# CoinNet: Copy Initialization Network for Multispectral Imagery Semantic Segmentation

Bin Pan, Zhenwei Shi, Xia Xu, Tianyang Shi, Ning Zhang and Xinzhong Zhu

## Abstract

Remote sensing imagery semantic segmentation refers to assigning label to every pixel. Recently, deep convolutional neural networks (CNNs) based methods have presented impressive performance in this task. Due to the lack of sufficient labeled remote sensing images, researchers usually utilized transfer learning strategies to fine tune networks which were pre-trained in huge RGB-scene data sets. Unfortunately, this manner may not work if the target images are multi/hyper-spectral. The basic assumption of transfer learning is that the low-level features extracted by the former layers are similar in most data sets, hence users only require to train the parameters in the last layers that are specific to different tasks. However, if one should use a pre-trained deep model in RGB data for multi/hyper-spectral imagery semantic segmentation, the structure of the input layer has to be adjusted. In this case, the first convolutional layer has to be trained using the multi/hyper-spectral data sets which are much smaller. Apparently, the feature representation ability of the first convolutional layer will decrease and it may further harm the following layers. In this paper, we propose a new deep learning model, COpy INitialization Network (CoinNet), for multispectral imagery semantic segmentation. The major advantage of CoinNet is that it can make full use of the initial parameters in the pre-trained network's first convolutional layer. Comparison experiments on a challenging multispectral data set have demonstrated the effectiveness of the proposed improvement. The demo and a trained network will be published in our homepage.

## Index Terms

Semantic segmentation, deep convolutional network, transfer learning, CoinNet

## I. Introduction

Semantic segmentation is a hot topic in computer version, especially after the breakthrough in deep convolutional networks (CNNs) [1]. In 2015 the CNN model is extended to a fully convolutional form (FCN) [2] which has further

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, and Shanghai Association for Science and Technology under the Grant SAST2017108. *(Corresponding author: Zhenwei Shi and Xia Xu.)*

Bin Pan, Zhenwei Shi (Corresponding Author), Xia Xu (Corresponding Author) and Tianyang Shi are with Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and with Beijing Key Laboratory of Digital Media, Beihang University, Beijing 100191, China, and also with State Key Laboratory of Virtual Reality Technology and Systems, School of Astronautics, Beihang University, Beijing 100191, China. (e-mail: panbin@buaa.edu.cn; shizhenwei@buaa.edu.cn; xuxia@buaa.edu.cn; shitianyang@buaa.edu.cn)

Bin Pan and Xia Xu are also with College of Geomatics, Shandong University of Science and Technology

Ning Zhang and Xinzhong Zhu are with Shanghai Aerospace Electronic Technology Institute, Shanghai, 201109, China

improved the performance. In the field of remote sensing imagery processing, semantic segmentation (also known as pixel-based classification in the past) has presented great significance not only in academic research but also in many applications [3]–[5]. For example, semantic segmentation is the basis forest-cover estimation, land-use investigation and urban planning. Traditional manual labeling manner is quite time-consuming, therefore it is necessary to develop automatic semantic segmentation methods.

Compared with common RGB natural scene images, some remote sensing sensors can provide images with many continuous bands, called multi/hyper-spectral imagery (to avoid ambiguity we assume multispectral imagery has more than 3 channels). The extra spectral information will definitely contribute to the semantic segmentation problem. Thanks to the excellent performance of FCN methods [2], [6], researchers have begun to study the semantic segmentation of multi-/hyperspectral remote sensing imagery based on FCN [7]–[10]. Generally, these methods adopted the same idea: transfer learning [11]. They used networks that were pre-trained in a huge RGB data set (ImageNet) for initialization, and used much smaller remote sensing data set for fine-tuning. Literatures [7], [8] used the VGG-based FCN [2], [12] as the pre-trained model, and literatures [9], [10] were ResNet-based [13]. Since there is no ImageNet-level data set for remote sensing images, this transfer learning strategy is reasonable and can avoid the overfitting problem to some extent. However, there is another critical issue that must be considered:

- Images in ImageNet only contains R-G-B bands, while multi/hyper-spectral images includes much more bands. Then how to transfer the pre-trained networks in ImageNet to the multi/hyper-spectral task in an appropriate way?

