DSSNet: A Simple Dilated Semantic Segmentation Network for Hyperspectral Imagery Classification

Bin Pan, Xia Xu, Zhenwei Shi, Ning Zhang, Huanlin Luo and Xianchao Lan

Abstract

Deep learning based methods have presented promising performance in the task of Hyperspectral Imagery Classification (HSIC). However, recent methods usually considered HSIC as a patchwise image classification problem and addressed it by giving a single label to the patch surrounding a pixel. In this paper, we propose a new semantic segmentation network which can directly label each pixel in an end-to-end manner. Compared with patchwise models, our method can significantly improve the training effectiveness and reduce some manual parameters. Another challenge in HSIC is that the spatial resolution of hyperspectral imagery is relatively low, in which case pooling operation may result in resolution and coverage loss. To address this issue, we introduce dilated convolution to our model, and construct a Dilated Semantic Segmentation Network (DSSNet). Different from some existing works, DSSNet is specially designed for HSIC without complicated architecture, and no pretrained models are required. The joint spatial-spectral information can be extracted via an end-to-end manner and thus avoid various preprocessing or postprocessing operations. Experiments on two public data sets have demonstrated the effectiveness of our improvements, when compared with some of the latest deep learning based HSIC models.

Index Terms

Dilated convolution, hyperspectral imagery classification, deep learning

I. Introduction

Hyperspectral Imagery (HSI) consists of tens to hundreds of continuous bands, which can provide abundant spectral information. Such high spectral resolution makes land-cover materials quite distinguishable in HSI data.

The work was supported by the National Key R&D Program of China under the Grant 2017YFC1405605, the National Natural Science Foundation of China under the Grants 61671037, the Beijing Natural Science Foundation under the Grant 4192034, and the Shanghai Association for Science and Technology under Grant SAST2018096. *(Corresponding author: Xia Xu)*

Bin Pan is with School of Statistics and Data Science, Nankai University, Tianjin 300350, China (e-mail: panbin@nankai.edu.cn).

Xia Xu (Corresponding author) is with College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: xuxia@nankai.edu.cn).

Zhenwei Shi is with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China (e-mail: shizhenwei@buaa.edu.cn).

Ning Zhang, Huanlin Luo and Xianchao Lan are with Shanghai Aerospace Electronic Technology Institute, Shanghai, 201109, China.

This characteristic brings in a hot spot of present research: Hyperspectral Imagery Classification (HSIC) [1]. HSIC refers to giving a label to each pixel in a hyperspectral image, which belongs to semantic segmentation technique.

Early research usually directly adopted Convolutional Neural Networks (CNNs) for feature extraction with patchwise strategies [2]–[6]. Recently, deep learning based semantic segmentation networks have been applied to the task of HSIC. Jiao *et al.* [7] introduced Fully Convolutional Networks (FCNs) to HSIC, where a pretrained network was transferred and the extracted features were combined with the spectral vectors via a weighting strategy. In literature [8], Lee and Kwon further enhanced FCN by multi-scale convolutions and residual learning. Some improvements based on FCN can also be observed in literature [9]–[12]. In literature [13], a conv-deconv network was proposed for unsupervised spectral-spatial feature learning. In literature [14], DeepLab was introduced by transfer learning where the input HSI data were reduced to 3 channels. In literature [15], a new initialization strategy was proposed for the FCN based HSIC problem.

Since the labeled samples for HSI are not sufficient, transfer learning is usually conducted, where dimension reduction is initially performed to fit the HSI data into well-pretrained backbones. However, hyperspectral data have presented unique characteristics when compared with natural scence data. There are at least two shortcomings that the networks for HSIC may have to overcome:

Firstly, the pooling operation in convolutional networks will significantly reduce the resolution of HSI data. Due to the low spatial resolution of HSI, some details may only occupy a few pixels. These details is likely to disappear after several times of pooling and thus harm the classification accuracy.

Secondly, in order to use some pretrained networks, the hyperspectral data have to be dimensionality-reduced to 3 channels, which will inevitably result in spectral information loss.

