Toulouse Hyperspectral Data Set: A benchmark data set to assess semi-supervised spectral representation learning and pixel-wise classification techniques

Romain Thoreau, Laurent Risser, Véronique Achard, Béatrice Berthelot, Xavier Briottet

TOULOUSE HYPERSPECTRAL DATA SET:
A BENCHMARK DATA SET TO ASSESS SEMI-SUPERVISED SPECTRAL REPRESENTATION LEARNING AND PIXEL-WISE CLASSIFICATION TECHNIQUES

A PREPRINT

Romain Thoreau[1,*], **Laurent Risser**[4], **Véronique Achard**[2], **Béatrice Berthelot**[3], **and Xavier Briottet**[2]

[1]CNES, FR-31401 Toulouse, France
[2]ONERA-DOTA, University of Toulouse, FR-31055 Toulouse, France
[3]Magellium, 31520 Ramonville Saint-Agne, France
[4]Toulouse Mathematics Institute (UMR 5219), CNRS, University of Toulouse, F-31062 Toulouse, France

ABSTRACT

Airborne hyperspectral images can be used to map the land cover in large urban areas, thanks to their very high spatial and spectral resolutions on a wide spectral domain. While the spectral dimension of hyperspectral images is highly informative of the chemical composition of the land surface, the use of state-of-the-art machine learning algorithms to map the land cover has been dramatically limited by the availability of training data. To cope with the scarcity of annotations, semi-supervised and self-supervised techniques have lately raised a lot of interest in the community. Yet, the publicly available hyperspectral data sets commonly used to benchmark machine learning models are not totally suited to evaluate their generalization performances due to one or several of the following properties: a limited geographical coverage (which does not reflect the spectral diversity in metropolitan areas), a small number of land cover classes and a lack of appropriate standard train / test splits for semi-supervised and self-supervised learning. Therefore, we release in this paper the Toulouse Hyperspectral Data Set that stands out from other data sets in the above-mentioned respects in order to meet key issues in spectral representation learning and classification over large-scale hyperspectral images with very few labeled pixels. Besides, we discuss and experiment self-supervised techniques for spectral representation learning, including the Masked Autoencoder [1], and establish a baseline for pixel-wise classification achieving 85% overall accuracy and 77% F1 score. The Toulouse Hyperspectral Data Set and our code are publicly available at www.toulouse-hyperspectral-data-set.com and www.github.com/Romain3Ch216/tlse-experiments/, respectively.

1 Introduction

Airborne hyperspectral images are a critical resource for the land cover mapping of large urban areas. As much as artificial impermeable surfaces impact watershed hydrology (particularly droughts and floods) [2, 3, 4, 5, 6], urban heat island effects [7, 8] and soil carbon uptake [9, 10, 11, 12], providing public authorities and scientists with accurate maps of land surface materials is a key issue to mitigate the effects of urban sprawl. Hyperspectral sensors measure the radiant flux reflected by the ground and by the atmosphere for several hundreds of narrow and contiguous spectral intervals in the visible and the reflective part of the infrared. While the spectral radiance measured at the sensor level partially depends on the atmosphere (*i.e.* its water vapor concentration, type and concentration of aerosols, etc.), atmospheric correction algorithms such as [13, 14] can estimate the pixel-wise reflectance at the ground level, which is the ratio of the reflected radiant flux on the incident radiant flux averaged over the pixel surface. Reflectance is intrinsic to the chemical composition of matter, and is therefore very informative of the land cover. In contrast, the spatial information brings little information (for instance, an orange tennis court could actually either be in porous concrete or in a synthetic

*Corresponding author: romain.thoreau@cnes.fr
Work done while at ONERA / Magellium

track, which would be indistinguishable with a conventional RGB image) though a large-scale context may sometimes raise ambiguities.

The main hindrance to the pixel-wise classification of hyperspectral images holds in the scarcity of labeled data. Labeling pixels of a hyperspectral image with land cover classes of low abstraction such as *gravel* or *asphalt* indeed requires expert knowledge and expensive field campaigns. Therefore, the ground truth that usually contains at most 1% of the pixels barely represents the spectral variability of the image.

To that extent, publicly available hyperspectral data sets have fueled a great deal of research in several directions including Active Learning [15, 16], unsupervised / self-supervised and semi-supervised learning [17, 18, 19, 20], to train machine learning models with few labeled samples that are robust to spectral intra-class variations. Nevertheless, if the Standardized Remote Sensing Data Website[1] of the IEEE Geoscience and Remote Sensing Society (GRSS) provides a set of community data sets and a tool to evaluate classifiers on undisclosed test samples, providing the ground truth of public data sets with standard training sets (divided in a subset for the supervised part and another subset for the unsupervised part) spatially disjoint to test sets would foster reproducible and fair evaluations of semi-supervised techniques. We emphasize that several works including [21, 22, 23] showed that random sampling of the training and test sets over-estimates the generalization performances of classifiers, which is partly explained by the fact that pixels belonging to the same semantic class but sampled in different geographical areas are obviously more likely to have different spectral signatures than neighboring pixels.

Therefore, we introduce the Toulouse Hyperspectral Data Set[2] that better represents the complexity of the land cover in large urban areas compared to currently public data sets, and provide standard training and test sets specifically defined to assess semi/self-supervised representation learning and pixel-wise classification techniques. First, we present the construction and the properties of the Toulouse data set in section 2. Second, we provide a qualitative comparison with the Pavia University[3] and Houston University [24] data sets in section 3. Third, we discuss and experiment[4] self-supervised techniques for pixel-wise classification in section 4. Finally, we conclude in section 5.

2 Construction and properties of the Toulouse Hyperspectral Data Set

In the context of the AI4GEO consortium[5] and the CAMCATT/AI4GEO field campaign [25], a hyperspectral image was acquired over the city of Toulouse the 15^{th} of June 2021 around 11am UTC with a AisaFENIX 1K camera (which has a spectral range from 0.4 µm to 2.5 µm with a 3.6 nm spectral resolution in the VNIR[6] and a 7.8 nm specral resolution in the SWIR[7], a swath of 1024 m and a ground sampling distance of 1 m) that was on-board a Safire aircraft that flew at 1,500 m above the ground level. The hyperspectral data was converted in radiance at aircraft level through radiometric and geometric corrections. Then, the radiance image was converted to surface reflectance with the atmospheric correction algorithm COCHISE [13]. Hyperspectral surface reflectances were also acquired on-ground with three ASD spectrometers in the range of 0.4 µm to 2.5 µm. Reflectance spectra of *clear paving stone*, *brown paving stone* and *red porous concrete* with pictures of the materials are shown in Fig. 1 as examples. These in-situ measurements have served as a basis to define a land cover nomenclature (several materials with in-situ measurements are not in our nomenclature because they were on walls or on small manhole covers for instance) and to build a ground truth by photo-interpretation, additional field campaigns as well as with the help of exogenous data. Precisely, we used the "Registre Parcellaire Graphique"[8], a geographical information system that informs the crop type of agricultural plots over France, to annotate cultivated fields. For a full description of the data acquired in the CAMCATT / AI4GEO campaign, we refer the reader to the data paper [25].

2.1 Land cover ground truth

In total, we define the land cover nomenclature with 32 classes, dividing into 16 impermeable materials and 16 permeable materials, that we organize in a hierarchical nomenclature as shown in Fig. 2. Approximately $380,000$ pixels are labeled with a land cover class. In contrast to conventional semantic segmentation data sets, our ground truth is made of sparse annotations, *i.e.* polygons that are disconnected from each other. We annotated the pixels with

[1] `http://dase.grss-ieee.org/index.php`

[2] The Toulouse Hyperspectral Data Set is available at www.toulouse-hyperspectral-data-set.com

[3] `https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes`

[4] Code to reproduce our experiments is available at www.github.com/Romain3Ch216/tlse-experiments/

[5] `https://www.ai4geo.eu/en`

[6] Visible and near infrared

[7] Short-wave infrared

[8] `https://artificialisation.developpement-durable.gouv.fr/bases-donnees/registre-parcellaire-graphique`