This is not a simple inputs-changing problem. According to the transfer learning theory, the former layers of a network tend to learn the low-level and general features such as angles and edges, while the latter layers may learn the high-level and specific features for different task [11]. Therefore transfer learning methods usually freeze or fine-tune the former layers and retrain the last ones [14]. However, in the multi/hyper-spectral task the condition is just opposite, which makes the usage of transfer learning not very reasonable. Although this problem looks small, it has not been widely studied.

In this paper, we propose a new end-to-end deep learning model, COpy INitialization Network (CoinNet), for multispectral imagery semantic segmentation. Our basic idea is to still initialize the first convolutional layer using pre-trained parameters while keeping the input data structure unchanged. Firstly, generate a pre-trained semantic segmentation network by ImageNet. Secondly, we redesign the structures of the input and the first convolutional layers so as to adapt to the multispectral data. Then, the added layers' parameters are initialized by the pre-trained network using a copy based strategy. Finally, the whole network is fine-tuned in a multispectral data set.

The major advantage of the CoinNet is that it can make better use of the ideal parameters in the pre-trained network, and thus the representation ability of the low-level features is significantly improved. The whole framework is still based on the idea of transfer learning, whereas we can avoid the retrain process for the low-level features.

To validate the effectiveness of our network, we will publish the demo and the trained model in our homepage[1].

[1] http://levir.buaa.edu.cn/Code.htm

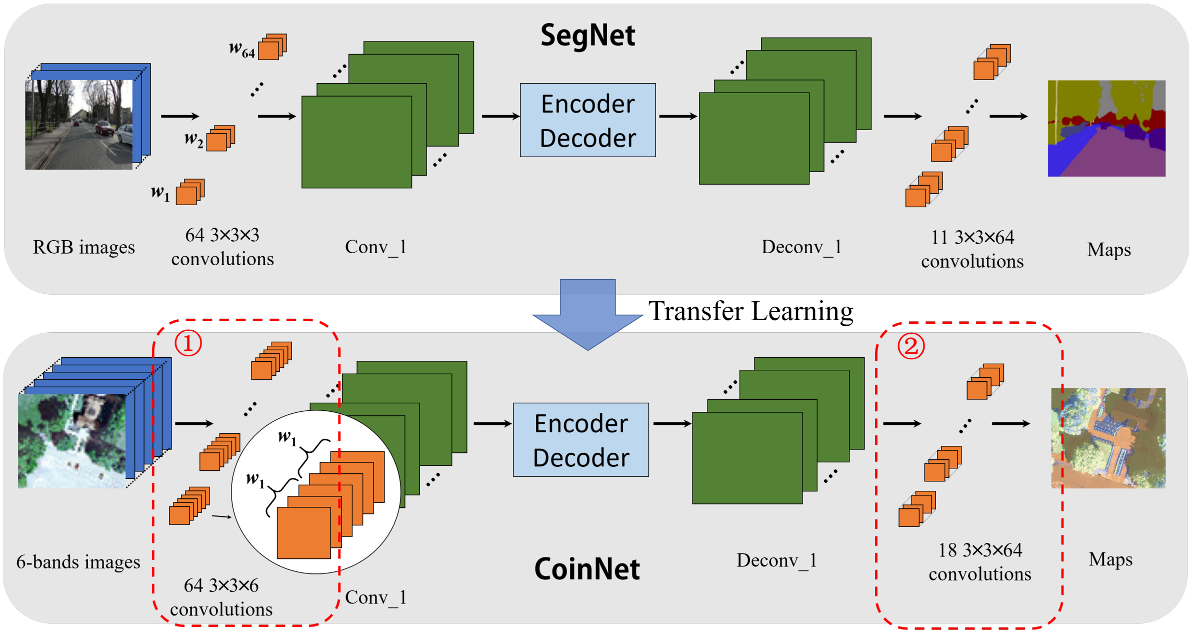


Fig. 1: The architecture of CoinNet. CoinNet is transferred from SegNet, which is also an encoder-decoder form. The two boxes indicate our modifications.

## II. CoinNet

### A. Problem Description

To transfer the ImageNet pre-trained networks to multi/hyper-spectral imagery semantic segmentation, there are mainly two strategies: 1) Reducing the dimension to 3 by PCA (or similar) and then directly using the pre-trained models [8]; 2) Using random initialization for the first layers and then training on the remote sensing data sets [10]. Both of these strategies may have some shortcomings. The former will result in spectral information loss. The latter is actually trying to learn low-level features from the remote sensing data sets. However, because the remote sensing data sets are generally too small, networks are difficult to learn good parameters and powerful features. Meanwhile, the poor representation ability of the low-level features may harm the following high-level representation. Overall, The learning of the first-layer parameters is a problem worthy of study for transfer learning.

