remote sensing

MDPI

Article

Effective Training of Deep Convolutional Neural Networks for Hyperspectral Image Classification through Artificial Labeling

Wojciech Masarczyk †, Przemysław Głomb †, Bartosz Grabowski *,† and Mateusz Ostaszewski †

Institute of Theoretical and Applied Informatics, Polish Academy of Sciences; 44-100 Gliwice, Poland; wmasarczyk@iitis.pl (W.M.); przemg@iitis.pl (P.G.); mostaszewski@iitis.pl (M.O.)

* Correspondence: bgrabowski@iitis.pl

† These authors contributed equally to this work.

Received: 22 July 2020; Accepted: 15 August 2020; Published: 17 August 2020

Abstract: Hyperspectral imaging is a rich source of data, allowing for a multitude of effective applications. However, such imaging remains challenging because of large data dimension and, typically, a small pool of available training examples. While deep learning approaches have been shown to be successful in providing effective classification solutions, especially for high dimensional problems, unfortunately they work best with a lot of labelled examples available. The transfer learning approach can be used to alleviate the second requirement for a particular dataset: first the network is pre-trained on some dataset with large amount of training labels available, then the actual dataset is used to fine-tune the network. This strategy is not straightforward to apply with hyperspectral images, as it is often the case that only one particular image of some type or characteristic is available. In this paper, we propose and investigate a simple and effective strategy of transfer learning that uses unsupervised pre-training step without label information. This approach can be applied to many of the hyperspectral classification problems. The performed experiments show that it is very effective at improving the classification accuracy without being restricted to a particular image type or neural network architecture. The experiments were carried out on several deep neural network architectures and various sizes of labeled training sets. The greatest improvement in overall accuracy on the Indian Pines and Pavia University datasets is over 21 and 13 percentage points, respectively. An additional advantage of the proposed approach is the unsupervised nature of the pre-training step, which can be done immediately after image acquisition, without the need of the potentially costly expert's time.

Keywords: hyperspectral image classification; deep learning; convolutional neural networks; transfer learning; unsupervised training sample selection

1. Introduction

The classification of hyperspectral images (HSI) has many potential applications, e.g., land cover segmentation [1], mineral identification [2], or anomaly detection [3]. The classification algorithms used include both general models, e.g., the support-vector machines (SVM) [4], and dedicated approaches, taking into account spectral properties or spatial class distribution [5]. Recently there have been attempts to use Deep Learning Neural Networks (DLNN) for HSI classification. The motivation is that such methods have gained attention after achieving state of the art in natural image processing tasks [6]. Their unique ability to process an image while using a hierarchical composition of simple features learned during training makes them a powerful tool in areas where manipulation of high-dimensional data is needed.

While DLNN can achieve very good accuracy scores, they have the drawback of requiring a large amount of training data for estimation of model parameters. Such data are not always available, as it is common to have a single HSI with just a handful of training labels available. To bridge a gap between this realistic scenario and DLNN network requirements, we propose an approach that trains the DLNN in two stages, with the first–pre-training–stage using artificial labels. In the remainder of this section, we discuss the relevant related works, and introduce the motivation of our approach and state the hypothesis that is the base of our method.

A number of DLNN architectures have been proposed, inspired by mathematical derivations and/or neuroscience studies. The Convolutional Neural Networks (CNN) [7] are a special case of deep neural networks that were originally developed to process images, but are also used for other types of data like audio. They combine traditional neural networks with biologically inspired structure into a very effective learning algorithm. They scan multidimensional input piece by piece with a convolutional window, which is a set of neurons with common weights. Convolution window processes local dependencies (features) in the input data. The output corresponding to one convolutional window is called a feature map and it can be interpreted as a map of activity of the given feature on the whole input. The CNN remain one of the most popular architectures for DLNN classification in use today.

Other approaches include the generative architectures, e.g., the Restricted Boltzmann Machine (RBM) [8,9], Autoencoders (AE) [10], or Deep Belief Network (DBN) [11,12]. Yet another popular architecture is the Recurrent Neural Network (RNN), which, through directed cycles between units, has the potential of representing the state of processed sequence. They are applicable e.g. for time series prediction or outlier detection. The most popular types of RNN are Long Short-Term Memory (LSTM) networks [13] and Gated Recurrent Units (GRUs) [14]. They improve the original RNN architecture by dealing with exploding and vanishing gradient problem.

For the classification of HSI data, the CNN is the most popular architecture chosen. In [15] the simple CNN architecture is adapted to HSI classification; the lack of training labels is mitigated by adding geometric transformations to training data points. In [16] authors use three kinds of convolutional windows: two of them are three-dimensional (3D) convolutions which analyse spatial and spectral dependencies in the input picture, while the third is the 1D kernel. Next, the feature maps from these three types of convolutions are stacked one after the other and create joint output of this first part of the network. The following layers only consist of the one dimensional convolutional kernels and residual connections. The authors of [17] introduce a parallel stream of processing with an original approach for spatial enhancement of hyperspectral data. The authors of [18] design a deep network that reduces the effect of Hughes phenomenon (curse of dimensionality) and use additional unlabelled sample pool to improve performance. In [19] authors propose an alternative architecture called RPNet based on prefixed convolutional kernels. It combines shallow and deep features for classification. Another architecture (MugNet) is proposed in [20] with a focus on the simplicity of processing for classification of hyperspectral data with few training samples and reduced number of hyperparameters. A yet another architecture approach is used in [21] where a multi-branch fusion network is introduced, which uses merging multiple branches on an ordinary CNN. An additional L2 regularization step is introduced in order to improve the generalization ability with limited number of training samples. The work [22] proposes a strategy based on multiple convolutional layers fusion. Two distinct networks, composed of similar modules, but different organization, are examined.

Other architectures are also used. For example, in [23] authors utilize the sequential nature of hyperspectral pixels and use some variations of recurrent neural networks—Gated Recurrent Unit (GRU) and Long-Short Term Memory (LSTM) networks. Moreover, in [24] one dimensional convolutional layers followed by LSTM units were used. Chen et. al. [25] use artificial neural networks for feature extraction. They utilize stacked autoencoders (SAE) for feature extraction from pixels, and Principal Component Analysis (PCA) for reduction of the spectral dimensionality of the training segments taken from the picture. Next, the logistic regression is performed on this spectral (SAE)

and spatial (PCA) extracted information. Another approach [26] uses stacked SAE for an application study—detection of a rice eating insect. In the general case RNN architectures are also employed, as they are suitable for processing the spectral vector data. The work [27] applies sequential spectral processing of hyperspectral data, using a RNN supported by a guided filter. In [28] authors use the multi-scale hierarchical recurrent neural networks (MHRNNs) to learn the spatial dependency of non-adjacent image patches in the two-dimension (2D) spatial domain. Another idea to analysing HSI is spatial–spectral method in which network takes information not only from spectrum bands but also from spatial dependencies of image [16].

A significant problem in practical hyperspectral classification is the small number of training samples. It is related to the difficulty of obtaining verified labels [1], as often each pixel must be individually evaluated before labelling. Therefore, a reference hyperspectral classification experiment may assume number as low as 1% available samples per class [2]. A number of approaches has been exploited in order to deal with this difficulty, e.g., including combining spatial and spectral features [29], additional training sample generation [30], extending the classification algorithm with segmentation [31], or employing Active Learning [32].

For the DLNN classification, the lack of high volume of training data is a serious complication, as they typically require a lot of data to achieve high efficiency. The optimal use of DLNN in HSI classification would require learning them with just a few labelled samples. This may be obtained by searching for well-tailored architecture for specific task [15], however such an approach requires relatively big validation set to obtain meaningful results. The other approach is to expand the available training set. It may be achieved either by artificially augmenting training set or using different dataset as a source for pre-training [33]. Another approach is to add a regularization step to improve the generalization ability with limited number of training samples [21]. A simplification of the network architecture for classification with few training samples is employed in the MugNet network [20]. Finally, where possible, the transfer learning approach is used, e.g. [34].

The transfer learning [35] uses training samples from two domains, which share common characteristics. A network is first pre-trained on the first domain, which has plentiful supply of training samples, but does not solve the problem at hand. Subsequently, the training is updated with the second domain, which adapts the weights to the actual problem.

Transfer learning is simple to apply in the case of convolutional neural networks (CNNs). In [36], the authors compared different versions of transfer learning for CNNs in the case of natural images classification. They studied its effects depending on the number of transferred layers and whether they were fine tuned or not as well as depending on the differences between the considered datasets. In [37], authors used transfer learning on CNNs to recognize emotions from the pictures of faces. Other uses include evaluating the level of poverty in a region given its remote sensing images [38] and computer-aided detection using CT scans [39].

There have been applications of transfer learning in the general remote sensing (not-hyperspectral) images. In [40] deep learned features are transferred for effective target detection; negative bootstrapping is used for improving the convergence of the detector. A similar approach is applied in [41] where RNN network trained on multispectral city images is used in order to derive features for studying urban dynamics across seasonal, spatial and annual variance. The authors of [42] study the performance of transfer learning in two remote sensing scene classifications. The results show that features generalize well to high resolution remote sensing images. As the work [43] shows, transfer learning can be applied in remote sensing using RNN architectures also.

Recently, transfer learning has been also applied to the HSI data. In [34], the authors applied it for CNNs originally used for classifying well known remote sensing hyperspectral images to classify images acquired from field-based platforms. The authors of [44] use a intermediate step of supervised similarity learning for anomaly detection in unlabelled hyperspectral image. A different approach to transfer learning is proposed in [45], which explores the high level feature correlation of two HSI. A new training method simultaneously processes both images, to estimate a common feature space for

both images. A yet another approach is proposed in [46], where HSI superresolution is achieved using supported high resolution natural image. This natural image is used as a training reference, which is later adapted to HSI domain. In [47], iterative process combines training and updating the currently used training label set. Two specialized architectures, for spatial and spectral processing, are used. The training iteratively extends the current label set, starting from the initial expert's labels.

