References

[1] Pablo F Alcantarilla, Simon Stent, German Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42:1301–1322, 2018.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Shizhen Chang and Pedram Ghamisi. Changes to captions: An attentive network for remote sensing change captioning. *arXiv preprint arXiv:2304.01091*, 2023.

[4] Hongruixuan Chen, Jian Song, Chengxi Han, Junshi Xia, and Naoto Yokoya. Changemamba: Remote sensing change detection with spatio-temporal state space model. *arXiv preprint arXiv:2404.03425*, 2024.

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

[6] Mengjin Dong, Long Xie, Sandhitsu R Das, Jiancong Wang, Laura EM Wisse, Robin DeFlores, David A Wolk, Paul A Yushkevich, Alzheimer's Disease Neuroimaging Initiative, et al. Deepatrophy: Teaching a neural network to detect progressive changes in longitudinal mri of the hippocampal region in alzheimer's disease. *Neuroimage*, 243:118514, 2021.

[7] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019.

[8] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. Neural naturalist: generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101*, 2019.

[9] Maoguo Gong, Tao Zhan, Puzhao Zhang, and Qiguang Miao. Superpixel-based difference representation learning for change detection in multispectral remote sensing images. *IEEE Transactions on Geoscience and Remote sensing*, 55(5):2658–2673, 2017.

[10] Zixin Guo, Tzu-Jui Julius Wang, and Jorma Laaksonen. Clip4idc: Clip for image difference captioning. *arXiv preprint arXiv:2206.00629*, 2022.

[11] Mehrdad Hosseinzadeh and Yang Wang. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2725–2734, 2021.

[12] Qingbao Huang, Yu Liang, Jielong Wei, Yi Cai, Hanyu Liang, Ho-fung Leung, and Qing Li. Image difference captioning with instance-level fine-grained feature representation. *IEEE transactions on multimedia*, 24:2004–2017, 2021.

[13] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*, 2018.

[14] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. Img-diff: Contrastive data synthesis for multimodal large language models. *arXiv preprint arXiv:2408.04594*, 2024.

[15] Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. Agnostic change captioning with cycle consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2095–2104, 2021.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018.

[19] Zhi Li, Siying Cao, Jiakun Deng, Fengyi Wu, Ruilan Wang, Junhai Luo, and Zhenming Peng. Stade-cdnet: Spatial–temporal attention with difference enhancement-based network for remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–17, 2024.

[20] R. Padilla, S. L. Netto, and E. A. B. da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242, 2020.

[21] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019.

[22] Julia Patriarche and Bradley Erickson. A review of the automated detection of change in serial imaging studies of the brain. *Journal of digital imaging*, 17:158–174, 2004.

[23] Hai Phan and Anh Nguyen. Deepface-emd: Re-ranking using patch-wise earth mover's distance improves out-of-distribution face identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20259–20269, 2022.

[24] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 421–429. Springer, 2018.

[25] Ragav Sachdeva and Andrew Zisserman. The change you want to see. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3993–4002, 2023.

[26] 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.

[27] Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-

change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693, 2019.

[28] Ken Sakurada and Takayuki Okatani. Change detection from a street image pair using cnn features and superpixel segmentation. In *British Machine Vision Conference*, 2015.

[29] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[30] Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 574–590. Springer, 2020.

[31] Simon Stent, Riccardo Gherardi, Björn Stenger, and Roberto Cipolla. Detecting change for multi-view, long-term surface inspection. In *BMVC*, pages 127–1, 2015.

[32] Yanjun Sun, Yue Qiu, Mariia Khan, Fumiya Matsuzawa, and Kenji Iwata. The stvchrono dataset: Towards continuous change recognition in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14120, 2024.

[33] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021.

[34] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. Expressing visual relationships via language. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1873–1883, Florence, Italy, July 2019. Association for Computational Linguistics.

[35] Congcong Wang, Wenbin Sun, Deqin Fan, Xiaoding Liu, and Zhi Zhang. Adaptive feature weighted fusion nested u-net with discrete wavelet transform for change detection of high-resolution remote sensing images. *Remote Sensing*, 13(24):4971, 2021.

