

SUPERVISING REMOTE SENSING CHANGE DETECTION MODELS WITH 3D SURFACE SEMANTICS

Isaac Corley Peyman Najafirad

Secure Artificial Intelligence Laboratory for Autonomy (AILA)
The University of Texas at San Antonio, Texas, USA
`{isaac.corley, peyman.najafirad}@utsa.edu`

ABSTRACT

Remote sensing change detection, identifying changes between scenes of the same location, is an active area of research with a broad range of applications. Recent advances in multimodal self-supervised pretraining have resulted in state-of-the-art methods which surpass vision models trained solely on optical imagery. In the remote sensing field, there is a wealth of overlapping 2D and 3D modalities which can be exploited to supervise representation learning in vision models. In this paper we propose Contrastive Surface-Image Pretraining (CSIP) for joint learning using optical RGB and above ground level (AGL) map pairs. We then evaluate these pretrained models on several building segmentation and change detection datasets to show that our method does, in fact, extract features relevant to downstream applications where natural and artificial surface information is relevant.¹

Index Terms— self-supervised learning, contrastive learning, remote sensing, above ground level maps

1. INTRODUCTION

Self-supervised learning methods have recently become the de facto pretraining method due to their conceptual simplicity and ability to make use of large unlabeled datasets. Due to the wealth of multiview and multimodal data, self-supervision methods have shown dramatic improvements in the remote sensing and earth observation fields. Manas et al. [2] showed improved performance while pretraining using Sentinel-2 satellite imagery of the same location from different seasons as multiple views. Heidler et al. [3] experimented with pretraining on an audiovisual dataset of aerial image and audio pairs and showed efficient learning and improved downstream performance compared to transfer learning from

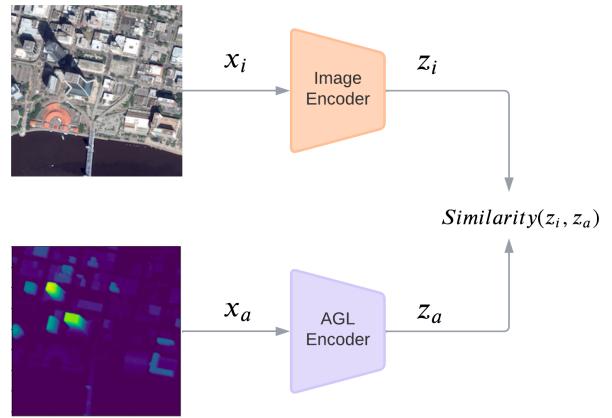


Fig. 1: Our Contrastive Surface-Image Pre-training (CSIP) model learns the combined latent distribution, 3D surface semantics while performing similarity matching of optical RGB and AGL map pairings.

ImageNet as well as other state-of-the-art self-supervised learning methods.

Change detection is an important problem in remote sensing. The development of large scale labeled change detection datasets has allowed for rapid humanitarian assistance and disaster response (HADR) applications for natural disaster relief. Improvements in performance of computer vision models for change detection is critical for planning and accurate detection of areas requiring assistance from the HADR organizations and communities.

The combination of self-supervised learning methods for change detection has been explored recently using mainly using pre and post imagery as multiple views of the same location. Leenstra et al. [4] explored using a large dataset of Sentinel-2 imagery sampled over the same geographic location at different times to train a model to predict whether a patch spatially overlaps another patch as a pretext task. Chen et al. [5] used pre and post imagery to train using a contrastive

* Corresponding author

¹Experimental code is available at <https://github.com/isaaccorley/contrastive-surface-image-pretraining> and model checkpoints are made available in the TorchGeo [1] library at <https://github.com/microsoft/torchgeo>

