Self-Supervised Pretraining for Scene Change Detection

Vijaya Raghavan T. Ramkumar, Prashant Bhat, Elahe Arani, Bahram Zonooz
Advanced Research Lab, NavInfo Europe, The Netherlands
{vijaya.ramkumar, prashant.bhat, elahe.arani}@navinfo.eu, bahram.zonooz@gmail.com

Abstract

High Definition (HD) maps provide highly accurate details of the surrounding environment that aids in the precise localization of autonomous vehicles. To provide the most recent information, these HD maps must remain up-to-date with the changes present in the real world. Scene Change Detection (SCD) is a critical perception task that helps keep these maps updated by identifying the changes of the scene captured at different time instances. Deep Neural Network (DNNs) based SCD methods hinge on the availability of large-scale labeled images that are expensive to obtain. Therefore, current SCD methods depend heavily on transfer learning from large ImageNet datasets. However, they induce domain shift which results in a drop in change detection performance. To address these challenges, we propose a novel self-supervised pretraining method for the SCD called D-SSCD that learns temporal-consistent representations between the pair of images. The D-SSCD uses absolute feature differencing to learn distinctive representations belonging to the changed region directly from unlabeled pairs of images. Our experimental results on the VL-CMU-CD and Panoramic change detection datasets demonstrate the effectiveness of the proposed method. Compared to the widely used ImageNet pretraining strategy that uses more than a million additional labeled images, D-SSCD can match or surpass it without using any additional data. Our results also demonstrate the robustness of D-SSCD to natural corruptions, out-of-distribution generalization, and its superior performance in limited label scenarios.

1 Introduction

Most autonomous driving systems require HD maps to help the vehicle localize itself more accurately in the surrounding environment. However, the physical environment is constantly susceptible to semi-static changes as features such as traffic signs, construction sites, and lane markings are constantly added or removed across time. HD maps, therefore, have to be continuously updated with these changes to provide the vehicle with reliable information to ensure safe and robust navigation. However, conventional HD map generation methods using specialized vehicles cannot reliably keep the maps up-to-date because of the low traversal frequencies. Scene change detection (SCD) is a critical perception task that helps alleviate the problem of efficient HD map maintenance and map update by identifying these semi-static changes from images of the scene captured at different times. It also plays a crucial role in other real-world applications such as ecosystem monitoring, urban expansion, and damage assessment.

SCD is a low-likelihood problem where the changed region is smaller than the unchanged region with uncertainty in change location and direction (11). Moreover, the changes that need to be detected depend on the nature of the application and are classified into semantic changes and noisy changes (3).

[1]The official code is available at: https://github.com/NeurAI-Lab/D-SSCD

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

The structural changes caused by the appearance or disappearance of objects present in a scene are considered as semantic changes, while the changes induced by the radiometric (illumination, shadows, seasonal changes) and geometric variations (viewpoint differences caused by camera rotation) are considered as noisy changes (3) (6) (1). A critical challenge in SCD is that these noisy changes are entangled with the semantic changes that alter the appearance of an image, thus degrading the change detection performance (1).

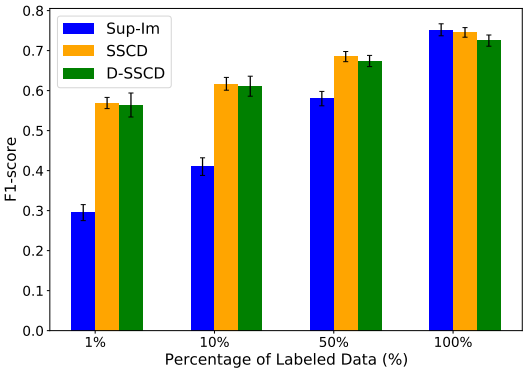

Figure 1: Performance of supervised and the proposed self-supervised pretraining (D-SSCD) evaluated using DR-TANet (2) on VL-CMU-CD dataset under limited label scenario.

Previous studies based on Deep Neural Networks (DNNs) have proposed to extract multi-level feature representations from the input images to improve the performance of SCD against noisy changes (1; 3; 9; 10). However, the success of these state-of-the-art methods (10; 2) often hinges on a large quantity of annotated data. Large-scale collection and annotation of SCD datasets are difficult to obtain as they are labor-intensive and time-consuming. For instance, on average, it takes around 20 minutes and 156 minutes to annotate a single pair of images in the panoramic change detection (PCD) (4) and panoramic semantic change detection (PSCD) (8) dataset, respectively. Therefore, the availability of large-scale labeled datasets for SCD is still scarce, and deficient (11). To address the dependency on labeled data, various SCD approaches initially pretrain their models on the large-scale ImageNet (27) in a supervised manner and later finetune with large amounts of pixel-level annotations on domain-specific dataset (1), (2). However, there still exists the problem of domain shift as the distribution of the ImageNet data widely differs from that of SCD datasets. This domain shift leads to the degradation of change detection performance in SCD methods.

To attenuate the reliance of SCD models on a large amount of dense pixel-level annotations and the additional large-scale labeled ImageNet data, we hypothesize a novel self-supervised pretraining methods that utilizes unlabeled data to learn representations useful for SCD. With extensive experiments, we show that our proposed self-supervised pretraining methods demonstrate remarkable performance compared to ImageNet pretraining under limited labels scenario as seen in Figure 1. To the best of our knowledge, this is the first work on SCD that relaxes the requirement of large-scale annotated datasets and the need to pretrain on additional large-scale labeled data. Our contribution can be summarized as follows:

1. We propose a novel self-supervised pretraining method called D-SSCD that learns temporal-consistent representations relevant for scene change detection.
2. We evaluated the proposed methods on two challenging SCD datasets. Our proposed method surpasses the widely used ImageNet pretraining without any additional data.
3. Current scene change detection models are vulnerable to severe performance impairments on images with natural corruptions, and the proposed self-supervised pretraining significantly enhances the robustness of the model to natural corruptions.
4. The effectiveness of the proposed self-supervised pretraining under limited labels and generalization to out-of-distribution data is verified.