[36] Junhui Wu, Yun Ye, Yu Chen, and Zhi Weng. Spot the difference by object detection. *arXiv preprint arXiv:1801.01051*, 2018.

[37] Quanfu Xu, Keming Chen, Guangyao Zhou, and Xian Sun. Change capsule network for optical remote sensing image change detection. *Remote Sensing*, 13(14):2646, 2021.

[38] Le Yang, Yiming Chen, Shiji Song, Fan Li, and Gao Huang. Deep siamese networks based change detection with remote sensing images. *Remote Sensing*, 13(17):3394, 2021.

[39] Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3108–3116, 2022.

[40] Zhuo Zheng, Yanfei Zhong, Liangpei Zhang, and Stefano Ermon. Segment any change. *arXiv preprint arXiv:2402.01188*, 2024.

Appendix for: Improving Zero-Shot Object-Level Change Detection by Incorporating Visual Correspondence

A. Upper bound accuracy of correspondence algorithm

Here, we want to estimate the correspondence component. Correspondence algorithm consists of alignment step before using the Hungarian algorithm. By using ground-truth boxes, we can evaluate the maximum accuracy of the matching algorithm.

Experiments To assess the effectiveness of the post-processing method we employ ground-truth boxes directly rather than utilising the change detector's projected box output as the feature extractor's input.

Results The findings presented in Tab. A1 upper bound indicate that our matching method demonstrates strong performance in F_1 score when applied to both the 📹(+100) and T(+99.96) algorithms. A gap persists in the availability of the 🧊(96.50) and 🟩(91.68) datasets. The efficacy of the transformation matrix is limited in certain challenging scenarios involving 🧊 or 🟩. The 🟩 dataset contains numerous artifacts, which hinder the accurate estimation of the transformation matrix.

	Change (F1 Score)				
Model	🟩	📹	🧊	T	🖨
Ground-truth Baseline	91.68	**100**	96.50	99.96	99.94

Table A1. **Correspondence Accuracy Upper Bound.** Using ground truth boxes as input for matching algorithm

B. Features of mean pooling provide more accurate correspondence than cropped images features

The proposed approach offers flexibility in selecting methods for assigning embeddings to predicted boxes. This section evaluates two methodologies for generating embeddings. To identify the optimal method, we conduct a comparative analysis using our fine-tuned model. The effectiveness of each approach is assessed based on the matching score (F1).

Experiments This section analyzes the impact of two embedding assignment methods: mean-pooling and region cropping on the correspondence score. The analysis is conducted based on the methodologies outlined in (Sec. 3.3).

Results We hypothesize that using only cropped images reduces the availability of contextual information surrounding the object, resulting in lower correspondence accuracy. The average feature method consistently outperforms the cropping method across all five datasets, with significant improvements observed in the T and 🖨 datasets. Consequently, we have adopted the average feature technique for all subsequent experiments. Detailed results are presented in Tab. A2.

Model	Average	Crop	Thres	🟩	📹	🧊	T	🖨
Our + ResNet-50		✓	0.25	44.10	56.29	68.10	67.73	62.25
Our + ResNet-50	✓		0.25	46.19	56.94	69.52	85.33	69.53
				+2.09	+0.65	+1.42	+17.60	+7.28

Table A2. Features obtained using the **average** method achieve higher F1 scores compared to those derived from the cropping method. This approach consistently produces reliable results across all datasets, with particularly notable performance on the T and 🖨 datasets.

C. Training hyperparameters

Results We follow the training hyperparameters in (Sec. 3.4). We investigate the impact of training parameters, including the number of epochs and learning rate, on model performance. Training for 500 epochs led to overfitting, reducing zero-shot accuracy on the 📹, 🔷, T, and 🖨 datasets (Tab. A4). Increasing the learning rate from 0.0001 to 0.0005 further degraded accuracy (Tab. A3). Additionally, using a deeper decoder did not improve accuracy (Tab. A5).