The above approaches do not apply to the arguably most popular practical scenario, where only a single HSI with a handful of labels is available. Moreover, getting the training labels often requires additional resources (e.g., expert consultation and/or site visit). Thus, it is desirable to have unsupervised methods for realization of the pre-training step. Authors of [48] use outlier detection and segmentation to provide candidates for training of target detector in HSI. This information is used to construct a subspace for target detection by transfer learning theory. This shows the potential of using an unsupervised approach, however limited to separation of target/anomaly points from the background. In the work of [33], a separate clustering step is used for the generation of pseudo-labels, using Dirichlet process mixture model. The network is trained on the pseudo-labels, then the all but last layers are extracted, and the final network is trained on the originally provided training labels. While this scheme is shown to be effective in the presented results, it relies on a complex non-neural preprocessing and tailoring the DLNN configuration to each dataset separately. Additionally, the effect of size of label areas and effects on different architectures are not investigated. We show that similar gains can be made with a simpler preprocessing, independent of the DLNN architecture chosen. The authors of [49] propose using a sparse coding to estimate high level features from unlabelled data from different sources. This approach does not require training data, but it is tailored to the case where multiple inputs are available, preferably with diverse contents.

To close the gap between data inefficient deep learning models and practical applications of HSI, we propose a method that takes advantage of abundant unlabelled data points that are present on HSI images. Precisely, we state a hypothesis: The spatial similarity of unlabelled data points can be utilized to gain accuracy in hyperspectral classification. To corroborate our hypothesis, we construct a simple spatial clustering method that assigns artificial label to each pixel on the image based on its spatial location. This artificial dataset is used to pre-train deep learning classifier. Next, the model is fine-tuned with original dataset. Through series of experiments we show superiority of the proposed approach over the standard learning procedure. Our approach is motivated by two known phenomena: cluster assumption [50] and regularization effect of noise in classes [51–53]. We note that many of remote sensing images share common properties, most notably the 'cluster assumption'—pixels that are close to one another or form a distinct cluster or group frequently share the class label. Additionally, due to the simplistic form of our clustering method, we purposefully introduce noise in labels used during pre-training phase; however, as shown in [51], this label noise has little to no effect on final accuracy, as long as number of properly labelled examples scales proportionally, which is our case.

2. Materials & Methods

Our method is to be applied in the following case:

1. classification of pixels from a remote sensing hyperspectral image;
2. neural networks used as a classifier; and,
3. few training labels available.

In such a situation, we propose to augment the training with a pre-training step that uses artificial labels, which are independent of the training labels. Inclusion of this pre-training step can be viewed as a modification of a transfer learning approach. Conventional transfer learning, in this case, would use a related dataset (source domain) with abundance of labels to pre-train, then the current dataset (target domain) to fine-tune. In our case, the source domain consists of every point in the hyperspectral image, while the target domain is composed of only the labelled samples. In the remainder of this Section, we discuss: the spatial structure of hyperspectral images and the characteristics of neural

network that make this approach feasible, and the details of its application. We also describe the experiments used to test the proposed approach.

2.1. Spatial Structure of Hyperspectral Images

It is well-known that remote sensing hyperspectral images contain spatial structure, which can be exploited to improve classification scores when only a few training samples are available [5,30,31,54,55]. A segmentation can be applied to propose candidate pixels for labelling with high confidence [54] or identify connected components for label assignment [31]. Class training samples can be extended through moderated region growing [5] or spatial filtering combined with spatial-spectral Label Propagation [55]. Finally, disagreement between spatial and spectral classifiers can be used to propose new samples [30]. A qualitative investigation of this phenomenon shows that hyperspectral pixels that are close to one another, whether spatially or spectrally, are likely to have the same class label, thus fulfilling the 'cluster assumption' [50]. This effect often leads to a blob-like structure of a hyperspectral dataset, observed in many hyperspectral classification problems (e.g., land cover labelling in remote sensing, paint identification in heritage science, scene analysis in forensics). A single class with samples in different parts of an image can be made of a number of blobs, which differ from each other because of, for example, non-uniformity in class structure (e.g., the same class can contain differing crop types), spectral variations (e.g., same crop in two areas can have differing properties due to sunlight exposure, soil type), or acquisition conditions (e.g., level of lighting, shadows).

2.2. Emergence of Data-Dependent Filters in Neural Network Training

During training, subsequent layers of a deep neural network form a representation of a local input data structure [56]. Given a data source, this representation, especially on lower layers, can be remarkably similar across different dataset. For example, in the problem of natural image classification, the learned kernels resemble a Gabor filter bank [6,57], independent of a class set. This form of a filter can be shown to arise independently when independent components [58] or an effective sparse code [59] for natural images is estimated. Another case where data-dependent filters emerge is the pretext task approach, e.g., [60,61], where the network first learns to predict the input sequence without class labels, which are introduced at a fine-tune stage for to obtain the final classification model. Apparently, the deep neural networks are able, at least in part, to extract an efficient class-independent data representation. This phenomenon has not been studied for hyperspectral images; however, it can be argued that similar class-independent but data-dependent representation is being learned in training for hyperspectral image classification.

2.3. Methods Used for Proposed Artificial Labelling Approach

Our method for creating artificial labels for the pre-training step is a simple segmentation algorithm that assumes the local homogeneity of samples' spectral characteristics. It works by dividing the considered image into k rectangles, where each of these rectangles has its own label. For an image of height h and width w, we divide its height into m roughly equal parts and its width into n roughly equal parts, so that $k = m \cdot n$. We then get k rectangles, where each one's height equals approximately h/m, while its width equals approximately w/n. Each of these rectangles define a different artificial class with a different label. Figure 1 presents a schematic.

The function of artificial labels is for the network to learn class-independent blob patterns present in the data. This focuses the network training in the fine tuning on the actual training labels, with the network 'oriented' towards the features of the current image. It can also be of advantage in situations when a class is composed of multiple blobs, and not all of them have samples in the training set. In that case, sufficiently correct labelling is unlikely to be obtained [62] with just the training samples, but the proposed grid structure forces the network to estimate features for the whole image. An additional advantage of this approach is to shift the potentially time consuming pre-training from the expert

labelling moment to the acquisition moment. In other words, network training does not need to be held back until the expert's labels are available, but can commence right after the image is recorded.

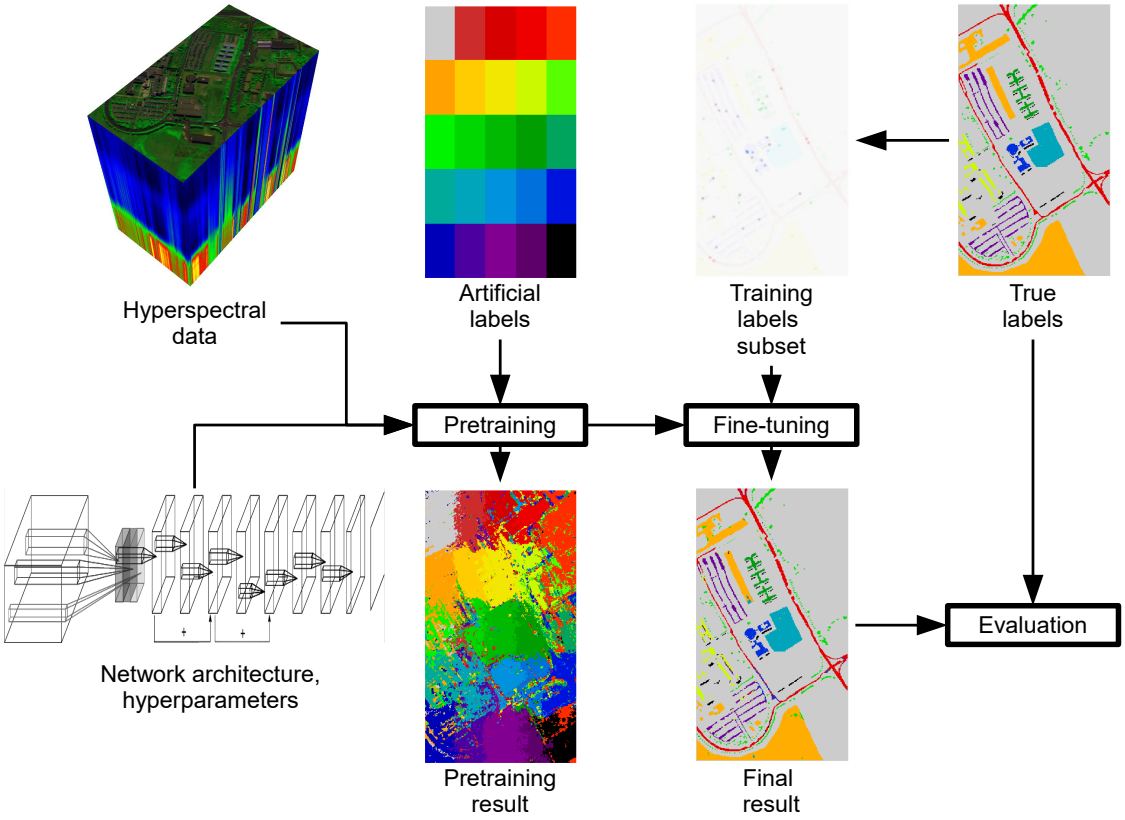


Figure 1. The overview of the unsupervised pretraining algorithm proposed in this work. First, the network is pretrained on grid-based scheme of artificially assigned labels. The network weights are then fine-tuned on a limited set of training samples selected from true labels, consistent with typical hyperspectral classification scenario.

2.4. Selected Network Architectures

In our experiments, three architectures were tested, based on [15,16,63]. All three share a common approach to exploit local homogeneity of hyperspectral images, however each one has its unique strengths and weaknesses making them an interesting testbed for the universality of the proposed method. The first architecture [16] features relatively high number of convolutional layers that might be helpful in transfer learning application. The second architecture [15], to the best of authors knowledge, is one of the best networks that are trained on limited number of samples per class. However due to its constrained capacity, it may not benefit as much from the pre-training phase. The last of the considered convolutional neural networks [63] is conceptually the simplest of the three, which allows us to test our approach using more conventional convolutional architecture.

2.5. Experiments

This subsection describes the experiments evaluating the proposed approach. We investigate the performance of the artificial label pre-training in the following four experiments:

1. Experiment 1 evaluates the accuracy improvement achieved by using the method.
2. Experiment 2 investigates the variability introduced by the size and shape of the patches used to define artificial classes, using one of the datasets from Experiment 1.