2 Related Work

Self-supervised Representation Learning. Contrastive learning has recently gained popularity because of their ability to learn useful representations from the unlabeled data (14; 17; 15; 19). InstDisc (18) proposed instance discrimination as a pretext task to learn a good feature representation by capturing the apparent similarity among instances. MoCo (19) proposed an idea of instance discrimination by utilizing momentum contrast and uses a queue-based dictionary for efficient sampling of negative samples. However, the model was able to discriminate the positive samples

with ease because of their weaker choice of data augmentations. SimCLR (14) overcomes this by exploring various augmentations in an end-to-end training network and uses a bigger batch size to collect negative samples instead of separate memory banks or dictionary queues. However, this approach is limited by large GPU memory size as the availability of negative samples is coupled with a bigger batch size. Moreover, this method suffers from trivial constant solutions wherein both the encoders learn similar representations from the input images. BYOL (20), SimSiam (21) avoid the problem of trivial solution and requirement of large negative samples by introducing asymmetric network architecture using 'predictor' network and asymmetric parameter updates using momentum encoder and stop-gradient. Unlike BYOL and SimSiam, where asymmetric network or parameter updates are required to avoid trivial solutions, Barlow Twins (13) uses an intuitive objective function that uses a cross-correlation matrix to maximize the correlation between the distorted views of samples while minimizing the redundancy between the components of these vectors. Owing to its advantages, we utilize the Barlow twins loss function to realize the goal.

Scene Change Detection (SCD). Recently, Deep Convolutional Neural Networks (CNNs) have demonstrated remarkable performance in SCD tasks when compared to the traditional change detection methods (1; 3; 9; 10; 8). Alcantarilla et al. (3) proposed a change detection method called CDNet that utilizes CNN to extract dense geometry and accurate registration to warp images from different times for the change detection. Guo et al. (1) proposed to learn the discriminative features with the customized feature distance metrics. Moreover, they proposed a threshold contrastive loss function to tackle significant viewpoint differences present in the input image pairs. Sakurada et al. (7), and (12) utilized dense optical flow and CNNs to model the spatial correspondences between images to minimize the noise due to significant viewpoint differences. Furthermore, Sakurada et al. (8) also proposed a method to capture the multi-scale feature information using hierarchically dense connections for semantic change detection. DR-TANet (2) proposes a lightweight network that utilizes a temporal self-attention mechanism to enhance the feature correlation between the two temporal images.

Despite the advances in SCD, all these methods hinge heavily on the availability of large-scale manually annotated datasets that are hard to obtain. When the labeled data is limited, they depend on transfer learning from models pretrained in a supervised manner on some other big datasets such as ImageNet (27). Yet, transfer learning from the ImageNet pretraining models induces domain shift and reduces the flexibility to use encoder architecture different from the one it is pretrained on. This work proposes self-supervised pretraining methods that utilize in-distribution unlabeled data to learn representations relevant to SCD. Moreover, our proposed approach is simple and can be easily adopted to any change detection methods.

3 Methodology

Inspired by the recent improvements in self-supervised representation learning, we propose novel pretraining method for the SCD task that exploits temporal consistency between the two images. At first, we employ self-supervised pretraining method based on Barlow twins (13) called SSCD to over come the dependency of SCD methods to large labeled data. We further propose a novel feature differencing based pretraining approach to learn better representations for the changed regions directly. These pretraining methods are described in detail in Sections 3.1 and 3.2, and the SCD algorithm used to evaluate the proposed pretraining methods is briefly presented in Section 3.3.

3.1 Self-supervised pretraining for Scene Change Detection (SSCD)

SCD aims to identify the changed region between the image pairs captured at different times. To achieve this, the alignment between the radiometric space of the image pairs and the low-level features of the change detection network is required. This alignment is challenging as the changed regions are easily affected by noisy changes caused due to seasonal variations and perspective differences. To facilitate the alignment of low-level features between two images, we employ a self-supervised pretraining method that utilizes Barlow twins (13) objective function that implicitly minimize the differences between the image pairs in the feature space by maximizing the cross-correlation of unchanged regions. The reason for choosing this objective function is due to its competitive performance against contrastive methods (14; 19) and its ability to learn robust representations without the requirement for large number of negative samples.