					Change (mAP Score)				
Model	LR	CenterNet	DETR	Contrastive	🟩	📹	🔷	T	🖨
Our	0.0005	✓	✓	✓	57.87	47.95	68.20	88.54	60.33
Our	0.0001	✓	✓	✓	**71.77**	**57.00**	**81.07**	**90.02**	**78.84**

Table A3. **Training with different learning rate (LR).** Using different learning rate in training

					Change (mAP Score)				
Model	Epochs	CenterNet	DETR	Contrastive	🟩	📹	🔷	T	🖨
Our	500	✓	✓	✓	**72.15**	54.17	78.64	89.01	78.65
Our	200	✓	✓	✓	71.77	**57.00**	**81.07**	**90.02**	**78.84**

Table A4. **Training with more epochs.** Training the model for 500 epochs decreases accuracy in zero-shot testing on the 📹, 🔷, T, and 🖨 datasets.

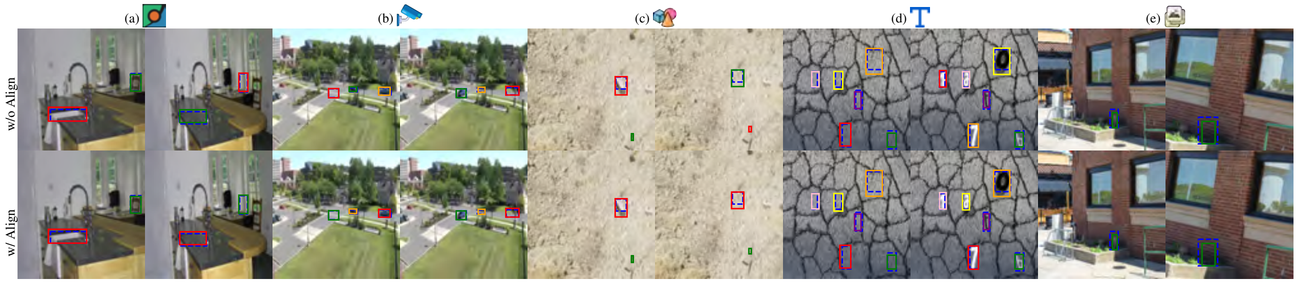

Figure A1. With the significant improvement in the 📹 and T datasets, the alignment stage is a crucial component in increasing correspondence accuracy. The second row's findings demonstrate how the alignment step aids in correcting every case's incorrect matching in the first row. You may view the improvement's specifics in the Tab. 6.

D. Training with a deeper decoder does not enhance model accuracy

In order to find the best change detection architecture, we added more layers to the decoder in this section.
Experiment We used [256, 128, 64] channels for each decoder layer in the prior configuration. We add two further layers with 32 and 8 channels, respectively, in this configuration.
Results The outcomes of employing deeper decoder layers are displayed in Tab. A5. The findings demonstrate that the final accuracy decreases with the number of decoder layers.

Change (mAP Score)									
Model	Epochs	CenterNet	DETR	Contrastive	🟩	📹	🔶	T	🖨
Our	200	✓	✓	✓	51.76	49.70	63.95	87.57	54.39
Our	200	✓	✓	✓	**71.77**	**57.00**	**81.07**	**90.02**	**78.84**

Table A5. **Training with a deeper decoder** does not enhance model accuracy

E. The alignment stage plays a crucial role in the success of the matching algorithm

The qualitative results Fig. A1 we present in this section demonstrate how well our alignment stage worked to enhance the matched pairs of modifications displayed in the Tab. 6.

F. Correspondence

We present qualitative results in this part that contrast our model with CYWS model in terms of matching qualitative. According to the qualitative results, our model outperforms CYWS model in the matching score, as indicated by the Tab. 7. See qualitative results in Fig. A2

G. Reduce false positive predicted box in no-change case

The output from CYWS model in the default situations is shown in the first row of Fig. A3. The outcomes of our post-processing procedure are shown in the row that follows.

H. Additional qualitative results

In this part, we present further qualitative comparison findings between our fine-tuned model and CYWS [25] model following the use of a detection threshold of 0.25 and a post-processing technique. CYWS findings are shown in the first row, while the results of our model are shown in the second row. For qualitative results, see Fig. A4.

