Self-Supervised Material and Texture Representation Learning for Remote Sensing Tasks

Peri Akiva
Rutgers University
peri.akiva@rutgers.edu

Matthew Purri
Rutgers University
matthew.purri@rutgers.edu

Matthew Leotta
Kitware Inc
matt.leotta@kitware.com

Abstract

*Self-supervised learning aims to learn image feature representations without the usage of manually annotated labels. It is often used as a precursor step to obtain useful initial network weights which contribute to faster convergence and superior performance of downstream tasks. While self-supervision allows one to reduce the domain gap between supervised and unsupervised learning without the usage of labels, the self-supervised objective still requires a strong inductive bias to downstream tasks for effective transfer learning. In this work, we present our material and texture based self-supervision method named MATTER (**MAT**erial and **TE**xture **R**epresentation Learning), which is inspired by classical material and texture methods. Material and texture can effectively describe any surface, including its tactile properties, color, and specularity. By extension, effective representation of material and texture can describe other semantic classes strongly associated with said material and texture. MATTER leverages multi-temporal, spatially aligned remote sensing imagery over unchanged regions to learn invariance to illumination and viewing angle as a mechanism to achieve consistency of material and texture representation. We show that our self-supervision pre-training method allows for up to 24.22% and 6.33% performance increase in unsupervised and fine-tuned setups, and up to 76% faster convergence on change detection, land cover classification, and semantic segmentation tasks. Code and dataset: https://github.com/periakiva/MATTER.*

1. Introduction

Automated understanding of remote sensing imagery has been a long standing goal of the computer vision community. Its broad applicability has driven research and development in construction phase detection [23], infrastructure mapping [36, 55, 71, 100], land use monitoring [41], post natural disaster damage assessment [42, 89, 97], urban 3D reconstruction [39, 57], population migration prediction [19], and climate change tracking [79]. Most of those methods require some degree of annotation effort, which is often expensive and/or time consuming. Satellite imagery is increasingly plentiful and accessible, with hundreds of satellites collecting images on a daily basis [1, 35, 81, 94]. However, annotating land cover, change, or similar labels often requires domain knowledge and/or extreme attention to detail, as labels in remote sensing imagery cover more numerous and smaller objects seen from unfamiliar view points. As a result, annotators require more domain expertise compared to standard benchmark datasets such as Pascal VOC [38], COCO [61], or similar.

Recent work in self-supervised learning aims to alleviate the requirement of labeled data by either detecting self-applied transformations, such as color or rotation change, or implicit metadata information, such as temporal order or geographical location. Those objectives are often achieved using contrastive learning methods [17, 45, 53], in which the distance between feature representations of original and transformed images is minimized. More advanced contrastive methods use triplet loss [10, 84] or quadruplet loss [18] which also include negative examples with which the distance between feature representations is maximized. Despite filling a significant need in the remote sensing domain, these approaches have been yet to be thoroughly investigated. Even methods that utilize contrastive approaches, such as SeCo [68] and the work of Ayush *et al.* [4], which learn seasonal change invariance or geographic-location consistency, still show weaker transfer-ability to downstream task learning, as demonstrated by inferior performance and convergence speeds shown in Tab. 1, 3.

Instead, we hypothesize that material and texture have a strong inductive bias to most downstream remote sensing tasks, with pre-training of surface representation to improve performance and convergence speeds (measured in epochs) for those tasks. Consider the task of change detection in remote sensing imagery: when semantic class changes (*i.e.*, soil to building, or forest to soil), change can also be expected in materials and texture, demonstrating the high correlation between material and texture and the change detection task. We show the effectiveness of our self-supervised pre-trained features in both raw and fine-tuned forms, obtaining state-of-the-art (SOTA) performance in change detection (unsupervised and fine-tuned), land cover segmentation (fine-tuned), and land cover classification (fine-tuned).

Here, we propose a novel self-supervised material and texture representation learning method which is inspired by classical and modern texton filter banks [58, 87, 113]. Textons [52, 58, 66] refer to the description of micro-structures in images often used to describe material and texture consistency [25, 27, 58, 101, 108]. Note that literature has only loosely defined what material, structures, texture, and surface refer to. Here, we define *material* as any single or combination of elements (soil, concrete, vegetation, etc.) corresponding to some multi-spectral signature, *structures* as gradients in intensity, *texture* as spatial distribution of structures, and *surface* as the combination of material and texture. Note that here we define the physical surface, rather than the geometric or algebraic surface, as described by its material and textural properties. By extension, we aim to jointly describe combinations of materials and textures in a single objective. For example, within a given image patch, a mixture of grass and concrete should be represented differently than patches with grass or concrete separately. In this example, the grass-concrete mixture may be associated to both grass and concrete material classes. To that end, we learn surface representations that describe the affinity, represented as residuals [48], to all pre-defined surface classes, represented as clusters. We achieve this by contrastively learning the similarity between the residuals of multi-temporal, spatially aligned imagery of unchanged regions to obtain consistent material and texture representations, regardless of illumination or viewing angle. This framework acts as a pre-training stage for downstream remote sensing tasks.

Overall, our contributions are: **1)** We present a novel material and texture based approach for self-supervised pre-training to generate features with high inductive bias for downstream remote sensing tasks. We propose a texture refinement network to amplify low level features and adapt residual cluster learning to characterize mixed materials and texture patches in a self-supervised, contrastive learning framework. **2)** We achieve SOTA performance on unsupervised and supervised change detection, semantic segmentation, and land cover classification using our pre-trained network. **3)** We provide our curated multi-temporal, spatially aligned, and atmospherically corrected remote sensing imagery dataset, collected over unchanged regions used for self-supervised learning.

2. Related Work

2.1. Downstream Remote Sensing Tasks

The main downstream tasks we investigate in this work are change detection, land cover segmentation, and land cover classification. The problem of change detection in satellite imagery has been thoroughly investigated over time [12, 13, 15, 24, 49, 77, 78, 83]. Notable examples include Daudt *et al*. [29], which predicts change by minimizing feature differences at every layer of the network from a given image pair input, and Chen *et al*. [13], which utilizes a spatial-temporal attention mechanism to detect anomalies in sequences of images. Land cover segmentation and classification have also seen a surge in interest, with growing repositories of annotated datasets [3, 32, 47, 91, 96] and methods [2, 5, 44, 82, 93, 96]. H20-Net [2] synthesizes multi-spectral bands and uses self-sampled points to generate pseudo-ground truth for flood and permanent water segmentation. VecRoad [93] sets the problem of road segmentation as iterative graph exploration. Multi3Net [82] learns fusion of multi-temporal, multi-spectral features from high resolution imagery to jointly predict pixels of floods and building.