### B. Data

In this paper, we use RIT-18 [10], [15] multispectral data set for validation. Compared with other more popular multi/hyper-spectral data sets, the main advantage of RIT-18 is that the training and testing sets are separated. This is an excellent characteristic. Current hyperspectral data sets usually consist of a single image that is randomly sampled for training and testing sets. Classification networks usually involve extracting spatial information from neighboring pixels [8], [16], which is called "spatial information". However, if the training/testing sets are randomly sampled, the training data could contaminate the test data, which would artificially inflate classification performance

[10], [15].

RIT-18 is a 6-band multispectral data set covering visible and near-infrared regions. Because the groundtruth for testing set is not published, here we use the original validating set for testing. The size of training data is 9394×5642, and that of testing data is 8833×6918. Since 2 materials are not included in the testing set, there are 16 classes remaining for classification. More details about this data can be found from literature [15].

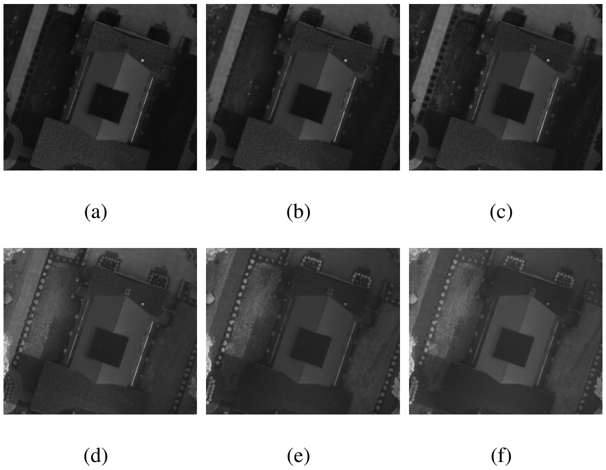


Fig. 2: A patch of the training set in different channels. (a) 490nm (b) 550nm (c) 680nm (d) 720nm (e) 800nm (f) 900nm. Although the reflectivity presents some differences, their overall textures are similar.

## C. Copy Initialization

Copy initialization is a strategy specially designed for the parameters (weights and bias) of the first convolutional layer. Assume that $\mathbf{W}_{\text{pre}} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_i, \cdots, \mathbf{w}_n]$ is the weight of the first convolutional layer in a pre-trained model. In CoinNet $\mathbf{w}_i \in \mathbb{R}^{3\times3\times3}$ and $n = 64$ (see Fig. 1 and Table I). Since the pre-trained network is learnt from ImageNet which is a RGB data set, the kernel $\mathbf{w}_i$ only includes 3 channels. However, multispectral semantic segmentation task requires $p$-channel inputs, where $p$ denotes the number of bands and $\{p \in \mathbb{N} \mid p > 3\}$. Let $\mathbf{W}^c_{\text{coin}} = [\mathbf{w}^c_1, \mathbf{w}^c_2, \cdots, \mathbf{w}^c_i, \cdots, \mathbf{w}^c_n]$ denotes the weights of the first convolutional layer in CoinNet with $\mathbf{w}^c_i \in \mathbb{R}^{3\times3\times6}$ and $n = 64$. Then we initialize $\mathbf{w}^c_i$ by

$$\mathbf{w}^c_i = [\overbrace{\mathbf{w}_i; \mathbf{w}_i; \cdots \mathbf{w}_i}^{p/3}], \tag{1}$$

where all the $\mathbf{w}_i$ are stacked. If $p$ is not divisible by 3, the remaining channels of $\mathbf{w}^c_i$ can be initialized by the first one or two channels of $\mathbf{w}_i$. In CoinNet, the inputs are 6 channels. Please note that the proposed copy initialization strategy works well when the bands number is relatively large. In some multispectral data sets with $p = 3$, the copy initialization strategy is unnecessary. We can stack two $\mathbf{w}_i$ to construct the $\mathbf{w}^c_i$ with size $3 \times 3 \times 6$, as shown in Fig. 1, box 1. Besides, another parameter, bias, is also transferred from the pre-trained model.