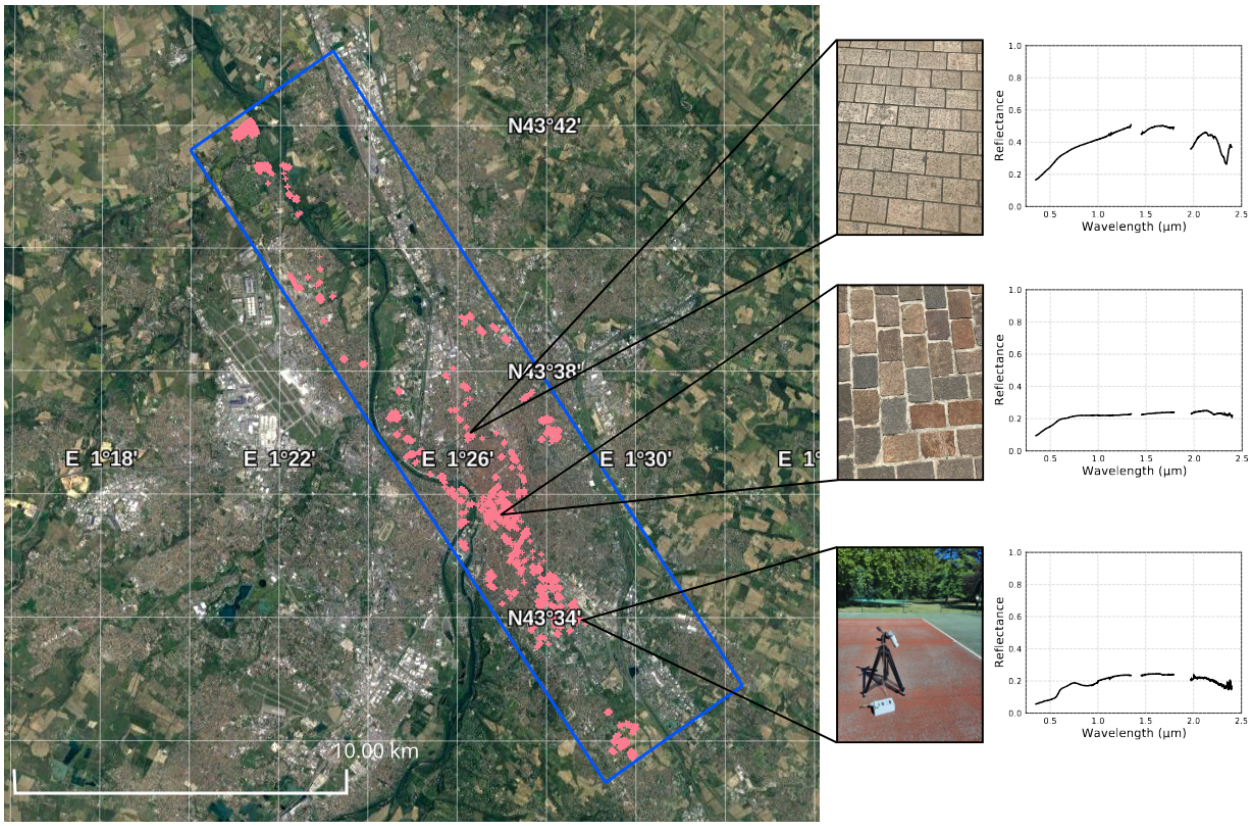


Figure 1: Area of Toulouse covered by the AI4GEO airborne hyperspectral image (in blue), our annotated ground truth (in red), and examples of reflectance spectra (clear paving stone, brown paving stone and red porous concrete, from top to bottom) measured on field with ASD spectrometers during the CAMCATT-AI4GEO field campaign [25].

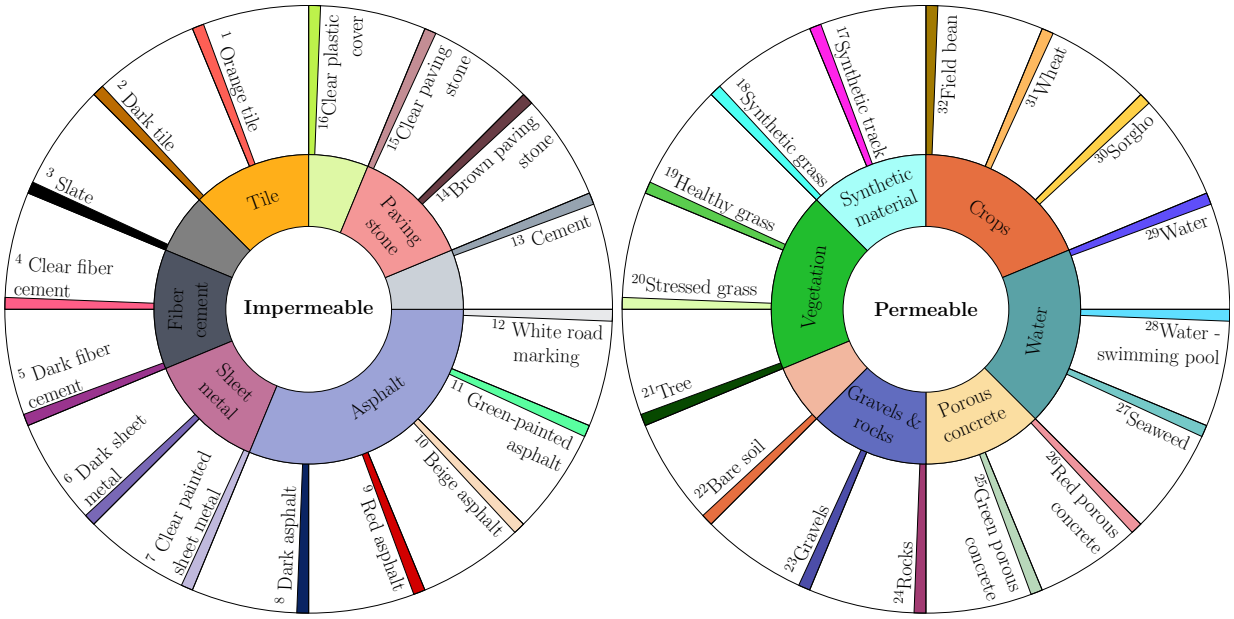


Figure 2: Land cover nomenclature of Toulouse Hyperspectral Data Set

particular attention to the exactness of the land cover labels. In particular, we omitted from the ground truth, as much as possible, mixed pixels (*i.e.* pixels with several materials) whose reflectance spectra are a combination of various spectra. Random spectra of the classes *orange tile* and *synthetic track* are shown in Fig. 3 while the other classes are shown in the appendices in Fig. 12.

2.2 Standard training and test sets for semi-supervised learning

We provide 8 spatially disjoint splits of the ground truth divided into:

- A labeled training set,
- An unlabeled training set divided itself in:
  - A set exclusively composed of pixels belonging to land cover classes defined in the nomenclature, called the *labeled pool*,
  - A larger set of truly unlabeled pixels that may not belong to known land cover classes, called the *unlabeled pool*,
- A validation set,
- A test set.

As much as the spectral intra-class variability of hyperspectral data comes from 1) variations in illumination conditions, 2) non-lambertian effects and slight variations in the material chemical composition (for instance due to variations in water or chlorophyll content of different trees from the same specie, variations of tar due to aging or to different environmental exposures) and 3) from larger variations in the material composition, due to the fact that the nomenclature gathers different materials under the same class (for instance different tree species gathered in a unique class) [26], there are high correlations between the intra-class spectral variability and the geographical location of pixels. Therefore, we suggest to foster the statistical independence of the training, validation and test sets by spatially separating them (see an example in Fig. 13 in the appendices).

To perform the spatially disjoint splits of the ground truth while ensuring that each class is distributed in appropriate proportions in the labeled training set, the validation set and the test set, we group neighboring polygons together in n_{groups} groups and define the ground truth split as mixed integer problem. Precisely, we aim to assign to each group of neighboring polygons a set, among the labeled training set (designated by index 1), the labeled pool (designated by index 2), the validation set (designated by index 3) and the test set (designated by index 4), while the unlabeled pool is left out. The main constraint of the problem is that each set $s \in \{1, 3, 4\}$ should contain, for each class, at least p_s percent of the total number of labeled pixels. We define the mixed integer problem as follows where u_{ij} is 1 if group i is in set j, $P[i, k]$ is the number of pixels of class k in group i and $\mathcal{S} = \{1, 3, 4\}$:

$$\min_{u} \sum_{s \in \mathcal{S}} \sum_{i=1}^{n_{groups}} \sum_{k=1}^{c} P[i,k] \cdot u_{is} \tag{1}$$

$$\text{subject to:} \sum_{j=1}^{4} u_{ij} = 1 \quad \textit{i.e. each group should be at least in one set} \tag{2}$$

$$\forall s \in \mathcal{S}, \forall k \in \{1, \dots, c\}, \sum_{i=1}^{n_{groups}} P[i,k] \cdot u_{is} \geq p_s \cdot \sum_{i=1}^{n_{groups}} P[i,k] \tag{3}$$

i.e. for each class k, the proportion of pixels in set s should be greater than the proportion p_s

In the standard splits that we provide, 13%, 29%, 14% and 46% of the labeled samples are in the labeled training set, the labeled pool, the validation set and the test set, respectively, in average (with regard to classes and to the 8 splits). In addition, the unlabeled pool contains nearly 2.6 million pixels. Hence, the labeled pixels used for training only represent 7% of all data. The decision to divide the ground truth into splits of 13%, 29%, 14% and 46% of the total number of labeled pixels stems from the following considerations, in order of priority: having a representative test set, having a representative validation set, having a sufficient number of samples in the training set for supervision to be relevant. However, the precise choice of the average proportions in each set is arbitrary and does not rely on a statistical analysis. In addition, we chose to provide only height splits of the ground truth because we could not find other solutions of the mixed integer problem that were significantly different from each other.

