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

Peri Akiva
Rutgers University
peri.akiva@rutgers.edu

Matthew Purri
Rutgers University
matthew.purri@rutgers.edu

Matthew Leotta
Kitware Inc
matt.leotta@kitware.com

1. Implementation Details

1.1. Surface Encoding Layer

```
# Q: learned clusters, shape: [n_eps, D, 1, 1]
# feats: input image, shape: [b, D, 1, 1]
# scale: float scaling factor
# weights: learnable cluster weights

Q = torch.empty(n_eps, D, dtype=torch.float), requires_grad=True)
weights = torch.empty(n_eps, dtype=torch.float).uniform_(-1, 0), requires_grad=True)
def encode(feats):
    res = weights*(feats-Q).pow(2).sum(dim=3)
    res = Softmax(res, dim=2)
    cluster_wise_residual_encoding = (res*(feats-Q)).sum(dim=1)
    residual_encoding = cluster_wise_residual_encoding.mean(dim=1)
    return residual_encoding
```

Listing 1. **Surface Residual Encoding Simplified Pseudo-code**

1.2. Texture Refinement Network

```
# x: input image
# feats: low level features
# num_iterations: number of refinement operations
# dilations: list of dilations to use

def refine(x, feats):
    K = [] # list of kernels for each locations
    for iter in range(num_iterations):
        for dilation in dilations:
            for all locations in x:
                # standard deviation of local kernel
                x_std = local_std(x, dilation) # float

                # cosine similarity of center pixel and neighboring pixels
                x_cos_similarity = cosine_similarity(x, dilation) # [h,w]

                # calculate weights in given location
                kernel = -x_cos_similarity/(0.00001 + x_std^2)
                kernel = Softmax(kernel.mean(dim=1))
                K.append(kernel)

        # reduce K to match features dimensions
        K.sum(dim=2)
        feats *= K

    return feats
```

Listing 2. **Texture Refinement Network Simplified Pseudo-code**

2. Additional Results

2.1. Running times

One of the challenges our method presents is the lengthy running time for each epoch during the pre-training stage. This stems from the fact that each pixels are now represented by $h \times w$ sized crops, which are stacked along the batch dimension. In practice, this is equivalent to running the method on input of size of $[h \times w, h, w]$, where $h \times w$ is the batch size, and h, w are the height and width of the crop. A given image of size 32×32 is typically represented by $32 \times 32 = 1024$ pixels, while in our method an image is represented by $32 \times 32 \times 7 \times 7 = 50176$ pixels. That constraint translates to roughly ten times longer training time, as shown in Tab. 1. It is important to note that this limitation only occurs during the pre-training stage. During fine-tuning or inference steps, the method does not employ the window sampling approach, making it significantly faster than the pre-training stage, and similar to other fully supervised methods.

Method	supervision	crop-size	time/10 epochs	peak memory/GPU
Onera Change Detection Experiment				
DeepLab-v3 [1] (ImageNet)	$\mathcal{F}$	96×96	0h24	24G
SeCo [3]	$\mathcal{S}$	96×96	0h19	24G
SeCo [3]	$\mathcal{S} + \mathcal{F}$	96×96	0h23	24G
Patch-wise Backbone	$\mathcal{S}$	32×32	3h15	24G
Ours	$\mathcal{S}$	32×32	5h48	24G
Ours (fine-tuned)	$\mathcal{S} + \mathcal{F}$	96×96	0h24	24G
SpaceNet Building Segmentation Task				
DeepLab-v3 [1] (random)	$\mathcal{F}$	128×128	0h21	24G
DeepLab-v3 [1] (ImageNet)	$\mathcal{F}$	128×128	0h21	24G
Ours (fine-tuned)	$\mathcal{S} + \mathcal{F}$	128×128	0h21	24G

Table 1. Training run time calculated for each model and its corresponding remote sensing task. The self-supervision timing is calculated when running on the Onera Satellite Change Detection (OSCD) dataset. During fine-tuning or inference steps, the method does not employ the window sampling approach, making it significantly faster than the pre-training stage, and similar to other fully supervised methods.

2.2. Architecture Selection

A natural direction given the findings of Tab. 1 is to reduce the size of the pre-training model. One consideration is the complexity of the learned downstream task once the pre-trained model is achieved. For complex downstream tasks, such as 3D reconstruction or semantic segmentation, an encoder with large learning capacity is needed. For simpler tasks, it may be feasible to use smaller encoder. Since the downstream tasks presented in this work vary in complexity, we selected to use ResNet-34 as our backbone, with the corresponding running times shown in Tab. 1. However, the pre-training step can be performed with smaller encoders, such as fully connected neural network, ResNet-18, or similar. In Tab. 2 we study the effects of model size on our method's running times.