TABLE I: The network configurations. The encoder layers are listed from top to bottom, and the decoder ones are adverse. The convolutional layer parameters are denoted as "conv-⟨receptive field⟩-⟨number of filters⟩-⟨number of convolutions⟩". The batch normalisation and ReLU activation layers are not shown for brevity

| Encoder ↓ | Decoder ↑ |
|---|---|
| | softmax |
| Input (6-band image) | conv3-16-1 |
| conv3-64-2 | conv3-64-1 |
| maxpool | upsample |
| conv3-128-2 | conv3-128-2 |
| maxpool | upsample |
| conv3-256-3 | conv3-256-3 |
| maxpool | upsample |
| conv3-512-3 | conv3-512-3 |
| maxpool | upsample |
| conv3-512-3 | conv3-512-3 |
| maxpool | upsample |

The motivations of the copy initialization mainly include two aspects: 1) We think that the networks trained in a large-scale data set (ImageNet) are much more powerful, and thus the corresponding parameters are more optimal; 2) Moreover, although the reflectivity in different channels presents some differences, their overall textures are similar, as shown in Fig. 2. CoinNet may ignore the color information represented by the pre-trained model, but the texture and spatial information can be utilized. The copy initialization strategy could make full use of the representative low-level features obtained by the pre-trained model, which is beneficial for the transfer learning tasks.

### *D. Architecture and Transfer Learning*

The architecture and configurations of CoinNet are shown in Table I. CoinNet is transferred from SegNet [6] whose basis is VGG16 [12]. However, the proposed approach is applicable to many networks besides SegNet. Other FCN models such as FCN-8s [2] and DeepLab [17] can also be improved by the proposed strategy. Here we take SegNet for example.

According to Table I, CoinNet is composed of encoder-decoder pairs. An encoder includes several convolutional, batch normalisation, ReLU layers as well as a maxpooling operation. The corresponding decoder has similar components, where the maxpooling is replaced by an unsampling. The indices of maxpooling locations were stored and passed to the decoder. The left column of Table I presents the encoding process, where totally 13 convolutional layers are observed. Different from the original SegNet, the last convolutional layer of the decoders contains 16 filters, and correspondingly they will generate 16 label maps for the test images. Additionally, pro-processing strategies such as Conditional Random Fields (CRFs) can also be used to improve CoinNet, but this is not our

TABLE II: Hyper-parameters in CoinNet.

| Hyper-parameters | Values |
|---|---|
| Max-epochs | 15 |
| Batch size | 12 |
| Regularization coefficient (L2) | 0.0001 |
| Learning rate | 0.01 |

focus.

In the training process, we first separate the input multispectral data into many patches. The size of the biggest receptive field can provide guidance for the patch size. The filter size of CoinNet is 3×3 and the pooling window is 2×2. Based on these configurations, the biggest receptive field is calculated as 285×285. In CoinNet we set the patch size the same as the biggest receptive.

We conduct transfer learning on SegNet, and the multispectral data set is used for fine-tuning. The first and the last convolutional layers are two focuses of our transfer learning based model. As discussed above, copy initialization approach is used to improve the connection between input and the first convolutional layers. Furthermore, because SegNet and CoinNet aim at different tasks, the last convolutional layer in the decoding process must be re-trained. Therefore the weights and bias in this layer is randomly initialized. Stochastic gradient descent (SGD) and cross-entropy loss are adopted for optimization.

CoinNet is a completely end-to-end model. Since the multispectral data set used here is not as small as hyper-spectral ones, it is not necessary to freeze parameters in the earlier layers.

## III. Experiments and Discussion

### A. Setups

In this section, we design comparison experiments to validate the following conclusions:

- Transfer learning is effective
- Our copy initialization approach is meaningful
- The proposed deep learning network outperforms traditional classifiers

The hyper-parameters in CoinNet are listed in Table II. This table may help readers reproduce our results. To further verify the effectiveness of the proposed method, we will publish the demo and a trained network in our homepage if this paper has the honor to be accepted.

### B. Comparison Methods

We design three variations of SegNet for comparison:

*1) SegNet-RI:* This is our self-built network which has the same architecture as SegNet. However, all the parameters are randomly initialized (RI). In other words, SegNet-RI is completely trained in RIT-18 data set without transfer learning.

Possible drawback: The network is trained in a small data set which may lead to overfitting.