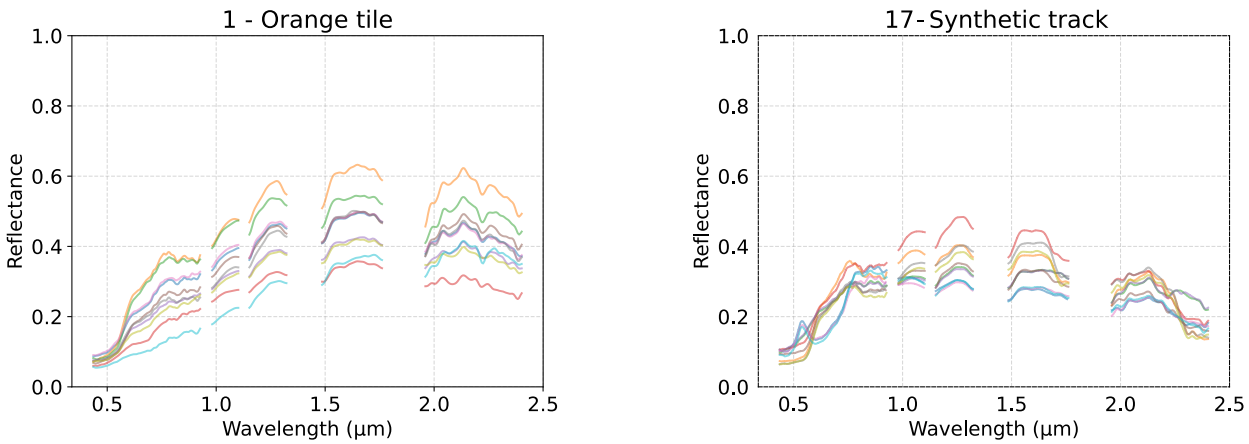


Figure 3: Random spectra of the classes *orange tile* and *synthetic track*.

2.3 Python package

Since the hyperspectral images of the Toulouse data set are too large to be loaded into memory all at once, we release TlseHypDataSet, a Python library whose main objective is to enable easy and rapid loading of the data into Pytorch[9] loaders.

```
import torch
from TlseHypDataSet.tlse_hyp_data_set import TlseHypDataSet
from TlseHypDataSet.utils.dataset import DisjointDataSplit

dataset = TlseHypDataSet('/path/to/dataset/', patch_size=1)

# Load the first standard ground truth split
ground_truth_split = DisjointDataSplit(dataset, split=dataset.standard_splits[0])

train_loader = torch.utils.data.DataLoader(ground_truth_split.sets_['train'],
                                           shuffle=True, batch_size=1024)

for epoch in range(100):
    for samples, labels in train_loader:
        ...
```

Figure 4: Minimal example of Python code to load data in Pytorch loaders with the TlseHypDataSet library

3 Comparison with publicly available data sets

In this section, we compare different properties of the Toulouse data set that are meaningful for machine learning applications, to those of the Pavia University data set and the Houston University data set. These two data sets are indeed widely used in the community, and cover urban or peri-urban areas as well.

Spectral and spatial information Tab. 1 recaps the spatial and spectral resolutions as well as the spectral domains of the compared data sets. While the spatial resolution is roughly the same, the Toulouse image has a much wider spectral domain which significantly brings more discriminating information for the mapping of the land cover, especially for mineral materials which are numerous in urban areas.

Spectral and spatial variability Comparing data sets with different spatial and spectral resolutions as well as different spectral domains is not straightforward. If the resolutions and spectral domains were the same, we could consider to learn representations of hyperspectral patches with an autoencoder and visualize the representations with a 2-dimensional t-SNE [27] transformation to qualitatively compare the data sets as in [28]. Instead, we suggest to represent 64×64

[9] https://pytorch.org/docs/stable/index.html

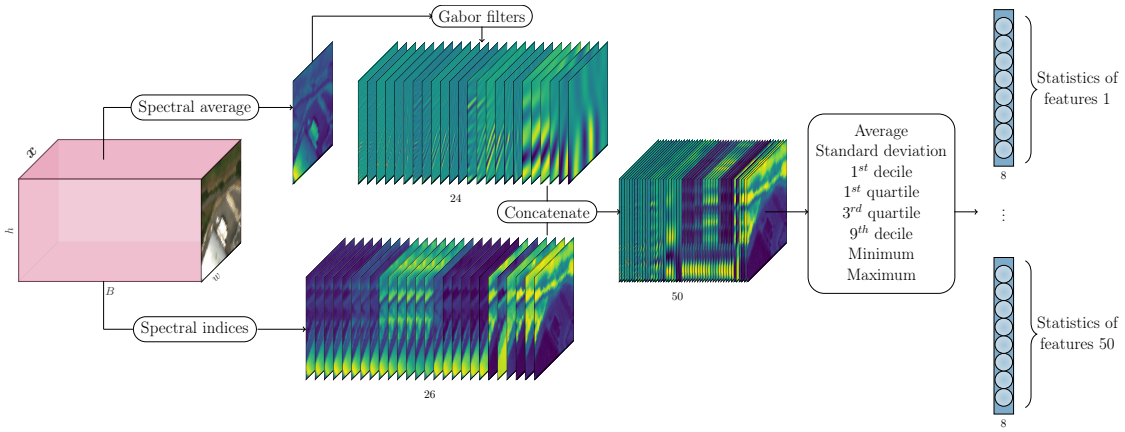


Figure 5: Illustration of our hand-crafted patch-wise feature extraction technique. The input is a 64 by 64 pixel hyperspectral patch. On one side, spectral indices (which include a selection of 20 spectral bands uniformly sampled along the spectral domain) are computed, resulting in 26 maps of 64 by 64 pixels. On the other side, the patch, averaged along the spectral dimension, is filtered by Gabor filters with 4 different frequencies (from 1 m^{-1} to 10 m^{-1}) and 6 different orientations, resulting in 24 maps. From every maps, spatial statistics are computed, resulting in a 400-dimensional feature.

pixel hyperspectral patches (that are, at least, partially labeled) with hand-crafted features. To summarize the spectral information of patches, we compute spectral indices (*i.e.* linear combinations of spectral bands, sometimes normalized) that use spectral channels included in the smallest spectral domain of the compared data sets (here Pavia University), precisely the NDVI [29], ANVI [30], CI [31], NDVI_RE [32], VgNIR_BI [33], SAVI [34], and uniformly sample 20 bands over the whole domain. To summarize the spatial information of patches, we compute 24 predefined Gabor filters on the spectral average of the patch (4 different frequencies (from 1 m^{-1} to 10 m^{-1}) and 6 different orientations). Then, we concatenate the spectral and spatial features and compute patch-wise statistics (the average, the standard deviation, the first and last deciles, the first and last quartiles, as well as the minimum and the maximum), yielding a 400-dimensional feature for each patch. The representation method is illustrated in Fig. 5.

Fig. 6 shows a t-SNE visualization of the hand-crafted features. First, the Toulouse data set clearly occupies more space than the Pavia and Houston data sets. Second, it seems that similar landscapes are projected in the same regions, as show a few false-color composition of hyperspectral patches close in the 2-dimensional space but taken from different data sets. In order to compare the contribution of the spatial information, spectral information and spectral information in the SWIR[10] (the spectral domains of Pavia and Houston are limited to the NIR[11]), we made additional comparisons presented in the appendices in Fig. 14. We found that the larger variability of Toulouse mainly comes from the spectral dimension and is mainly a consequence of a larger variability of the land cover.

Class distribution Tab. 2 recaps statistics about the class distribution of the data sets, including the imbalance ratio, which is the ratio of the number of samples in the largest class over the number of samples in the smallest class. The Houston and Toulouse data sets are particularly imbalanced: the biggest class of Houston accounts for 43% of the samples while the biggest class of Toulouse accounts for 24% of the samples. Yet, Fig. 7 shows that the Toulouse data set particularly exhibits a long-tailed class distribution, which is representative of life-like scenarios. Data sets with a long-tailed class distribution are data sets where a small number of classes account for a large part of samples while a large number of classes have only few examples [35]. The difference between usual class imbalance and long-tailed class imbalance mainly lies in the number of classes with few samples.

Noisy labels Compared to the Houston data set, we argue that the Toulouse data set contains less noise in the ground truth. Although the Houston data set contains more than a half million labeled pixels, many pixels are wrongly labeled, or are at least misleading as there are a mix of several materials. This noise in the ground truth is detrimental to classification models that put more emphasize on the spectral information rather than the spatial information. We show in the appendices a few examples of noisy labels in Fig. 15 and illustrate in Fig. 16 the care we took to avoid noisy labels in the Toulouse ground truth.