Backbone	parameters	crop-size	time/10 epochs	peak memory/GPU
3-layer NN	805,536	32×32	0h18	24G
ResNet-18	15,428,544	32×32	4h40	24G
ResNet-34	25,536,704	32×32	5h48	24G
ResNet-50	39,163,328	32×32	6h41	24G
ResNet-101	58,155,456	32×32	9h19	24G

Table 2. **Training** run time calculated for various backbones when using the Onera Satellite Change Detection (OSCD) dataset.

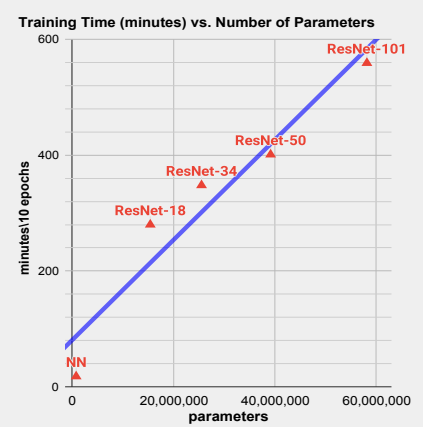

Figure 1. Plot of training time (in minutes per 10 training epochs) with respect to number of learnable model parameters.

2.3. Qualitative Results

We include additional qualitative results on Onera in Fig. 3 and SpaceNet in Fig 2.

2.4. Evaluating Visual Word Maps and Cluster Scatters

As mentioned in the paper, we measure the effectiveness of our approach to describe materials and textures through qualitative evaluation of 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 also desire to achieve distinctly mapped clusters with minimal overlaps. We visually compare classical textons, a patch-wise backbone, and our method with corresponding scatter plots (t-SNE, perplexity=100, steps=5000) in Fig. 4 and 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 the main paper. 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 our method is able to retain texture-essential features, and generalize surface representation which translates to superior surface-based visual word maps. The superiority in surface-representation learning can also be seen in the scatter plots, in which our method generates distinct cluster scatters corresponding to different texture and material combinations.

