrueck.de
[2] German Aerospace Center (DLR), Germany - peter.rainartz@dlr.de

Keywords: Building Change Detection, Vision-Language Model, Satelllite Imagery, Transformer Model, Grounding Dino.

Abstract

Accurate building change detection in high-resolution satellite imagery is critical for urban planning, disaster response, and smart city applications. Existing methods often rely on large labeled datasets or handcrafted features, limiting scalability across diverse geographic regions. In this paper, we propose a hybrid framework that integrates a pretrained Vision-Language Model (Grounding DINO) with a lightweight CNN-Transformer architecture to perform text-guided building change detection. Without any fine-tuning, Grounding DINO generates semantic building masks from bi-temporal image pairs using the text prompt "building," which are used to amplify structural features in a ResNet18 backbone. A custom Transformer encoder with dual spatial and channel attention refines these features to capture both local details and global context. On the LEVIR-CD dataset, our framework improves Recall by +3.98%, F1-Score by +3.01%, and Intersection over Union (IoU) by +4.70% compared to a CNN-Transformer baseline. These results highlight the potential of vision-language models to enhance remote sensing workflows without extensive domain-specific fine-tuning.

1. Introduction

Urban landscapes evolve rapidly, making the detection of building changes in high-resolution satellite imagery critical for urban planning, smart city development, and disaster response. Accurate and timely change detection enables monitoring of infrastructure growth, assessing post-disaster impacts, and supporting sustainable urban development initiatives. However, traditional approaches to building change detection, such as those relying on handcrafted features or shallow learning algorithms, often fail to capture the complex spatio-temporal patterns present in remote sensing data.

Recent advancements in deep learning, particularly convolutional neural networks (CNNs) and Transformer architectures, have significantly improved the ability to model these intricate relationships. At the same time, Vision-Language Models (VLMs) have emerged as powerful tools for integrating textual cues with visual data, enabling improved contextual understanding and focusing on task-relevant features. Despite these advances, adapting pretrained VLMs to remote sensing tasks remains challenging due to differences between natural images and satellite data, the need for domain-specific fine-tuning, and the scarcity of labeled datasets for supervised learning.

To address these challenges, this paper introduces a novel framework that integrates a pretrained Vision-Language Model (Grounding DINO) with a lightweight CNN-Transformer architecture for building change detection in satellite imagery. Grounding DINO generates semantic building masks from bi-temporal image pairs using the text prompt "building." These masks are then used to amplify structural features extracted by a ResNet18 backbone (He et al., 2016), while a custom Transformer encoder with dual spatial and channel attention mechanisms captures both local details and global context.

This study investigates whether integrating a pretrained Vision-Language Model (VLM), specifically Grounding DINO, with a lightweight CNN-Transformer framework can enhance building change detection in high-resolution satellite imagery without extensive fine-tuning or large labeled datasets. Specifically, we ask:

Does incorporating VLM-guided masking and dual attention mechanisms into a CNN-Transformer framework improve building change detection accuracy compared to a baseline model?

We address this question through experiments on the LEVIR-CD dataset, including ablation studies and threshold sensitivity analyses. Results show that incorporating VLM-guided masking improves Recall, F1-Score, and IoU over a CNN-Transformer baseline, demonstrating the potential of Vision-Language Models to enhance change detection pipelines without domain-specific fine-tuning.

1.1 Contributions

The main contributions of this work are summarized as follows:

We propose a text-informed preprocessing strategy that employs Grounding DINO's pretrained weights to generate building masks without domain-specific fine-tuning, adapting Vision-Language Models for remote sensing change detection.

We design an input amplification approach that enhances building regions in satellite imagery, guiding a hybrid ResNet18-Transformer network to focus on structural changes.

We conduct extensive evaluations, including ablation studies and sensitivity analyses, demonstrating that the proposed framework achieves better results on the LEVIR-CD dataset while minimizing reliance on large labeled datasets.

This approach provides a scalable and adaptable solution for urban monitoring and Earth Observation(EO) applications.