[10] short-wave infrared
[11] near infrared

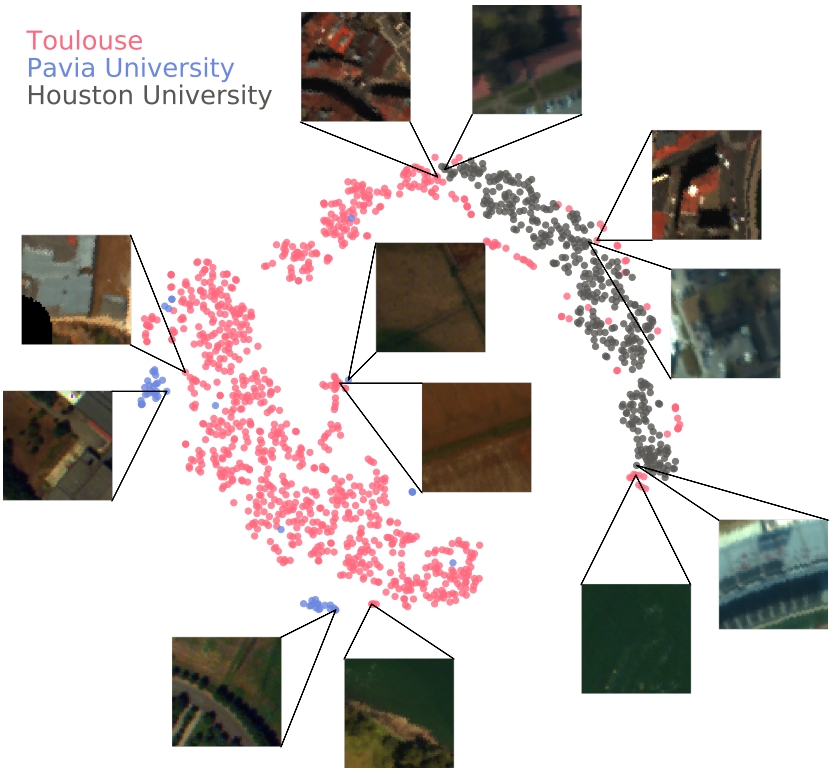


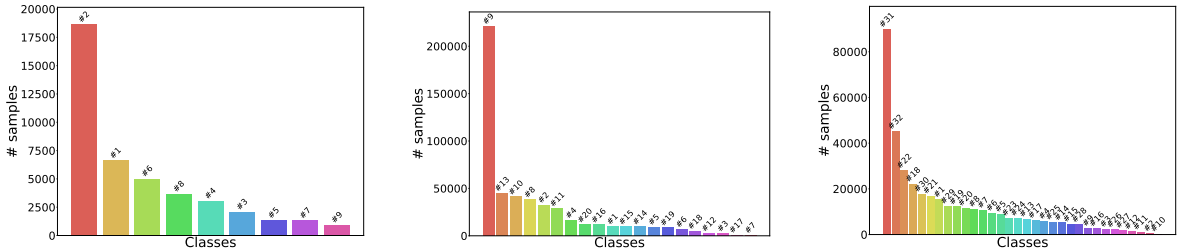
Figure 6: t-SNE visualization of hand-crafted representations of 64×64 pixel hyperspectral patches from the Pavia University, Houston University and Toulouse data sets.

Table 1: Spectral and spatial characteristics of the Pavia, Houston and Toulouse hyperspectral data sets.

Data set	**GSD**	**Spectral domain**	**Spectral resolution**	**Spectral dimension**
Pavia University	1.3 m	0.43 µm - 0.86 µm	4 nm	103
Houston University	1 m	0.38 µm - 1.0 µm	3.5 nm	48
Toulouse	1 m	0.4 µm - 2.5 µm	3.6 nm (VNIR) & 7.8 nm (SWIR)	310

Table 2: Statistics of the Pavia, Houston and Toulouse hyperspectral data sets.

Data set	**# classes**	**# data samples**	**Imbalance ratio**
Pavia University	9	43,000	20
Houston University	20	510,000	329
Toulouse	32	380, 000	225



(a) Pavia University (b) Houston university (c) Toulouse

Figure 7: Number of samples by class sorted from the most to the less represented.

4 Self-supervision for spectral representation learning

In this section, we review state-of-the-art self-supervised techniques, discuss their applicability to hyperspectral data and establish a first semi-supervised baseline on the Toulouse Data Set that shall serve as comparison for future works.

4.1 Self-supervised learning: an overview

Common semi-supervised learning techniques jointly optimize machine learning models on a supervised task and on an auxiliary unsupervised task. A common choice for the auxiliary task is the reconstruction of the (high dimensional) data from a (low dimensional) representation. However, a wide range of new approaches, known as self-supervised learning techniques, have recently emerged by introducing more useful auxiliary tasks. Self-supervision consists in training the model on a supervised pretext task for which labels are automatically generated.

In computer vision, a variety of pretext tasks have emerged in order to learn similar visual representations for different views of the same data. Those tasks include rotation self-supervision [36], exemplar self-supervision [37], contrastive learning [38, 39, 40, 41, 42] or self-supervised knowledge distillation [43]. Rotation self-supervision aims to predict the rotation (0°, 90°, 180° or 270°) applied to an image. Exemplar self-supervision gathers transformations of the same data sample under one class and trains the model on the subsequent classification task. Contrastive learning consists in learning similar representations of automatically-generated pairs of data samples with common semantic properties (positive pairs), while learning dissimilar representations of unrelated data samples (negative pairs). In particular, the seminal framework of [42] is based on stochastic data augmentation (specifically the combination of random cropping, random color distortions and random Gaussian blur) and a contrastive loss function that aims to identify a positive pair within a batch of samples. Self-supervised knowledge distillation takes inspiration from knowledge distillation [44] by taking a teacher network to supervise a student network that sees different transformations of the input data. Self-supervised learning has been integrated in semi-supervised learning frameworks, for instance in [45] or [46] that combine cluster-based self-supervision with class prototype learning.

To summarize, these strategies of self-supervision generate different views of the data (far apart in the input space) with the same semantic content thanks to data augmentation techniques, and rely on various tricks to prevent representation collapse (when the encoder learns useless representations that nevertheless minimize the training objective) [47]. They have experimentally demonstrated benefits on RGB natural images in term of robustness to various spatial contexts (*e.g.* pose, orientation, background...), which are the main cause of intra-class variability (*e.g.* a car in a parking lot and the same car on a freeway). In contrast, the spectral intra-class variability of hyperspectral data does not depend on context and divides into *physics*, *intrinsic* and *semantic* intra-class variabilities, as discussed in section 2.2.

Many self-supervised learning techniques have been directly applied to hyperspectral data, such as [48] that augment hyperspectral patches with random cropping and random color distortions, and [49] that apply random rotations as well as spectral random noise and spectral mirroring, both in the framework of self-supervised contrastive learning. Here we shall note that the physical soundness of random color distortions and spectral mirroring should be questioned, and that those augmentations are unlikely to preserve the semantic information of the data. Other works have introduced data augmentation techniques that are specific to hyperspectral data, such as [50] that creates positive pairs of data samples by sampling monochromatic images from neighboring spectral channels, or [51] that pairs spectrally close samples. All in all, most attention has been put on learning spatial-spectral representations with self-supervised techniques, often based on data augmentation, though we believe that finding a data augmentation technique that is faithful to realistic *physics*, *intrinsic* and *semantic* intra-class variations is not trivial, if not impossible. As a matter of fact, the true illumination conditions (that depend on topography) should be known to simulate realistic *physics* variations, while *intrinsic* and *semantic* spectral variations are, by nature, intrinsic to the chemical composition of matter.

Therefore, we focus on two self-supervised techniques that do not rely on data augmentation: Deep Clustering [52] and Masked Autoencoders (MAE) [1].
Deep Clustering is a seminal technique in cluster-based self-supervision that uses pseudo-labels derived from a clustering algorithm to supervise the training. The core idea behind Deep Clustering is that the use of convolutional layers for learning visual representations is a very strong inductive bias about the data structure. As a matter of fact, a dense network fine-tuned on top of the frozen features computed by a randomly initialized CNN, namely AlexNet, achieved 12% accuracy on ImageNet which is far above chance (*i.e.* the performance of a classifier with a uniform predictive distribution) [53]. Deep Clustering leverages this prior on the input signal to iteratively learn representations from the supervision of pseudo-labels obtained by a standard clustering algorithm performed in the feature space.
Concerning MAE, it strongly masks the input data and learns to reconstruct its missing parts. For reflectance spectra, as much as the combination of spectral features (absorption peaks, spectral inflexion, etc.) at different wavelengths is closely related to the chemical composition of the land surface, the *masked reconstruction* task of MAE seems particularly relevant to learn discriminating features (of the materials) without class supervision. While MAE were