3. Experiment 3 is an additional investigation of the observed phenomenon that big patches (larger area, smaller total number of classes) perform worse than little ones (smaller area, larger total number of classes), done using a different dataset.
4. Experiment 4 is an examination of a claim about the emergence of data-independent representations during neural network training using proposed artificial labelling scheme.

In the following subsections, the detailed descriptions of the conducted experiments are given, while, in Section 3, we present the results of the experiments.

2.5.1. Experiment 1

In this experiment, the proposed approach is evaluated using different hyperspectral images and neural network architectures to prove its robustness.

For the experiment, we have used two well-known hyperspectral datasets: Indian Pines and Pavia University.

The Indian Pines dataset was collected by the AVIRIS sensor over the Northwest Indiana area. The image consists of 145×145 pixels. Each pixel has 220 spectral bands in the frequency range 0.4–2.5 $\times 10^{-6}$ m. Channels affected by noise and/or water absorption were removed (i.e., 104–108, 150–163, 220), bringing the total image dimension to 200 bands. The reference ground truth contains 16 classes representing mostly different types of crops. To be consistent with the experiments performed in [16], we only choose 8 classes.

The Pavia University dataset was collected by the ROSIS sensor over the urban area of the University of Pavia in Italy. This image consist of 610×340 pixels. It has 115 spectral bands in the frequency range from 0.43 to 0.86 $\times 10^{-6}$ m. The noisiest 12 bands were removed, and the remaining 103 were utilized in the experiments. Ground truth includes 9 classes, corresponding mostly to different building materials.

The two datasets were subjected to a feature transformation. For a given dataset, the mean m_b of each hyperspectral band b were calculated. In the case of each dataset, and for each given pixel x and band b, the corresponding mean m_b was subtracted, $x(b) := x(b) - m_b$.

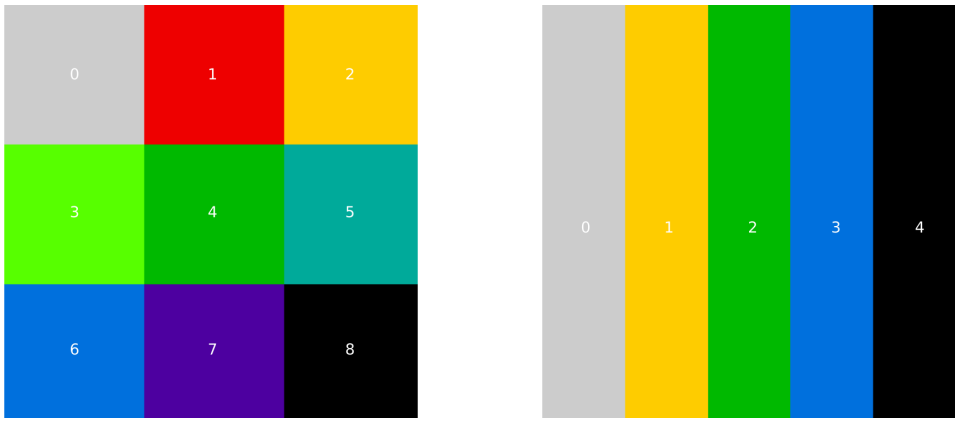
For this experiment, all three of the previously introduced neural network architectures were used. As discussed previously, the training was divided into pre-training and fine-tuning stages. In pre-training, the data were labelled through assigning an artificial class to each block within a grid of dimensions 5×5. No ground truth data was used at this stage. In the fine-tuning stage, a selected number of ground truth labels was used. The number of training samples from each class was set at $n = 5, 15, 50$. This allowed observing the performance both in typical hyperspectral scenarios (small number of classes used) and deep network scenarios (larger number of samples per class available). Because the classification accuracy depends on the training set used in fine-tuning each experiment was repeated $n = 15$ times for error reporting.The performance is reported in Overall Accuracy (OA) after fine tuning. Additionally, Average Accuracy (AA) and κ coefficient were inspected and improvements verified with statistical tests.

2.5.2. Experiment 2

The second experiment investigates the variability introduced by the size and shape of the patches used in artificial labelling.

For this experiment, only the Indian Pines image introduced in Experiment 1 was used, as it is the more challenging of the two introduced datasets. The mean was subtracted, as was the case with the previous experiment. The network investigated is the architecture based on [16], chosen because it has the most potential to be affected by the transfer learning process.

In this experiment, first the grid size was investigated. The dimensions of the patches, varies from 2×2, which equals 4 artificial classes, up to 72×72, 5184 artificial classes. Furthermore, another way of creating artificial labels is considered. The image is divided into the given number of vertical stripes. The visualisation of different artificial labelings is presented in Figure 2.



(**a**) Grid 3×3 (**b**) Stripes 5

Figure 2. Example scheme of creating artificial classes on Indian Pines dataset. From left: grid of 9 artificial classes (**a**), vertical stripes with 5 artificial classes (**b**). Artificial classes for Pavia University dataset were created analogically.

The investigated patches were created by dividing horizontal and vertical side of an image into $w = 2, 3, 5, 7, 9, 15, 19, 25, 29, 36, 39, 48, 72$ equal parts. The vertical stripes were created by dividing horizontal side of an image into $s = 2, 5, 9, 16, 25, 36, 49, 81$ equal parts (so, in the case of $s = 2$, there are only 2 classes located to the right and left of a single vertical line). The vertical stripes were included to observe whether the pixel distance affects the performance—for patches, all of the pixels share similar neighbourhood; for stripes, the top and bottom pixels have a notable spatial separation and, arguably, the distant pixels should not be marked with the same class label without prior knowledge of spatial class distribution. Note that, in the case of patches made by dividing each side of an image into $w = 29, 36, 39, 48, 72$ equal parts the size of a square patch is smaller than the size of a processing window 5×5 in tested architecture. That means no sample fed to a network during pre-training phase has a coherent class representation (i.e. a single class present in the window). This experiment was performed with 5 training samples per class and 50 experiment runs for each grid density and the number of stripes.

2.5.3. Experiment 3

In this experiment, we test the hypothesis that the more numerous patches' division produces a better pre-training set than the less numerous ones. We investigate this using a specially designed hyperspectral test image.

In this experiment, we use the image of paints from museum's collection. This dataset [64] was collected by the SPECIM hyperspectral system in the Laboratory of Analysis and Nondestructive Investigation of Heritage Objects (LANBOZ) in National Museum in Kraków. This image consists of 455×310 pixels. Each pixel has 256 spectral bands in the frequency range from 1000 to 2500 nm. Ground truth consists of manual annotations of different green pigments used in the mixture of paints for various painting regions. The image of oil paints on paper was used, selected from four available, as it was considered one of the more challenging of the images.

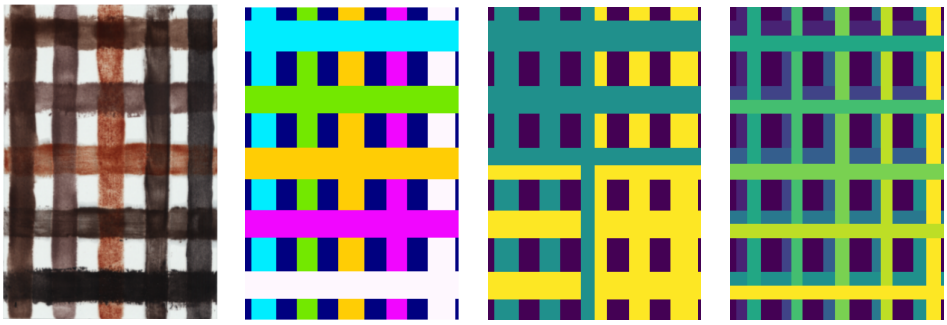
The layout of classes present in this image was especially designed to verify hyperspectral classifiers. The different chemical compositions of the paints used introduce variations of class spectra, yet at the same time all paints are variations of the green pigment with more or less greenish hue. The classification problem is thus difficult, but not exceedingly so. Regular grid layout, with different thickness of paints and fragments where one pigment overpaints another, introduce spatial diversity in the spectra. Because the image is artificially created, ground truth can be precisely marked. The original purpose of the image was to evaluate identification of copper pigments, difficult to differentiate by other (non-hyperspectral) sensors. Here we take advantage of its regularity by complementing the original ground truth ($n_{GT} = 5$ classes) with a joined set ($n_{GT-2} = 2$ classes) and split set ($n_{GT-10} = 10$ classes)

(see Figure 3). Those two sets of modified ground truth allow us to compare the proposed grid scheme, as tested in experiments 1 and 2, with a ground truth based pre-training with more and less classes than the original set. We argue that the regular layout of this image is more suited for this experiment than e.g., Indian Pines or Pavia University images; usage of additional dataset allows us to further verify the generalization potential of our approach.

In the case of this dataset, the mean was subtracted, as in the case of the previous experiments. Additionally, the standard deviation σ_b of each hyperspectral band b was calculated and then all of the pixels were divided by the corresponding standard deviation value σ_b, $x(b) := \frac{x(b)}{\sigma_b}$. In this experiment, as in the previous one, the neural network based on [16] was used. Training size was equal to five training samples per class and there were 50 experimental runs for each examined case.

The following cases were investigated:

1. The performance of DLNN with pre-training performed with 2 classes prepared from joining the ground truth classes (GT-2).
2. The performance of DLNN with pre-training performed with 10 classes prepared by splitting the ground truth classes (GT-10).
3. The performance of DLNN without pre-training (GT).
4. The performance of DLNN with pre-training with artificial patches of size $5 \times 5, 20 \times 20, 30 \times 30$.



(**a**) Painting (**b**) GT (**c**) GT-2 (**d**) GT-10

Figure 3. Scheme of creating artificial classes on Pigment dataset. From left: false-colour RGB (bands 50, 27, 17) image of the painting (**a**), original class labels (**b**), classes artificially joined into 2 sets (**c**), and classes artificially split into 10 sets (**d**). Dark rectangles denote background, excluded from the experiment.