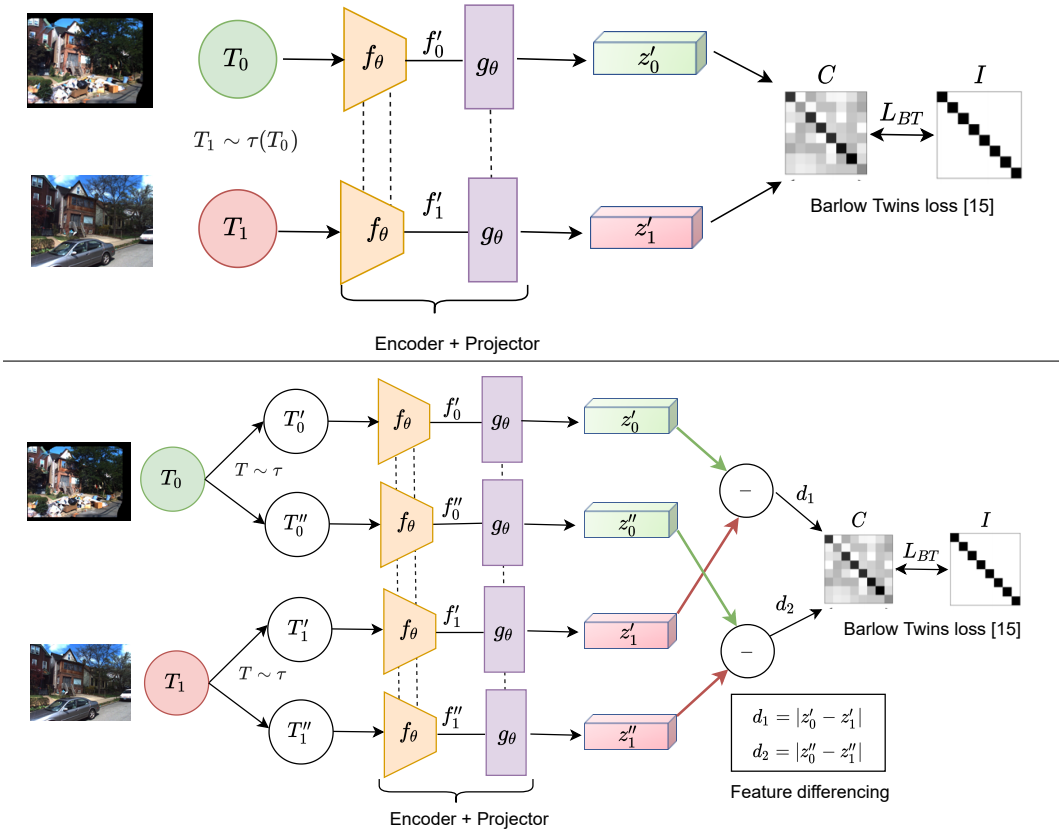

Figure 2: Schematics of proposed self-supervised pretraining methods for scene change detection. Top: **SSCD** learns representation of the unchanged regions by maximizing cross-correlation between two images in the feature space. Bottom: **D-SSCD** uses absolute feature differencing to learn the representation of the changed region directly. Both models are trained in a self-supervised manner using Barlow twins objective function (13).

Although the image pair captured at a different instances of time (T_0 & T_1) are semantically different from each other, they both represent the same scene at two different times. Therefore, we consider T_0 as an augmented version of T_1 image and vice versa. These input pairs are fed into the Siamese encoder (f_θ) with shared parameters θ producing feature vectors f'_0 and f'_1. Then, a non-linear projection (g_θ) is applied over the encoded feature vectors to get representations z'_0 and z'_1 (Figure 2, Top plot). The model is trained in a self-supervised manner using the objective function (13) as follow,

$$L_{BT} \triangleq \underbrace{\sum_i (1 - C_{ii}^2)}_{\text{Invariance term}} + \underbrace{\lambda \sum_i \sum_{j \neq i} C_{ij}^2}_{\text{Redundancy reduction term}} \tag{1}$$

$$C_{ij} \triangleq \frac{\sum_b (z'_0)_{b,i} (z'_1)_{b,j}}{\sqrt{\sum_b ((z'_0)_{b,i})^2} \sqrt{\sum_b ((z'_1)_{b,j})^2}} \tag{2}$$

where λ is a trade-off constant, C is the cross-correlation matrix calculated between the representations of the input image pairs (z'_0 and z'_1) along the batch samples b and i, j index the vector dimension of the network outputs. This objective function consists of two components: (1) the invariance term that makes the representations of the input image pair (T_0, T_1) invariant to the presence of noisy changes (e.g., seasonal variations) by maximizing the diagonal components of the cross-correlation matrix close to identity matrix. (2) the redundancy reduction term tries to decorrelate the off-diagonal components of the cross-correlation matrix and thus, aligning the representations of the input image pairs to be similar. Therefore, the SSCD model learns temporal-consistent representations that are useful for the downstream task of scene change detection.

3.2 Differencing based Self-supervised pretraining for Scene Change Detection (D-SSCD)

Intuitively, maximizing the correlation between two images may affect the discrimination ability of the model in the downstream SCD task, because the representations of the two images along with that of the changed regions is forced to be closer together. Therefore, in contrast to SSCD, where we maximize the correlation between the image pair (T_0, T_1), we propose a differencing based self-supervised pretraining called D-SSCD that maximizes the correlation of the changed regions to learn distinctive representations that are vital for SCD.

D-SSCD gets an image pairs (T_0, T_1) from different time instances as inputs. Random transformations such as color distortions and Gaussian blur are applied to this input image pair to obtain two pairs of augmented images ($T_0 \rightarrow T_0'$, T_0''; $T_1 \rightarrow T_1'$, T_1''). These augmented pairs are passed into the Siamese encoder (f_θ) and projection head (g_θ) to output the corresponding feature representations. The model parameters (θ) are shared. To learn the representation of the changed features between the pair of images, absolute feature differencing is applied over the projection outputs;

$$\begin{aligned} d_1 &= |g(f(\mathbf{T_0'})) - g(f(\mathbf{T_1'}))| \\ d_2 &= |g(f(\mathbf{T_0''})) - g(f(\mathbf{T_1''}))| \end{aligned} \tag{3}$$

Then, Barlow twins objective function (Eq 1) is applied on the difference representations d_1 and d_2 to maximize the cross-correlation of the changed features. In this way, the model will pursue to learn the non-redundant information about the relevant changes that occur between the image pairs. After the pretraining step, the parameters of the encoder f_θ are transferred to the downstream task of change detection.