Considering the above two issues, in this paper we develop a new Dilated Semantic Segmentation Network (DSSNet) for hyperspectral imagery classification. Our motivations can be summarized by two aspects. Firstly, the resolution loss by pooling should be avoided so as to preserve the details in HSI data. Secondly, the network should be trained without transfer learning whereas some tricks should be adopted to avoid overfitting. Based on these motivations, we construct a new network which includes the following contributions:

- We utilize the dilated convolution strategy to avoid the influence of resolution reduction and aggregate multi-scale contextual information from HSI data.
- We construct a new semantic segmentation network with simple architecture and end-to-end training manner, which is able to make full use of the abundance spectral information of HSI data.

II. METHODOLOGY

Dilated convolutions [16] are the core of DSSNet. In this section we first give an detailed analysis about the advantage of dilated convolutions in the task of HSIC, and then we describe the overall architecture of DSSNet.

A. Dilated Convolutions

In CNN based works, discrete convolution with pooling are widely used. Let $f(\cdot)$ denote a nonlinear function, then the i-th feature map in the $(l$+1$)$-th layer can be described by

$$\boldsymbol{F}_i^{l+1} = f(\boldsymbol{F}^l * \boldsymbol{K}_i^{l+1} + b_i^{l+1}), \tag{1}$$

where $\boldsymbol{K}_i^{l+1} \in \mathbb{R}^{k\times k\times s_l}$ is the i-th convolution kernel in the $(l$+1$)$-th layer, s_l denotes the number of feature maps in the l-th layer, and here we assume that the kernel is squared. After convolution block, pooling operation follows. Take average pooling for example, the i-th feature map after pooling can be described by

$$\boldsymbol{FP}_i^{l+1} = \frac{1}{{k_\mathrm{p}}^2}\boldsymbol{F}_i^{l+1} * \boldsymbol{KP} \tag{2}$$

where $k_\mathrm{p} \times k_\mathrm{p}$ is the size of pooling window, $\boldsymbol{KP}$ is an all-one matrix with size $k_\mathrm{p} \times k_\mathrm{p}$, and we set stride=$k_\mathrm{p}$.

Semantic segmentation is a dense prediction task where the final outputs should share the same size as the inputs, thus deconvolutions or similar operations are required. However, the spatial resolution of HSI data is usually quite limited, and some materials may only occupy a few pixels. After several layers of pooling these materials may "disappear". In this paper, we use dilated convolutions to avoid this problem.

Let $\boldsymbol{FD}_i^l$ denote the i-th feature map in the l-th layer obtained by dilated convolution, then $\boldsymbol{FD}_i^{l+1}$ can be calculated by

$$\boldsymbol{FD}_i^{l+1} = f(\boldsymbol{F}^l * \boldsymbol{KD}_i^{l+1} + b_i^{l+1}), \tag{3}$$

where $\boldsymbol{KD}_i^{l+1} \in \mathbb{R}^{(k+2n-2)\times(k+2n-2)\times s_l}$ is the i-th convolution kernel in the $(l$+1$)$-th layer. n is a dilation factor which dominates the number of 0 elements in $\boldsymbol{KD}$. For example, the following kernel is called 2-dilated convolution:

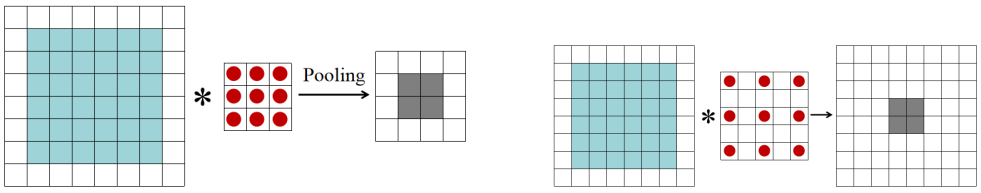
$$\boldsymbol{KD} = \begin{bmatrix} w_{11} & 0 & w_{12} & 0 & w_{13} \\ 0 & 0 & 0 & 0 & 0 \\ w_{21} & 0 & w_{22} & 0 & w_{23} \\ 0 & 0 & 0 & 0 & 0 \\ w_{31} & 0 & w_{32} & 0 & w_{33} \end{bmatrix}. \tag{4}$$