2. Related Works

2.1 Deep Learning Based Change Detection

Building change detection using remote sensing has been a challenging and hot topic in the field of earth observation. With

the advent of deep learning, Convolutional Neural Networks (CNNs) have markedly improved building change detection by learning hierarchical feature representations. In particular, Siamese CNN architectures—such as FC-Siam-Diff (Daudt et al., 2018) and U-Net based models have been successfully applied to remote sensing change detection tasks (Peng et al., 2019), (Chen and Shi, 2020).

Despite these advances, CNNs struggle to capture long-range dependencies and global context due to their limited receptive fields, hindering comprehensive change analysis. Attention-based Transformer architectures, including Vision Transformers (ViT), address this by blending local and global information (Vaswani and ..., 2017), (Dosovitskiy et al., 2020). Using attention to model distant relationships, they improve image understanding and feature extraction, enabling more effective change detection (Chen et al., 2022).

Hybrid models that combine CNNs with Transformers (Bazi et al., 2021) effectively fuse local features with global context, thereby improving multi-scale feature extraction. In addition, hierarchical transformers such as the Swin Transformer (Liu et al., 2021), (Zhang et al., 2022) have further advanced remote sensing applications by efficiently modeling multi-scale contextual features through shifted window-based attention. Despite these advances, existing methods often require large amounts of labeled data, extensive fine-tuning, and may lack the object-specific focus needed for precise change detection.

2.2 Vision Language Models in Remote Sensing

Vision-language models (VLMs) have transformed many computer vision tasks by effectively combining visual and textual information. For instance, CLIP (Radford et al., 2021) aligns images with text descriptions in natural scenes, while Grounding DINO (Liu et al., 2023) uses a transformer-based design to directly incorporate text prompts for open-set object detection. Despite these advances, remote sensing applications have mostly relied on vision-only methods for tasks such as satellite change detection (Chen and Shi, 2020) and land type classification (Bazi et al., 2021).

In remote sensing, models like SAM (Kirillov et al., 2023), developed by Meta AI, have been adapted for segmentation tasks. SAM achieves zero-shot generalization through promptable interactive inputs, such as points, bounding boxes, and masks. For example, it has been used to segment land cover using simple prompts (e.g., “forest” or “urban”) without retraining, as seen in tools like samgeo (Wu and Osco, 2023). Unlike SAM’s interactive prompts, Grounding DINO directly leverages text for zero-shot detection, suiting automated change tasks. However, adapting vision foundation models (VFMs) to remote sensing data is challenging because they are originally trained on natural images. These models often require extensive fine-tuning due to lower resolution and inconsistent datasets, making the process computationally expensive and data-intensive (Diab et al., 2025).

Recent advancements have started to leverage the multimodal strengths of VLMs for change detection in satellite imagery. For example, ChangeCLIP (Dong et al., 2024) introduced a framework that uses CLIP’s multimodal features to enhance building change detection via text prompts with minimal fine-tuning. Qiu et al. (Qiu et al., 2024) further refined this approach by integrating a transformer-based fusion of optimized prompts and image features, and SegCLIP (Zhang et al., 2024) demonstrates how CLIP can be incorporated into semantic segmentation tasks to effectively merge textual and visual information.

While methods such as ChangeCLIP and SegCLIP have shown promising results in leveraging multimodal features, they often require domain-specific adaptation or optimized prompt engineering for remote sensing applications. In contrast, our approach employs pretrained Grounding DINO in a zero-shot setting, demonstrating its feasibility without task-specific fine-tuning.

To our knowledge, our work is the first to employ Grounding DINO in remote sensing building change detection. By using its zero-shot capabilities, our method avoids domain-specific retraining and introduces a simple hybrid approach that uses pretrained VFMs to guide a Transformer model in detecting building changes in satellite imagery.

3. Methodology

In this section, we first introduce the base model, a simple and lightweight hybrid Transformer with a ResNet18 backbone (He et al., 2016), designed for building change detection using benchmark EO dataset. Next, we provide a detailed description of our proposed method to employ pretrained Grounding DINO model (Liu et al., 2023). This will enhance the base model and improve results without requiring fine-tuning or domain adaptation.