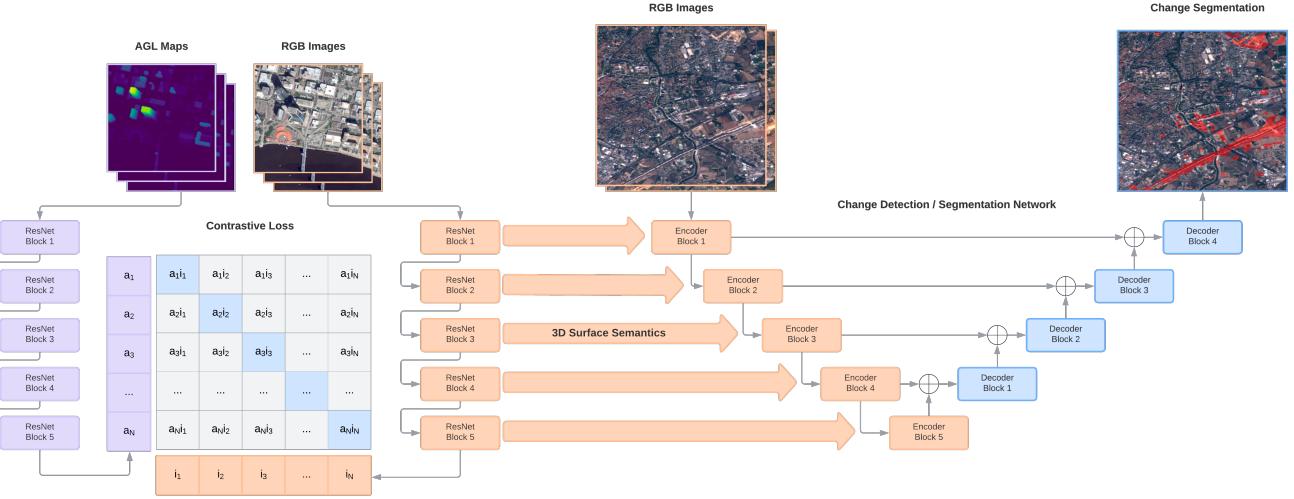


Fig. 2: An overview of the CSIP architecture we propose. Using optical RGB and AGL map pairs, the multimodal self-supervised pretraining encoders contrastively learn to extract features relevant to 3D surfaces, after which a downstream model provides change detection segmentations based on the learned 3D surface semantics.

framework. However, to our knowledge, no other methods have contrastively trained using overlapping remotely sensed imagery and 3D surface maps as a pretext task.

Our contributions can be described as following:

- We propose a contrastive multimodal framework for pretraining deep neural networks to learn representations of remotely sensed imagery which extract features relevant to 3D surface information.
- We perform thorough investigation through extensive experiments on downstream tasks to validate our hypothesis that our pretraining framework does in fact improve performance on tasks which require surface or height extraction to perform image segmentation.

2. METHODS

Inspired by the Contrastive Language-Image Pretraining (CLIP) [6] framework, we propose Contrastive Surface-Image Pretraining (CSIP) framework which utilizes dual encoders trained using a contrastive loss to encode imagery and above ground level (AGL) maps into similar latent space representations. Given an image x_i and AGL map x_a pair, an image encoder network f_i , and AGL encoder network f_a , we encode the image and AGL map into feature vectors z_i and z_a , respectively. Each encoder network is composed of a backbone and a projection multi-layer perceptron (MLP) head. We then minimize the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss [7] with a learned temperature parameter τ which is a variant of the InfoNCE contrastive loss

[8] and is described in Equation 1. By minimizing the NT-Xent loss we seek to maximize the cosine similarity between latent RGB and AGL vectors from the same pair relative to other latent vectors in a minibatch. Within a minibatch, we utilize other randomly sampled pairs as negative samples. An overview of the CSIP architecture is provided in Figure 1.

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}, \quad (1)$$

The end result of our framework is to take the pretrained image encoder network, f_i , and use as a backbone or encoder in segmentation or change detection networks. An overview of this is provided in Figure 2.