2.2. Self-Supervision

In order to effectively utilize large amounts of unlabeled data, recent methods have focused on obtaining good feature representations without explicit annotation efforts. This is done by deriving information from the data itself or learning sub-tasks within data instances without changing the overall objective. The first is often used when high confidence labels can be obtained and trained on, similar to [2, 4], where the method infers weak supervision about input images through provided meta-data or classical methods. The second, and more common approach, leverages metric learning objectives to learn generalizable features for the same data instance or class. Recent methods involve learning invariance to color and geometric transformations [9, 50, 70], temporal ordering [6, 40], sub-patch relative location prediction [34], frame interpolation [73], colorization [33, 56, 110], patch and background filling [102], and point cloud reconstruction [105].

More relevant to the remote sensing domain, SeCo [68] has taken a step toward utilizing the potential in the abundance of satellite imagery by contrastively learning seasonal invariance as a pre-text self-supervision task. It then fine-tunes the pre-trained network on downstream tasks such as change detection and land cover classification. Ayush *et al*. [4] also proposes a self-supervised approach enforcing geographical-location consistency as a pre-training objective used for downstream tasks such as land-cover segmentation and classification. While both methods show improved results on benchmark datasets when compared to random weights initialization, we show that their inductive bias is still significantly weaker than that of our material and texture consistency based pre-trained weights, which learn an illumination and viewing angle invariance to achieve consistency of material and texture representation.

2.3. Material and Texture Identification

Early material and texture recognition methods relied on hand crafted filter banks, whose combinatorial output are also referred to as textons [58], to encode statistical representations of image patches [7, 8, 26, 28, 58, 95, 114].

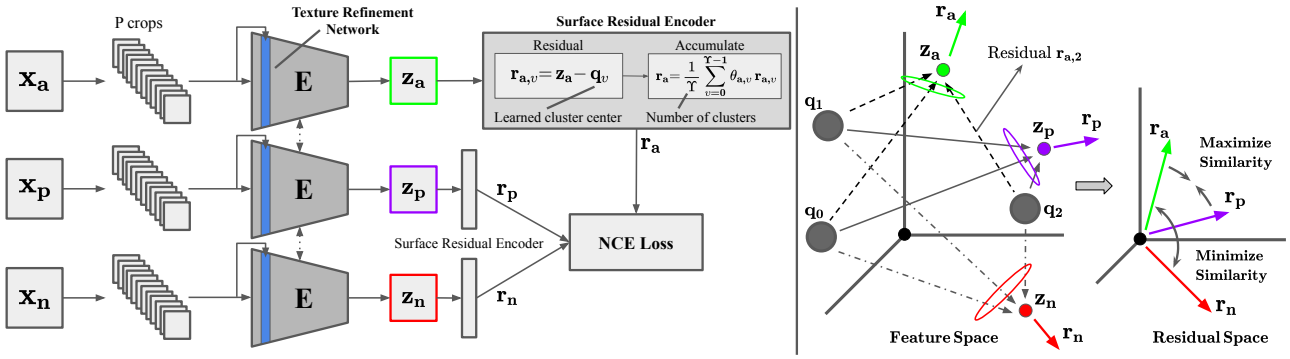

Figure 1. **(Left) MATTER**: anchor, positive, and negative images x_a, x_p, and x_n are densely windowed to P crops which are fed to encoder E, and correspond to output features z_a, z_p, and z_n. Crops are also fed to the Texture Refinement Network (shown in blue) which amplifies activation of low-level features to increase their impact in deeper layers. The encoder's output is then fed to the Surface Residual Encoder to generate patch-wise cumulative residuals, which represents affinity between input data and all learned clusters. A residual vector between feature output z_a and cluster v is denoted as $r_{a,v}$. Output learned residuals, cluster weights, and number of clusters are noted as r, θ, and Υ respectively. **(Right)** Simplified example of the contrastive objective with $\Upsilon = 3$. Residuals from learned clusters are extracted and averaged for all crops as representations of correlation between inputs and all clusters. Best viewed in zoom and color.

Later works investigated the use of clustering and inter-patch statistics as replacement for pre-defined filter banks [98, 99], at the cost of defining the feature space it operates in. Most notable feature spaces include color intensity [14], texture homogeneity [43, 58, 69], multi-resolution features [74, 88], and feature curvature [63, 86]. More recent, deep learning approaches have translated the problem of texture representation to focus on explicit identification of materials through texture encoding [20, 109, 112], differential angular imaging [106], 3D surface variation estimation [31], auxiliary tactile property [85], and radiometric properties estimation such as the bidirectional reflectance distribution function (BRDF) [11, 62, 103] and the bidirectional texture function (BTF) [104]. Those methods seek to learn low-level features that are key to material classification and segmentation. Some methods choose to add skip connections [60, 107, 115] to supply low-level features in deep layers, and others choose explicit concatenation of texture related information [59, 85]. Many of these elements are meant to reduce the receptive field or increase impact of low-level features of the network while keeping it sufficiently deep. FV-CNN [22] aims to generate texture descriptions of densely sampled windows. Since the features describe regions removed from global spatial information, it explicitly constrains the receptive field of the network to the size of the window. DeepTEN [109] learns residual representations of material images in an end-to-end pipeline using material labels. Our approach combines elements from FV-CNN and DeepTEN in two ways. First, we densely sample windows and refine low level features as receptive field constraints. Then, we contrastively learn implicit surface residual representations without the usage of material labels or auxiliary information. To our knowledge, we are the first to employ self-supervised material and texture based objectives for pre-training steps.

3. Methodology

The goal of **MAT**erial and **TE**xture **R**epresentation Learning (MATTER) is to learn a feature extractor that generates illumination and viewing angle invariant material and texture representations from given multi-temporal satellite imagery sampled over unchanged regions. To this end, to train our model, we utilize our self-collected dataset described in Sec. 4.1, which samples multi-temporal imagery of rural and remote regions, in which little to no change is assumed between every consecutive pair of sampled images. See Fig. 2 for an overview of our approach.

Given an anchor reference image $x_a \in \mathcal{R}^{B\times H\times W}$ sampled over an unchanged region, we obtain a positive, temporally succeeding image $x_p \in \mathcal{R}^{B\times H\times W}$ over the same region, and a negative image $x_n \in \mathcal{R}^{B\times H\times W}$ sampled over a different region. B, H, and W correspond to number of channel bands, height, and width of input images. We tile all images into P equally sized, corresponding patches of size $h \times w$, with spatially aligned reference and positive patches, c_a and c_p, and negative patches, c_n, randomly sampled from regions other than the reference region. The usage of densely sampled crops aims to restrict the receptive field by removing features from the global spatial context, and to prevent the model from learning higher level features ineffective in describing surfaces. We study the effects of receptive field variation in Sec. 5.1.