Here we also assume that the kernel is squared. In Eq. (4), only the weights w_{ij} are learnable parameters. n-dilated convolution means there are (n-1) zeros between w_{ij} and $w_{(i+1)j}$/$w_{i(j+1)}$. Obviously, 1-dilated convolution is the same as conventional 3×3 convolution.

Fig. 1 is an illustration for the difference between conventional discrete convolution and dilated convolution. The receptive field of convolution can be obtained by

$$r_\mathrm{out} = (r_\mathrm{in} - 1) \cdot \mathrm{stride} + k, \tag{5}$$

where r_in is the receptive field size of the input layer, and r_out is that of the output layer. In Fig. 1(a), the convolution kernel size is 3×3, pooling size is 2×2, convolutional stride is 1 and pooling stride is 2. If assume that the input is the original image, then after convolution and pooling the receptive field should be (((1-1)×1+3)-1)×2+2=6.



(a) (b)

Fig. 1: An illustration for (a) discrete convolution with pooling and (b) dilated convolution. The red points are the parameters that should be learned, the green squares are the pixels participating the convolution, and the grey squares are the outputs of the convolution on green squares.

In Fig. 1(b), although there are still only 9 weights required optimization, the kernel size in dilated convolution is not 3×3. Actually, if there are 3×3 weights to be optimized, the equivalent kernel size in dilated convolution should be:

$$k_{\mathrm{d}} = n \times 2 + 1, \tag{6}$$

and thus the receptive field of dilated convolution can be calculated by

$$r_{\mathrm{out}} = (r_{\mathrm{in}} - 1) \cdot \mathrm{stride} + (n \times 2 + 1). \tag{7}$$

According to Eq. (7), the receptive field of Fig. 1(b) is (1-1)×1+(2×2+1)=5. We can see that dilated convolution can expand the receptive field without increasing the learnable parameters. More importantly, there is no pooling operation, which means there is no resolution loss. Based on dilated convolution we can simply extract the multi-scale deep features as well as preserve the details in HSI data.

B. The Architecture of DSSNet

Fig. 2 describes the overall architecture of DSSNet. DSSNet is composed of 6 layers, including 4 convolution blocks and 2 fully connected blocks. Although DSSNet is not a deep network, recent studies have indicated that shadow networks may also perform well in HSIC [6], [9], [13].

1) Convolution block: This block contains 3 operation: convolution, batch normalization and ReLU activation function.

The convolution in DSSNet is achieved by $\boldsymbol{F}^l * \boldsymbol{K}\boldsymbol{D}_i^{l+1}$. In Fig. 2, the convolutional layers are described by conv-(kernel size)-(filters number)-(dilation factor). For example, conv-3-32-2 indicates that the kernel contains 3×3 non-zero weights, there are 32 filters and the dilation factor is 2. Please note that when k=3 and n=2, the real filter size is 5×5. However, because only 3×3 weights require optimization, here we record the kernel size as 3×3. Besides the kernel $\boldsymbol{K}\boldsymbol{D}$, the bias $\boldsymbol{b}$ should also be optimized. After the convolution, Batch Normalization (BN) and Rectified Linear Unit (ReLU) activation function are followed. As a practice, padding operation is conducted to make sure the feature maps keep the same size before and after convolutions.