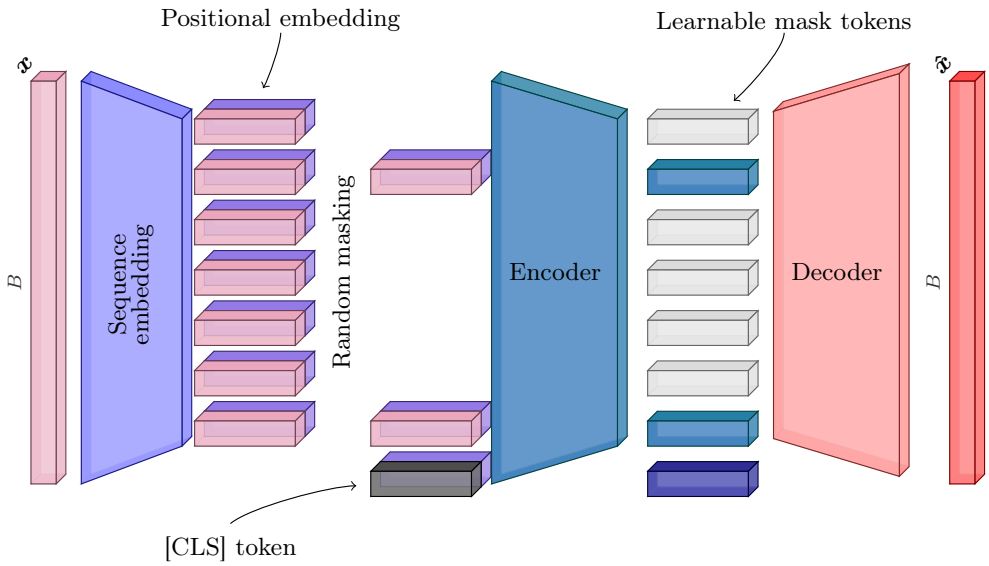


Figure 8: Illustration of the Masked Autoencoders [1] for 1D data. The $1 \times B$ dimensional input spectrum $\boldsymbol{x}$ is divided in small sequences. A large part of the sequences are randomly masked. The visible spectral sequences (with positional encoding) are encoded by the transformer. Then, learnable mask tokens are concatenated with the $1 \times d$ dimensional embeddings of the sequences, that are mapped to the reconstructed data $\hat{\boldsymbol{x}}$ by the light-weight decoder.

used on 8×8 pixel spaceborne hyperspectral patches (> 10 m GSD) in [54], we are interested in this paper on spectral representations only as the spectral information is much more discriminating of the materials on the ground surface than the spatial information in the case of airborne hyperspectral images. The Masked Autoencoder for 1D data is illustrated in Fig. 8. Originally, MAE processes RGB images that are divided in small patches which are encoded and decoded by vision transformers [55]. While transformers [56] have been adapted for hyperspectral data in [57], we use a simpler architecture and make as few changes as possible from the original MAE, keeping in mind two important points: 1) in contrast to words that are very abstract concepts (or to small image patches that can contain high-level information), reflectance values are not meaningful by themselves, 2) in contrast to words or to image patches, the relative distance between spectral channels does not contain semantic information. By this we mean that the position of a word in a sentence contains valuable semantic information, while the distance between two spectral features simultaneously observed on a spectrum is not informative[12]. This is why we believe that the transformer architecture is not particularly relevant for hyperspectral data, but is very convenient for the masked reconstruction task.

4.2 Experiments on the Toulouse Hyperspectral Data Set

4.2.1 Experimental protocol

In this section, we experimentally evaluated the potential of the spectral representations learned by Deep Clustering and MAE for a downstream classification task.

We compared the representation learning potential of the MAE against a conventional autoencoder in terms of classification accuracy. We trained a k-nearest neighbor algorithm (KNN) and a Random Forest classifier (RF) applied on the latent space of the autoencoder, on the [CLS] tokens of the masked autoencoder, and on the raw data as a baseline.

The rationale behind the use of the KNN classifier is that a self-supervised pretrained model should output grouped representations of semantically similar data [58]. Recent and prominent works have used the KNN classifier to evaluate the discriminating potential of representations learned by self-supervised models [43, 59, 38]. Other common validation protocols are to learn a linear classifier on top of the frozen encoder or to fine-tune the self-supervised model on the classifcation task [43, 60]. We found that a RF classifier on top of the frozen features outperformed both a linear classifier and a fine-tuned model by large margins, therefore we only reported the RF performances.

[12]For instance, a random permutation of the spectral channels would not decrease the performance of dense neural networks on a classification task

In order to assess whether Deep Clustering could be relevant for learning spectral representations, we evaluated whether using spectral convolutions or dense layers would provide as strong prior on the input signal as convolutions for images. To this end, we evaluated the performances of a multi-layer perceptron trained on top of a randomly initialized 1) spectral CNN and 2) dense network.

Hyperparameters, including architecture details, learning rate, and weight decay, were selected through a random search on the validation set. Concerning the MAE that has a high number of hyperparameters, the masking ratio, the number of attention heads and the embedding dimension were selected through an ablation study presented in section 4.2.3.

4.2.2 Experimental results

Table 3: Average overall accuracy and F1 score over the 8 standard ground truth splits

Model	OA	F1 score	Model	OA	F1 score
KNN	0.78	0.69	RF	0.75	0.65
AE + KNN	0.82	0.73	AE + RF	0.81	0.73
MAE + KNN	**0.84**	**0.76**	MAE + RF	**0.85**	**0.77**

MAE Results on Tab. 3 show that the representations learned with an MAE combined with a KNN and a RF have led to a significant increase compared to a KNN (+7% F1-score) and a RF (+12% F1-score) applied on raw data, and to a standard autoencoder baseline, but by smaller margins.

Deep Clustering A dense classifier trained on top of a random CNN and a random dense feature extractor achieved an average[13] F1-score of 0.038 and 0.024, respectively. For comparison, the average F1-score that a random classifier (with a uniform predictive distribution) would achieve on the Toulouse data set is 0.024. Therefore, the performances reached with a randomly initialized CNN and dense network are barely above chance and as good as chance, respectively. In conclusion, the use of spectral convolutions and dense layers to extract spectral representations do not provide useful priors on the spectral information, in contrast to convolutions for the spatial information of natural images. Thus, the pseudo-labels that we could derive from a clustering algorithm applied on the spectral features would not provide relevant semantic information. The iterative algorithm would not converge to useful spectral representations. This conclusion is confirmed by numerical experiments for which Deep Clustering combined with a KNN led to much worse accuracy than a KNN on the raw data.

4.2.3 Ablation study

We studied the impact of the masking ratio, the number of attention heads, and the dimension of the latent space on the MAE performance. Note that we speak indifferently of the latent space dimension and the total embedding dimension. We shall also precise that each element in the sequence is represented by an embedding whose actual dimension is the total embedding dimension divided by the number of attention heads. The experiments were conducted on different ground truth splits and for several random initializations of the MAE parameters. The mean and standard deviation of the validation loss and validation accuracy are reported for 20 epochs on Figs. 9, 10 and 11.

[13] F1-score averaged over the 8 ground truth splits.

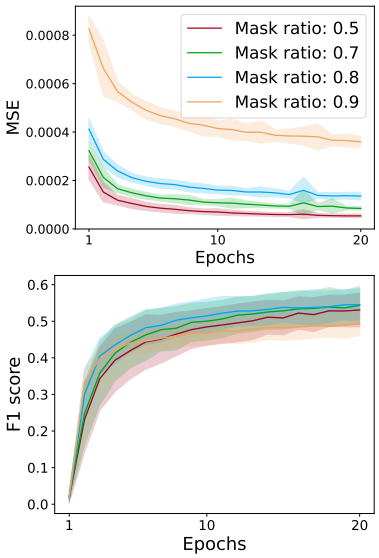


Figure 9: Influence of the masking ratio for a 32-dimensional latent space and 8 attention heads. Top: validation loss. Bottom: validation accuracy.

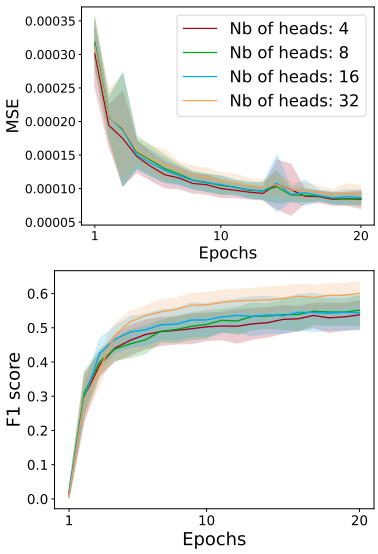


Figure 10: Influence of the number of attention heads for a 32-dimensional latent space and a 0.7 masking ratio. Top: validation loss. Bottom: validation accuracy.

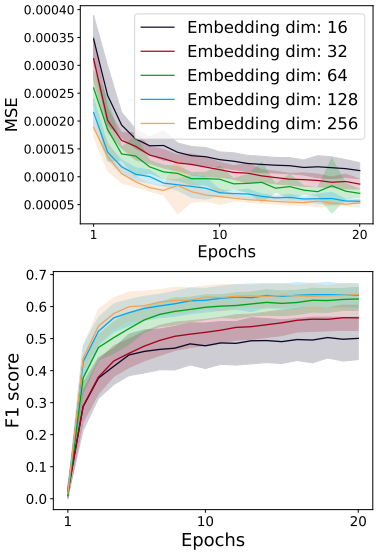


Figure 11: Influence of the latent space dimension for a 0.7 masking ratio and as many attention heads as latent space dimensions. Top: validation loss. Bottom: validation accuracy.