3.1 Base Model Architecture

Our base model is a lightweight CNN-Transformer hybrid, integrating a ResNet18 backbone (He et al., 2016) with a custom Transformer encoder for efficient building change detection. The ResNet18, pretrained on ImageNet, extracts multi-scale spatial features from image pairs (epoch1, epoch2) through its convolutional layers (up to conv4_x), yielding feature maps with 512, 256, 128, 64, and 3 channels. These features are differenced to capture temporal changes, then reshaped into sequences for Transformer processing. The custom Transformer encoder employs multi-head self-attention (8 heads) to model spatial dependencies and a channel attention mechanism inspired by the Convolutional Block Attention Module (CBAM) (Woo et al., 2018) to refine feature importance across channels.

The spatial attention, via multi-head self-attention, computes weights over spatial positions. For input $x \in \mathbb{R}^{L\times B\times C}$ (where L is sequence length, B is batch size, C is feature dimension), it calculates:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where $Q = xW_Q$, $K = xW_K$, $V = xW_V$ are projections ($W_Q, W_K, W_V \in \mathbb{R}^{C\times d_k}$), and $d_k = C/8$ is the head dimension. This focuses on spatially relevant change regions.

The channel attention weights features using global pooling. For input $x \in \mathbb{R}^{C\times(H\cdot W)}$, it computes:

$$\text{CA}(x) = \sigma(a+m), \tag{2}$$

$$a = \text{FC}_2(\text{ReLU}(\text{FC}_1(\text{AvgPool}(x)))), \tag{3}$$

$$m = \text{FC}_2(\text{ReLU}(\text{FC}_1(\text{MaxPool}(x)))), \tag{4}$$

where $\text{FC}_1 : C \to C/16$, $\text{FC}_2 : C/16 \to C$, and σ is the sigmoid function. These weights enhance informative channels.

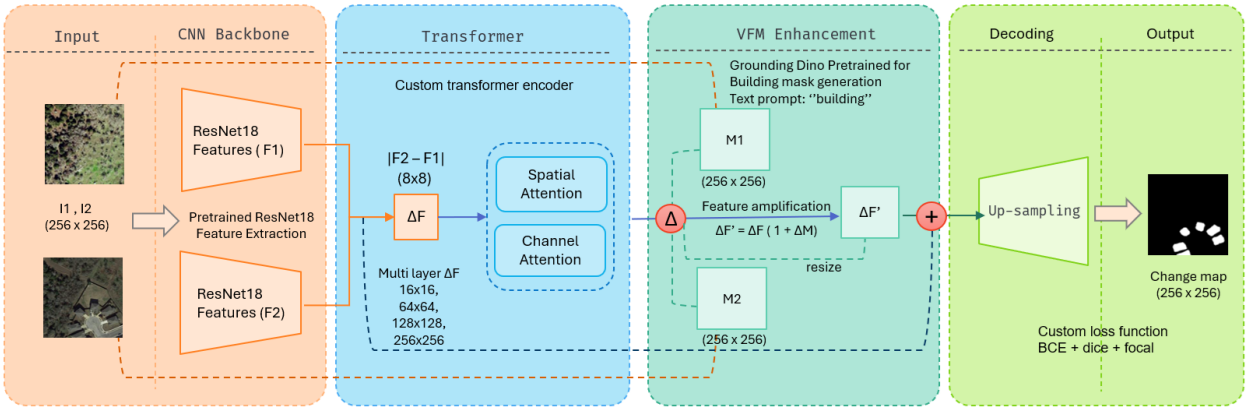

Figure 1. Flowchart of the proposed methodology for building change detection.

Features are upsampled and refined to output a 256x256 change map.

3.2 VFM Enhancement

To enhance our base CNN-Transformer, we employ Grounding DINO (Liu et al., 2023), which uses a Swin-T backbone to extract multi-scale image features, a BERT text encoder for "building" prompt embeddings, and a DINO-based decoder with cross-modal attention (Kirillov et al., 2023) to produce bounding boxes and logits. For each image pair (I_1, I_2) from epoch1 and epoch2, we process the images with Grounding DINO to generate building masks, converting these outputs into binary masks M_1 and M_2 based on confidence scores exceeding 0.5. This 0.5 threshold was chosen empirically to balance precision and recall, following standard practice in object detection. These masks are defined as:

$$M_i(x,y) = \begin{cases} \max(\ell_j) & \text{if } (x,y) \in B_j,\ \ell_j > 0.5, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where B_j are bounding boxes, ℓ_j are logits for "building," and $i = 1, 2$ denotes epochs. The change mask is computed as:

$$M_\Delta = |M_1 - M_2|, \tag{6}$$

resized to match ResNet18 feature dimensions (e.g., 8x8) using bilinear interpolation. Feature differences from the CNN ($F_\Delta = F_1 - F_2$) are amplified:

$$F'_\Delta = F_\Delta \cdot (1 + M'_\Delta), \tag{7}$$

where M'_Δ is the resized change mask. This amplification emphasizes building-related changes, leveraging pretrained VFM knowledge to boost performance without domain adaptation.