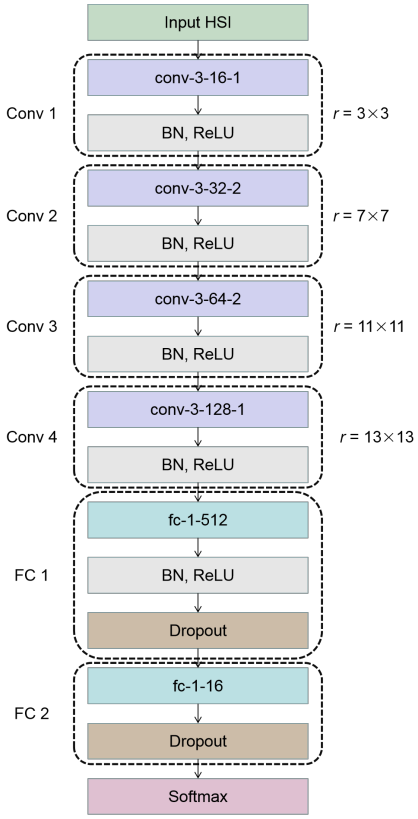


Fig. 2: The architecture of DSSNet.

The receptive field of each convolutional block is shown in the right side of Fig. 2. Because the spatial resolution of HSI is usually low, 13×13 receptive field has already covered a large region in the image, where abundant contextual information has been included. Therefore, it is not necessary as well as not appropriate to further enlarge the receptive field.

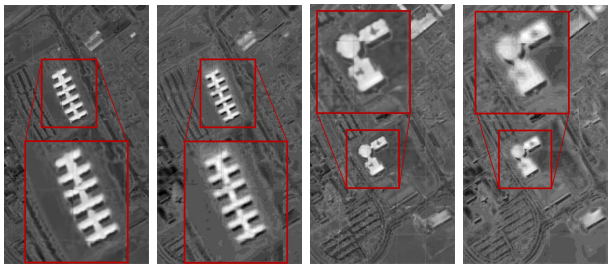
2) Fully connected block: There are two Fully Connected (FC) blocks in DSSNet. The first one (FC1) can generate features with 512 dimensions for all the pixels, and the second one (FC2) corresponds to the class number. FC operation in DSSNet is implemented by 1×1 convolution. To relieve the risk of overfitting, we conduct 0.5 dropout in FC1 and FC2.

3) Outputs and Optimization: The outputs are mapped to [0,1] by softmax. Assume $\boldsymbol{x}_p \in \mathbb{R}^{1\times C}$ is the output vector of pixel p after FC2, C is the number of classes in the data set, then we can obtain the following equation:

$$\hat{y}_c = \frac{e^{\boldsymbol{x}_p(c)}}{\sum_{t=1}^{C} e^{\boldsymbol{x}_p(t)}}, \tag{8}$$

where $\hat{y}_c$ is the probability that $\boldsymbol{x}_p$ belongs to class c, $\boldsymbol{x}_p(c)$ denotes the c-th element in $\boldsymbol{x}_p$.

Finally, we select cross entropy as the loss function, stochastic gradient descent with momentum as the optimization strategy and weight decay (L2) for regularization. Based on Fig. 2, we can calculate the number of parameters that require optimization. There are totally ∼199K parameters, which requires ∼778K memory.



(a) (b) (c) (d)

Fig. 3: Feature maps of Pavia University data by (a) DSSNet on Metal Sheets, (b) PSSNet on Metal Sheets, (c) DSSNet on Bitumen and (d) PSSNet on Bitumen. The main differences are highlighted.

C. Analysis about DSSNet

To clarify the effectiveness of the dilated convolution strategy, here we design a Pooling-based Semantic Segmentation Network (PSSNet) which has very similar architecture to DSSNet. PSSNet also contains 4 convolutional layers and 2 fully connected layers, and all the hyper-parameters are nearly the same as DSSNet. The only difference is that there are max-pooling layers (size 2×2, stride=1) at the end of every convolutional layers. Correspondingly, the kernel size in each layer is set as 3×3 and n=1. Overall, PSSNet uses pooling instead of dilated convolutions to expand the receptive field.