First, experiments on the masking ratio (Fig. 9) showed that a masking ratio around 0.7 and 0.8 leads to a trade-off between a trivial task and one that is too hard for the model to learn useful representations. Precisely, the lowest validation reconstruction error is obtained with a 0.5 masking ratio but the highest validation accuracies are obtained with 0.7 and 0.8 masking ratios.

Second, experiments on the number of attention heads (Fig. 10) showed that best performances were reached when the number of attention heads equaled the embedding dimension, *i.e* when each element in the sequence had a 1-dimensional (scalar) representation. While the validation reconstruction errors were bearly unchanged, a significant increase was reached with 32 attention heads. This result is consistent with our prior belief that spectral data, in contrast to words, contain very low-level information, and therefore do not require high-dimensional embeddings.

Third, experiments on the latent space dimension (Fig. 11) showed that best validation loss and validation accuracies were reached for 128 and 256 dimensions, despite the increase of parameters in the encoder.

5 Conclusions and perspectives

We have introduced the Toulouse Hyperspectral Data Set, a large benchmark data set designed to assess semi/self-supervised representation learning and pixel-wise classification techniques. Quantitative and qualitative comparisons have shown that several properties of the Toulouse data set better reflect the complexity of the land cover in large urban areas, compared to currently public data sets. In order to facilitate fair and reproducible experiments, we released a Python library to easily load PyTorch[14] data loaders and we hope that the standard train / test ground truth splits will foster the evaluation and comparison of new model architectures and learning algorithms.

The numerical experiments showed that the masked autoencoding task is very appropriate for learning useful spectral representations. The baseline established in this paper for pixel-wise classification based on a MAE [1] and Random Forests [61] reached a 85% overall accuracy and 77% F1 score. Besides, the ablation study about the MAE hyperparameters showed that a higher number of attention heads (and a lower embedding dimension for each element in the sequence) than for vision transformers is beneficial to extract spectral features. We argue that this result can be explained by the fact that reflectance spectra contain low-level information, compared to RGB images. Experiments also demonstrated that the use of spectral convolutions do not provide as strong prior on reflectance spectra as convolutions do for RGB images, which explains why clustering-based techniques are inadequate for spectral representation learning.

[14] https://pytorch.org/docs/stable/index.html

We focused in this paper on the spectral information because the reflectance, that is intrinsic to the chemical composition of matter, is by nature highly discriminating of the land cover. In some cases, however, we believe that large-scale contextual information could prevent some confusions (for instance, pixels predicted as *wheat* in a green urban area or pixels predicted as *orange tile* in a cultivated field). Precisely, we argue that hyperspectral patches of at least 64×64 pixels are necessary with a ≈ 1 m GSD (while common semantic segmentation models applied on RGB satellite images rather use 256×256 pixel patches with a ≈ 50 cm GSD). Processing such big hyperspectral patches with macine learning models though raises memory issues due to the large spectral dimension, which hinders the direct application of state-of-the-art vision models, especially large models such as the recent foundation model SAM [62]. Nevertheless, combining deep vision models to extract contextual information (*i.e.* abstract land use information) with shallow models to extract spectral information (*i.e.* low-level land cover information) is a promising research direction. Besides, we believe that the hierarchical nomenclature, the land use annotations in addition to the land cover annotations and the long-tailed class distribution open the path towards important research areas, respectively (hierarchical) multi-label classification [63] and long-tailed learning [35], that have been little discussed for pixel-wise hyperspecral image classification. Lastly, the Toulouse Hyperspectral Data Set is also particularly suited to evaluate Active Learning algorithms thanks to the provided *labeled pools* from which pixels to label can be sampled.

Acknowledgement

We thank Philippe Déliot for providing the hyperspectral images in ground-level reflectance and geometrically rectified.

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[2] M. Labbas. *Modélisation hydrologique de bassins versants périurbains et influence de l'occupation du sol et de la gestion des eaux pluviales. Application au bassin de l'Yzeron (130 km2)*. Theses, Doctorat, spécialité : Océan, Atmosphère, Hydrologie, Université de Grenoble, 2015.

[3] Perkins F.E. Bras R.L. Effects of urbanization on catchment response. *Journal of Hydraulics Division*, 101:451–466, 1975.

[4] M. Desbordes. Principales causes d'aggravation des dommages dus aux inondations par ruissellement superficiel en milieu urbanisé. *Bulletin hydrologie urbaine*, 4:2–10, 1989.

[5] Furkan Dosdogru, Latif Kalin, Ruoyu Wang, and Haw Yen. Potential impacts of land use/cover and climate changes on ecologically relevant flows. *Journal of Hydrology*, 584:124654, 2020.

[6] Subhasis Giri, Zhen Zhang, Daryl Krasnuk, and Richard G. Lathrop. Evaluating the impact of land uses on stream integrity using machine learning algorithms. *Science of The Total Environment*, 696:133858, 2019.

[7] Akio Onishi, Xin Cao, Takanori Ito, Feng Shi, and Hidefumi Imura. Evaluating the potential for urban heat-island mitigation by greening parking lots. *Urban Forestry and Urban Greening*, 9(4):323 – 332, 2010.

[8] H.-0 Pörtner, D.C. Roberts, M. Tignor, Poloczanska E.S., K. Mintenbeck, A. Alegría, S. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, B. Rama, and eds. *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, volume 1. Cambridge University Press, 2022.

[9] Marlon Correa Pereira, Roisin O'Riordan, and Carly Stevens. Urban soil microbial community and microbial-related carbon storage are severely limited by sealing. *Journal of Soils and Sediments*, 21:1455–1465, 2021.

[10] R. O'Riordan, J. Davies, C. Stevens, and J. N. Quinton. The effects of sealing on urban soil carbon and nutrients. *SOIL*, 7(2):661–675, 2021.

[11] EU: European Commission. Guidelines on best practice to limit, mitigate or compensate soil sealing. *Luxembourg: European Union SWD (2012) 101*, 2012.

[12] Riccardo Scalenghe and Franco Ajmone Marsan. The anthropogenic sealing of soils in urban areas. *Landscape and urban planning*, 90(1-2):1–10, 2009.

[13] C. Miesch, L. Poutier, V. Achard, X. Briottet, X. Lenot, and Y. Boucher. Direct and inverse radiative transfer solutions for visible and near-infrared hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(7):1552–1562, 2005.

[14] Bo-Cai Gao, Marcos J Montes, Curtiss O Davis, and Alexander FH Goetz. Atmospheric correction algorithms for hyperspectral remote sensing data of land and ocean. *Remote sensing of environment*, 113:S17–S24, 2009.

[15] Devis Tuia, Michele Volpi, Loris Copa, Mikhail Kanevski, and Jordi Munoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing*, 5(3):606–617, 2011.

[16] Romain Thoreau, Veronique Achard, Laurent Risser, Beatrice Berthelot, and Xavier Briottet. Active learning for hyperspectral image classification: A comparative review. *IEEE Geoscience and Remote Sensing Magazine*, pages 2–24, 2022.

[17] Gustavo Camps-Valls, Tatyana V Bandos Marsheva, and Dengyong Zhou. Semi-supervised graph-based hyperspectral image classification. *IEEE transactions on Geoscience and Remote Sensing*, 45(10):3044–3054, 2007.

[18] Hao Wu and Saurabh Prasad. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270, 2017.

[19] Shrutika S Sawant and Manoharan Prabukumar. A review on graph-based semi-supervised learning methods for hyperspectral image classification. *The Egyptian Journal of Remote Sensing and Space Science*, 23(2):243–248, 2020.

[20] Jun Yue, Leyuan Fang, Hossein Rahmani, and Pedram Ghamisi. Self-supervised learning with adaptive distillation for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.

[21] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Deep learning for classification of hyperspectral data: A comparative review. *IEEE geoscience and remote sensing magazine*, 7(2):159–173, 2019.

[22] Julius Lange, Gabriele Cavallaro, Markus Götz, Ernir Erlingsson, and Morris Riedel. The influence of sampling methods on pixel-wise hyperspectral image classification with 3d convolutional neural networks. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2087–2090. IEEE, 2018.

[23] Christian Geiß, Patrick Aravena Pelizari, Henrik Schrade, Alexander Brenning, and Hannes Taubenböck. On the effect of spatially non-disjoint training and test samples on estimated model generalization capabilities in supervised classification with spatial features. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2008–2012, 2017.

[24] Saurabh Prasad, Bertrand Le Saux, Naoto Yokoya, and Ronny Hansch. 2018 ieee grss data fusion challenge – fusion of multispectral lidar and hyperspectral data. 2020.