3.3 Loss Function

To train our model effectively for building change detection, we employ a combined loss function that balances binary cross-entropy (BCE), Dice loss, and focal loss to address class imbalance in the LEVIR-CD dataset (Chen and Shi, 2020). The BCE loss is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(p_i) + (1-y_i)\log(1-p_i)\Big], \tag{8}$$

where $p_i = \sigma(\hat{y}_i)$ is the sigmoid output and N is the total number of pixels.

The Dice loss, which enhances overlap accuracy, is given by:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} p_i y_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} y_i + \epsilon}, \tag{9}$$

with $\epsilon = 10^{-6}$ ensuring stability.

The total loss is:

$$\mathcal{L} = w_{\text{BCE}}\mathcal{L}_{\text{BCE}} + w_{\text{Dice}}\mathcal{L}_{\text{Dice}}, \tag{10}$$

with $w_{\text{BCE}} = w_{\text{Dice}} = 1.0$.

Figure 1 illustrates the complete methodology, from bi-temporal image input through ResNet18 feature extraction, VFM enhancement with Grounding DINO, feature amplification, Transformer processing, and decoder upsampling, including the training loop with costume loss function.

4. Experiments

4.1 Dataset

We use the LEVIR-CD dataset (Chen and Shi, 2020), a benchmark for building change detection in high-resolution satellite imagery. It comprises 637 pairs of RGB images captured at different times between 2002 and 2018, covering various regions in Texas, USA, to study building-related changes. Each pair is originally 1024x1024 pixels, with corresponding binary change masks. To align with our model's input requirements and enhance sample diversity, we resize images and masks to 256x256 pixels. Assuming each 1024x1024 pair is cropped into four 256x256 sub-pairs, this yields 2548 pairs, split into training (70%, 1780 pairs), validation (10%, 256 pairs), and test (20%, 512 pairs) sets. This resizing enhances the detail level for detecting fine building changes.

4.2 Training Setup

Training was performed using PyTorch on an NVIDIA GeForce RTX 3090 GPU with 24GB memory for efficient batch processing. We applied data augmentation, including random horizontal and vertical flips (probability 0.5), to enhance training robustness and generalization. The model was trained with

Adam optimization (`lr=0.001`, `weight_decay=0.00001`) and a StepLR scheduler (step size 5, gamma 0.1), using the combined loss from Section 3. Training ran for 60 epochs, with the best model selected based on validation loss.

4.3 Evaluation Metrics

We assess performance using standard change detection metrics: Overall Accuracy (OA), Precision, Recall, F1-Score, and Intersection over Union (IoU). For predictions $\hat{y}$ and ground truth y (both $\{0, 1\}$ per pixel), with true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), the formulations are:

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{11}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{12}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{13}$$

$$\text{F1} = 2 \cdot \frac{\text{Pre} \cdot \text{Rec}}{\text{Pre} + \text{Rec}}, \tag{14}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}. \tag{15}$$

5. Results and Discussion

We evaluated the impact of integrating Grounding DINO's text-guided enhancement into our CNN-Transformer framework by comparing our full VFM model to a baseline CNN-Transformer on the LEVIR-CD test set. As shown in Table 1, the baseline model achieved an Overall Accuracy (OA) of 98.80%, Precision of 88.11%, Recall of 82.80%, F1-Score of 85.37%, and an Intersection over Union (IoU) of 74.48%. With the addition of Grounding DINO's enhancement, our full model improved these metrics to an OA of 99.03%, Precision of 90.04%, Recall of 86.78%, F1-Score of 88.38%, and IoU of 79.18%. These improvements demonstrate that text-guided feature amplification significantly boosts the detection of building changes, with F1-Score and IoU serving as key metrics that balance precision and recall while accurately measuring spatial overlap.