Fig. 3 compares the feature maps obtained by DSSNet and PSSNet on FC2 of Pavia University data set. Some stitching traces are observed because we have separated the original imagery to many 64×64 subfigures for training. Fig. 3(a)(c) are the feature maps by DSSNet, and Fig. 3(b)(d) are those by PSSNet. Different materials have generated strong responses in certain feature maps. However, we can see that the feature maps obtained by PSSNet are a little blurry, while those by DSSNet include more details. This is the very motivation of DSSNet: pooling may harm the details in HSI, and thus we hope to refine the feature maps by dilated convolutions.

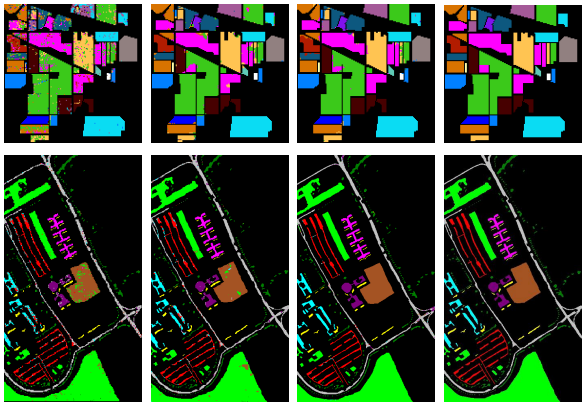
Overall, the superiority of DSSNet results from the full usage of the abundant spectral information. Although the transfer learning frameworks can utilize some powerful pretrained models, they have to sharply reduce the dimensionality of HSI data so as to adapt the inputs of the pretrained networks.

III. Experiments

In this section, we use 2 public HSI data sets to validate the effectiveness of DSSNet: Indian Pines (IndianP) and Pavia University (PaviaU). Both of them are widely used in many HSIC works [1]. The proposed methods are conducted 10 times and the average accuracies are reported. The hyper-parameters of DSSNet are listed in Table I.

A. The comparison of classification accuracies

4 recently proposed HSIC methods are selected for comparison, namely 3D-CNN [5], MugNet [6], HiFi [4] and FEFCN [9]. 3D-CNN and MugNet were constructed based on convolutional network with patchwise classifica-



(a) (b) (c) (d)

Fig. 4: Classification maps on by (a) 3D-CNN, (b) FEFCN, (c) DSSNet and (d) gournd truth. Top: Maps on IndianP. Bottom: Maps on PaviaU.

TABLE I: The list for hyper-parameters.

Mini-batch Size	15	Max-Epoches	20
Initial Learning Rate η	0.01	Learning Rate Drop	10
L2 Regularization λ	0.0001	Momentum α	0.9

tion. HiFi was a traditional machine learning based method. FEFCN was a new semantic segmentation network. The comparison with FEFCN aims at validating the effectiveness of our dilated convolution strategy as well as demonstrating that our new-developed architecture seems better than a transfer learning framework. Support Vector Machine (SVM) was used as the baseline. Furthermore, the results of PSSNet are also shown to verify the advantage of dilated convolutions in DSSNet. Overall Accuracy (OA), Average Accuracy (AA) and KAappa (κ) are used as the metrics.

Fig. 4 and Table II-III display the classification maps and accuracies on IndianP and PaviaU, respectively. As an early research, 3D-CNN performs not well, but it is an exploratory work which has provided meaningful guidance to the follows. Furthermore, the accuracies by 3D-CNN are much better than that of SVM. This result indicates that convolutional neural networks are quite promising. MugNet is a simplified deep neural networks which combines some preprocessing approaches. The accuracies reported by HiFi are close to MugNet. However, as a traditional machine learning based method, the improvement space of HiFi is limited. FEFCN is also a semantic segmentation network, but we can see that its accuracies are not very high. The reason may be that there are spectral loss as well as spatial resolution loss. In FEFCN, the input HSI data have to be reduced to 3 channels which has weakened the useful spectral information. Besides, the pooling operation has harmed the spatial resolution, which may lead to details loss. The gaps between PSSNet and DSSNet are about 2%. The reasons may include two parts: 1) The spatial resolution of IndianP is low, in which case pooling operation will bring more land-cover loss; 2) The

architecture of PSSNet is not well-designed, and we simply replace the dilated convolutions by pooling. Overall, DSSNet outperforms others by about 0.4%-2%.