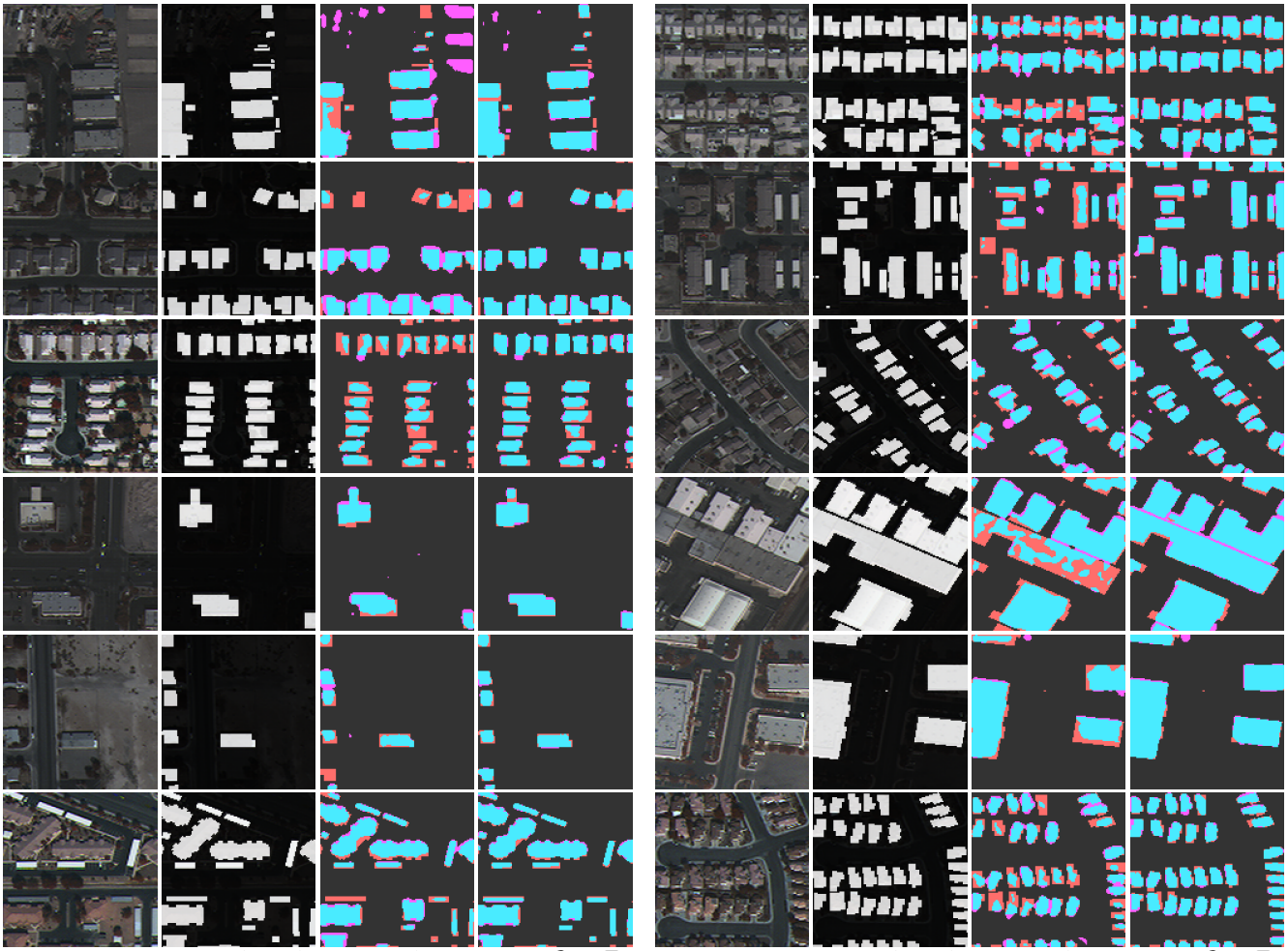

Figure 2. **Qualitative results** of our method on SpaceNet dataset [4]. Cyan, magenta, gray, and red colors represent true positive, false positive, true negative, and false negative respectively. Best viewed in zoom and color.