I. Number of predicted box after applying detection threshold

For both the ground-truth and our refined model with different thresholds, CYWS, we display the average number of boxes per image. You can view the detail in the Tab. A6

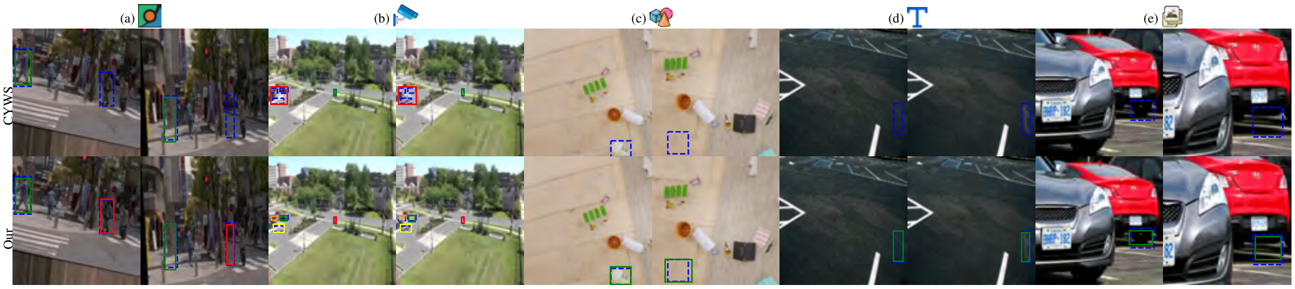

Figure A2. CYWS model, as seen in (a), (b), (c), (d) and (e), is unable to identify every difference between two images. Conversely, our model is able to identify every change in the two images. CYWS model can only identify one change for the entire region in the 📹 example, where three changes appear at nearly the same location. Our model, on the other hand, can identify each of the three changes independently. We hypothesise that the model learns the number of changes implicitly based on information gleaned from the contrastive matching loss. Check Tab. 7 for quantitative results.

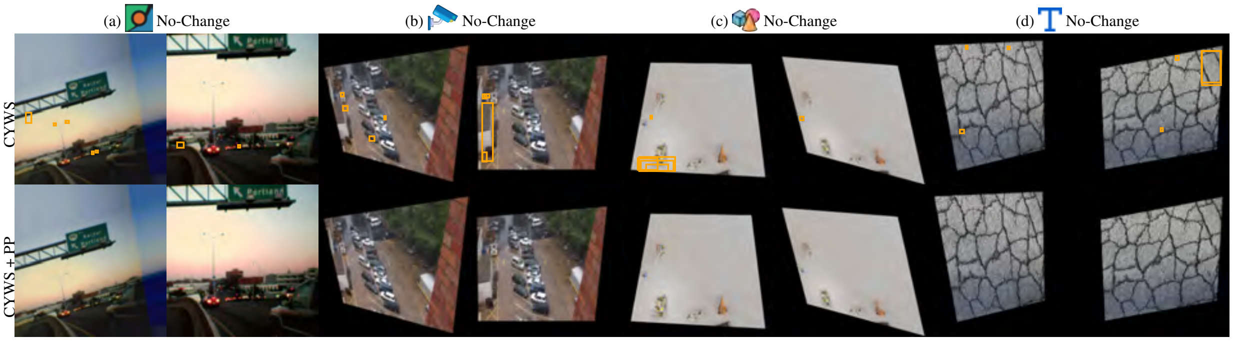

Figure A3. In no-change scenarios, our post-processing approach reduces false positive predicted boxes.

Change						
Avg Predicted Box Per Image						
Model	Thres	🟩	📹	🧊	T	🖨
Ground-Truth	n/a	**1.93**	**5.85**	**1.80**	**1.10**	**1.0**
CYWS	n/a	100	100	100	100	100
Our	n/a	100	100	100	100	100
CYWS	0.1	3.63	6.55	2.27	2.55	3.54
Our	0.1	3.21	7.23	2.37	2.25	2.64
CYWS	0.2	1.75	4.38	1.95	1.23	1.17
Our	0.2	1.85	4.90	1.96	1.19	1.14
CYWS	0.3	0.98	2.81	1.75	0.80	0.54
Our	0.3	1.20	3.13	1.79	0.84	0.67
CYWS	0.4	0.55	1.38	1.50	0.49	0.26
Our	0.4	0.70	1.48	1.56	0.63	0.37
CYWS	0.5	0.29	0.59	1.08	0.42	0.10
Our	0.5	0.45	0.60	1.18	0.44	0.18