TABLE II: Classification accuracies by different methods on IndianP data set (%).

Class	Samples		Methods						
	Train	Test	SVM	3D-CNN	HiFi	MugNet	FEFCN	PSSNet	DSSNet
Alfalfa	5	41	40.43	82.93	98.01	97.84	92.50	82.31	98.06
Corn-notill	143	1285	75.28	77.43	94.36	93.75	97.03	94.50	93.94
Corn-mintill	83	747	67.33	79.52	95.50	95.23	97.67	98.29	95.45
Corn	24	213	42.65	89.25	88.16	93.00	87.32	97.88	94.36
Grass-pasture	48	435	90.86	76.67	98.04	97.25	98.20	96.75	99.11
Grass-trees	73	657	96.02	98.63	99.94	99.79	98.51	99.70	99.97
Grass-pasture-mowed	3	25	10.16	48.00	76.09	84.97	100.0	87.57	91.82
Hay-windrowed	48	430	98.90	91.38	99.68	99.82	99.31	100.0	99.62
Oats	2	18	4.12	50.00	100.0	86.32	50.00	75.64	80.00
Soybean-notill	97	975	70.19	70.40	94.88	93.99	96.04	95.14	96.35
Soybean-mintill	246	2209	85.01	91.76	98.47	99.07	96.23	96.85	98.84
Soybean-clean	59	534	72.42	76.50	95.84	96.46	95.54	91.21	96.91
Wheat	21	184	95.26	99.46	99.24	99.13	99.57	98.85	99.56
Woods	127	1138	96.09	99.03	99.84	99.86	97.94	99.86	99.86
Buildings-Grass	39	347	50.70	99.43	96.58	97.68	94.60	80.25	97.98
Stone-Steel-Towers	9	84	86.74	86.75	98.34	97.59	97.61	86.14	99.05
OA			80.50	82.32	97.10	97.20	96.70	95.95	**97.61**
AA			67.63	86.62	95.81	95.73	93.69	92.56	**96.31**
κ			77.63	84.66	96.69	96.81	95.45	95.34	**97.27**

B. Ablation Study

DSSNet is designed according to our experience and attempts. Here we conduct ablation experiments by modifying the architecture of DSSNet so as to evaluate the influence of structural parameters.

Table IV shows some variants of DSSNet. The sign "conv-p-convd2-fc128" refers to "convolution-maxpooling-dilated convolution with n=2-fully connected layer with 128 neurons". Other hyper-parameters keep the same as DSSNet. `net1` removes one convolution layer from DSSNet, while `net2` adds one. `net3` combines dilated convolutions and pooling. `net4` increases the number of neurons in fully connected layers.

The classification accuracies by the variant networks are also shown in Table IV where κ is selected for evaluation. We can see that slight variations of the networks structure will not significantly affect the accuracies. Shallow networks (*e.g.*, `net1`) may lack representative ability. On the other hand, increasing the number of neurons in FC layers (*e.g.*, `net4`) will introduce much more parameters that may aggravate overfitting.

IV. Conclusion

In this paper, we propose a simple semantic segmentation network called DSSNet for hyperspectral imagery classification. During the process of designing of DSSNet, we always hold the opinion that if a simple network can

TABLE III: Classification accuracies by different methods on PaviaU data set (%).