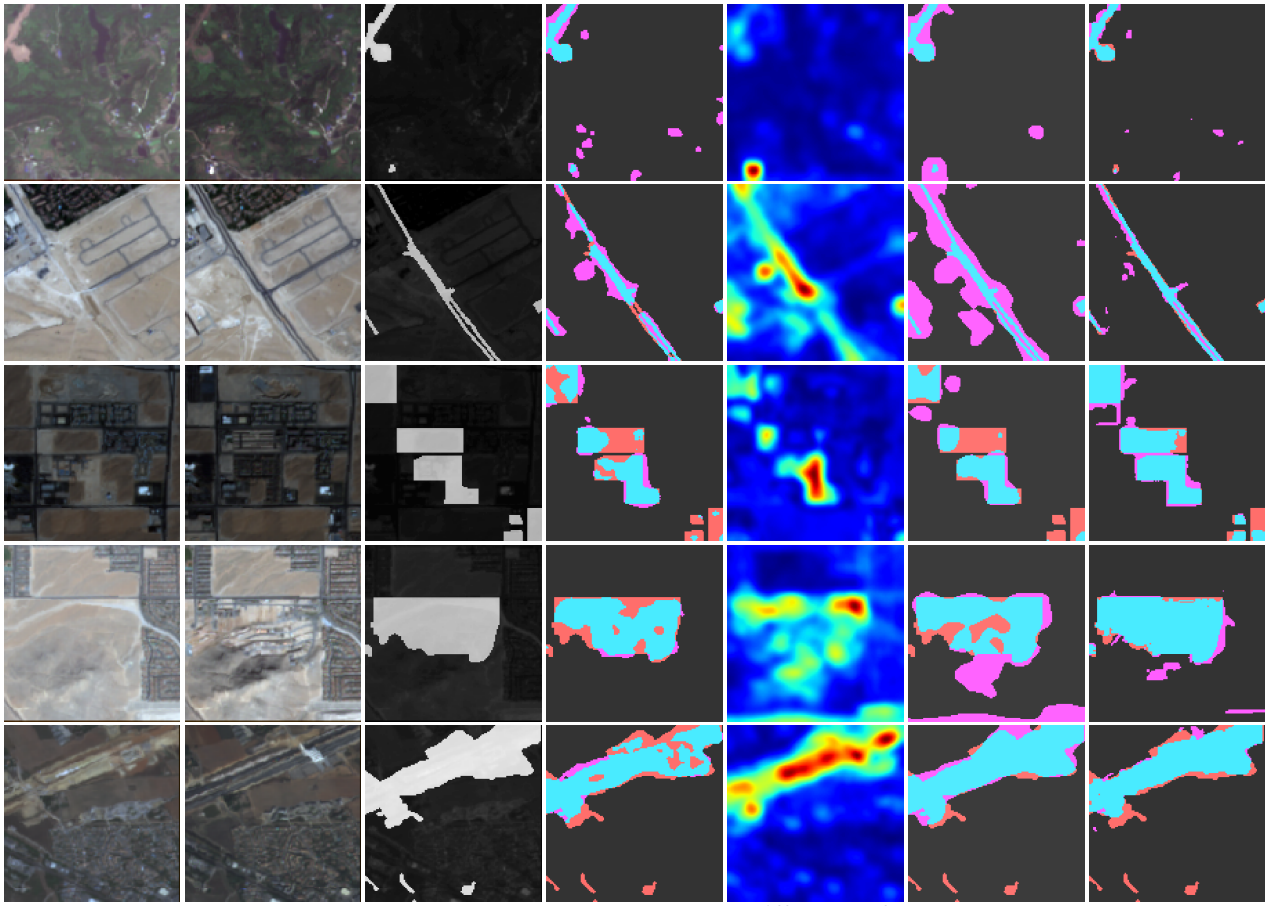

Figure 3. **Qualitative results** of our method on Onera Satellite Change Detection (OSCD) dataset [2]. It can be seen that our self-supervised alone is capable of detecting change only by inferring on the change of material and texture. The fine-tuned model is able to utilize the pre-trained material 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.