3.3 SCD Method

We evaluate the performance of self-supervised pretrained model by finetuning it using exsisting SCD method. DR-TANet (2) is selected because of its ability to achieve state-of-the-art results on SCD datasets. It employs an encoder-decoder architecture that incorporates a temporal attention module to exploit the similarity and dependency of feature maps at two temporal channels. Additional details and parameter settings used for finetuning were mentioned in the Section A.3 in Appendix.

4 Experiments

4.1 SCD Datasets

To train and validate the proposed framework, we considered two SCD datasets subjected to noisy changes such as illumination, shadows, seasonal variations, and camera viewpoint differences.

VL-CMU-CD dataset (3): It consists of 152 perspective image sequences taken at different time instances. Each image sequence contains approximately nine pairs of softly co-registered images taken at different times. Therefore, 1362 RGB image pairs of 1024×768 are generated with their manually labeled pixel-level change masks. This dataset portrays the typical macroscopic changes that occur in an urban scenario. The changes in this dataset are not detailed clearly as the images are predominantly affected by noisy changes.

Panoramic Change Detection (PCD) dataset (4): It contains two subsets of data, namely 'TSUNAMI' and 'GSV'. Each subset has 100 pairs of non-registered panoramic images (224×1024 pixels) along with the manually labeled change masks. TSUNAMI subset contains image pairs representing the aftermath of tsunami-affected areas in Japan, whereas the GSV subset contains image pairs belonging to Google street view. Compared to VL-CMU-CD dataset, the image pairs in this dataset are less affected by noisy changes and contain large view-point differences.

In both the datasets, the structural changes such as the emergence/vanishing of buildings and cars are considered relevant, and the noisy changes are deemed irrelevant and excluded from the ground truth change map.

Table 1: Performance (F1-score) of DR-TANet model trained on VL-CMU-CD and PCD datasets using different pretraining methods.

Methods	VL-CMU-CD	PCD Dataset		
		Tsunami	GSV	Average
Rand Init	$0.708_{\pm 0.051}$	$0.634_{\pm 0.031}$	$0.407_{\pm 0.021}$	$0.535_{\pm 0.024}$
Sup-Im	$\mathbf{0.752}_{\pm 0.015}$	$0.687_{\pm 0.013}$	$0.465_{\pm 0.012}$	$0.576_{\pm 0.012}$
SSCD	$0.745_{\pm 0.012}$	$0.709_{\pm 0.018}$	$0.456_{\pm 0.022}$	$0.583_{\pm 0.021}$
D-SSCD	$0.725_{\pm 0.014}$	$\mathbf{0.712}_{\pm 0.014}$	$\mathbf{0.558}_{\pm 0.019}$	$\mathbf{0.642}_{\pm 0.017}$

4.2 Evaluation Criteria

We use the F1-score metric to evaluate the change detection performance after finetuning. The value of the F1-score ranges from 0 to 1. The higher the F1-score, the better the precision and recall.

$$F1 - score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{4}$$

5 Results and Discussion

Experiments are conducted by finetuning the state-of-the-art SCD model with three sets of pretraining strategies on PCD and VL-CMU-CD datasets. (1) Random initialized (Rand Init), (2) Supervised ImageNet pretraining (Sup-Im), (3) Randomly initialized self-supervised pretraining (SSCD, & D-SSCD). In addition, we provide the results on using Supervised ImageNet initialization for self-supervised pretraining (Sup-Im + SSCD, & Sup-Im + D-SSCD) step in Appendix Section A.1.1).

5.1 Evaluation on VL-CMU-CD and PCD Datasets

Table 1 shows the performance of proposed pretraining methods evaluated using DR-TANet on the VL-CMU-CD and PCD dataset. The results show that our proposed pretraining method (SSCD) can match the widely-used ImageNet pretraining (Sup-Im) that utilizes millions of images without the use of any additional data. In comparison with the SSCD, the performance of the D-SSCD pretraining drops by 2% on VL-CMU-CD. Since the D-SSCD method learns the representation of changed regions directly which leads to the performance decrease as these changed regions are not well distinguished in the VL-CMU-CD dataset. This shows that pretraining by SSCD helps to learn temporal-consistent representations when the image pairs are predominantly subjected to noisy changes in illuminations and seasonal variations.

We also evaluate the performance of our proposed methods on the PCD dataset using DR-TANet (Table 1). Similar to the results observed on the VL-CMU-CD dataset, the proposed pretraining outperforms the supervised ImageNet pretraining comfortably by a large margin. In the PCD dataset, D-SSCD exceeds SSCD pretraining by 6%. Thus, learning the representation of the change region using differencing is likely to improve the change detection performance when the changes between the image pairs are subjected to large view-point differences and less affected by the noisy changes caused by illumination, seasonal variations, and view-point differences.

Overall, the evaluation on VL-CMU-CD and PCD datasets shows that our proposed methods on unlabeled data can match or surpass the widely used ImageNet pretraining that uses more than a million labeled images. Moreover, it also alleviates the problem of domain shift caused by transfer learning the ImageNet weights pretrained on datasets vastly different from that of SCD datasets.

5.2 Generalization on Out-of-distribution Data

In practice, the SCD model has to perform in challenging scenarios where the testing distribution is unknown and drastically different from the one it is trained on. Therefore, it is necessary that the learned representations generalize well across out-of-distribution data. The PCD dataset is considered OOD data for a model pretrained and finetuned on the VL-CMU-CD dataset and vice-versa. Table 2 shows the F1-score of different pretraining methods on DR-TANet model to out-of-distribution

Table 2: Out-of-distribution performance evaluation (F1-score) of DR-TANet model using pretraining methods.