To learn material and texture centric features, we present the Texture Refinement Network (TeRN) (Sec. 3.1), and patch-wise Surface Residual Encoder (Sec. 3.2). TeRN aims to amplify the activation of lower level features essential for texture representation (as seen in Fig. 3), and Surface Residual Encoder is our patch-wise adaptation of Deep-TEN [109] to learn surface-based residual representations. We train our network to minimize the feature dis-

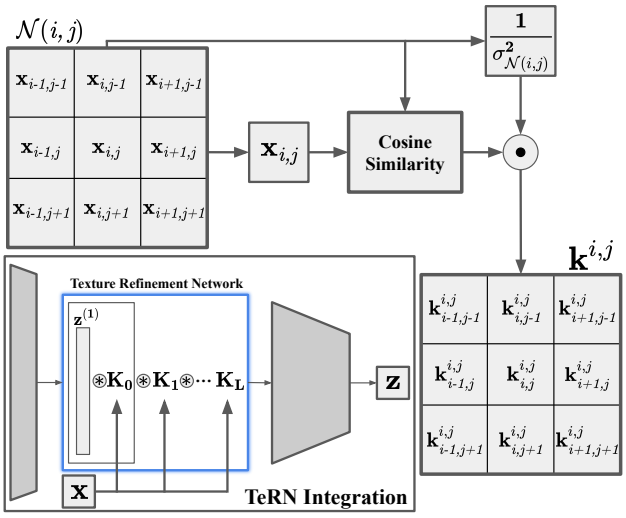

Figure 2. **Texture Refinement Network (TeRN)** assigns convolution weights based on the cosine similarity of the kernel's center pixel and its neighbors, divided by the standard deviation of that kernel. We then convolve features $z^{(1)}$ to refine low-level features essential for texture and material centric learning. The symbols $\circledast$ and $\odot$ correspond to convolution and element-wise multiplication operations. Best viewed in zoom.

tance of positive patch pairs, c_a and c_p, and maximize the feature distance of negative patch pairs, c_a and c_n, where the features are the learned residual representations. For our learning objective, we use the Noise Contrastive Estimation loss [75]:

$$\mathcal{L}_{NCE} = -\mathbb{E}_C \left[\log \frac{\exp(f(c_a) \cdot f(c_p))}{\sum_{c_j \in C} \exp(f(c_a) \cdot f(c_j))} \right], \quad (1)$$

where $f(c_j)$ is the output features of input patch c_j, and C is the set of positive and negative patches.

3.1. Texture Refinement Network

Capturing texture details is difficult in low resolution images, and is especially challenging when considering satellite images that have low contrast. As a result, texture will be less visible and have less impact on the final extracted features. We address this challenge by using our Texture Refinement Network (TeRN) to refine lower level texture features to increase their impact in deeper layers. TeRN utilizes the recently introduced pixel adaptive convolution layer [90], in which the convolution kernel weights are a function of the features locally confined by the kernel. Here, our kernel considers the corresponding local pixels in the original image as follows: given kernel $k^{i,j}$ centered at location (i,j), we calculate the cosine similarity between pixel $x_{i,j}$ and all of its neighboring pixels $\mathcal{N}(i,j)$. We note that while this can be achieved with any similarity metric, we observe that orientation based functions (such as cosine similarity) produce better results than magnitude based functions (such as Euclidean distance). The output

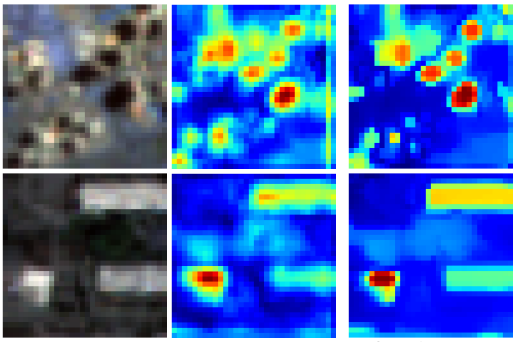

Figure 3. **Qualitative results** of our Texture Refinement Network (TeRN). It can be seen that similar textured pixels obtain similar feature activation intensity in the refined output. Notice how the building in the second row obtains similar activation throughout the concrete building pixel locations compared to the raw features output. Best viewed in zoom and color.

matrix is then divided by the squared standard deviation of all pixels within $\mathcal{N}(i,j)$, noted as $\sigma_{\mathcal{N}(i,j)}$.

$$k^{i,j} = -\frac{1}{\sigma^2_{\mathcal{N}(i,j)}} \frac{x_{i,j} \cdot x_{p,q}}{||x_{i,j}||_2 \cdot ||x_{p,q}||_2}, \forall\, p,q \in \mathcal{N}(i,j). \quad (2)$$

The output matrix of those operations describes both the similarity of the center pixel to its surroundings, and the intensity gradients within the kernel. As previously defined, texture is the spatial distribution of structures, which are represented as intensity gradients. Since we want to emphasize texture, we explicitly polarize the feature activation in regions with high variance or low similarity, with kernel weights decreasing with high variance and/or low cosine similarity. When convolved over our low level features, it highlights edges, and encourages representation consistency for pixels with similar material signatures, as seen in figure 3. The described operation constitutes a single kernel location of a single refinement layer. We define a single refinement layer, K, when the operation is repeated for all image locations. We construct an L-layer refinement network, where each layer is able to utilize different kernel sizes, dilations, and strides. Since the network has deterministicly defined weights, it does not have learned parameters. A base TeRN kernel, and its integration in the overall network, are visually depicted in Fig. 2, and sample refined features in Fig. 3.

3.2. Learning Consistency of Surface Residuals

The task of residual encoding is tightly related to classical k-means clustering [64] and bag-of-words [51], in which some hard cluster assignment is learned based on the data instance proximity to cluster centers. Given the cluster centers, the residual is calculated as the distance of any data instance from its corresponding cluster center. In practice, we can use the residual to measure how similar a given data instance is to its assigned cluster, and all other clusters. Our

method adapts the work presented in Deep-TEN [109] to learn patch-wise residual encodings without explicit handcrafted clustering through a differentiable pipeline. Traditionally, in Deep-TEN [109] and other classical and deep clustering methods [9,17,21,45,65], the objective is to cluster image-wise inputs to corresponding class-wise cluster centers. In contrast, as we employ a patch-wise approach. A given patch containing some material and texture may be associated with multiple clusters (*i.e.* if a patch captures multiple material elements), so it requires a soft representation depicting an affinity to all learned clusters, and not only to its closest cluster.