Table A6. **Average Predicted Box Per Image for Change with Different Thresholds.** Evaluate the influence of detection threshold on the number of predicted boxes per image in change case with CYWS model and our fineturned model

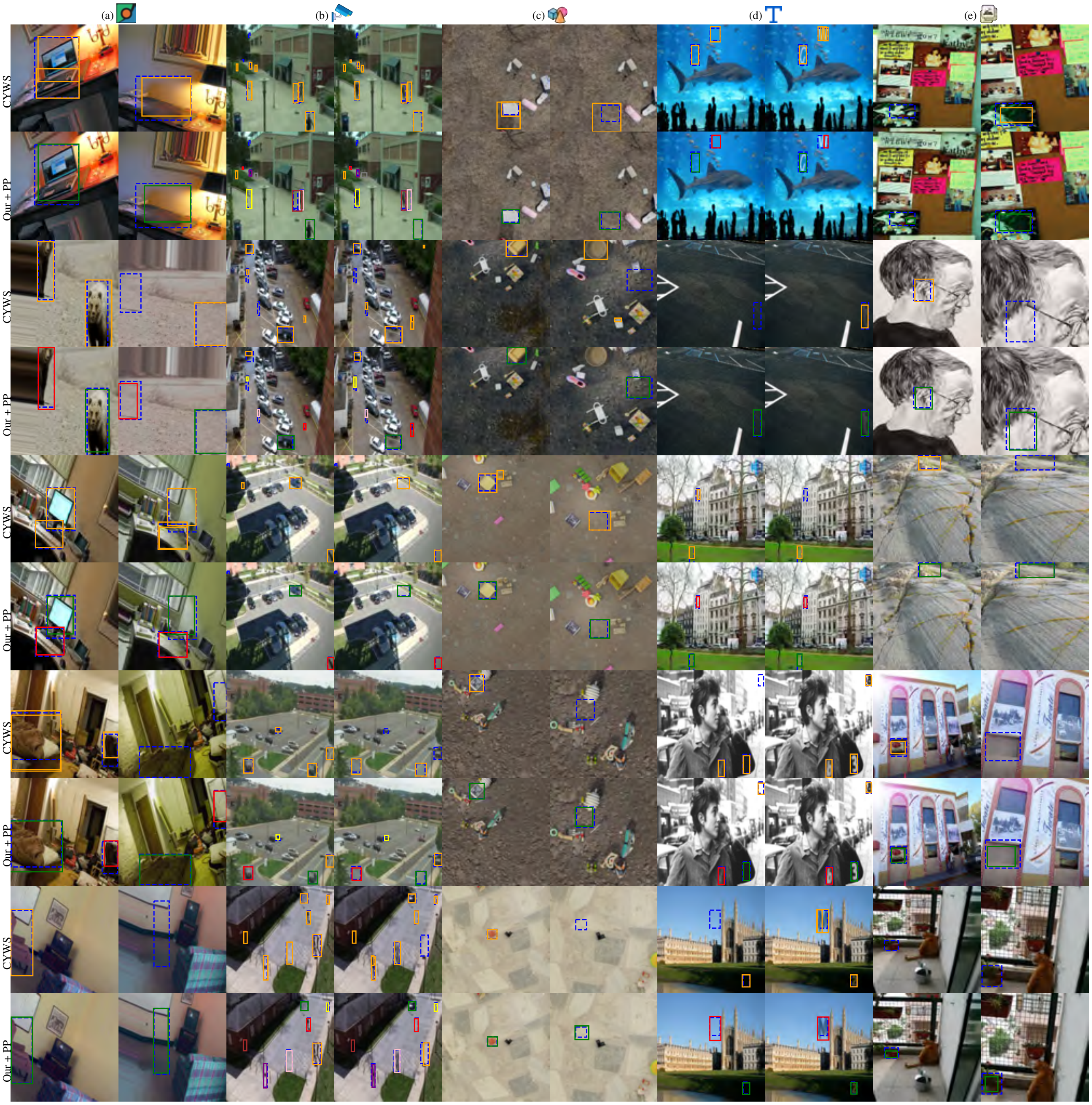

Figure A4. When comparing our model's change detection output to that of CYWS model, it is evident that our contrastive matching loss enhances the model's accuracy. Additionally, our post-processing technique can apply in many situations with multiple modifications