[25] L. Roupioz, X. Briottet, K. Adeline, A. Al Bitar, D. Barbon-Dubosc, R. Barda-Chatain, P. Barillot, S. Bridier, E. Carroll, C. Cassante, A. Cerbelaud, P. Déliot, P. Doublet, P.E. Dupouy, S. Gadal, S. Guernouti, A. De Guilhem De Lataillade, A. Lemonsu, R. Llorens, R. Luhahe, A. Michel, A. Moussous, M. Musy, F. Nerry, L. Poutier, A. Rodler, N. Riviere, T. Riviere, J.L. Roujean, A. Roy, A. Schilling, D. Skokovic, and J. Sobrino. Multi-source datasets acquired over toulouse (france) in 2021 for urban microclimate studies during the camcatt/ai4geo field campaign. *Data in Brief*, 48:109109, 2023.

[26] Charlotte Revel, Yannick Deville, Véronique Achard, Xavier Briottet, and Christiane Weber. Inertia-constrained pixel-by-pixel nonnegative matrix factorisation: A hyperspectral unmixing method dealing with intra-class variability. *Remote Sensing*, 10(11):1706, 2018.

[27] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[28] Javiera Castillo-Navarro, Bertrand Le Saux, Alexandre Boulch, Nicolas Audebert, and Sébastien Lefèvre. Semi-supervised semantic segmentation in earth observation: The minifrance suite, dataset analysis and multi-task network study. *Machine Learning*, pages 1–36, 2021.

[29] Piers J Sellers. Canopy reflectance, photosynthesis and transpiration. *International journal of remote sensing*, 6(8):1335–1372, 1985.

[30] José Manuel Peña-Barragán, Francisca López-Granados, Montserrat Jurado-Expósito, and Luis García-Torres. Mapping ridolfia segetum patches in sunflower crop using remote sensing. *Weed Research*, 47(2):164–172, 2007.

[31] WC Bausch and RAJIV Khosla. Quickbird satellite versus ground-based multi-spectral data for estimating nitrogen status of irrigated maize. *Precision Agriculture*, 11:274–290, 2010.

[32] Anatoly Gitelson and Mark N Merzlyak. Spectral reflectance changes associated with autumn senescence of aesculus hippocastanum l. and acer platanoides l. leaves. spectral features and relation to chlorophyll estimation. *Journal of plant physiology*, 143(3):286–292, 1994.

[33] Ronald C Estoque and Yuji Murayama. Classification and change detection of built-up lands from landsat-7 etm+ and landsat-8 oli/tirs imageries: A comparative assessment of various spectral indices. *Ecological indicators*, 56:205–217, 2015.

[34] Alfredo R Huete. A soil-adjusted vegetation index (savi). *Remote sensing of environment*, 25(3):295–309, 1988.

[35] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[36] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[37] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.

[38] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[39] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, pages 4182–4192. PMLR, 2020.

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

[41] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[42] 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.

[43] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[44] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[45] Xiaohua Zhai, Avital Oliver, Alexander Kolesnikov, and Lucas Beyer. S4l: Self-supervised semi-supervised learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1476–1485, 2019.

[46] Enrico Fini, Pietro Astolfi, Karteek Alahari, Xavier Alameda-Pineda, Julien Mairal, Moin Nabi, and Elisa Ricci. Semi-supervised learning made simple with self-supervised clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3187–3197, 2023.

[47] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[48] Lin Zhao, Wenqiang Luo, Qiming Liao, Siyuan Chen, and Jianhui Wu. Hyperspectral image classification with contrastive self-supervised learning under limited labeled samples. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[49] PuHong Duan, ZhuoJun Xie, XuDong Kang, and ShuTao Li. Self-supervised learning-based oil spill detection of hyperspectral images. *Science China Technological Sciences*, 65(4):793–801, 2022.

[50] Yuntao Qian, Honglin Zhu, Ling Chen, and Jun Zhou. Hyperspectral image restoration with self-supervised learning: A two-stage training approach. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

[51] Yao Qin, Yuanxin Ye, Yue Zhao, Junzheng Wu, Han Zhang, Kenan Cheng, and Kun Li. Nearest neighboring self-supervised learning for hyperspectral image classification. *Remote Sensing*, 15(6), 2023.

[52] 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.

[53] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[54] Lingxuan Zhu, Jiaji Wu, Wang Biao, Yi Liao, and Dandan Gu. Spectralmae: Spectral masked autoencoder for hyperspectral remote sensing image reconstruction. *Sensors*, 23(7):3728, 2023.

[55] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[57] Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

[58] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[59] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

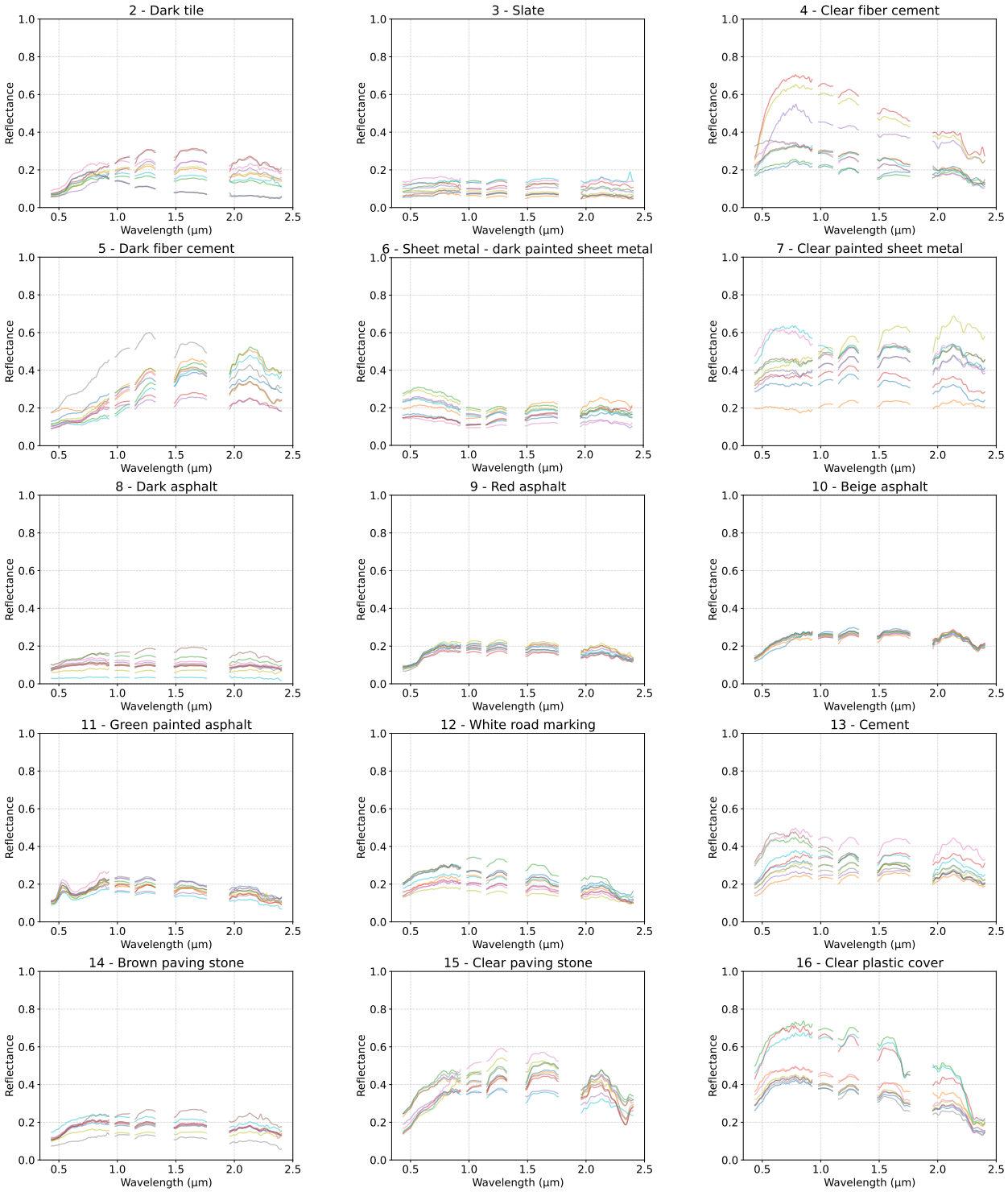
[60] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[61] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[62] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[63] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine learning*, 73:185–214, 2008.