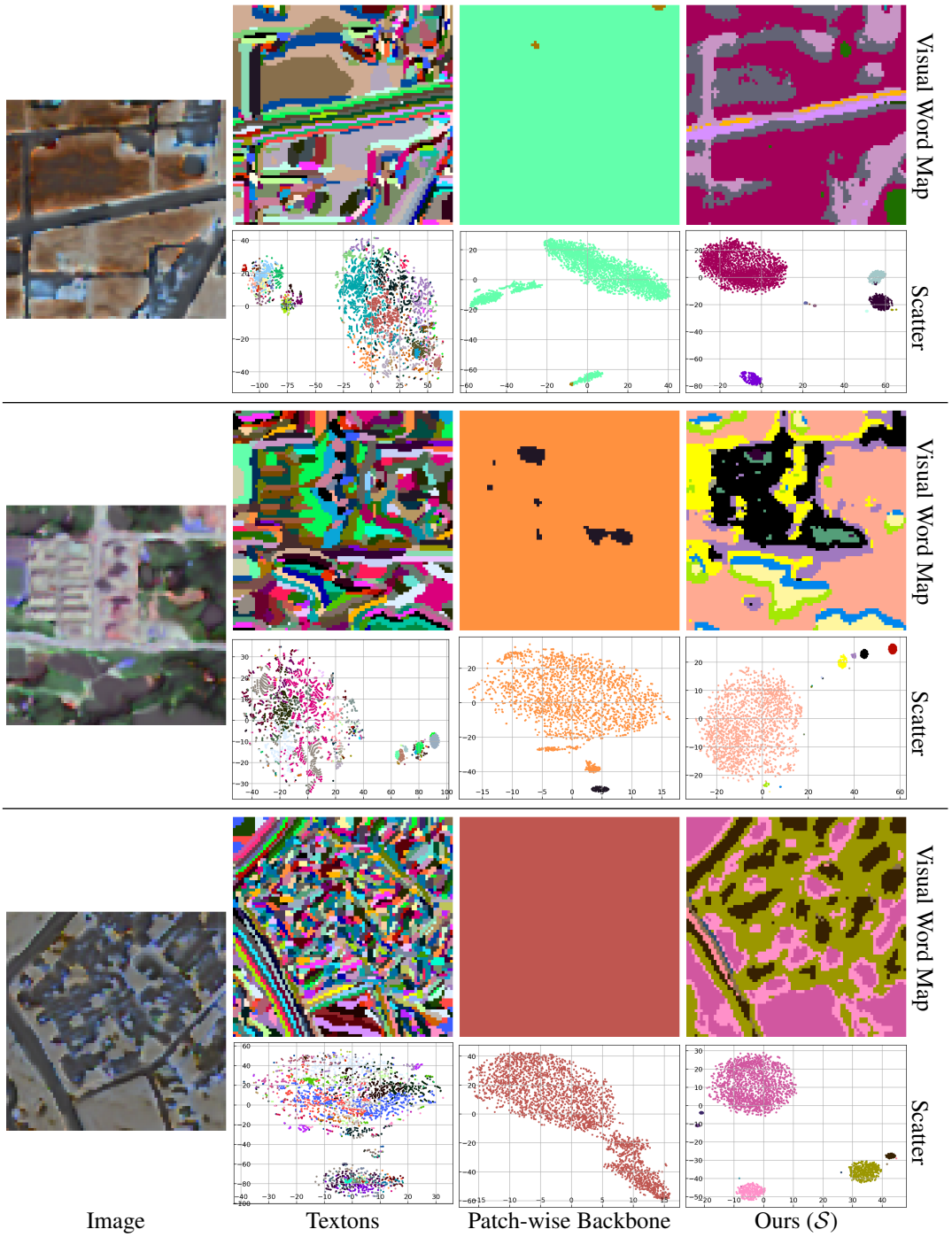

Figure 4. **Qualitative evaluation** of our generated material and texture based visual word maps and corresponding clusters. It can be seen that our method provides more descriptive surface-based features that are not highly sensitive to small texture irregularities like textons, or under-sensitive to structure changes like the patch-wise backbone. It can also be seen that our method provides more distinct cluster centers, better differentiating material and texture combinations. We randomly sample 30% of the pixels for the scatter plot visualization. Best viewed in zoom and color. Colors are random.