Methods	VL-CMU-CD→PCD	PCD→VL-CMU-CD
Rand Init	$0.234_{\pm 0.052}$	$0.186_{\pm 0.032}$
Sup-Im	$0.286_{\pm 0.021}$	$0.228_{\pm 0.012}$
SSCD	$0.366_{\pm 0.017}$	$\mathbf{0.306}_{\pm 0.015}$
D-SSCD	$\mathbf{0.417}_{\pm 0.016}$	$0.250_{\pm 0.020}$

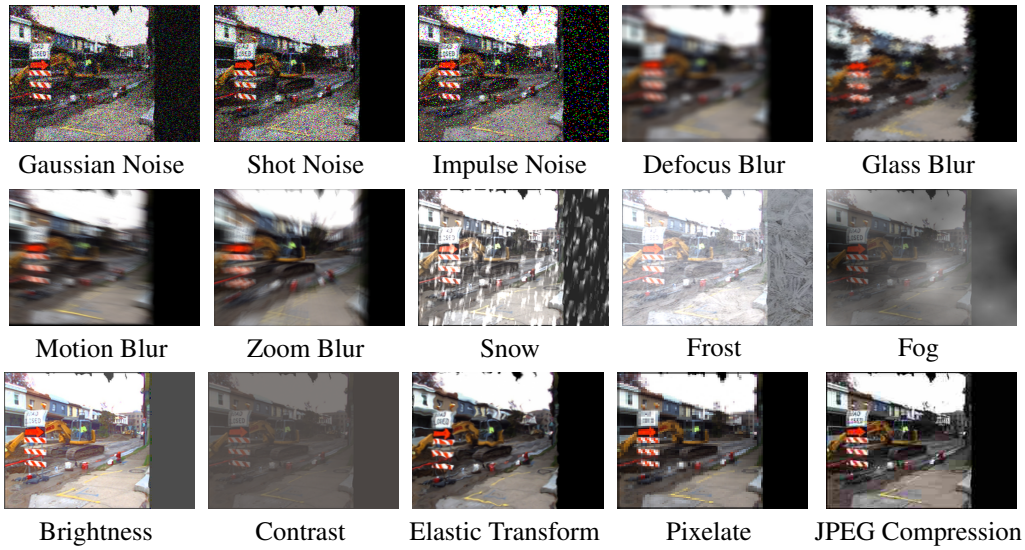

Figure 3: VL-CMU-CD test set is exposed to 15 types of artificially generated natural corruptions (16) with five levels of corruption severity. This figure shows a randomly selected image from the VL-CMU-CD test set with severity 3—best viewed on color.

data. Results show that the model initialized with proposed self-supervised pretraining (SSCD and D-SSCD) generalizes well to the OOD dataset compared to the model initialized with random weights and ImageNet pretrained weights, indicating that self-supervised pretraining learns better and more generalizable feature representations. Results show that the proposed pretraining helps the SCD model to generalize well across datasets with different distributions.

5.3 Robustness under Natural Corruption

The task of SCD is often applied to the outdoor environment, where the images are subjected to seasonal variations. Therefore, the SCD model must be robust to the common natural corruptions such as illumination, noise and blur. Here, we evaluate the robustness of SCD models to natural corruptions which has not been addressed in this domain previously. Following Hendrycks et al. (16), we use 15 different natural corruptions applied on VL-CMU-CD test set to generate VL-CMU-CD-C (examples are shown in Figure 3). These corruptions are categorized in 4 categories as noise, blur, weather, and digital. Each corruption category is subjected to five severity levels obtained by varying intensities of corruption. The DR-TANet model initialized with different pretraining methods is

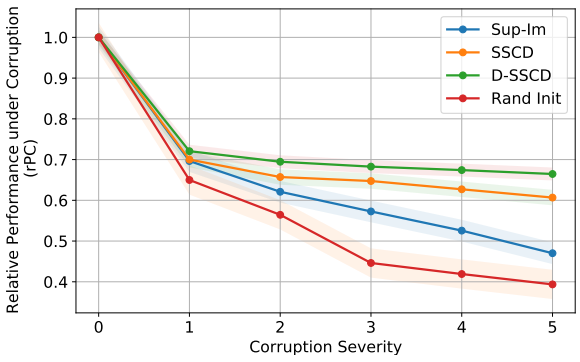

Figure 4: Relative Performance degradation on corrupted images with increasing levels of corruption severity, best viewed on color.

Table 3: Performance (F1-score) of pretraining methods evaluated using DR-TANet on VL-CMU-CD dataset under varying label availability.

Methods	Label Fraction			
	1%	10%	50%	100%
Rand Init	$0.252_{\pm 0.05}$	$0.423_{\pm 0.032}$	$0.545_{\pm 0.072}$	$0.708_{\pm 0.051}$
Sup-Im	$0.295_{\pm 0.02}$	$0.411_{\pm 0.018}$	$0.601_{\pm 0.018}$	$\mathbf{0.752}_{\pm 0.015}$
SSCD	$\mathbf{0.569}_{\pm 0.014}$	$\mathbf{0.617}_{\pm 0.016}$	$\mathbf{0.685}_{\pm 0.015}$	$0.745_{\pm 0.012}$
D-SSCD	$0.564_{\pm 0.03}$	$0.611_{\pm 0.025}$	$0.674_{\pm 0.014}$	$0.725_{\pm 0.014}$

trained using clean VL-CMU-CD while being tested on VL-VMU-CD-C. The metrics mean performance under corruption (mPC) (26) and relative performance under corruption (rPC) (26) are used to evaluate the robustness of models (Eqs (5) and (6), respectively). rPC measures the relative degradation of performance on corrupted data with respect to clean data.