2.5.4. Experiment 4

In this experiment, we examine the claims from Section 2.2 about the emergence of data-depenedent representations during neural network training using proposed artificial labelling scheme with noisy labels. To this end, we visualised internal network parameters resulting from network training while using t-SNE algorithm [65]. In the experiment, we used neural network architecture based on [16] and the Indian Pines dataset described in Section 2.5.1. We trained the network on the dataset using the following scenarios:

1. The network was trained using 1600 labelled samples, with 200 samples per class. This scenario represents the neural network trained with abundant information about the data—unrealistic, but convenient from the point of network's requirements.
2. The network was trained using 40 labelled samples, with 5 samples per class. This scenario represents the neural network trained with very limited information about the data–realistic, but difficult learning problem.
3. The network was trained using only the artificial labels created as explained in Section 2.3. Therefore, the network did not 'see' the true labels and could create the internal represenations only based on the noisy labels provided for training.

4. The network was trained using the complete pretraining-finetuning scheme introduced in this section. That is, first it was pretrained using artificial labels, as in point 3, and then all layers, except the last, was finetuned using the training set analogous to the one from point 2. This scenario was introduced to help explain the impact of the finetuning step in our approach.

As a result, we obtain 4 trained neural networks. As a next step, using validation dataset, we extract the activations of the next-to-last layers of the considered networks, and use the t-SNE algorithm, which is used to visualise high-dimensional data, in order to learn whether the layers right before the classification layers of the networks did learn useful data representations.

3. Results

This Section presents the results of experiments introduced in Section 2.5.

3.1. Experiment 1

Table 1 presents the first experiment's results. Each column presents the result for one type of network, each row for a set dataset and the number of training examples. Each table cell presents the results with and without pre-training, in percent of Overall Accuracy, including the standard deviation of the result. The results from Table 1 were computed from a batch of $n = 15$ independent runs for each case. The specific value of n was chosen to provide robust result, after a set of preliminary runs with different n values. A Mann–Whitney U test was performed on the results in order to confirm statistical significance of the improvement gained with the proposed method. As Overall Accuracy can be sensitive to class imbalances, Average Accuracy and κ coefficient were computed for additional verification, and were inspected for negative performance.

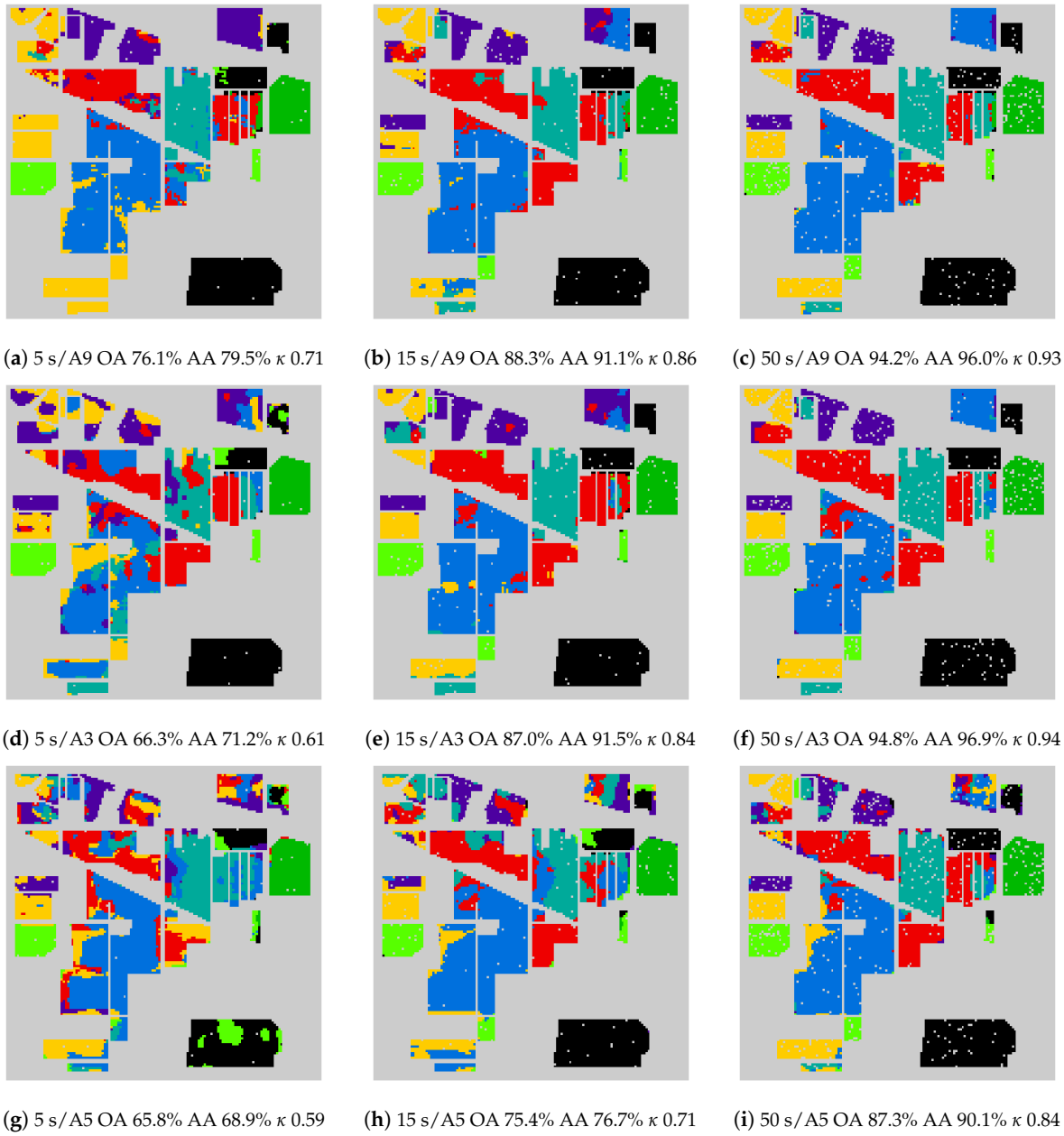
Table 1. The result of the first experiment. Each row presents Overall Accuracy (OA), Average Accuracy (AA), and Cohen's kappa (κ) for given scenario. IP denotes the Indian Pines dataset, PU the Pavia University; further differentiation is for number of samples per class in fine-tuning. Accuracies are given as averages with standard deviations with and without pretraining for the three investigated network architectures.

		Architecture [16]		**Architecture [15]**		**Architecture [63]**	
		No Pretraining	**Pretraining**	**No Pretraining**	**Pretraining**	**No Pretraining**	**Pretraining**
IP 5/class	OA:	52.62 ± 4.4	74.04 ± 4.1 †	66.15 ± 4.5	72.80 ± 3.2 †	50.05 ± 5.1	63.52 ± 4.2 †
	AA:	58.15 ± 2.9	78.83 ± 2.6 †	71.42 ± 3.9	78.60 ± 3.0 †	53.66 ± 3.0	65.86 ± 3.6 †
	κ:	0.45 ± 0.04	0.69 ± 0.05 †	0.60 ± 0.05	0.68 ± 0.04 †	0.41 ± 0.05	0.57 ± 0.05 †
IP 15/class	OA:	67.58 ± 3.2	87.04 ± 2.4 †	82.61 ± 2.8	87.04 ± 2.1 †	64.18 ± 2.8	75.30 ± 1.7 †
	AA:	73.82 ± 2.7	90.41 ± 1.5 †	87.07 ± 1.8	90.97 ± 1.5 †	67.54 ± 2.5	78.40 ± 2.1 †
	κ:	0.62 ± 0.03	0.85 ± 0.03 †	0.79 ± 0.03	0.85 ± 0.02 †	0.58 ± 0.03	0.71 ± 0.02 †
IP 50/class	OA:	80.51 ± 4.8	93.66 ± 1.3 †	93.75 ± 1.2	94.65 ± 1.0	81.39 ± 1.1	87.06 ± 0.9 †
	AA:	87.48 ± 2.6	95.81 ± 0.8 †	95.86 ± 0.7	96.62 ± 0.8 †	85.10 ± 0.9	90.38 ± 1.1 †
	κ:	0.77 ± 0.05	0.92 ± 0.02 †	0.92 ± 0.01	0.94 ± 0.01	0.78 ± 0.01	0.85 ± 0.01 †
PU 5/class	OA:	67.47 ± 6.5	80.08 ± 7.0 †	73.31 ± 4.1	80.33 ± 5.2 †	65.55 ± 3.8	74.34 ± 7.0 †
	AA:	76.56 ± 3.1	87.66 ± 2.9 †	84.67 ± 2.6	88.86 ± 3.2 †	64.39 ± 2.4	76.92 ± 3.7 †
	κ:	0.60 ± 0.07	0.75 ± 0.08 †	0.67 ± 0.05	0.76 ± 0.06 †	0.56 ± 0.04	0.68 ± 0.08 †
PU 15/class	OA:	83.63 ± 2.7	91.87 ± 3.3 †	88.21 ± 2.9	91.96 ± 2.6 †	75.50 ± 2.4	89.33 ± 3.4 †
	AA:	89.48 ± 1.1	94.65 ± 1.0 †	93.40 ± 1.1	95.01 ± 0.8 †	77.59 ± 1.2	89.95 ± 1.8 †
	κ:	0.79 ± 0.03	0.90 ± 0.04†	0.85 ± 0.04	0.90 ± 0.03 †	0.69 ± 0.03	0.86 ± 0.04 †
PU 50/class	OA:	93.40 ± 1.4	97.86 ± 0.5 †	96.08 ± 0.9	96.84 ± 1.2 ‡	87.79 ± 1.7	96.55 ± 0.5 †
	AA:	95.47 ± 0.7	98.13 ± 0.3 †	97.09 ± 0.5	97.90 ± 0.4 †	89.05 ± 0.9	96.37 ± 0.3 †
	κ:	0.91 ± 0.02	0.97 ± 0.01 †	0.95 ± 0.01	0.96 ± 0.02 ‡	0.84 ± 0.02	0.95 ± 0.01 †

†,‡ Statistically significant improvement, evaluated with Mann–Whitney U test, with $P < 0.01$ (†) or $P < 0.05$ (‡).