Consequently, we learn residuals of small patches and enforce multi-temporal consistency between corresponding patch residuals, imposing similarity of cluster affinity. Given an output feature vector $z_i^{1\times D}$ for some crop c_i, and a set of Υ learned cluster centers $Q = \{q_0, q_2, ..., q_{\Upsilon-1}\}$, each of shape $1 \times D$, we can find the residual corresponding to the feature vector z_i and learned cluster center q_v using $r_{i,v}^{1\times D} = z_i - q_v$. We repeat this for all cluster centers and take the weighted average of residuals from each cluster to obtain the cumulative residual vector,

$$r_i = \frac{1}{\Upsilon}\sum_{v=0}^{\Upsilon-1}\theta_{i,v}r_{i,v}, \tag{3}$$

with learned cluster weight θ_v. By combining the residuals of a given crop, we represent its affinity with all learned clusters. When maximizing or minimizing similarity between residuals, we effectively enforce a consistent cluster affinity between input crops.

4. Experiments

4.1. Self-Supervised Pre-Training

Pre-Training Dataset. To train our self-supervised task, we collect a large amount of freely available, orthorectified, atmospherically corrected Sentinel-2 imagery of regions with limited human development. Regions of interest were manually selected to cover a variety of climates. Given spatial and temporal ranges, we use the PyStac library [37] to fetch imagery from the AWS Sentinel-2 catalog closest to our points of interest. Imagery within the spatial-temporal constraints containing over 20% cloud cover and less than 80% data coverage were removed. A maximum of 100 images meeting these constraints were collected per region. The collected images were divided into 14,857, 1096×1096 px^2 sized tiles for training. The resultant dataset contains 27 regions of interest spanning 1217 km^2 over three years. We provide all points of interest (Lat., Long.) in the supplementary material, and will release the dataset upon publication.

Implementation Details. We adopt a standard ResNet-34 backbone, with TeRN inserted after the first layer, and the Surface Residual Encoder as the output layer. TeRN is constructed with 10 blocks, each containing three layers of kernel size 3×3 and dilations of 1-1-2. For the Surface Residual Encoder, we use $\Upsilon = 64$. We use training patch size of 7×7, batch size of 32, learning rate of 0.01, momentum of 0.6, and weight decay of 0.001 for training. For the Noise Contrastive Estimation loss, we use a temperature scaling of 0.05. We pre-train the network for 110,000 iterations or until convergence. Note that the self-supervised baselines SeCo [68] and Ayush *et al.* [4] use 1 million and 543,435 images respectively for pre-training, while we use only 14,857 images.

Dataset		OSCD [30]		
Method	Sup.	Precision (%)	Recall (%)	F-1 (%)
Full Supervision				
U-Net [80] (random)	$\mathcal{F}$	**70.53**	19.17	29.44
U-Net [80] (ImageNet)	$\mathcal{F}$	70.42	25.12	36.20
MoCo-v2 [45]	$\mathcal{S}+\mathcal{F}$	64.49	30.94	40.71
SeCo [68]	$\mathcal{S}+\mathcal{F}$	65.47	38.06	46.94
DeepLab-v3 [16] (ImageNet)	$\mathcal{F}$	51.63	51.06	53.04
Ours (fine-tuned)	$\mathcal{S}+\mathcal{F}$	61.80	**57.13**	**59.37**
Self-Supervision only				
VCA [67]	$\mathcal{S}$	9.92	20.77	13.43
MoCo-v2 [45]	$\mathcal{S}$	29.21	11.92	16.93
SeCo [68]	$\mathcal{S}$	**74.70**	15.20	25.26
Ours	$\mathcal{S}$	37.52	**72.65**	**49.48**

Table 1. Precision, recall, and F-1 (%) accuracies (higher is better) of the "change" class on Onera Satellite Change Detection (OSCD) dataset validation set [30]. $\mathcal{F}$, and $\mathcal{S}$ represent full and self-supervision respectively. $\mathcal{S}$ + $\mathcal{F}$ refer to self-supervised pre-training followed by fully supervised fine-tuning. Random and ImageNet denote the type of backbone weight initialization that method uses.

4.2. Change Detection

Implementation Details. This task is evaluated on the Onera Satellite Change Detection (OSCD) dataset [30], and performed in two ways: self-supervised, and supervised fine-tuning. The self-supervised approach utilizes *only* the pre-trained backbone to extract patch-wise residual features from both images, with each 9×9 patch representing its center pixel. We calculate the euclidean distance as a change metric between corresponding residual features, which are thresholded using Otsu thresholding [76] to predict change pixels when residual distance is large. For the fine-tuned approach, we use image-wise inputs to a DeepLab-v3 [16] with skip-connections network with our pre-trained backbone, fine-tuning the decoder for 30 epochs while freezing the backbone's weights. We use channel-wise concatenations of image pairs as input to the network, with the output features optimized using the cross entropy loss and ground truth change masks. For evaluation, we report precision, recall, and F-1 score of the "change" class in Tab. 1. We use batch-size of 32, learning rate of 0.001, momentum of 0.6, and weight decay of 0.001. For the self-supervised baseline methods, we use the publicly available model weights and follow the same previously described self-supervised change prediction pipeline. The fully-supervised baselines

Dataset	BigEarthNet [91]		
Method	Sup.	Fine-Tune Epochs	mAP (%)
Inception-v2 [92]	$\mathcal{F}$	-	48.23
InDomain [72]	$\mathcal{S}+\mathcal{F}$	90	69.70
S-CNN [91]	$\mathcal{F}$	-	69.93
ResNet-50 [46] (random)	$\mathcal{F}$	-	78.98
ResNet-50 [46] (ImageNet)	$\mathcal{F}$	-	86.74
MoCo-v2 [45]	$\mathcal{S}+\mathcal{F}$	100	86.05
SeCo [68]	$\mathcal{S}+\mathcal{F}$	100	87.81
Ours (fine-tuned)	$\mathcal{S}+\mathcal{F}$	24	**87.98**

Table 2. Mean average precision accuracy (higher is better) on BigEarthNet land cover multi-label classification dataset validation set [91]. $\mathcal{F}$, and $\mathcal{S}$ represent full and self-supervision respectively. $\mathcal{S}+\mathcal{F}$ refer to self-supervised pre-training followed by fully supervised fine-tuning.

follow the same steps as our fine-tuned approach, without the pre-trained weight initialization.

Results Discussion. In Tab. 1 and Fig. 6 we compare our method to SOTA baselines for both self-supervised and fine-tuned approaches. We present common semantic segmentation networks initialized with weights that are random, or pre-trained with ImageNet [54], MoCo-v2 [45], and SeCo [68]. We hypothesized that change in material and texture corresponds to actual change in the scene. Hence by learning good material and texture representation and comparing representations of image pairs, we can reliably locate regions of change. As evident by Tab. 1, our self-supervised approach learns sufficiently good material and texture representation to outperform other fine-tuned methods, surpassing self-supervised SeCo by 24.22%, and fine-tuned SeCo by 2.08%. When considering our fine-tuned method, we outperform our baselines even further, with 12.43% performance increase compared to our self-supervision based baseline, and 6.33% performance increase compared to the fully supervised baseline. Additionally, we show that the inductive bias of material and texture representation to the task of change detection is significant as evidenced by the quicker convergence speed (measured in epochs), with our method converging within only 30 epochs, compared to 100 epochs reported by SeCo.