Table 1. Performance on LEVIR-CD test set (%) Base: CNN-Transformer, Full: VFM-Enhance.

Model	OA	P	R	F1	IoU
FC-Siam-Diff	98.11	74.38	80.00	86.55	66.68
STANet	98.46	91.21	85.79	80.99	75.12
CDNet	98.67	85.08	86.70	88.38	76.52
Our Base model	98.80	88.11	82.80	85.37	74.48
Our Full model	99.03	90.04	86.78	88.38	79.18

Furthermore, when compared to other reported methods in the literature (Qiu et al., 2024), our approach exhibits competitive performance. For example, while FC-Siam-Diff achieves an IoU of 66.68%, our full model improves IoU by over 12 percentage points. Although STANet shows high precision (91.21%), its overall F1-Score (80.99%) and IoU (75.12%) are lower than those of our method. Similarly, CDNet's performance, with an F1-Score of 88.38% and an IoU of 76.52%, is comparable, yet our model provides a balanced enhancement across all metrics. These comparisons highlight the effectiveness of our text-guided enhancement in achieving robust and competitive results for building change detection.

An ablation study (Table 2) further isolates the contributions of individual components. Removing the VFM enhancement drops the IoU to 74.47%, nearly matching the baseline performance, while disabling the channel attention module reduces the IoU further to 68.31%. These results confirm that both the text-guided enhancement and the channel attention module are critical to achieving higher detection accuracy.

Table 2. Ablation study on LEVIR-CD test set to evaluate the effect of channel attention and VFM enhancement (%).

Configuration	IoU	F1
Full Model	79.18	88.38
w/o Channel Attention	74.47	85.38
w/o VFM Enhancement	68.31	81.16

Qualitative results support these quantitative findings too. As illustrated in Figure 2, our framework effectively highlights building-related changes while suppressing irrelevant background differences. White pixels correspond to true positives, while red and green pixels denote false positives and false negatives, respectively. Notably, the model demonstrates strong performance in suburban areas with sparse construction but exhibits minor inaccuracies in dense urban cores where small rooftop structures are partially occluded.

Leveraging pretrained models such as Grounding DINO and SAM for building change detection in satellite imagery introduces challenges due to domain mismatch between their general-purpose pretraining (e.g., on datasets like COCO or SA-1B) and the specific characteristics of remote sensing data. Without fine-tuning, these models may produce inaccurate building masks—such as missed detections or false positives—which can propagate errors into the downstream change detection pipeline and degrade overall accuracy. To mitigate these issues, we apply strategies like weighted integration of VLM outputs and confidence thresholding, which limit the influence of potentially noisy masks on feature differences. This approach enhances robustness and reduces dependency on fine-tuning, offering a balance between computational efficiency and detection performance, albeit with some trade-offs in precision compared to fully customized models.

Overall, our experimental results validate the effectiveness of incorporating pretrained vision-language models, specifically Grounding DINO, into a CNN-Transformer framework for change detection in satellite imagery. The integration of text-guided feature amplification not only enhances overall performance but also contributes to more reliable and interpretable change maps.

6. Conclusion

In this paper, we presented a hybrid change detection framework that integrates a pretrained Vision-Language Model (Grounding DINO) with a CNN-Transformer architecture for building change detection in high-resolution satellite imagery. By leveraging text-guided semantic masks, our approach amplifies structural features and focuses the network on task-relevant regions without requiring extensive domain-specific fine-tuning. Experiments on the LEVIR-CD dataset demonstrate notable improvements over a baseline CNN-Transformer model, with gains in Recall, F1-Score, and IoU. Ablation studies further confirm the critical contributions of both the vision-language enhancement and the dual attention mechanisms to these improvements.