Class	Samples		Methods						
	Train	Test	SVM	3D-CNN	HiFi	MugNet	FEFCN	PSSNet	DSSNet
Asphalt	346	6285	90.87	94.97	93.31	97.11	93.58	99.63	99.45
Meadows	936	17713	96.28	96.44	99.87	99.55	96.70	99.81	99.52
Gravel	89	2010	64.66	84.69	94.38	92.43	90.78	92.16	93.05
Trees	151	2913	90.55	97.39	97.16	95.71	96.72	93.36	99.75
Painted metal sheets	75	1270	98.13	99.14	99.89	99.16	88.55	99.73	99.73
Bare	245	4784	69.74	94.77	98.79	96.96	97.33	99.94	98.64
Bitumen	65	1265	62.53	88.90	95.80	96.29	92.38	83.32	97.12
Self-Blocking Bricks	189	3493	81.80	84.11	97.23	96.89	90.83	96.28	97.35
Shadows	43	904	86.85	100.0	97.37	98.07	88.20	97.19	99.02
OA			87.91	94.35	97.86	97.87	95.11	96.84	**98.43**
AA			82.38	93.38	97.09	96.91	92.79	95.71	**98.17**
κ			83.78	92.49	97.17	97.17	93.22	95.79	**97.90**

TABLE IV: Variants of DSSNet and their κ on test data sets.

Networks	Architecture	IndianP	PaviaU
`net1`	convd1-convd2-convd3 -fc512-fc	95.73	96.14
`net2`	convd1-convd2-convd3 -convd4-convd4-fc512-fc	96.21	95.61
`net3`	convd1-convd2-convd2 -conv1-p-fc512-fc	96.28	96.17
`net4`	convd1-convd2-convd3 -convd4-fc2048-fc	97.01	95.13

work well, it should not use a complex one. DSSNet tries to take full advantages of the HSI characteristics, and thus the integral HSI data are considered as the input of the network. To overcome the challenge of resolution loss, dilated convolution is introduced, and experiments are conducted to validate the effectiveness of our improvements. Moreover, DSSNet is an end-to-end network, which is different from some joint spectral-spatial methods where spectral and spatial information is combined via pre/postprocessing strategies. Actually, pre/postprocessing methods such as morphological attribute profiles can also be used to improve the accuracies achieved by DSSNet.

REFERENCES

[1] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

[2] Q. Wang, Z. Meng, and X. Li, "Locality adaptive discriminant analysis for spectralcspatial classification of hyperspectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2077–2081, Nov 2017.

[3] Y. Yuan, J. Lin, and Q. Wang, "Hyperspectral image classification via multitask joint sparse representation and stepwise mrf optimization," *IEEE Transactions on Cybernetics*, vol. 46, no. 12, pp. 2966–2977, Dec 2016.

[4] B. Pan, Z. Shi, and X. Xu, "Hierarchical guidance filtering-based ensemble classification for hyperspectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4177–4189, 2017.

[5] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 6232–6251, 2016.

[6] B. Pan, Z. Shi, and X. Xu, "MugNet: Deep learning for hyperspectral image classification using limited samples," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 108–119, 2018.

[7] L. Jiao, M. Liang, H. Chen, S. Yang, H. Liu, and X. Cao, "Deep fully convolutional network–based spatial distribution prediction for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5585–5599, 2017.

[8] H. Lee and H. Kwon, "Going deeper with contextual CNN for hyperspectral image classification," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4843–4855, 2017.

[9] J. Li, X. Zhao, Y. Li, Q. Du, B. Xi, and J. Hu, "Classification of hyperspectral imagery using a new fully convolutional neural network," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 292–296, 2018.

[10] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 60 – 77, 2018.

[11] X. Ma, A. Fu, J. Wang, H. Wang, and B. Yin, "Hyperspectral image classification based on deep deconvolution network with skip architecture," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 8, pp. 4781–4791, 2018.

[12] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 59–69, 2019.

[13] L. Mou, P. Ghamisi, and X. X. Zhu, "Unsupervised spectral–spatial feature learning via deep residual conv–deconv network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 391–406, 2018.

[14] Z. Niu, W. Liu, J. Zhao, and G. Jiang, "Deeplab-based spatial feature extraction for hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 251–255, 2019.

[15] B. Pan, Z. Shi, X. Xu, T. Shi, N. Zhang, and X. Zhu, "CoinNet: Copy initialization network for multispectral imagery semantic segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 5, pp. 816–820, May 2019.

[16] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.