4.3. Land Cover Classification

Implementation Details. We evaluate our pre-trained backbone on the BigEarthNet [91] dataset for the task of multi-label land cover classification. The dataset provides 590,326 multi-spectral images of size 120×120 annotated with multiple land-cover labels, split into train and validation sets (95%/5%). We fine-tune a classifier head added to our frozen pre-trained backbone network for 24 epochs using given ground truth labels. We use SGD optimizer, batch-size of 128, learning rate of 0.0005, momentum of 0.6, and weight decay of 0.001. For performance, we report the mean average precision of all classes (19).

Dataset	SpaceNet [96]		
Method	Sup.	Fine-Tune Epochs	mIoU (%)
DeepLab-v3 [16] (random)	$\mathcal{F}$	-	69.44
DeepLab-v3 [16] (ImageNet)	$\mathcal{F}$	-	72.22
MoCo-v2 [45]	$\mathcal{S}+\mathcal{F}$	100	78.05
Ayush *et al.* [4]	$\mathcal{S}+\mathcal{F}$	100	78.51
Ours (fine-tuned)	$\mathcal{S}+\mathcal{F}$	24	**81.12**

Table 3. Mean intersection over union (higher is better) on SpaceNet building segmentation dataset validation set [96]. $\mathcal{F}$, and $\mathcal{S}$ represent full and self-supervision respectively. $\mathcal{S}+\mathcal{F}$ refer to self-supervised pre-training followed by fully supervised fine-tuning. Random and ImageNet denote the type of backbone weight initialization that method uses.

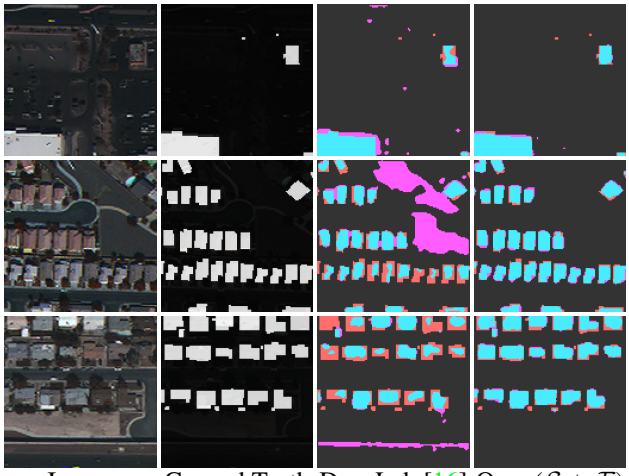

Figure 4. **Qualitative results** of our method on SpaceNet dataset [96]. Cyan, magenta, gray, and red colors represent true positive, false positive, true negative, and false negative respectively. Best viewed in zoom and color.

Results Discussion. Tab. 2 reports the mean average precision performance of baseline and our methods after fine-tuning. While our method only outperforms our baseline by 0.18%, we note that our method converges within 24 epochs, which is significantly faster than our best-performing baseline which reports convergence within 100 epochs.

4.4. Semantic Segmentation

Implementation Details. We use the SpaceNet building segmentation dataset for this task. The dataset provides 10,593 multi-spectral images of size 163×163 labeled with pixel-wise building/no-building masks, split into train and validation sets (90%/10%). We use a DeepLab-v3 [16] with skip-connections network with our frozen pre-trained backbone and fine-tune it for 24 epochs with batch-size of 32, learning rate of 0.0085, momentum of 0.6, and weight decay of 0.001. We report mean intersection over union (mIoU) of the best performing epoch in Tab. 3. The fully-supervised baselines follow the same steps as our fine-tuned approach, without the pre-trained weight initialization.

Results Discussion. Tab. 3 and Fig 4 compare the quantitative and qualitative results of baselines and our method. For our baselines, we report Ayush *et al*. [4] and MoCo-v2 [45], which use PSANet [111] with backbones pre-trained on geography consistency objective. We also report the performance of DeepLab-v3 initialized with random and ImageNet [54] pre-trained weights. As shown in Tab. 3, our method requires significantly less epochs to obtain superior performance on the SpaceNet building segmentation dataset. We outperform our self-supervised based baseline by 2.61%, and fully supervised based baseline by 8.90%, with 76% convergence speed reduction.

5. Results

Evident by our qualitative and quantitative results, our method provides both superior performance and convergence time (measured in epochs) for the evaluated downstream tasks. It is shown that material and texture are strongly associated with common remote sensing downstream tasks, and the ability to represent material and texture effectively improves performances on those tasks. Since quantitatively measuring the ability to represent material without material labels is difficult, we analyze and showcase qualitative texture and material results in the form of visual word maps (pixel-wise cluster assignments). We also discuss limitations, running time, pseudo-code, and additional qualitative results in the supplementary material.

Visual Word Maps Generation. In order to measure the effectiveness of our approach to describe materials and textures, we qualitatively evaluate the visual word maps (pixel-wise cluster assignments) generated by our method. Ideally, we expect similar material and textures to be mapped to the same clusters, without over or under grouping of pixels. We visually compare classical textons, a patch-wise backbone, and our method in Fig. 5. The patch-wise backbone has the same base architecture as MATTER, but without TeRN and surface residual encoding modules. Both methods were trained on the same dataset, with the same hyperparameters, and number of iterations, as described in Sec. 4.1. It can be seen that the textons and patch-wise backbone approaches generate two extreme cases of over-sensitivity and under-sensitivity to changes in material and texture. Since textons operate on raw intensity values, the inter-material variance is small, making it highly sensitive to small texture changes. This can be seen in the textons-generated visual word map, in which small irregularities on the road results in mapping to different visual words. On the other hand, the patch-wise backbone, even with the receptive field constraints through patched inputs, still loses crucial low-level details essential for texture representation. This is indicated by the grouping of obviously different textures to a single visual word. In contrast, as demonstrated in Fig. 5, our Texture Refinement Network and Surface Residual Encoder boost the impact of low-level features, generating surface-based visual word maps. Our method is able to retain texture-essential features, and generalize surface representation which translates to superior surface-based visual word maps.

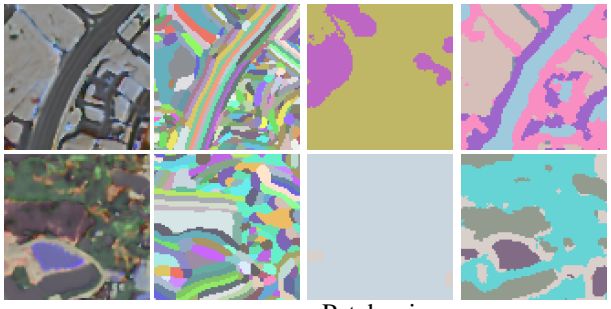

Figure 5. **Qualitative evaluation** of our generated material and texture based visual word maps. It can be seen that our method provides more descriptive surface-based features that are not highly sensitive to small texture irregularities like textons, or under-sensitive to structure changes like the patch-wise backbone. Best viewed in zoom and color. Colors are random.