The presented results show that the application of the proposed method leads to definite and consistent improvement in accuracy across different images, number of ground truth labels used and network architectures. In all but one case, the improvement is statistically significant, and in some cases

approaches 20 percentage points. The most challenging is scenario with 5 training samples per class. Even average overall accuracy achieved by architecture originally examined on small training set [15] does not exceed 67% on Indian Pines dataset. After the application of the proposed method, performance improves up to 72.8% OA. The most improvement is seen in the architecture [16], namely on IP dataset with only five training samples per class in fine-tuning procedure, it improves from average 52.62 OA to 74.04 OA. This is to be expected as this architecture has the most potential to benefit from additional training samples. When considering these improvements, it can be summarized that the results of the experiment support stated hypothesis and the validity of the proposed approach. The qualitative evaluation of selected realizations (corresponding to the median score) is presented in Figures 4 and 5.



(**a**) 5 s/A9 OA 76.1% AA 79.5% κ 0.71 (**b**) 15 s/A9 OA 88.3% AA 91.1% κ 0.86 (**c**) 50 s/A9 OA 94.2% AA 96.0% κ 0.93

(**d**) 5 s/A3 OA 66.3% AA 71.2% κ 0.61 (**e**) 15 s/A3 OA 87.0% AA 91.5% κ 0.84 (**f**) 50 s/A3 OA 94.8% AA 96.9% κ 0.94

(**g**) 5 s/A5 OA 65.8% AA 68.9% κ 0.59 (**h**) 15 s/A5 OA 75.4% AA 76.7% κ 0.71 (**i**) 50 s/A5 OA 87.3% AA 90.1% κ 0.84

Figure 4. Sample results from experiment one, Indian Pines dataset. Rows present the three examined architectures, where A9, A3 and A5 corresponds to architectures [15,16,63], respectively. Columns present the three cases of number of true training samples per class in fine-tuning (5 s, 15 s, and 50 s). For each result, the Overall Accuracy (OA), Average Accuracy (AA), and κ coefficient are reported. Isolated grey points mark locations of the training samples, and are excluded from the evaluation.

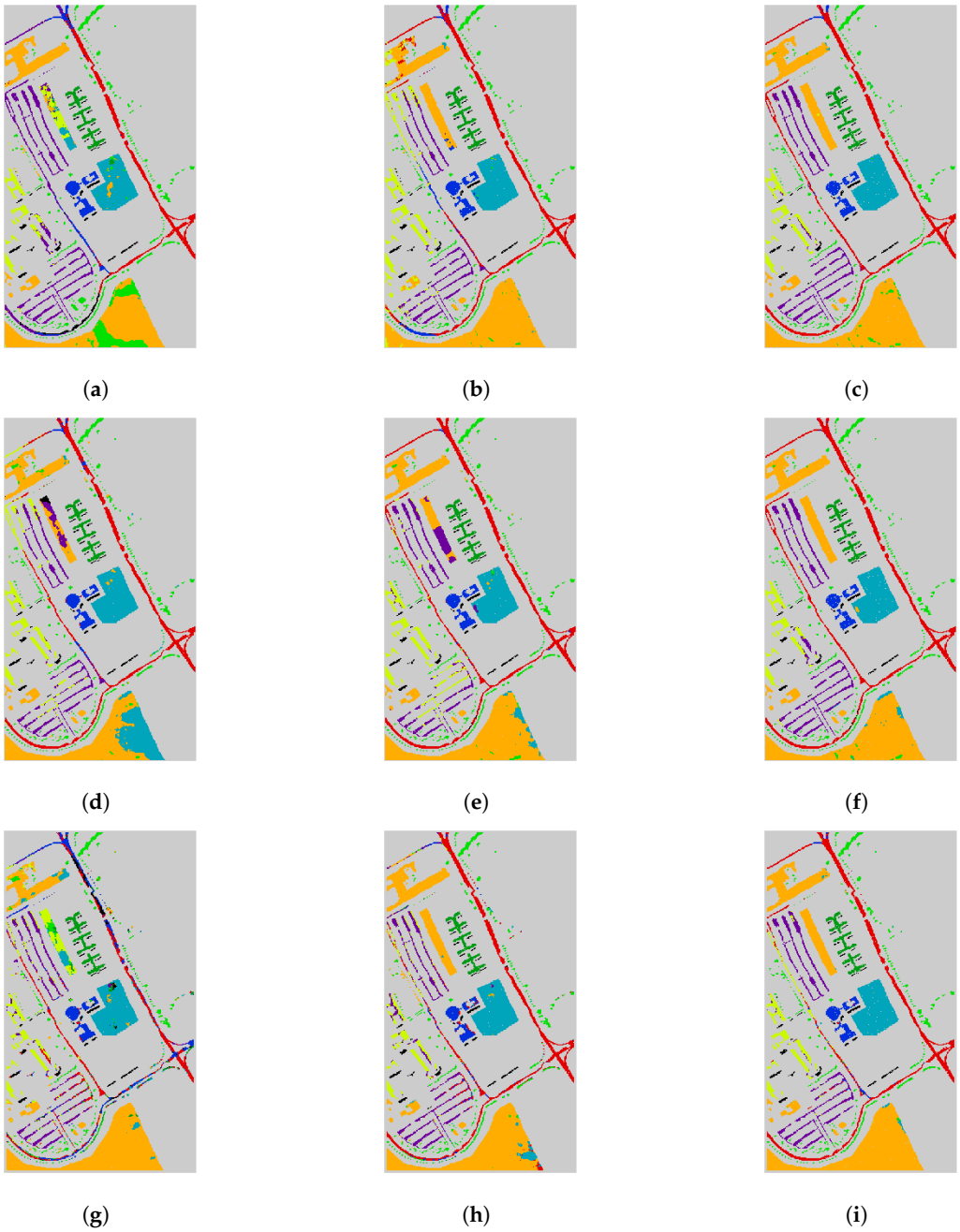


Figure 5. Sample results from experiment one, Pavia University dataset. The scheme is identical to the Figure 4. (**a**) 5 s/A9 OA 79.7% AA 88.4% κ 0.75; (**b**) 15 s/A9 OA 91.3.3% AA 93.8% κ 0.89; (**c**) 50 s/A9 OA 97.8% AA 98.2% κ 0.97; (**d**) 5 s/A3 OA 81.7% AA 91.0% κ 0.77; (**e**) 15 s/A3 OA 92.3% AA 94.7% κ 0.90; (**f**) 50 s/A3 OA 97.4% AA 97.3% κ 0.97; (**g**) 5 s/A5 OA 77.1% AA 79.8% κ 0.71; (**h**) 15 s/A5 OA 90.4% AA 88.6% κ 0.88; (**i**) 50 s/A5 OA 96.6% AA 96.4% κ 0.96.

3.2. Experiment 2

Table 2 presents the results of the experiment. For each grid size or the number of stripes, the overall accuracy and the standard deviation are given. These statistics are based on 50 experiment runs for each artificial labelling scheme.

Table 2. The second experiment results. Grid density describes number of rectangular patches which represent artificial labels for pre-training phase. Num of stripes denotes number of vertical stripes that represent artificial labels for pre-training phase. Accuracies are given as Overall Accuracy for learning of the network based on [16] with transfer learning on the Indian Pines dataset.

Grid Density/Model	OA	Num of Stripes/Model	OA
(2 × 2)	61.88 ± 4.5	2	58.53 ± 4.9
(3 × 3)	64.45 ± 4.1	5	67.68 ± 4.4
(5 × 5)	75.05 ± 4.3	9	68.58 ± 3.9
(7 × 7)	72.33 ± 4.4	16	69.25 ± 3.7
(9 × 9)	74.06 ± 3.6	25	69.09 ± 3.6
(14 × 14)	74.13 ± 3.7	36	69.23 ± 3.3
(19 × 19)	73.24 ± 3.9	49	70.08 ± 4.0
(24 × 24)	73.43 ± 4.0	81	68.41 ± 4.3
(29 × 29)	70.19 ± 4.6		
(36 × 36)	69.88 ± 4.3		
(39 × 39)	68.69 ± 4.4		
(48 × 48)	69.25 ± 3.2		
(72 × 72)	66.70 ± 3.3		

It can be seen that the the score rises sharply until the number of artificial classes reaches approximately the number of original classes (at 5 × 5, note that the original IP ground truth leaves a sizeable portion of background unmarked, which most probably would contribute some additional classes if marked). After that value, there's a declining trend. It can be noted that the scores are higher with smaller patches. It seems viable to form a conclusion that, when the original class number is unknown, it is better to overestimate than underestimate their number. In the latter case, it is possible that even a chance guess would provide a satisfactory performance. The stripes do not form as good a training set as rectangular grid, which confirms the initial supposition that artificial classes should be confined to local areas. Some improvement however is still seen, which supports our overall proposition, that general artificial labelling can be used for improving the DLNN performance without precise estimation of the artificial class patch size.

3.3. Experiment 3

Table 3 presents the results of the third experiment. The overall accuracy was calculated based on $n = 50$ runs for each examined scenario.

Table 3. The third experiment results. Evaluation of pretraining on Pigments dataset using the proposed approach and classes created from ground truth. The objective was to collate the performance of artificial labels of different sizes with those created through splitting or joining the ground truth.

Experiment Setting	GT [a]	(5 × 5) [b]	(20 × 20) [b]	(30 × 30) [b]	GT-2 [c]	GT-10 [c]
OA	61.15%	68.35%	75.70%	73.99%	75.70%	85.40%

[a] No pretraining. [b] Pretraining with artificial classes (proposed method). [c] Pretraining with modified ground truth classes (verification).

Here, the original performance (GT) can be significantly improved by the grid-based artificial labelling (see results for 5 × 5, 20 × 20, 30 × 30). However, in this case, the performance gain can be confronted with a label dataset created from ground truth data (GT-2, GT-10). As can be expected, the ground truth data provides a higher performance; however, the artificial labelling provides half of that gain with no prior information needed. The ground truth experiments GT-2 and GT-10 also confirm the observation that classes split is a better option than joining. The latter observation provides an additional support to the conclusion that more small classes (dense grid) is preferable than few large ones (sparse grid).

3.4. Experiment 4

The results of the experiment are presented in Figure 6. As expected, the network trained using 1600 true-labeled samples generated good internal representations, which can be seen by the good separability of the classes. In contrast, neural network trained using only 5 samples per class did not generate representations allowing the separation of samples of different classes. In the case of scenario 3, we can clearly see that the classes were better separated when compared with scenario 2, though of course not as good as in scenario 1. Moreover, the authors did not observe any noticeable differences between scenarios 3 and 4.