3. EXPERIMENTS

To assess the value of the feature representations learned during pretraining, we compare against backbones pretrained on ImageNet [9] on several downstream tasks. In all experiments, we freeze the pretrained backbones and only fine-tune the model heads for each downstream task. In all downstream experiments we use a batch size of 16, 25 number of epochs, AdamW optimizer [10] with a learning rate of $\alpha = 0.001$, patch sizes of 512x512, and a training/validation split of 0.2 for datasets which did not have a predefined validation split. For measuring downstream change detection and segmentation performance we evaluate using the mean Intersection-over-Union (mIoU), F1-score, and Average Accuracy metrics.



Fig. 3: Visual comparisons of ImageNet and CSIP weight initializations to ground truth on randomly sampled test set predictions.

3.1. Contrastive Pretraining

During pretraining, we use the Overhead Geopose Challenge dataset [11] which is an extension to the Urban Semantic 3D (US3D) dataset [12] used in the 2019 IEEE GRSS Data Fusion Contest (DFC19) [13]. The dataset is composed of pairs of 2048x2048 2D orthorectified RGB Maxar WorldView satellite imagery and 3D above ground level (AGL) maps. Furthermore, the dataset consists of oblique or significantly off-nadir imagery which adds additional complexity during pretraining. We pretrain separate ResNet-18 [14] RGB and AGL encoders using the contrastive framework proposed in Section 2 for 500 epochs. We then utilize the RGB encoder as a backbone in the following experiments.

3.2. Change Detection

We perform experiments for change detection on the following datasets:

OSCD The Onera Satellite Change Detection (OSCD) [15] dataset is a dataset for building change detection. The

dataset contains 24 multispectral image pairs of various sizes taken by the Sentinel-2 satellites [16] along with binary change masks representing building change between pre and post imagery.

xBD The xBD [17] is a dataset for change detection and HADR applications utilized in the xView2 Challenge. The labels consist of building polygons categorized into 4 damage severities in imagery from before and after natural disasters.

S2Looking The S2Looking [18] dataset is a building change detection dataset consisting of 5,000 1024x1024 0.5-0.8 m resolution image pairs taken by the Gaofen, SuperView, and BeiJing-2 satellites. The dataset contains imagery with significantly off-nadir angles of rural locations along with labels for both newly built and demolished buildings.

During training we fine-tune the Fully-Convolutional Siamese Difference (FC-Siam-Diff) [19] architecture which is based on the U-Net architecture with the exception of taking the difference between skip connections after passing

Dataset	Weight Init.	mIoU	F1	Acc.
OSCD	ImageNet	0.302	0.674	0.643
	CSIP (ours)	0.454	0.692	0.707
xView2	ImageNet	0.315	0.526	0.480
	CSIP (ours)	0.367	0.529	0.533
S2Looking	ImageNet	0.415	0.602	0.609
	CSIP (ours)	0.490	0.568	0.657

Table 1: Experimental results on three benchmark change detection datasets comparing our CSIP method with ImageNet pretrained weights in the encoder of FCSiamDiff. Best results are marked in bold.

both pre and post imagery through the U-Net encoder. To compare the learned semantic surface features, we freeze the encoder of the network and only fine-tune the decoder on each dataset.

3.3. Semantic Segmentation

To compare performance on tasks not focused solely on building change or mapping, we conduct experiments on the LandCover.ai land cover semantic segmentation dataset [20]. The dataset consists of 10,674 512x512 RGB aerial images with 0.25-0.55 m spatial resolution. The imagery is annotated with semantic masks of 4 land cover categories. For benchmarking purposes, we utilize the original splits provided with the dataset. During training, we fine-tune the U-Net [21] architecture with the pretrained ResNet backbone as the encoder. Similarly to the change detection experiments, we freeze the encoder of the network and only fine-tune the decoder on the dataset.

Weight Init.	mIoU	F1	Acc.
ImageNet	0.833	0.865	0.860
CSIP (ours)	0.766	0.843	0.809

Table 2: Experimental results on the LandCover.ai land cover semantic segmentation dataset comparing our CSIP method with ImageNet pretrained weights in the encoder of U-Net. Best results are marked in bold.