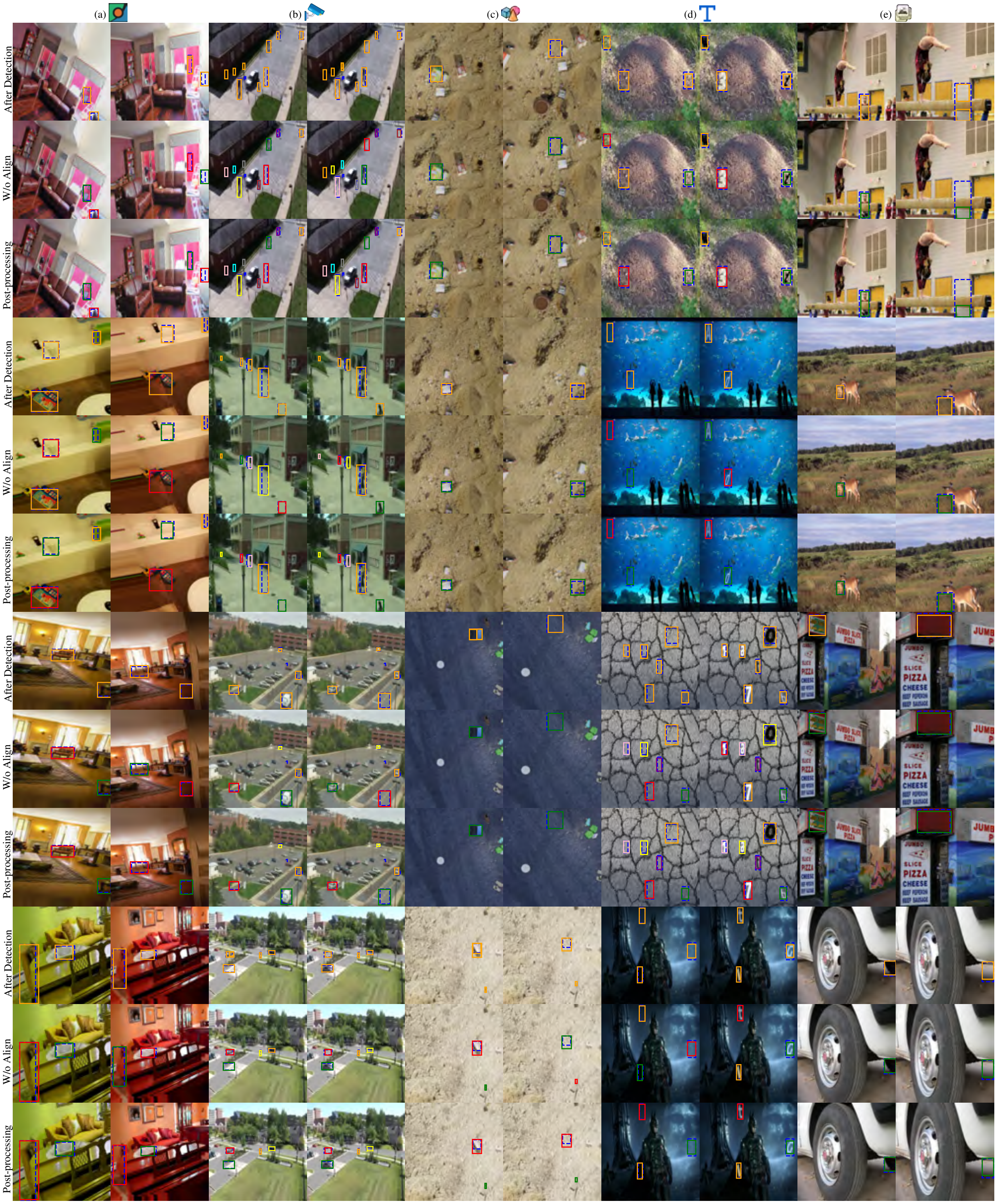

Figure A5. Contrasting the results following change detection and using our post-processing both with and without the alignment step. Evaluation of the findings in , , , T, and demonstrates the significance of the alignment stage