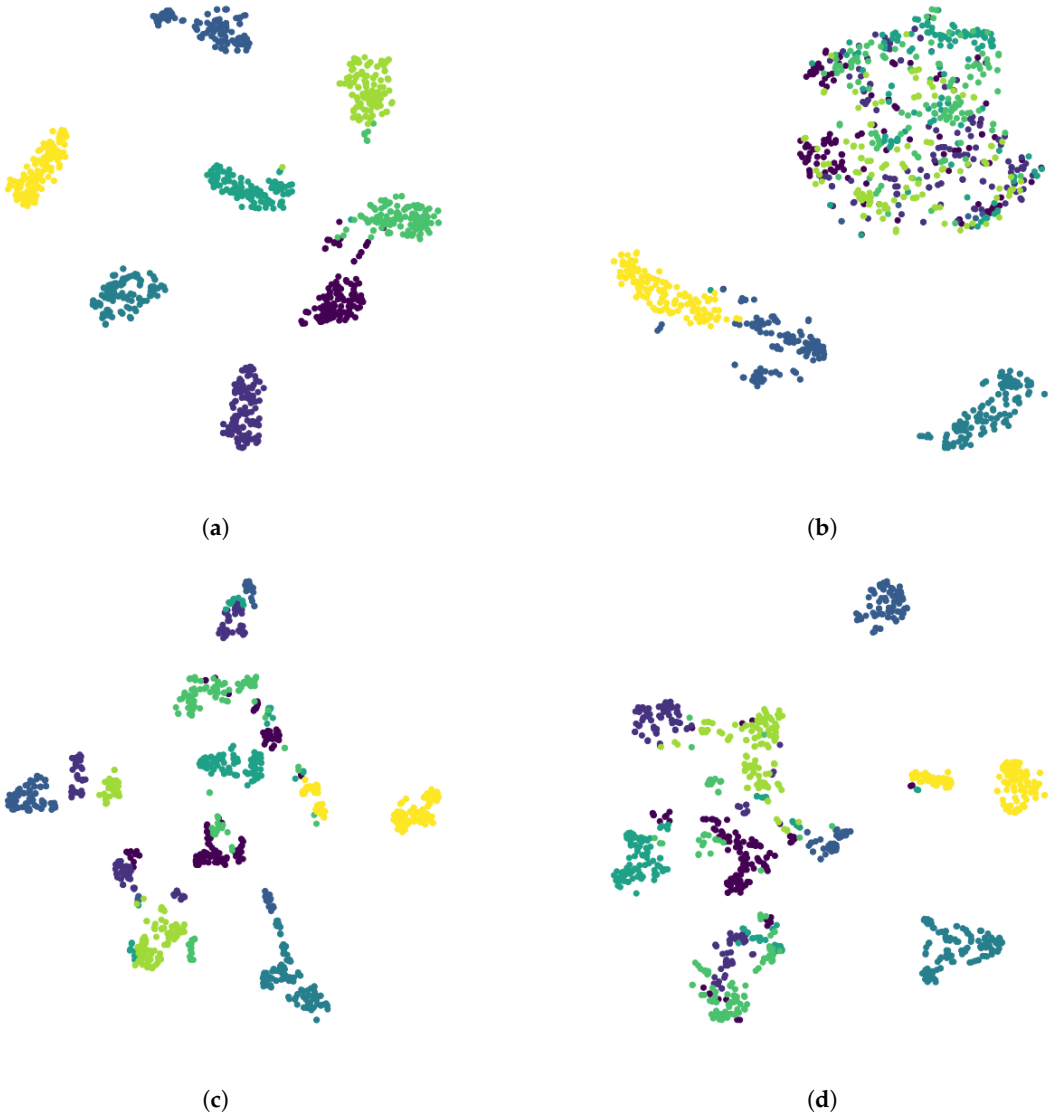


Figure 6. The visualisation of the learned parameters for four networks introduced in Section 2.5.4. Each point represents given sample's activations transformed to the two-dimensional space using t-SNE algorithm. Different colors represent different classes present on the image. (**a**) Activations from network trained using 200 samples/class; (**b**) Activations from network trained using 5 samples/class; (**c**) Activations from network trained using artificial labels only; (**d**) Activations from network trained with fine-tuning (artificial followed by training labels)

We argue that the presented results provide some suggestion that, during neural network training using proposed artificial labelling scheme, there is an emergence of useful data-related representations, even before the fine-tuning step.

4. Discussion

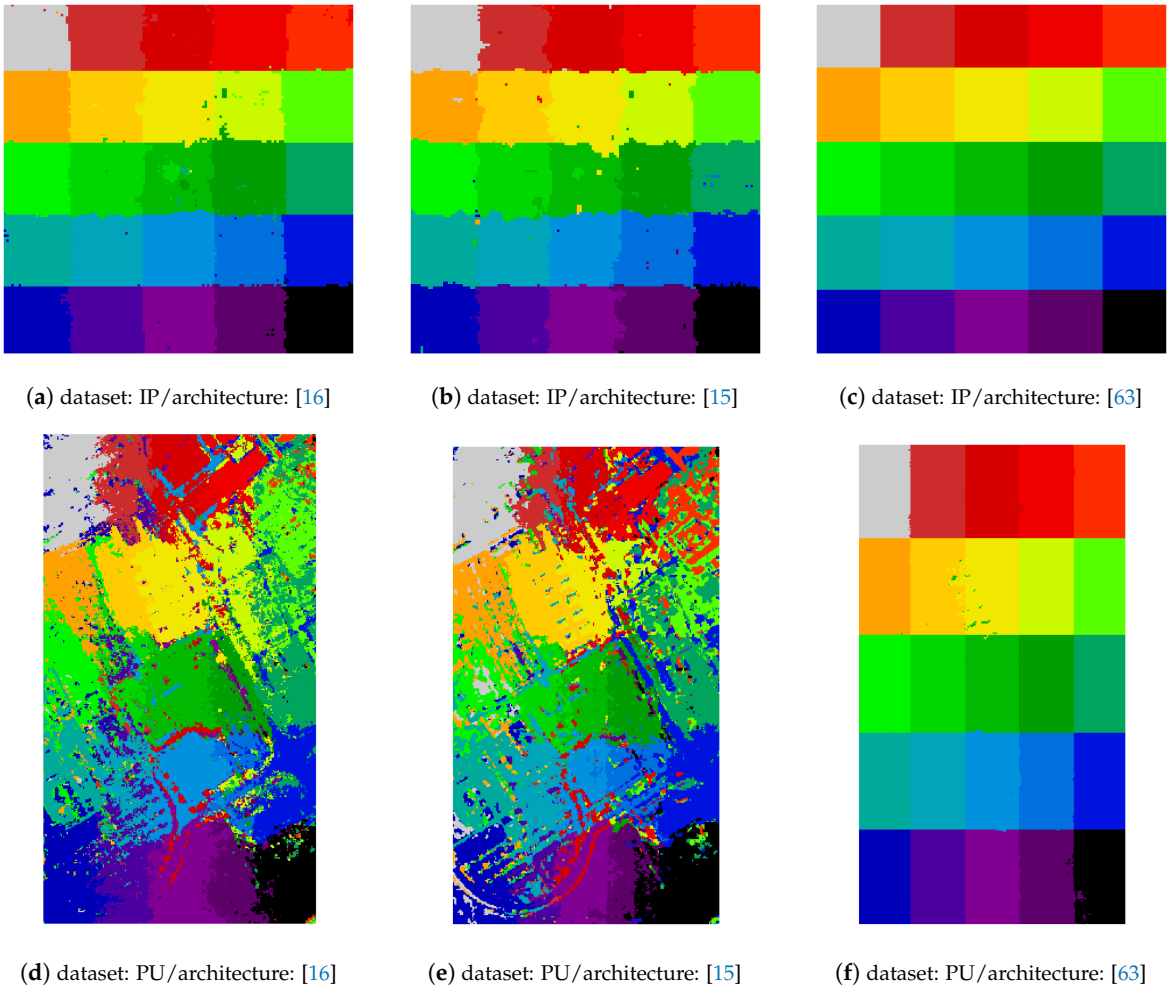
Our results confirm the validity of our proposition: a simple artificial labelling through grouping of the samples based on a local neighbourhood provides an efficient transfer learning scheme. It brings significant improvements of accuracy across datasets and DLNN configurations. The results for different datasets, which have distinctive ground truth layouts, suggest that it is not the random alignment with the regularity of a particular ground truth pattern. It is also seen that the local structure is important, as seen in the advantage of grid division over stripe division. The generally better performance of higher over lower number of artificial classes suggests an explanation in that for transfer learning, it is not as important to locate the exact number of classes, but to isolate and learn their components, perhaps for better internal feature representation.

We view the main advantage of the proposed method as enhancing the training of a neural network for hyperspectral remote sensing classification. The proposed pre-training offers a number of benefits:

1. Enhance the training of neural networks in hyperspectral classification scenario. With low number of training samples in typical scenarios (e.g., 5–15/class, sometimes even less) the number of network free parameters can be several orders of magnitude higher than the training data, which poses a risk of overtraining.
2. Through splitting the training into two phases, it can be used to shift some of the computational burden of network training to the time before an expert is called in to perform labelling, and to make more effective use of his or her time.
3. Larger number of training samples available can be of use in case different network architectures are compared for the same problem, or during the searching the hyperparameter space.

An open question is whether a clustering algorithm, like [33] or outlier segmentation [48] could be adapted here, leading to greater efficiency. It is probable that a more complex artificial labelling algorithm could outperform the proposed solution; however, even in that case, a simple, generally applicable heuristic that improves performance can be of value. Our approach has common motivation with self-taught learning [49], where we want the classifier to derive high-level input representation from the unlabelled data; however, we use the same data for both training stages and instead change the label set. It also avoids combining neural and non-neural approaches, and it prevents introducing additional assumptions through the manual selection of the latter.