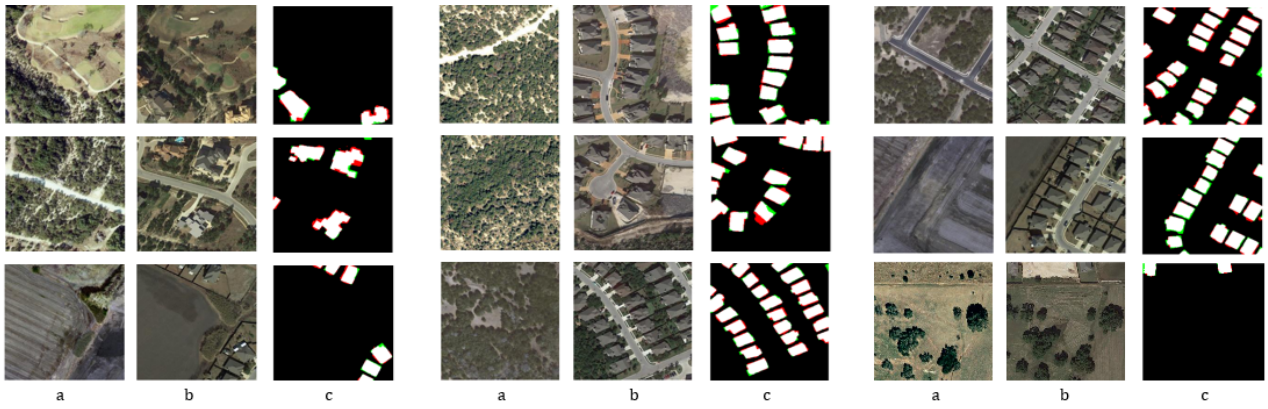

Figure 2. Qualitative results: (a) epoch1 image, (b) epoch2 image, (c) ground truth overlay on the resulting change map. white: TP, red: FP, green: FN

While the proposed framework shows promising results, it also has some limitations. The current evaluation is limited to a single benchmark dataset, and additional experiments are necessary to validate generalizability across diverse geographic regions, imaging conditions, and sensor types. Furthermore, relying on pretrained Vision-Language Models introduces potential domain mismatch challenges, particularly in dense urban areas or under varying illumination, which may affect mask accuracy and downstream performance.

Future work will explore strategies for fine-tuning VLMs on remote sensing imagery to mitigate domain gaps, assess computational efficiency on larger-scale datasets, and extend the framework to other change detection tasks, such as infrastructure monitoring and land cover analysis. Additionally, incorporating temporal sequences and ancillary geospatial information could further improve robustness and enable more comprehensive urban monitoring applications.

References

Bazi, Y. et al., 2021. Vision transformers for remote sensing image analysis: A survey. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Chen, H., Qi, Z., Shi, Z., 2022. Remote Sensing Image Change Detection With Transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14.

Chen, H., Shi, Z., 2020. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sensing*, 12(10), 1662. https://www.mdpi.com/2072-4292/12/10/1662.

Daudt, R. C., Le Saux, B., Boulch, A., Gousseau, Y., 2018. Fully convolutional siamese networks for change detection. *Proceedings of the 2nd IEEE International Workshop on Change Detection*.

Diab, M., Kolokoussis, P., Brovelli, M. A., 2025. Optimizing zero-shot text-based segmentation of remote sensing imagery using SAM and Grounding DINO. *Artificial Intelligence in Geosciences*, 100105.

Dong, S., Wang, L., Du, B., Meng, X., 2024. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208, 53–69.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D. et al., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment Anything. *arXiv preprint arXiv:2304.02643*. https://arxiv.org/abs/2304.02643.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L., 2023. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*. https://arxiv.org/abs/2303.05499.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002.

Peng, D., Zhang, Y., Guan, H., 2019. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sensing*, 11(11), 1382.

Qiu, J., Liu, W., Zhang, H., Li, E., Zhang, L., Li, X., 2024. A Novel Change Detection Method Based on Visual Language from High-Resolution Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

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

Vaswani, A., ..., 2017. Attention is All You Need. *Advances in Neural Information Processing Systems*.

Woo, S., Park, J., Lee, J.-Y., Kweon, I. S., 2018. Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Wu, Q., Osco, L., 2023. samgeo: A Python package for segmenting geospatial data with the Segment Anything Model (SAM). *Journal of Open Source Software*, 8(89), 5663. https://doi.org/10.21105/joss.05663.

Zhang, C., Wang, L., Cheng, S., Li, Y., 2022. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-13.

Zhang, S., Zhang, B., Wu, Y., Zhou, H., Jiang, J., Ma, J., 2024. Segclip: Multimodal visual-language and prompt learning for high-resolution remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*.