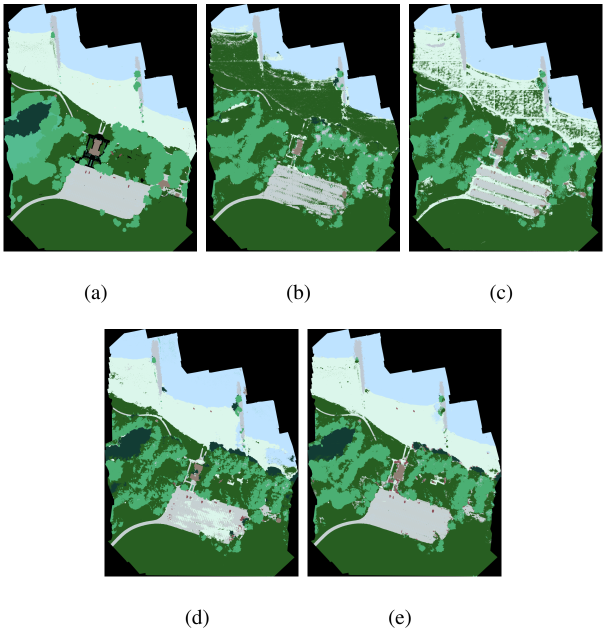
*2) SegNet-TL:* This network shares the initialized parameters from SegNet which was trained in ImageNet. Therefore, SegNet-TL is a standard transfer learning (TL) framework. However, because of the different image forms between ImageNet and RIT-18, the filter size of the first convolutional layer in SegNet-TL is $3\times3\times6$. In SegNet-TL the weights and bias in this layer is randomly initialized.

Possible drawback: The low-level features may lack representation ability.

*3) SegNet-TL-PCA:* The difference between this network and SegNet-TL is that SegNet-TL-PCA firstly conduct dimension reduction (from 6 to 3) on training and testing sets so as to adapt the inputs of SegNet. In this case the parameters in the pre-trained SegNet can be directly transferred to SegNet-TL-PCA.

Possible drawback: Spectral information loss.

Besides, some classical methods are compared, *i.e.*, support vector machine (SVM), K-nearest neighbor (k-NN) and stacked convolutional autoencoders (SCAE). Note that SCAE is also a deep learning model.



(a) (b) (c)

(d) (e)

Fig. 3: Semantic segmentation maps by different networks. (a) The ground truth, (b) SegNet-RI, (c) SegNet-TL, (d) SegNet-TL-PCA, and (e) CoinNet.

### *C. Results and Analysis*

Fig. 3 displays the semantic segmentation maps of different networks. Due to the limited paper length the maps by traditional methods are not shown. Generally, all these networks present acceptable results. These methods perform

TABLE III: Semantic segmentation accuracies for RIT-18 data set (%).

| Class | Methods | | | | | | |
|---|---|---|---|---|---|---|---|
| | SVM | k-NN | SCAE | SegNet-RI | SegNet-TL | SegNet-TL-PCA | CoinNet |
| Road Markings | 51.0 | 65.1 | 37.0 | 71.0 | 63.1 | 21.7 | 85.1 |
| Tree | 43.5 | 71.0 | 62.0 | 79.2 | 77.7 | 71.7 | 77.6 |
| Building | 1.5 | 0.3 | 11.1 | 42.0 | 60.1 | 52.8 | 52.3 |
| Vehicle | 0.2 | 15.8 | 11.8 | 0.0 | 4.7 | 59.1 | 59.8 |
| Person | 19.9 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Lifeguard Chair | 22.9 | 1.0 | 29.4 | 0.0 | 0.0 | 0.0 | 0.0 |
| Picnic Table | 0.8 | 0.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Orange Pad | 15.2 | 14.6 | 82.6 | 0.0 | 0.0 | 0.0 | 0.0 |
| Buoy | 0.7 | 3.6 | 7.2 | 0.0 | 0.0 | 0.0 | 0.0 |
| Rocks | 20.8 | 34.0 | 36.0 | 76.3 | 96.9 | 89.3 | 84.8 |
| Low Vegetation | 0.4 | 2.3 | 1.1 | 1.0 | 12.5 | 1.3 | 4.1 |
| Grass/Lawn | 71.0 | 79.2 | 84.7 | 98.0 | 93.0 | 96.1 | 96.7 |
| Sand/Beach | 89.5 | 56.1 | 85.3 | 6.4 | 73.9 | 77.6 | 92.1 |
| Water (Lake) | 94.3 | 83.6 | 97.5 | 90.4 | 93.9 | 98.4 | 98.4 |
| Water (Pond) | 0.0 | 0.0 | 0.0 | 1.7 | 0.5 | 87.6 | 92.7 |
| Asphalt | 82.7 | 80.0 | 59.8 | 73.1 | 42.4 | 61.2 | 90.4 |
| OA | 61.3 | 66.1 | 67.4 | 67.5 | 77.0 | 81.8 | 88.8 |
| AA | 30.3 | 31.2 | 36.1 | 33.7 | 38.7 | 44.8 | 52.1 |