A qualitative examination of the pre-training results shows that some class structure is visible after pre-training (see examples in Figure 7). No identifiable features of this structure have been noticed when investigating pre-training images when associated with better or worse final (after fine-tuning) results. However, the general level of structure visible after pre-training relates to the final performance. The network architecture based on the work [63] is best in learning the artificial classes grid and also the worst at the final classification. The other two networks that are based on the works [15,16] have more complex pre-training results and correspondingly better final results. This suggests that the training scheme and/or network architecture functions as a form of regularization that prevents overtraining, and that the pre-training classification result can be possibly used to control the pre-training process and avoid overtraining too. The emergence of partial class structure in the pre-training phase—which does not use ground truth, hence can be viewed as unsupervised processing—also suggests that this approach can be adapted to solve unsupervised tasks, e.g., clustering or anomaly detection.



(**a**) dataset: IP/architecture: [16] (**b**) dataset: IP/architecture: [15] (**c**) dataset: IP/architecture: [63]

(**d**) dataset: PU/architecture: [16] (**e**) dataset: PU/architecture: [15] (**f**) dataset: PU/architecture: [63]

Figure 7. Sample pre-training results. Top row Indian Pines, bottom row Pavia University datasets. Columns present the three architectures studied (based on the works [15,16,63]). Some class structure is visible depending on the dataset and network selected.

In order to provide additional verification, we have analysed per-class classification scores for both datasets, using the data from experiment one, and the same Mann–Whitney U with $P < 0.05$. As could be expected, performance gains are unequal, as classes differ with their overlap and general difficulty of classification. However, the individual classes showed improvement in most of the cases. Across 198 tests, in 104 cases the improvement was statistically significant; for the remaining cases, in 39 cases the accuracy of 100% was achieved irrespective of pre-training, in 32 cases pre-training improved the mean of the class score. In the remaining cases where pre-training score mean was lower that the reference, the average difference was below two percentage points. The proposed method thus can be viewed as 'not damaging' to individual class scores.

Additionally, a batch of experiments were performed for sensitivity analysis of small variations of hyperparameter setting; the results were very similar to those presented. A separate experiment was conducted analysing time-requirements when training the networks. The results of the experiments are presented in Figure 8. The results show that it is more important to train the network during pre-training stage than during the fine-tuning stage (one can clearly see the results getting better when moving vertically within a grid from Figure 8, as opposed to moving horizontally). As one can see, in the case of the lower number of pre-training iterations (10 k–50 k), even moderate increase leads to definite improvement in the accuracy of the classification. The results also suggest that it could be possible to reduce the time of training in both of the stages without sacrificing the effectiveness of classification. Moreover, it can be presumed that choosing a different number of iterations of the

pre-training and fine-tuning stages could lead to achieving even better results than the ones presented in this work (for example, when training the network for 90k iterations in the pre-training stage, and 20 k iterations in the fine-tuning stage, it was possible to achieve the accuracy of 81.34%).

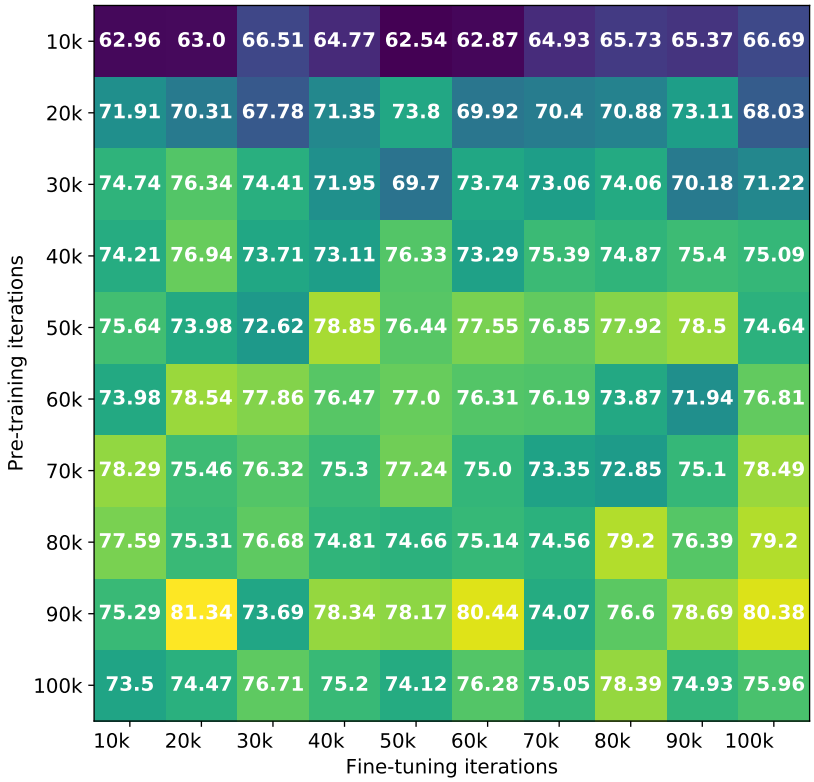


Figure 8. The classification accuracies of the networks trained on 5 samples/class and tested with the rest of the image using the Indian Pines dataset and neural network based on [16]. On the y-axis, the number of epochs for the pre-training stage is written, while on the x-axis the number of epochs for the fine-tuning stage is written. The resuts suggest higher relative importance of the pre-training stage in comparison to the fine-tuning stage.

Analysing the results from the Table 1, one can notice that pre-training improves the accuracy in some networks more than in others. We suspect that an important factor determining such differences is the capacity of neural networks. We argue this with fact that artificial neural network with greater number of parameters is able to better process information contained in the entire image, which we utilize in the pre-training phase. Therefore, the architecture [15] with the smallest number of parameters achieves a smaller increment of the accuracy in comparison to other two networks.

However, one must to be aware that there are a number of other factors that affect network performance. In particular, architectures [15,16] were designed for the task of HSI classification. With an emphasis on the architecture [15], which has been studied on a small training data sets and, therefore, has competitive accuracy even without pre-training. On the other hand, architecture [63] was designed for a slightly different training regimen, which may explain the fact that it achieves worse results than the other two.

Our approach could be used for semi-automatic systems, like [66], which only use a part of the annotation, and could be made fully unsupervised. Furthermore, we believe this is one approach for self-taught learning [49], which can be helpful in diverse application of deep learning models. We note, however, that optimization would require further studies to address the issue of which layers benefit most of this scheme, i.e. similar to [36]. Our experiments show that the proposed scheme is largely resistant to the incorrect estimation of the number of classes, hence its parametrization can

be considered low-cost. It can be also viewed as a confirmation of traditional software development principle of 'divide and conquer', as of even older proverb, 'divide et impera'.

5. Conclusions

We have presented and verified a simple method pre-training of DLNN for hyperspectral classification based on the hypothesis that spatial similarity of unlabelled data points can be utilized in order to gain accuracy in hyperspectral classification. In the first experiment we showed that, for all three neural network architectures tested and for the all two reference datasets, the proposed procedure leads to an improvement of classification efficiency for small number of training samples. In the second and third experiments, we analysed the properties of the proposed method; the obtained results suggest that the number and shape of the pixel blobs have an impact on the effectiveness of the method. Specifically, we conclude from the second experiment that it is safer to underestimate the size of a label cluster rather than overestimate and simultaneously reduce chance of joining separate classes. This conclusion is in line with results of the third experiment, from which we also conclude that it is better to split ground truth classes than join them. In the fourth experiment, we confirmed the claim about the emergence of data-independent representations during neural network training using artificial labelling.

The absence of training labels requirement provides an important advantage: it shifts the need of expert's participation and data labelling from the start of the data analysis process to its late stages. This allows for the use of the potentially long time from the acquisition to the start of data interpretation stage for pre-training the network, and decreases the delay between expert's labelling to getting the classification result. Considering the length of time required to train deep neural networks, this is a significant advantage for their applications. An additional benefit is that multiple unannotated images can be used in the pre-training stage, potentially increasing the robustness of the result.

Author Contributions: Conceptualization, W.M., P.G., B.G., M.O.; methodology, W.M., P.G., B.G., M.O.; software, W.M., B.G., M.O.; validation, W.M., B.G.; investigation, W.M., P.G., B.G., M.O.; writing—original draft preparation, W.M., P.G., B.G., M.O.; writing—review and editing, W.M., P.G., B.G., M.O.; visualization, W.M.; supervision, P.G.; funding acquisition, P.G., B.G., M.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the projects: 'Representation of dynamic 3D scenes using the Atomic Shapes Network model' financed by the National Science Centre, decision DEC-2011/03/D/ST6/03753 and 'Application of transfer learning methods in the problem of hyperspectral images classification using convolutional neural networks' funded from the Polish budget funds for science in the years 2018-2022, as a scientific project under the "Diamond Grant" program, no. DI2017 013847. M.O. acknowledges support from Polish National Science Center scholarship 2018/28/T/ST6/00429.