Train Crop Size \ Inference Crop Size	-	7	9	11	13	15	17	19	21
5	46.68	47.38	46.94	46.94	45.74	44.51	43.74	42.95
7	48.52	**49.48**	49.01	49.02	47.76	46.64	45.92	44.69
9	48.58	47.60	48.02	47.83	46.51	45.57	45.45	43.27
11	48.98	47.83	47.32	46.65	45.64	44.51	44.16	42.45
13	47.46	47.14	46.35	46.99	44.65	43.79	43.09	41.61
15	47.63	47.15	47.30	46.10	45.55	44.68	44.10	41.85
17	46.74	46.81	46.49	45.92	44.69	43.60	43.19	41.09

Table 4. **Receptive Field Constraint Analysis**. F-1 score (%) performance for the unsupervised change detection task. Reported values are of the "change" class with respect to training and inference crop sizes (without fine-tuning). It can be seen that the method benefits from smaller receptive field, achieving superior performance when using smaller train and inference crop sizes.

5.1. Ablation Study

Constraining Receptive Field. In Tab. 4 we study the effects of varying receptive field constraints on our method. As mentioned before, as the receptive field increases, the impact of low-level features diminishes, along with the quality of material and texture representation. Unlike traditional methods, which resort to the usage of smaller networks to reduce receptive field, we explicitly constrain the method by feeding crops to the network, removing them from any global context. Recall that the objective of our method is to learn representation of material and spatial distribution of micro-structures, which are largely affected by low level features which are diminished in larger receptive field methods. In practice, the largest possible receptive field of our network during training is $7 \times 7 = 49$ pixels, which is significantly smaller than the receptive fields of ResNet-50, and ResNet-101, and ResNet-152 with sizes of 483, 1027, and 1507 pixels respectively. It can be seen in

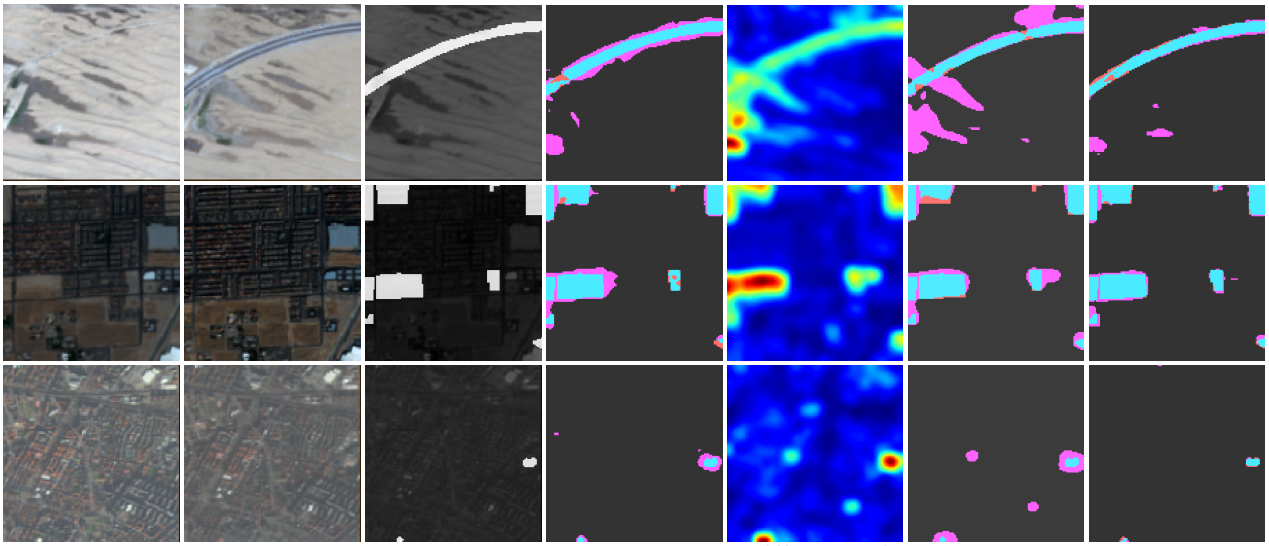

Figure 6. **Qualitative results** of our method on Onera Satellite Change Detection (OSCD) dataset [30]. It can be seen that our self-supervised alone is capable of detecting change only by inferring on the change of material and texture. The fine-tuned model is able to utilize the pre-trained material and texture based weights and achieve significantly better results than models with ImageNet weight initialization. Cyan, magenta, gray, and red colors represent true positive, false positive, true negative, and false negative respectively. Best viewed in zoom and color.

Tab. 4 that in fact, the method benefits from removal from global spatial context and smaller receptive fields, helping it learn better representation for material and texture and achieve better performance on the unsupervised change detection task. Our best results are achieved with train crop size of 7×7, and inference crop size of 9×9, while worst performance is achieved with largest training and inference receptive fields.

Impact of Modules. In Tab. 5 we study the impact of each module in our proposed method. We evaluate performance of the self-supervised change detection task (F-1 score of "change" class) as an ablation metric since it has strong transferability to material and texture representation learning. We consider all possible network combinations with patch-wise backbone, TeRN, and Surface Residual encoder. The patch-wise backbone corresponds to the network fed by patch-wise inputs, without TeRN or Surface Residual encoder. We then selectively add TeRN and Surface Residual Encoder to the network and record its performance. Every network combination was trained and evaluated with the same hyperparameters and procedure described in Sec. 4.1 and 4.2. It can be seen that each module provides incremental performance boost, with best performance achieved when both modules are implemented in the network.

Patch-wise Backbone	Texture Refinement	Surface Residual	F-1 Score (%)
✓			37.42
✓		✓	41.84
✓	✓		43.23
✓	✓	✓	**49.48**

Table 5. **Ablation study.** F-1 score of the "change" class of the Onera Satellite Change Detection dataset using the self-supervised approach with respect to modules used.

6. Conclusion

In this work, we present MATTER, a novel self-supervised method that learns material and texture based representation for multi-temporal, spatially aligned satellite imagery. By utilizing patch-wise inputs and our refinement network, we constrain the receptive field and enhance texture-essential features. Those are then mapped to residuals of learned clusters as an affinity measurement, which represents the material and texture composition of the sampled patch. Through our self-supervision pipeline, MATTER learns discriminative features for various material and texture surfaces, which are shown to have strong correlation to change (change of surface infers actual change), or can be used as pre-trained weights for other remote sensing tasks.