well when materials are conjoint to large areas. However, the results in some small-scale classes are unwarranted. For example, class “Person” and “Buoy” are seldom correctly classified. This drawback is not difficult to explain: After 5 times pooling the small patches can hardly be observed before upsampling. Since the spatial resolution of RIT-18 is 4.7cm, a single “Person” may account for only dozens of pixels. The unsmooth edges are also found, but they can be significantly improved by CRF or multi-scale operations. Overall, CoinNet is superior according to the visible maps. SegNet-RI and SegNet-TL performs poor even in some large-scale classes such as “Pond”. This comparison indicates that the initialization parameters in the earlier layers are important.

Table III shows the objective evaluation for different methods. Overall accuracy (OA) and average accuracy (AA) are used for evaluation. The results of SVM, k-NN and SCAE are originally reported by literature [10]. Compared with deep learning methods, traditional classifiers perform poor, especially in AA. SCAE and SegNet-RI slightly outperform SVM and k-NN, but the gaps are not significant. Because SegNet-RI is randomly initialized, it may suffer overfitting. By comparison, transfer learning based semantic segmentation networks (SegNet-TL, SegNet-TL-PCA and CoinNet) have achieved 10-20% improvements. Such an apparent advantage has demonstrated that transfer learning is a powerful strategy for small-scale classification problems such as multispectral semantic segmentation. Although there is spectral information loss, SegNet-TL-PCA still presents about 5% advantages over SegNet-TL.

This result also indicates the effectiveness of the initial parameters learnt from a large data set.

CoinNet achieves higher accuracies, especially in OA. This is because the “Sand” and “Asphalt” are better classified, which account for a large area of the whole scene. However, several classes with very limited samples are still missing. This is the common drawback of all the reported deep networks. Overall, CoinNet has apparent advantage with sufficient training samples, but performs poor if the samples are limited.

## IV. Conclusion

Transfer learning has become a popular approach in utilizing the powerful deep networks to various applications. The basic assumption of transfer learning is that images in different tasks usually present similar low-level features, and thus users only require to retrain the parameters in the last layers. However, in the multispectral imagery semantic segmentation task, this assumption may not meet, since the first layer has to be randomly initialized and completely retrained. In this paper, we propose a copy initialization strategy which could take full advantage of the initial parameters pre-trained in a much larger data set. Based on this strategy we develop a new deep learning model, CoinNet. From the experiments we can draw three conclusions about multispectral imagery semantic segmentation: 1) Deep learning is superior; 2) Transfer learning is promising; 3) The proposed copy initialization approach works in some conditions. In our future work, we will further consider how to represent the color information in the initialization process.

## References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[2] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[3] L. Zhang, L. Zhang, and B. Du, “Deep learning for remote sensing data: A technical tutorial on the state of the art,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, June 2016.

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

[5] Q. Wang, J. Gao, and Y. Yuan, “A joint convolutional neural networks and context transfer for street scenes labeling,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1457–1470, 2018.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for scene segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[7] Q. Wang, J. Gao, and Y. Yuan, “Embedding structured contour and location prior in siamesed fully convolutional networks for road detection,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 230–241, Jan 2018.

[8] 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, Oct 2017.

[9] W. Sun and R. Wang, “Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM,” *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474–478, March 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*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271618301229

[11] Y. Bengio, G. Guyon, V. Dror, G. Lemaire, and D. Taylor, “Deep learning of representations for unsupervised and transfer learning,” *Workshop on Unsupervised and Transfer Learning*, pp. 17–37, 2012.

[12] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint:1409.1556v6*, 2015.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, “A review on deep learning techniques applied to semantic segmentation,” *arXiv preprint:1704.06857*, 2017.

[15] R. Kemker, C. Salvaggio, and C. Kanan, “High-resolution multispectral dataset for semantic segmentation,” *arXiv preprint:1703.01918*, 2018.

[16] B. Pan, Z. Shi, and X. Xu, “MugNet: Deep learning for hyperspectral image classification using limited samples,” *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271617303416

[17] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, April 2018.