Acknowledgments: This research was supported in part by PLGrid Infrastructure. The authors would like to thank Laboratory of Analysis and Nondestructive Investigation of Heritage Objects (LANBOZ) in National Museum in Kraków for providing the pigments dataset, in particular to Janna Simone Mostert for her help in the preparation of paintings and Agata Mendys for acquisition of the dataset. The authors also thank Zbigniew Puchała for help in carrying out statistical analysis of the results. Additionally authors thank Yu et al [15] for sharing the code.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Bioucas-Dias, J.; Plaza, A.; Camps-Valls, G.; Scheunders, P.; Nasrabadi, N.; Chanussot, J. Hyperspectral Remote Sensing Data Analysis and Future Challenges. *IEEE Geosci. Remote Sens. Mag.* **2013**, *1*, 6–36. [CrossRef]
2. Ghamisi, P.; Plaza, J.; Chen, Y.; Li, J.; Plaza, A.J. Advanced Spectral Classifiers for Hyperspectral Images: A review. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–32. [CrossRef]
3. Chang, C.I. *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*; Springer: Boston, MA, USA, 2003. [CrossRef]

4. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with Support Vector Machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
5. Romaszewski, M.; Głomb, P.; Cholewa, M. Semi-supervised hyperspectral classification from a small number of training samples using a co-training approach. *ISPRS J. Photogramm. Remote Sens.* **2016**, *121*, 60–76. [CrossRef]
6. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: USA, 2012; pp. 1097–1105.
7. LeCun, Y.; Bengio, Y.; others. Convolutional networks for images, speech, and time series. *Handb. Brain Theory Neural Netw.* **1995**, *3361*, 1995.
8. Smolensky, P. *Information Processing in Dynamical Systems: Foundations of Harmony Theory*; Technical Report; Colorado University at Boulder Department of Computer Science: Boulder, CO, USA, 1986.
9. Ackley, D.H.; Hinton, G.E.; Sejnowski, T.J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **1985**, *9*, 147–169. [CrossRef]
10. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
11. Hinton, G.E. Deep belief networks. *Scholarpedia* **2009**, *4*, 5947. [CrossRef]
12. Hinton, G.E.; Osindero, S.; Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
13. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
14. Dey, R.; Salemt, F.M. Gate-variants of gated recurrent unit (GRU) neural networks. In Proceedings of the 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, USA, 6–9 August 2017; pp. 1597–1600.
15. Yu, S.; Jia, S.; Xu, C. Convolutional neural networks for hyperspectral image classification. *Neurocomputing* **2017**, *219*, 88–98. [CrossRef]
16. Lee, H.; Kwon, H. Going Deeper With Contextual CNN for Hyperspectral Image Classification. *IEEE Trans. Image Proc.* **2017**, *26*, 4843–4855. [CrossRef] [PubMed]
17. Han, M.; Cong, R.; Li, X.; Fu, H.; Lei, J. Joint spatial-spectral hyperspectral image classification based on convolutional neural network. *Pattern Recognit. Lett.* **2018**. [CrossRef]
18. Zhou, X.; Liu, N.; Tang, F.; Zhao, Y.; Qin, K.; Zhang, L.; Li, D. A deep manifold learning approach for spatial-spectral classification with limited labeled training samples. *Neurocomputing* **2019**, *331*, 138–149. [CrossRef]
19. Xu, Y.; Du, B.; Zhang, F.; Zhang, L. Hyperspectral image classification via a random patches network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 344–357. [CrossRef]
20. Pan, B.; Shi, Z.; Xu, X. MugNet: Deep learning for hyperspectral image classification using limited samples. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 108–119. [CrossRef]
21. Gao, H.; Yang, Y.; Lei, S.; Li, C.; Zhou, H.; Qu, X. Multi-branch fusion network for hyperspectral image classification. *Knowl.-Based Syst.* **2019**, *167*, 11–25. [CrossRef]
22. Zhao, G.; Liu, G.; Fang, L.; Tu, B.; Ghamisi, P. Multiple convolutional layers fusion framework for hyperspectral image classification. *Neurocomputing* **2019**. [CrossRef]
23. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
24. Wu, H.; Prasad, S. Convolutional recurrent neural networks for hyperspectral data classification. *Remote Sens.* **2017**, *9*, 298. [CrossRef]
25. Chen, Y.; Lin, Z.; Zhao, X.; Wang, G.; Gu, Y. Deep learning-based classification of hyperspectral data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 2094–2107. [CrossRef]
26. Fan, Y.; Zhang, C.; Liu, Z.; Qiu, Z.; He, Y. Cost-sensitive stacked sparse auto-encoder models to detect striped stem borer infestation on rice based on hyperspectral imaging. *Knowl.-Based Syst.* **2019**, *168*, 49–58. [CrossRef]
27. Guo, Y.; Han, S.; Cao, H.; Zhang, Y.; Wang, Q. Guided filter based Deep Recurrent Neural Networks for Hyperspectral Image Classification. *Procedia Comput. Sci.* **2018**, *129*, 219–223. [CrossRef]
28. Shi, C.; Pun, C.M. Multi-scale hierarchical recurrent neural networks for hyperspectral image classification. *Neurocomputing* **2018**, *294*, 82–93. [CrossRef]

29. Plaza, A.; Benediktsson, J.A.; Boardman, J.W.; Brazile, J.; Bruzzone, L.; Camps-Valls, G.; Chanussot, J.; Fauvel, M.; Gamba, P.; Gualtieri, A.; et al. Recent advances in techniques for hyperspectral image processing. *Remote Sens. Environ.* **2009**, *113*, S110–S122. [CrossRef]
30. Cholewa, M.; Głomb, P.; Romaszewski, M. A Spatial-Spectral Disagreement-Based Sample Selection with an Application to Hyperspectral Data Classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 467–471. [CrossRef]
31. Tarabalka, Y.; Chanussot, J.; Benediktsson, J.A. Segmentation and classification of hyperspectral images using minimum spanning forest grown from automatically selected markers. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2010**, *40*, 1267–1279. [CrossRef]
32. Dópido, I.; Li, J.; Plaza, A.; Gamba, P. Semi-supervised classification of urban hyperspectral data using spectral unmixing concepts. In Proceedings of the Urban Remote Sensing Event (JURSE), Sao Paulo, Brazil, 21–23 April 2013; pp. 186–189.
33. Wu, H.; Prasad, S. Semi-Supervised Deep Learning Using Pseudo Labels for Hyperspectral Image Classification. *IEEE Trans. Image Proc.* **2018**, *27*, 1259–1270. [CrossRef]
34. Windrim, L.; Melkumyan, A.; Murphy, R.J.; Chlingaryan, A.; Ramakrishnan, R. Pretraining for Hyperspectral Convolutional Neural Network Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2798–2810. [CrossRef]
35. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
36. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How Transferable Are Features in Deep Neural Networks? In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; MIT Press: Cambridge, MA, USA, 2014; Volume 2, pp. 3320–3328.
37. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15), Seattle, WA, USA, 9–13 November 2015; pp. 443–449. [CrossRef]
38. Xie, M.; Jean, N.; Burke, M.; Lobell, D.; Ermon, S. Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. *arXiv* **2015**, arXiv:cs.CV/1510.00098
39. Shin, H.C.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *arXiv* **2016**, arXiv:cs.CV/1602.03409
40. Zhou, P.; Cheng, G.; Liu, Z.; Bu, S.; Hu, X. Weakly supervised target detection in remote sensing images based on transferred deep features and negative bootstrapping. *Multidimens. Syst. Signal Process.* **2016**, *27*, 925–944. [CrossRef]
41. Lyu, H.; Lu, H. A deep information based transfer learning method to detect annual urban dynamics of Beijing and Newyork from 1984–2016. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 1958–1961. [CrossRef]
42. Hu, F.; Xia, G.S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
43. Lyu, H.; Lu, H.; Mou, L. Learning a Transferable Change Rule from a Recurrent Neural Network for Land Cover Change Detection. *Remote Sens.* **2016**, *8*, 506. [CrossRef]
44. Li, W.; Wu, G.; Du, Q. Transferred Deep Learning for Anomaly Detection in Hyperspectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 597–601. [CrossRef]
45. Lin, J.; Ward, R.; Wang, Z.J. Deep Transfer Learning for Hyperspectral Image Classification. In Proceedings of the 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), Vancouver, BC, Canada, 29–31 August 2018; pp. 1–5. [CrossRef]
46. Yuan, Y.; Zheng, X.; Lu, X. Hyperspectral Image Superresolution by Transfer Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 1963–1974. [CrossRef]
47. Fang, B.; Li, Y.; Zhang, H.; Chan, J.C.W. Semi-Supervised Deep Learning Classification for Hyperspectral Image Based on Dual-Strategy Sample Selection. *Remote Sens.* **2018**, *10*. [CrossRef]
48. Du, B.; Zhang, L.; Tao, D.; Zhang, D. Unsupervised transfer learning for target detection from hyperspectral images. *Neurocomputing* **2013**, *120*, 72–82. Image Feature Detection and Description. [CrossRef]

49. Raina, R.; Battle, A.; Lee, H.; Packer, B.; Ng, A.Y. Self-taught Learning: Transfer Learning from Unlabeled Data. In Proceedings of the 24th International Conference on Machine Learning (ICML'07), Brno, Czech Republic, 28–30 June 2007; pp. 759–766. [CrossRef]
50. Chapelle, O.; Scholkopf, B.; Zien, A. *Semi-Supervised Learning*; The MIT Press: Cambridge, MA, USA, 2006.
51. Rolnick, D.; Veit, A.; Belongie, S.; Shavit, N. *Deep Learning is Robust to Massive Label Noise*; *arXiv* **2018**, arXiv:cs.LG/1705.10694
52. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:stat.ML/1503.02531
53. Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.; Chen, X. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: USA, 2016; pp. 2234–2242.
54. Tan, K.; Li, E.; Du, Q.; Du, P. An efficient semi-supervised classification approach for hyperspectral imagery. *ISPRS J. Photogramm. Remote Sens.* **2014**, *97*, 36–45. [CrossRef]
55. Wang, L.; Hao, S.; Wang, Q.; Wang, Y. Semi-supervised classification for hyperspectral imagery based on spatial-spectral Label Propagation. *ISPRS J. Photogramm. Remote Sens.* **2014**, *97*, 123–137. [CrossRef]
56. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: https://www.deeplearningbook.org (accessed on 17 August 2020).
57. Lee, H.; Grosse, R.; Ranganath, R.; Ng, A.Y. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09), Montreal, QC, Canada, 14–18 June 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 609–616. [CrossRef]
58. Bell, A.J.; Sejnowski, T.J. The "independent components" of natural scenes are edge filters. *Vis. Res.* **1997**, *37*, 3327–3338. [CrossRef]
59. Olshausen, B.A.; Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **1996**, *381*, 607–609. [CrossRef]
60. Dai, A.M.; Le, Q.V. Semi-supervised Sequence Learning. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2015; pp. 3079–3087.
61. Howard, J.; Ruder, S. Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 328–339.
62. Wang, W.; Zhou, Z.H. A new analysis of co-training. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel, 21–24 June 2010; pp. 1135–1142.
63. Liu, B.; Yu, X.; Zhang, P.; Yu, A.; Fu, Q.; Wei, X. Supervised Deep Feature Extraction for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 1909–1921. [CrossRef]
64. Grabowski, B.; Masarczyk, W.; Głomb, P.; Mendys, A. Automatic pigment identification from hyperspectral data. *J. Cult. Herit.* **2018**, *31*, 1–12. [CrossRef]
65. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
66. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14), Columbus, OH, USA, 24–27 June 2014; pp. 580–587. [CrossRef]