Acknowledgement This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2021-2011000005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied , of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] World meteorological organization (wmo) observing systems capability analysis and review (oscar) tool, https://space.oscar.wmo.int/satellites. 1

[2] Peri Akiva, Matthew Purri, Kristin Dana, Beth Tellman, and Tyler Anderson. H2o-net: Self-supervised flood segmentation via adversarial domain adaptation and label refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 111–122, 2021. 2

[3] Hamed Alemohammad and Kevin Booth. Landcovernet: A global benchmark land cover classification training dataset. *arXiv preprint arXiv:2012.03111*, 2020. 2

[4] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 1, 2, 5, 6, 7

[5] Seyed Majid Azimi, Corentin Henry, Lars Sommer, Arne Schumann, and Eleonora Vig. Skyscapes fine-grained semantic understanding of aerial scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7393–7403, 2019. 2

[6] Nadine Behrmann, Jurgen Gall, and Mehdi Noroozi. Unsupervised video representation learning by bidirectional feature prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1670–1679, 2021. 2

[7] Ioannis Brilakis, Lucio Soibelman, and Yoshihisa Shinagawa. Material-based construction site image retrieval. *Journal of computing in civil engineering*, 19(4):341–355, 2005. 2

[8] Ioannis K Brilakis and Lucio Soibelman. Shape-based retrieval of construction site photographs. *Journal of Computing in Civil Engineering*, 22(1):14–20, 2008. 2

[9] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 2, 5

[10] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3), 2010. 1

[11] Chao Chen, Yong-qiang Zhao, Li Luo, Dan Liu, and Quan Pan. Robust materials classification based on multispectral polarimetric brdf imagery. In *International Symposium on Photoelectronic Detection and Imaging 2009: Advances in Imaging Detectors and Applications*, volume 7384, page 73840T. International Society for Optics and Photonics, 2009. 3

[12] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020. 2

[13] Hongruixuan Chen, Chen Wu, Bo Du, and Liangpei Zhang. Deep siamese multi-scale convolutional network for change detection in multi-temporal vhr images. In *2019 10th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, pages 1–4. IEEE, 2019. 2

[14] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikäinen, Xilin Chen, and Wen Gao. Wld: A robust local image descriptor. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1705–1720, 2009. 3

[15] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1194–1206, 2020. 2

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5, 6, 8

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 5

[18] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 1

[19] Xi Chen. Nighttime lights and population migration: Revisiting classic demographic perspectives with an analysis of recent european data. *Remote Sensing*, 12(1):169, 2020. 1

[20] Zhile Chen, Feng Li, Yuhui Quan, Yong Xu, and Hui Ji. Deep texture recognition via exploiting cross-layer statistical self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5231–5240, 2021. 3

[21] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995. 5

[22] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3828–3836, 2015. 3

[23] Joseph Paul Cohen, Wei Ding, Caitlin Kuhlman, Aijun Chen, and Liping Di. Rapid building detection using machine learning. *Applied Intelligence*, 45(2):443–457, 2016. 1

[24] Pol R Coppin and Marvin E Bauer. Digital change detection in forest ecosystems with remote sensing imagery. *Remote sensing reviews*, 13(3-4):207–234, 1996. 2

[25] Oana G Cula and Kristin J Dana. Compact representation of bidirectional texture functions. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 2

[26] Oana G Cula and Kristin J Dana. Recognition methods for 3d textured surfaces. In *Human Vision and Electronic Imag-*

ing VI, volume 4299, pages 209–220. International Society for Optics and Photonics, 2001. 2

[27] Kristin J Dana. Computational texture and patterns: From textons to deep learning. *Synthesis Lectures on Computer Vision*, 8(3):1–113, 2018. 2

[28] Kristin J Dana, Bram Van Ginneken, Shree K Nayar, and Jan J Koenderink. Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG)*, 18(1):1–34, 1999. 2

[29] 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)*, pages 4063–4067. IEEE, 2018. 2

[30] 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*, pages 2115–2118, 2018. 5, 8

[31] Joseph DeGol, Mani Golparvar-Fard, and Derek Hoiem. Geometry-informed material recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1554–1562, 2016. 3

[32] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 172–181, 2018. 2

[33] Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6837–6845, 2017. 2

[34] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[35] 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*, 120:25–36, 2012. 1

[36] Liuyun Duan and Florent Lafarge. Towards large-scale city reconstruction from satellites. In *European Conference on Computer Vision*, pages 89–104. Springer, 2016. 1

[37] Rob Emanuele, Jon Duckworth, Victor Engmark, Simon Kassel, Kurt Schwehr, Víctor Olaya, Matthew Hanson, Pete Gadomski, Emmanuel Mathot, and Christopher Helm. Pystac. https://github.com/stac-utils/pystac, 2021. 5

[38] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010. 1

[39] Gabriele Facciolo, Carlo De Franchis, and Enric Meinhardt-Llopis. Automatic 3d reconstruction from multi-date satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 57–66, 2017. 1

[40] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2

[41] Giles M Foody. Remote sensing of tropical forest environments: towards the monitoring of environmental resources for sustainable development. *International journal of remote sensing*, 24(20):4035–4046, 2003. 1

[42] Thomas W Gillespie, Jasmine Chu, Elizabeth Frankenberg, and Duncan Thomas. Assessment and prediction of natural hazards from satellite imagery. *Progress in Physical Geography*, 31(5):459–470, 2007. 1

[43] Yimo Guo, Guoying Zhao, and Matti Pietikäinen. Discriminative features for texture description. *Pattern Recognition*, 45(10):3834–3843, 2012. 3

[44] Rohit Gupta and Mubarak Shah. Rescuenet: Joint building segmentation and damage assessment from satellite imagery. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4405–4411. IEEE, 2021. 2

[45] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 5, 6, 7

[46] 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*, pages 770–778, 2016. 6

[47] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 2

[48] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3304–3311. IEEE, 2010. 2

[49] Huiwei Jiang, Xiangyun Hu, Kun Li, Jinming Zhang, Jinqi Gong, and Mi Zhang. Pga-siamnet: Pyramid feature-based attention-guided siamese network for remote sensing orthoimagery building change detection. *Remote Sensing*, 12(3):484, 2020. 2

[50] Longlong Jing, Xiaodong Yang, Jingen Liu, and Yingli Tian. Self-supervised spatiotemporal feature learning via video rotation prediction. *arXiv preprint arXiv:1811.11387*, 2018. 2

[51] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998. 4

[52] Bela Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. 2

[53] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 1

[54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 6, 7

[55] Florent Lafarge, Xavier Descombes, Josiane Zerubia, and M-P Deseilligny. An automatic 3d city model: a bayesian approach using satellite images. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 2, pages II–II. IEEE, 2006. 1