$$mPC = \frac{1}{N_c N_s} \sum_{c=1}^{N_c} \sum_{s=1}^{N_s} P_{c,s} \tag{5}$$

where $P_{c,s}$ is the F1-score measure evaluated on VL-CMU-CD-C under cth corruption with severity level s. while N_c= 15 and N_s= 5 indicate the number of corruptions and severity levels, respectively.

$$rPC = \frac{mPC}{P_{clean}} \tag{6}$$

Figure 4 shows that the drop in performance increases with an increase in the severity of the applied corruption. Moreover, (1) there is a large degradation in performance on the model initialized with supervised Imagenet pretraining (Sup-Im) when subjected to the corrupted test set. (2) Unlike the ImageNet pretrained models that suffer severe performance loss on corrupted images, the proposed pretraining methods (SSCD and D-SSCD) are more robust to increase in corruption severity as the degradation in performance is gradual. (3) D-SSCD is slightly more robust than SSCD as it likely learns robust representations from an extra pair of augmented images during model pretraining. Therefore, it is evident that the proposed self-supervised pretraining brings discernible benefits in terms of robustness to natural corruptions when compared to the ImageNet pretraining.

5.4 Efficiency under Limited Labels

The availability of large annotated data remains a critical challenge in SCD due to the high cost of acquiring manual annotations. Therefore, the SCD model needs to demonstrate steady performance when the availability of labeled data is limited. Table 3 shows the performance of different pretraining under limited labels setting. Different percentages of labeled data (1%, 10%, 50%, and 100%) are sampled in a class-balanced manner from the training split of VL-CMU-CD. The finetuning performance of proposed pretraining methods with varying quantities of labeled data are evaluated using the VL-CMU-CD test set. Our proposed pretraining methods (SSCD and D-SSCD) outperform the widely used Imagenet pretraining by a large margin across all limited label scenarios. The performance drop of Sup-Im is more significant when the amount of labeled data is 10% or less compared to our proposed methods, and then the gap decreases as the availability of labeled data increases. Thus, the proposed pretraining methods increase the performance of the change detection model to a greater extent when the availability of the labeled data is scarce.

Overall, compared to SSCD the proposed D-SSCD method increases the robustness and generalization of the SCD model to a larger extent in many real-world scenarios where the images are affected by challenging conditions. This can be attributed to the learned representations of the changed features through differencing which helps D-SSCD to learn more generalizable task agnostic features that further contributes to the increase in robustness of the SCD model.

6 Conclusion and Future work

We proposed a novel self-supervised pretraining method (D-SSCD), specifically for SCD, that learn temporal-consistent features inherent to the data in an unsupervised manner. We demonstrated that our method can be easily extended to existing state-of-the art SCD methods. With extensive experiments on two challenging SCD datasets, we demonstrated the superiority of the D-SSCD over the widely used ImageNet pretraining without any additional data. Our results also demonstrate the robustness of D-SSCD to natural corruptions, out-of-distribution generalization and its efficiency under limited annotations. Therefore, we believe that our findings in this work can be harnessed to increase the performance and robustness of SCD where obtaining the labeled data is scarce and expensive. Although our approach reduces the dependency of the SCD models to large-scale labeled data, one possible limitation is that the task of SCD is not entirely unsupervised. In the future, we intend to extend the proposed self-supervised approach to tackle the problem of unsupervised change detection.

7 Broader Impact

The findings outlined in this work can potentially be exploited to enhance the performance and robustness in any application of change detection where the labeled data is scarce and expensive to obtain. Some real-world applications such as ecosystem monitoring, urban expansion, damage assessment, and autonomous HD map maintenance have an immensely positive impact on society. For instance, in the application of urban expansion and disaster assessment, it is important to identify the damages caused due to natural disasters such as tsunami. Owing to the difficulty in obtaining labeled data, our model can help in estimating the damages and help the government in making crucial decisions. Additionally, our model can also be used in medical applications to estimate disease severity by detecting the changes from medical images. This can potentially help clinicians to make timely decisions, thus saving a patient's life. Furthermore, we do not foresee any negative implications of our model on society. Finally, we would also like to point out that the scene change task in computer vision is less explored compared to other tasks such as object detection and semantic segmentation. We believe that this research can kindle further developments in the direction of unsupervised scene change detection for the above-mentioned applications which are clearly beneficial to society.

References

[1] Guo, E., Fu, X., Zhu, J., Deng, M., Liu, Y., Zhu, Q. and Li, H., 2018. Learning to measure change: Fully convolutional siamese metric networks for scene change detection. arXiv preprint arXiv:1810.09111.

[2] Chen, S., Yang, K. and Stiefelhagen, R., 2021. DR-TANet: Dynamic Receptive Temporal Attention Network for Street Scene Change Detection. arXiv preprint arXiv:2103.00879.

[3] Alcantarilla, P.F., Stent, S., Ros, G., Arroyo, R. and Gherardi, R., 2018. Street-view change detection with deconvolutional networks. Autonomous Robots, 42(7), pp.1301-1322.

[4] Sakurada, K., Okatani, T. and Deguchi, K., 2013. Detecting changes in 3D structure of a scene from multi-view images captured by a vehicle-mounted camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 137-144).

[5] S. H. Khan, X. He, F. Porikli, M. Bennamoun, F. Sohel, and R. Togneri, "Learning deep structured network for weakly supervised change detection," arXiv preprint arXiv:1606.02009, 2016.

[6] K. Sakurada and T. Okatani, "Change detection from a street image pair using cnn features and superpixel segmentation." in BMVC, 2015, pp.61–1.

[7] Sakurada, K., Wang, W., Kawaguchi, N. and Nakamura, R., 2017. Dense optical flow based change detection network robust to difference of camera viewpoints. arXiv preprint arXiv:1712.02941.