Appendices


2 - Dark tile
3 - Slate
4 - Clear fiber cement
5 - Dark fiber cement
6 - Sheet metal - dark painted sheet metal
7 - Clear painted sheet metal
8 - Dark asphalt
9 - Red asphalt
10 - Beige asphalt
11 - Green painted asphalt
12 - White road marking
13 - Cement
14 - Brown paving stone
15 - Clear paving stone
16 - Clear plastic cover
Reflectance
Wavelength (µm)

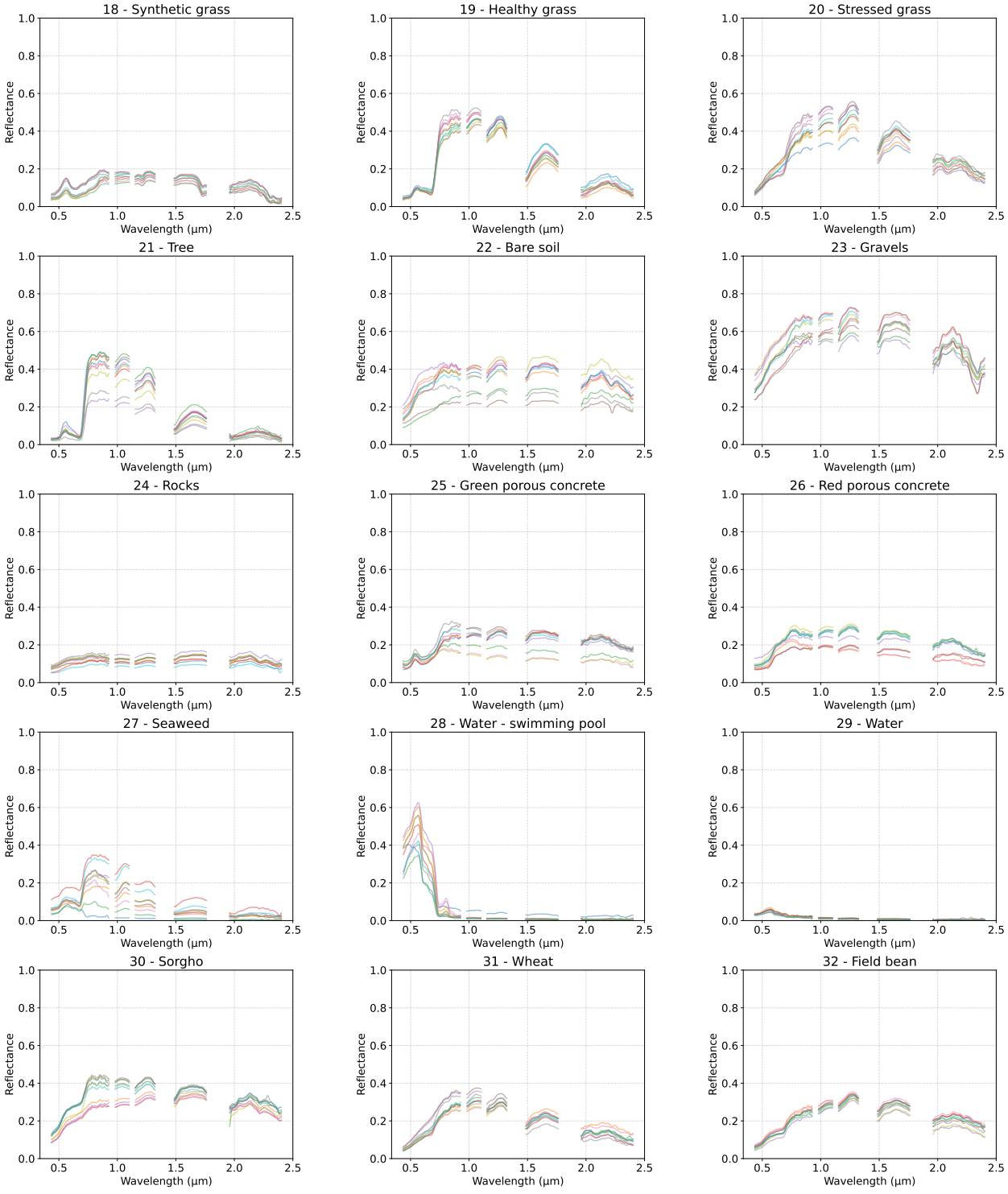


Figure 12: Additional random spectra of the Toulouse Hyperspectral Data Set

In addition to the land cover, we define a land use nomenclature which gathers more abstract semantic classes, listed in Tab. 4. Besides, we provide the direct and diffuse irradiance at ground level, as well as the solar zenith angle which is of 22.12°.

Table 4: Land use nomenclature of the Toulouse Hyperspectral Data Set

#1	Roads	#7	Lakes / rivers / harbors
#2	Railways	#8	Swimming pools
#3	Roofs	#9	Forests
#4	Parking lots	#10	Cultivated fields
#5	Building sites	#11	Boats
#6	Sport facilities	#12	Open areas

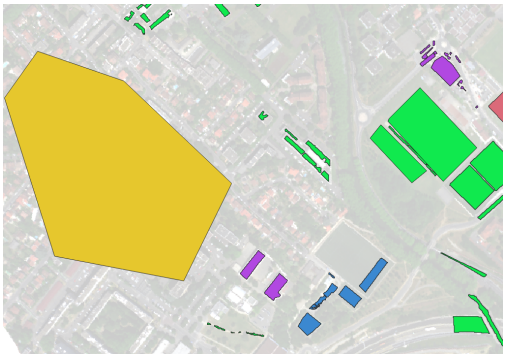


Figure 13: Example of annotations of the hyperspectral image. Polygons in red, green, blue, purple and yellow belong to the train labeled set, the labeled pool, the validation set, the test set and the unlabeled pool, respectively.

(a) Spatial features

(b) Spectral (visible to NIR) features

Figure 14: t-SNE projections of hand-crafted representations of 64×64 pixel hyperspectral patches from the Pavia University, Houston University and Toulouse data sets. (a) corresponds to the 2D projection of the spatial features only, *i.e.* only the Gabor filters were used to represent the data and then the t-SNE projection was performed, and (b) corresponds to the 2D projection of the spectral features only, *i.e.* only the spectral indices were used to represent the data and then the t-SNE projection was performed. Moreover, only the smallest spectral domain, *i.e.* the spectral domain of the Pavia University image which covers the 0.4 µm - 0.86 µm range, was used.

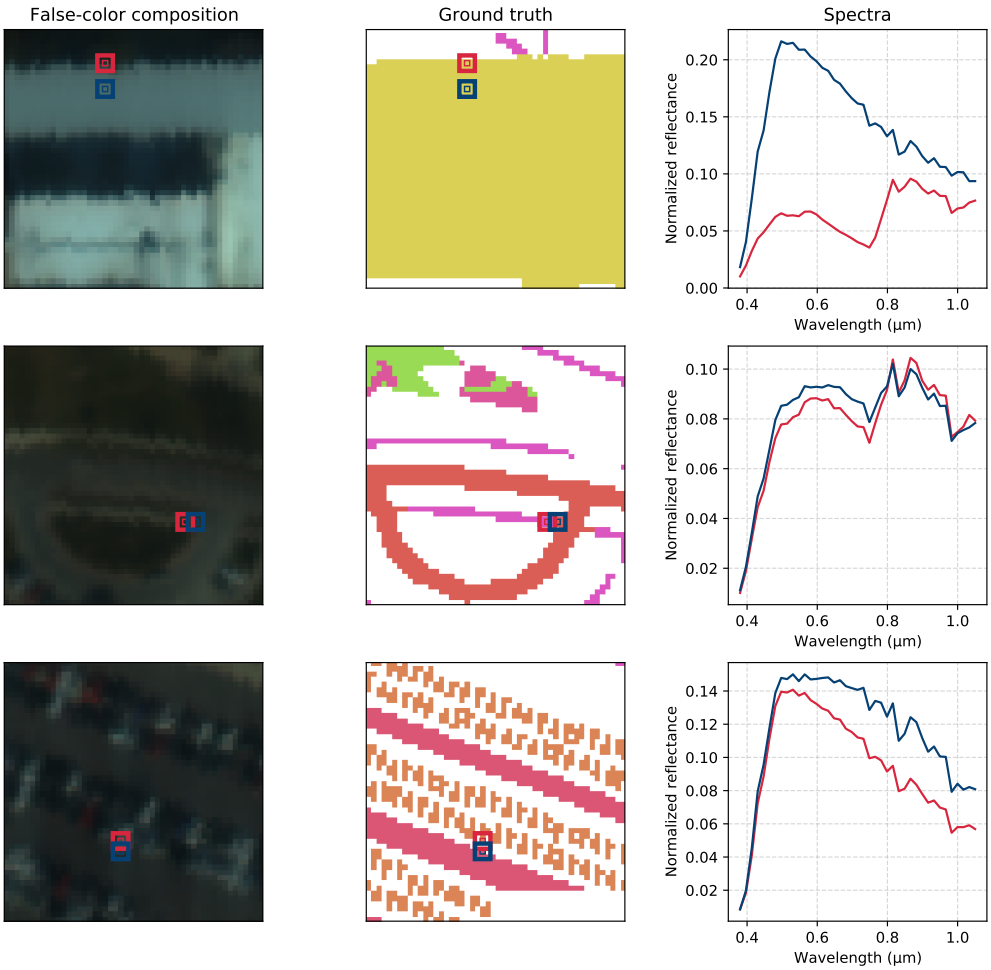


Figure 15: Illustration of noisy labels in the Houston university data set. The first row shows an example of two spectra labeled as *Non-residential buildings*. The pixel at the edge of the roof is mixed with the ground vegetation, which provides erroneous spectral information. The second and third rows show neighboring pixels labeled with different classes though they are actually a mixed of both classes.

Figure 16: Examples of annotations in the Toulouse data set. We paid attention to ignore mixed pixels such as pixels at edges of roofs or pixels mixed with small objects such as pipes on roofs.