4. DISCUSSION AND FUTURE WORK

The experimental results in Table 1 provide evidence that pretraining using the CSIP framework does in fact improve downstream performance for building change detection datasets in comparison to ImageNet initialization. Additionally, the results on the LandCover.ai segmentation dataset display a limitation to the framework where surface and

height information may not be as important and ImageNet feature extraction has a distinct advantage where texture and color is more important in the delineating between land cover categories. Visual comparisons on randomly sampled test set predictions is provided for all datasets in Figure 3.

While our experiments only explore pretraining of ResNet backbones, we note that investigation of pretraining Vision Transformers (ViT) [22] would be a direct next step in this research. ViTs have shown improved performance in segmentation [23] and change detection [24] applications in comparison to fully-convolutional based architectures. Furthermore, pretraining of ViT backbones with self-supervision [25] has recently shown promising results.

5. CONCLUSION

In this paper, we propose a contrastive framework for supervising optical RGB image feature extraction of 3D surface information using AGL maps. We then presented experimental results supporting the hypothesis that pretraining vision models with self-supervision from 3D surface maps would improve downstream performance for building change detection which has direct applications in HADR efforts.

6. REFERENCES

- [1] Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee, “Torchgeo: deep learning with geospatial data,” *arXiv preprint arXiv:2111.08872*, 2021.
- [2] Oscar Mañas, Alexandre Lacoste, Xavier Giro-i Nieto, David Vazquez, and Pau Rodriguez, “Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.
- [3] Konrad Heidler, Lichao Mou, Di Hu, Pu Jin, Guangyao Li, Chuang Gan, Ji-Rong Wen, and Xiao Xiang Zhu, “Self-supervised audiovisual representation learning for remote sensing data,” *arXiv preprint arXiv:2108.00688*, 2021.
- [4] Marrit Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia, “Self-supervised pre-training enhances change detection in sentinel-2 imagery,” in *International Conference on Pattern Recognition*. Springer, 2021, pp. 578–590.
- [5] Yuxing Chen and Lorenzo Bruzzone, “Self-supervised remote sensing images change detection at pixel-level,” *arXiv preprint arXiv:2105.08501*, 2021.
- [6] 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,” *arXiv preprint arXiv:2103.00020*, 2021.
- [7] Kihyuk Sohn, “Improved deep metric learning with multi-class n-pair loss objective,” *Advances in neural information processing systems*, vol. 29, 2016.
- [8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [10] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [11] Gordon Christie, Rodrigo Rene Rai Munoz Abujder, Kevin Foster, Shea Hagstrom, Gregory D Hager, and Myron Z Brown, “Learning geocentric object pose in oblique monocular images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14512–14520.
- [12] Kevin Foster, Gordon Christie, and Myron Brown, “Urban semantic 3d dataset,” 2020.
- [13] Bertrand Le Saux, Naoto Yokoya, Ronny Hänsch, and Myron Brown, “Data fusion contest 2019 (dfc2019),” 2019.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [15] Rodrigo Caye Daudt, Bertr Le Saux, Alexandre Boulch, and Yann Gousseau, “Urban change detection for multispectral earth observation using convolutional neural networks,” in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 2115–2118.
- [16] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al., “Sentinel-2: Esa’s optical high-resolution mission for gmes operational services,” *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.
- [17] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hostfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston, “Creating xbd: A dataset for assessing building damage from satellite imagery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 10–17.
- [18] Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang, “S2looking: A satellite side-looking dataset for building change detection,” *Remote Sensing*, vol. 13, no. 24, pp. 5094, 2021.
- [19] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch, “Fully convolutional siamese networks for change detection,” in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 4063–4067.
- [20] Adrian Boguszewski, Dominik Batorski, Natalia Ziemb-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka, “Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2021, pp. 1102–1110.
- [21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [23] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun, “Vision transformers for dense prediction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12179–12188.
- [24] Hao Chen, Zipeng Qi, and Zhenwei Shi, “Remote sensing image change detection with transformers,” *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [25] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9650–9660.