[8] Sakurada, K., Shibuya, M. and Wang, W., 2020, May. Weakly supervised silhouette-based semantic scene change detection. In 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6861-6867).

[9] Varghese, A., Gubbi, J., Ramaswamy, A. and Balamuralidhar, P., 2018. ChangeNet: A deep learning architecture for visual change detection. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 0-0).

[10] Lei, Y., Peng, D., Zhang, P., Ke, Q. and Li, H., 2020. Hierarchical Paired Channel Fusion Network for Street Scene Change Detection. IEEE Transactions on Image Processing, 30, pp.55-67.

[11] Shi, W., Zhang, M., Zhang, R., Chen, S. and Zhan, Z., 2020. Change Detection Based on Artificial Intelligence: State-of-the-Art and Challenges. Remote Sensing, 12(10), p.1688.

[12] Bu, S., Li, Q., Han, P., Leng, P. and Li, K., 2020. Mask-CDNet: A mask based pixel change detection network. Neurocomputing, 378, pp.166-178.ccf

[13] Zbontar, J., Jing, L., Misra, I., LeCun, Y. and Deny, S., 2021. Barlow twins: Self-supervised learning via redundancy reduction. arXiv preprint arXiv:2103.03230.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.

[15] Chen, T., Kornblith, S., Swersky, K., Norouzi, M. and Hinton, G., 2020. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029.

[16] Hendrycks, D. and Dietterich, T., 2019. Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261.

[17] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," arXiv preprint arXiv:1906.05849, 2019.

[18] Wu, Z., Xiong, Y., Yu, S. and Lin, D., 2018. Unsupervised feature learning via non-parametric instance-level discrimination. arXiv preprint arXiv:1805.01978.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722, 2019.

[20] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G. and Piot, B., 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733.

[21] Chen, X. and He, K., 2021. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15750-15758).

[22] Loshchilov, I. and Hutter, F., 2016. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983.

[23] You, Y., Gitman, I. and Ginsburg, B., 2017. Scaling sgd batch size to 32k for imagenet training. arXiv preprint arXiv:1708.03888, 6, p.12.

[24] Barlow, H.B., 1961. Possible principles underlying the transformation of sensory messages. Sensory communication, 1(01).

[25] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks.

[26] Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M. and Brendel, W., 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484.

[27] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255).

[28] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[29] Newell, A. and Deng, J., 2020. How useful is self-supervised pretraining for visual tasks?. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7345-7354).

A Appendix

A.1 Additional Experiments

A.1.1 Combination of Supervised and Self-supervised pretraining

In many computer vision tasks, studies have shown that self-supervised pretraining benefits from prolonged training on large amounts of unlabeled datasets (14) (19). However, the task of SCD demands registered/aligned pairs of temporal images to identify the changes in a scene. Because of this, it is also hard to obtain these registered pair of unlabeled images as it is expensive and time-consuming. We seek to address this issue by conducting additional experiments with a two-stage pretraining approach, where we initialize our proposed self-supervised pretraining models with the supervised ImageNet weights (Sup-Im + SSCD & Sup-Im + D-SSCD) instead of random initialization and train it with a fewer amount of unlabeled data.

Similar to the experiments shown in Table 1 in the main paper, Table 5 shows the performance of two-stage pretraining methods evaluated using DR-TANet on the VL-CMU-CD and PCD datasets. We discern that the proposed pretraining in conjunction with the ImageNet pretraining (Sup-Im + SSCD) surpasses the Sup-Im and SSCD by 1.3% and 2% on VL-CMUCD, respectively. Similarly, Sup-Im +D-SSCD pretraining outperforms Sup-Im and D-SSCD by 8.6% and 2% on the PCD dataset, respectively. Besides improving the performance of self-supervised pretraining, the combined pretraining also reduces the requirement of registered pair of unlabeled images in SCD. Finally, we highlight that pretraining in this way minimizes the problem of domain shift caused by transfer learning from the ImageNet weights directly.

Evaluation under Natural corruption. In Figure 5, we show results similar to the performance of different pretraining methods on corrupted images with increasing levels of corruption severity but with an additional set of experiments using the two-stage pretraining approach. The combination of supervised and self-supervised pretraining (Sup-Im + SSCD & Sup-Im + D-SSCD) is more robust than the Sup-Im. However, when compared with the SSCD and D-SSCD pretrained models, the robustness of the model is slightly affected as the representations learned during self-supervised pretraining are highly influenced by the representations learned by the supervised ImageNet pretrained weights (trained on million images).

Generalization to out-of-distribution data. We also evaluate the generalization ability of the combined pretraining to out-of-distribution data (See Table 4) with similar experimental settings discussed in Section 5.2. Similar to the results obtained when evaluating the combined pretraining on natural corruptions, SSCD and DSSCD pretraining outperform the SSCD+Im and DSSCD+Im, indicating that supervised pretraining affects the generalization ability of the representations learned during the self-supervised phase when combined.

Table 4: Performance of pretraining methods evaluated using DR-TANet (2) on out-of-distribution dataset

Methods	VL-CMU-CD→PCD	PCD→VL-CMU-CD
Sup-Im	$0.286_{\pm 0.021}$	$0.228_{\pm 0.012}$
SSCD	$0.366_{\pm 0.017}$	$\mathbf{0.306}_{\pm 0.015}$
D-SSCD	$\mathbf{0.417}_{\pm 0.016}$	$0.250_{\pm 0.020}$
Sup-Im + SSCD	$0.308_{\pm 0.018}$	$0.265_{\pm 0.026}$
Sup-Im + D-SSCD	$0.328_{\pm 0.011}$	$0.232_{\pm 0.018}$

Evaluation under varying label quantities. Table 3 shows the performance of the combined pretraining under varying label quantities. The experimental set-up is the same as discussed in Section 5.4. In comparison to the Sup-Im, SSCD, and D-SSCD, combined pretraining (Sup-Im + SSCD & Sup-Im + D-SSCD) demonstrates superior performance when finetuned on 100% labeled data. However, when compared to the SSCD and D-SSCD pretraining, there is a slight decrease in performance when the quantities of labeled data are limited.