[56] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 2

[57] Matthew J Leotta, Chengjiang Long, Bastien Jacquet, Matthieu Zins, Dan Lipsa, Jie Shan, Bo Xu, Zhixin Li, Xu Zhang, Shih-Fu Chang, et al. Urban semantic 3d reconstruction from multiview satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[58] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43(1):29–44, 2001. 2, 3

[59] Xiangtai Li, Xia Li, Li Zhang, Guangliang Cheng, Jianping Shi, Zhouchen Lin, Shaohua Tan, and Yunhai Tong. Improving semantic segmentation via decoupled body and edge supervision. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 435–452. Springer, 2020. 3

[60] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3

[61] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[62] Chao Liu and Jinwei Gu. Discriminative illumination: Per-pixel classification of raw materials based on optimal projections of spectral brdf. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):86–98, 2013. 3

[63] Ce Liu, Lavanya Sharan, Edward H Adelson, and Ruth Rosenholtz. Exploring features in a bayesian framework for material recognition. In *2010 ieee computer society conference on computer vision and pattern recognition*, pages 239–246. IEEE, 2010. 3

[64] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 4

[65] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 5

[66] Jitendra Malik, Serge Belongie, Jianbo Shi, and Thomas Leung. Textons, contours and regions: Cue integration in image segmentation. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 918–925. IEEE, 1999. 2

[67] William A Malila. Change vector analysis: an approach for detecting forest changes with landsat. In *LARS symposia*, page 385, 1980. 5

[68] Oscar Manas, 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 (ICCV)*, pages 9414–9423, October 2021. 1, 2, 5, 6

[69] Junhua Mao, Jun Zhu, and Alan L Yuille. An active patch model for real world texture and appearance classification. In *European Conference on Computer Vision*, pages 140–155. Springer, 2014. 3

[70] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[71] Shailesh Nayak. Use of satellite data in coastal mapping. *Indian Cartographer*, 22:147–157, 2002. 1

[72] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019. 6

[73] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1710, 2018. 2

[74] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002. 3

[75] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[76] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979. 5

[77] Maria Papadomanolaki, Sagar Verma, Maria Vakalopoulou, Siddharth Gupta, and Konstantinos Karantzalos. Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 214–217. IEEE, 2019. 2

[78] Daifeng Peng, Yongjun Zhang, and Haiyan Guan. End-to-end change detection for high resolution satellite images using improved unet++. *Remote Sensing*, 11(11):1382, 2019. 2

[79] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha

Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019. 1

[80] 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*, pages 234–241. Springer, 2015. 5

[81] David P Roy, Michael A Wulder, Thomas R Loveland, Curtis E Woodcock, Richard G Allen, Martha C Anderson, Dennis Helder, James R Irons, David M Johnson, Robert Kennedy, et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote sensing of Environment*, 145:154–172, 2014. 1

[82] Tim GJ Rudner, Marc Rußwurm, Jakub Fil, Ramona Pelich, Benjamin Bischke, Veronika Kopačková, and Piotr Biliński. Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 702–709, 2019. 2

[83] Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection. In *2020 IEEE International conference on robotics and automation (ICRA)*, pages 6861–6867. IEEE, 2020. 2

[84] Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16:41–48, 2004. 1

[85] Gabriel Schwartz and Ko Nishino. Visual material traits: Recognizing per-pixel material context. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 883–890, 2013. 3

[86] Gaurav Sharma, Sibt ul Hussain, and Frédéric Jurie. Local higher-order statistics (lhs) for texture categorization and facial analysis. In *European conference on computer vision*, pages 1–12. Springer, 2012. 3

[87] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European conference on computer vision*, pages 1–15. Springer, 2006. 2

[88] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1233–1240, 2013. 3

[89] Sergii Skakun, Nataliia Kussul, Andrii Shelestov, and Olga Kussul. Flood hazard and flood risk assessment using a time series of satellite images: A case study in namibia. *Risk Analysis*, 34(8):1521–1537, 2014. 1

[90] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019. 4

[91] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 2, 6

[92] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6

[93] Yong-Qiang Tan, Shang-Hua Gao, Xuan-Yi Li, Ming-Ming Cheng, and Bo Ren. Vecroad: Point-based iterative graph exploration for road graphs extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8910–8918, 2020. 2

[94] Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, BjÖrn Rommen, Nicolas Floury, Mike Brown, et al. Gmes sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012. 1

[95] Kimmo Valkealahti and Erkki Oja. Reduced multidimensional co-occurrence histograms in texture classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):90–94, 1998. 2

[96] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 2, 6

[97] CJ Van Westen. Remote sensing for natural disaster management. *International archives of photogrammetry and remote sensing*, 33(B7/4; PART 7):1609–1617, 2000. 1

[98] Manik Varma and Andrew Zisserman. Texture classification: Are filter banks necessary? In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–691. IEEE, 2003. 3

[99] Manik Varma and Andrew Zisserman. A statistical approach to material classification using image patch exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):2032–2047, 2008. 3

[100] Stefan Voigt, Fabio Giulio-Tonolo, Josh Lyons, Jan Kučera, Brenda Jones, Tobias Schneiderhan, Gabriel Platzeck, Kazuya Kaku, Manzul Kumar Hazarika, Lorant Czaran, et al. Global trends in satellite-based emergency mapping. *Science*, 353(6296):247–252, 2016. 1

[101] Jing Wang and Kristin J Dana. Hybrid textons: modeling surfaces with reflectance and geometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004. 2

[102] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11804–11813, 2021. 2

[103] Oliver Wang, Prabath Gunawardane, Steve Scher, and James Davis. Material classification using brdf slices. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2805–2811. IEEE, 2009. 3

[104] Michael Weinmann, Juergen Gall, and Reinhard Klein. Material classification based on training data synthesized using a btf database. In *European Conference on Computer Vision*, pages 156–171. Springer, 2014. 3

[105] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020. 2

[106] Jia Xue, Hang Zhang, Kristin Dana, and Ko Nishino. Differential angular imaging for material recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–773, 2017. 3

[107] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 3

[108] Hang Zhang, Kristin Dana, and Ko Nishino. Reflectance hashing for material recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3071–3080, 2015. 2

[109] Hang Zhang, Jia Xue, and Kristin Dana. Deep ten: Texture encoding network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 708–717, 2017. 3, 5

[110] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[111] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 267–283, 2018. 7

[112] Lanyun Zhu, Deyi Ji, Shiping Zhu, Weihao Gan, Wei Wu, and Junjie Yan. Learning statistical texture for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12537–12546, 2021. 3

[113] Song-Chun Zhu, Cheng-En Guo, Yizhou Wang, and Zijian Xu. What are textons? *International Journal of Computer Vision*, 62(1):121–143, 2005. 2

[114] Song Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (frame): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, 1998. 2

[115] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 593–602, 2019. 3