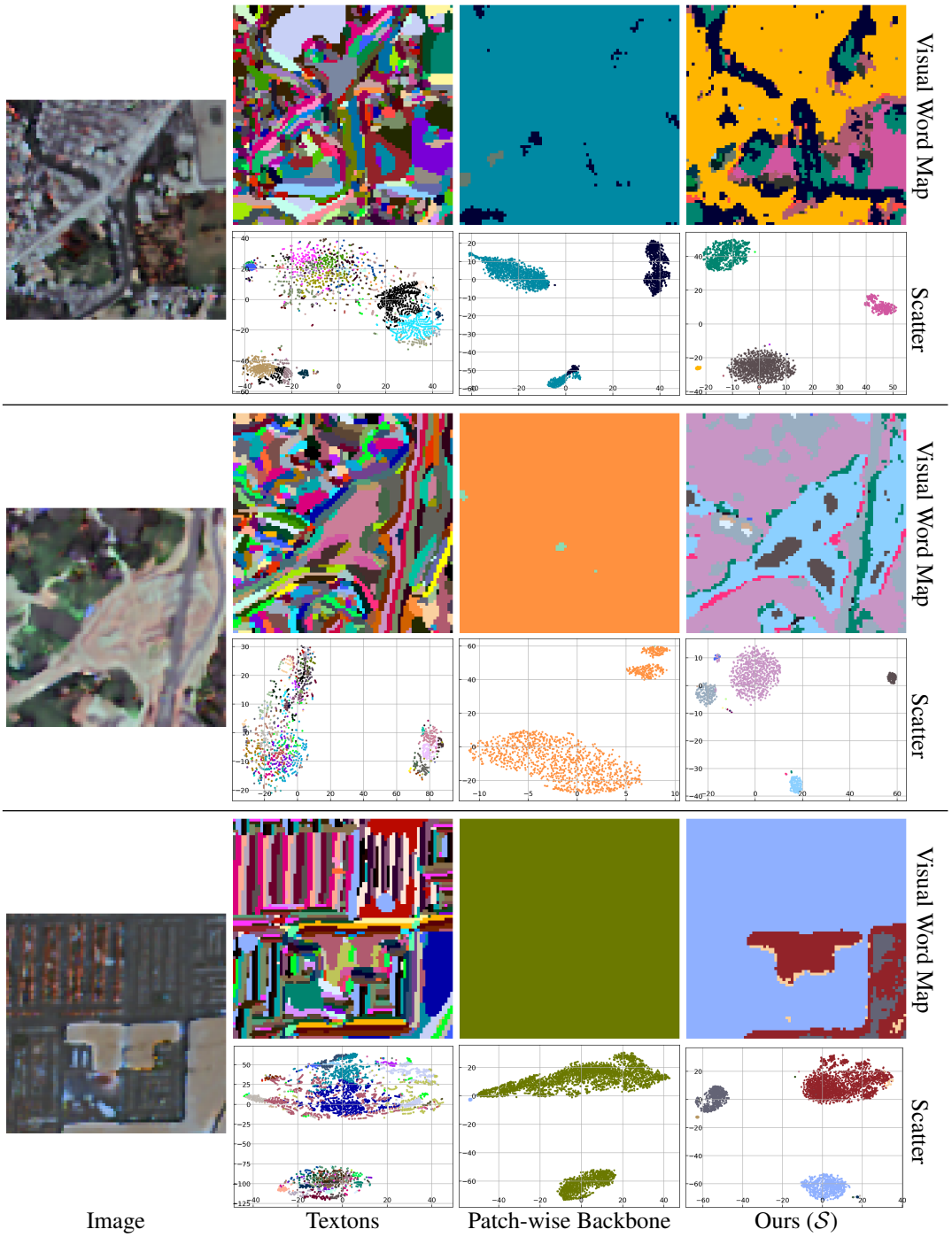

Figure 5. **Qualitative evaluation** of our generated material and texture based visual word maps and corresponding clusters. It can be seen that our method provides more descriptive surface-based features that are not highly sensitive to small texture irregularities like textons, or under-sensitive to structure changes like the patch-wise backbone. It can also be seen that our method provides more distinct cluster centers, better differentiating material and texture combinations. We randomly sample 30% of the pixels for the scatter plot visualization. Best viewed in zoom and color. Colors are random.

3. Self-Collected Dataset

3.1. Points of Interest


```
{
    "change_region": "no_change",
    "region_name": "usa_wyoming_shoshone",
    "geo_coords": [
        "43.943477_-109.492681"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "vietnam_00",
    "geo_coords": [
        "19.783545_105.355797"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "yemen_asadi-al-faya",
    "geo_coords": [
        "14.928610_50.284421"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "bulgaria_ozizare",
    "geo_coords": [
        "42.637424_25.953441"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
```

```
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "chile_aqi-keqiaole",
    "geo_coords": [
        "40.947455_85.205265"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "china_yuxiancun",
    "geo_coords": [
        "42.939327_124.852949"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "india_sundarban",
    "geo_coords": [
        "22.044656_88.757307"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "maludam_indonesia",
    "geo_coords": [
        "1.506800_111.165672"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
```

```
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "iran_naghab",
    "geo_coords": [
        "32.383612_59.689489"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "kazakhstan_kapaxar_karazhal",
    "geo_coords": [
        "47.981633_70.952032"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "madagascar_00",
    "geo_coords": [
        " -18.969615_49.046689"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "mali_fodebougou",
    "geo_coords": [
        "12.533564_-9.805519"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
```

```
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "mexico_santa-elena",
    "geo_coords": [
        "23.636026_-104.527631"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "mongolia_ronin-hur",
    "geo_coords": [
        "48.459630_110.512820"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "morocco_brarha",
    "geo_coords": [
        "34.453855_-4.287849"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "myanmar_00",
    "geo_coords": [
        "18.233579_97.030426"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
```

```
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "north-korea_00",
    "geo_coords": [
        "40.282847_128.339383"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "qatar_quasil",
    "geo_coords": [
        "25.282863_50.784150"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "russia_burukan",
    "geo_coords": [
        "53.061472_136.035468"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "saudi-arabia_bir-jaydah",
    "geo_coords": [
        "36.139782_37.488645"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
```

```
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "south-sudan_boma",
    "geo_coords": [
        "6.077807_34.132967"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "spain_urda",
    "geo_coords": [
        "39.455428_-3.730460"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "usa_nevada_tonopah",
    "geo_coords": [
        "38.240222_-117.320516"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "usa_new-york_fabius",
    "geo_coords": [
        "42.808692_-75.985309"
    ],
    "collection": "sentinel-s2-l2a-cogs",
```

```
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "usa_north-carolina_mt-mitchell",
    "geo_coords": [
        "35.765959_-82.254093"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "usa_oklahoma_tahlequah",
    "geo_coords": [
        "36.035096_-94.811176"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}

{
    "change_region": "no_change",
    "region_name": "usa_texas_dumont",
    "geo_coords": [
        "33.800483_-100.563941"
    ],
    "collection": "sentinel-s2-l2a-cogs",
    "start_date": "01/01/2017",
    "end_date": "01/01/2020",
    "crop_size": 1096,
    "radius": 0.001,
    "max_cloud_cover": 20,
    "min_data_coverage": 80,
    "min_num_images": 2,
    "max_num_images": 100,
}
```


References

[1] 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. 2, 4, 5

[2] 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

[3] 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. 2

[4] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 4