Finally, we can conclude that although the combination of supervised and self-supervised pretraining shows performance improvement over self-supervised pretraining (under normal settings), the robust-

Table 5: Performance (F1-score) of DR-TANet model trained on VL-CMU-CD and PCD datasets using different pretraining methods.

Methods	VL-CMU-CD	PCD Dataset		
		Tsunami	GSV	Average
Sup-Im	0.752±0.015	0.687±0.013	0.465±0.012	0.576±0.012
SSCD	0.745±0.012	0.709±0.018	0.456±0.022	0.583±0.021
D-SSCD	0.725±0.014	0.712±0.014	0.558±0.019	0.642±0.017
Sup-Im + SSCD	**0.765**±0.017	0.716±0.021	0.466±0.019	0.591±0.020
Sup-Im + D-SSCD	0.748±0.012	**0.727**±0.018	**0.605**±0.016	**0.662**±0.017

Table 6: Performance (F1-score) of pretraining methods evaluated using DR-TANet on VL-CMU-CD dataset under varying label availability.

Methods	Label Fraction			
	1%	10%	50%	100%
Sup-Im	0.295±0.02	0.411±0.018	0.601±0.018	0.752±0.015
SSCD	**0.569**±0.014	**0.617**±0.016	**0.685**±0.015	0.745±0.012
D-SSCD	0.564±0.03	0.611±0.025	0.674±0.014	0.725±0.014
Sup-Im + SSCD	0.552±0.024	0.598±0.019	0.661±0.021	**0.765**±0.015
Sup-Im + D-SSCD	0.556±0.02	0.604±0.011	0.664±0.017	0.748±0.015

ness and generalization of the model are slightly reduced under natural corruptions, out-of-distribution data, and limited labeled scenarios.

A.2 Self-supervised pre-training setup

Dataset pre-processing. For PCD dataset, original images are cropped into 224×224. By sliding 56 pixels in width and data augmentation of plane rotation, each image pair is expanded into 60 patches with a 224×224 resolution. In total, 12000 image pairs are generated. As the input, the image pairs will be resized into 256×256. For VL-CMU-CD dataset, we follow the random training and testing splits in (1; 2). Nine hundred thirty-three image pairs (98 sequences) for training and 429 (54 sequences) for testing are resized into a 256×256 resolution. Note that only images belonging to the train set (without labels) are used to train the model.

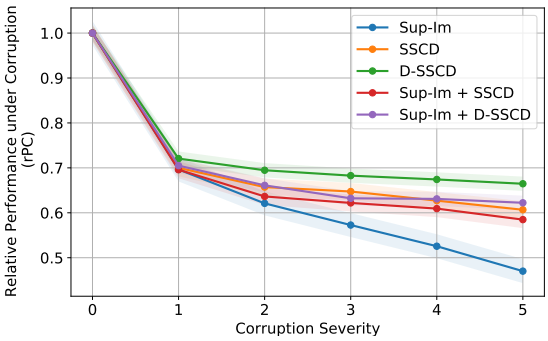

Figure 5: Relative Performance degradation on corrupted images with increasing levels of corruption severity—best viewed on color.

Architecture. We use Barlow Twins (13) as our baseline model. It consists of ResNet50 (28) (without the final classification layer) as a feature extractor followed by a projector network. The projector network has two linear layers, each with a hidden layer size of 512 output units. Owing to the high computational requirements, the output of the projector network was modified to generate embeddings of size 256 compared to the Barlow twins network, which generates embeddings of size 8192. The first layer of the projector is followed by a batch normalization layer and rectified linear units. The architecture for the proposed SSCD and D-SSCD method remains the same except for a differencing layer after the projection network in D-SSCD.

Data Augmentation. Our image augmentation pipeline consists of the following transformations: Image resizing to 256 × 256, color jittering, converting to grayscale, Gaussian blurring. Except resizing, the other transformations are applied randomly, with some probability. Random crop is not considered when pre-processing the change detection datasets as the presence of changed regions between an image pair taken at different times is much smaller and random compared to the unchanged regions.

Training and Optimization. We follow the optimization protocol described in Barlow Twins. We use the LARS optimizer (23) and train for 400 epochs with a batch size of 16 on two NVIDIA RTX-2080 Ti GPU. We use a learning rate of 0.003, multiply the learning rate by the batch size, and divide it by 256. The learning rate is reduced by a factor of 1000 using a cosine decay schedule (22). We use a weight decay parameter of $1x10^{-}6$. The architecture, augmentation, hyper-parameters, and training procedures mentioned above remain the same for both proposed pretraining methods(SSCD and D-SSCD).

A.3 Change Detection setup

We evaluate the proposed self-supervised pretraining methods by finetuning them to a downstream task of SCD. DR-TANet (2), a state-of-the art SCD network is considered for finetuning. To keep the consistency throughout the experiments, we used ResNet50 (28) as a feature extractor for finetuning the pre-trained model on both of these networks. During finetuning, the data pre-processing, training, and testing protocols followed by DR-TANet were replicated. We considered a batch size of 8 while training the DR-TANet on VL-CMU-CD and PCD datasets. While training on PCD, we reduced the dependency-scope size of the DR-TANet to 1x1 and trained the model in lowest setting owing to limited GPU memory and longer training time.