SEMANTIC SEGMENTATION OF AERIAL IMAGERY VIA MULTI-SCALE SHUFFLING CONVOLUTIONAL NEURAL NETWORKS WITH DEEP SUPERVISION

Kaiqiang Chen[1,2], Michael Weinmann[3], Xian Sun[1], Menglong Yan[1], Stefan Hinz[4], Boris Jutzi[4], Martin Weinmann[4]

[1] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Institute of Electronics, Chinese Academy of Sciences, Beijing, P.R. China - chenkaiqiang14@mails.ucas.ac.cn, sunxian@mail.ie.ac.cn, yanmenglong@gmail.com
[2] University of Chinese Academy of Sciences, Beijing, P.R. China
[3] Institute of Computer Science II, University of Bonn, Bonn, Germany - mw@cs.uni-bonn.de
[4] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany - (stefan.hinz, boris.jutzi, martin.weinmann)@kit.edu

Commission I, ICWG I/IV

KEY WORDS: Semantic Segmentation, Aerial Imagery, Multi-Modal Data, Multi-Scale, CNN, Deep Supervision

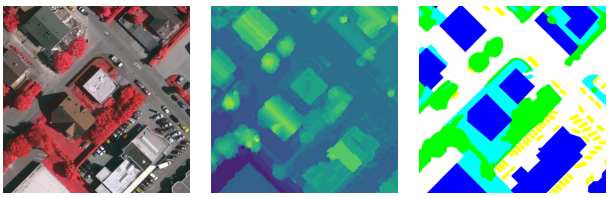
ABSTRACT:

In this paper, we address the semantic segmentation of aerial imagery based on the use of multi-modal data given in the form of true orthophotos and the corresponding Digital Surface Models (DSMs). We present the Deeply-supervised Shuffling Convolutional Neural Network (DSCNN) representing a multi-scale extension of the Shuffling Convolutional Neural Network (SCNN) with deep supervision. Thereby, we take the advantage of the SCNN involving the shuffling operator to effectively upsample feature maps and then fuse multi-scale features derived from the intermediate layers of the SCNN, which results in the Multi-scale Shuffling Convolutional Neural Network (MSCNN). Based on the MSCNN, we derive the DSCNN by introducing additional losses into the intermediate layers of the MSCNN. In addition, we investigate the impact of using different sets of hand-crafted radiometric and geometric features derived from the true orthophotos and the DSMs on the semantic segmentation task. For performance evaluation, we use a commonly used benchmark dataset. The achieved results reveal that both multi-scale fusion and deep supervision contribute to an improvement in performance. Furthermore, the use of a diversity of hand-crafted radiometric and geometric features as input for the DSCNN does not provide the best numerical results, but smoother and improved detections for several objects.

1. INTRODUCTION

The semantic segmentation of aerial imagery refers to the task of assigning a semantic label (e.g. *Building*, *Impervious Surface*, *Car* or *Vegetation*) to each pixel and thereby providing meaningful segments. Over the last few years, this kind of image interpretation has become a topic of great interest not only in remote sensing (Volpi and Tuia, 2017; Chen et al., 2018a; Maggiori et al., 2017; Marmanis et al., 2016; Paisitkriangkrai et al., 2016) but also in the field of computer vision (Chen et al., 2016; Zhao et al., 2016; Liu et al., 2015; Badrinarayanan et al., 2017). Meanwhile, some benchmarks such as the *ISPRS Benchmark on 2D Semantic Labeling* (Rottensteiner et al., 2012) have been initiated to foster research on the semantic segmentation of aerial imagery. Thereby, the given data consists of true orthophotos and the corresponding Digital Surface Models (DSMs) as shown in Figure 1.

Given data in the form of true orthophotos and the corresponding DSMs, the semantic segmentation of aerial imagery has been addressed by extracting hand-crafted features and using standard classifiers such as Random Forests (Weinmann and Weinmann, 2018; Gerke and Xiao, 2014) or Conditional Random Fields (CRFs) (Gerke, 2014). In recent years, however, the use of modern deep learning techniques has become increasingly popular, as such techniques are famous for their representation capacity and their ability of learning features. As one of the most successful representatives of deep learning, Convolutional Neural Networks (CNNs) (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2016a) have become the most popular method in image classification. In the context of semantic image segmentation, a class label should be predicted for each pixel. To achieve this, a rich variety of deep networks has been proposed (Long et al., 2015; Noh et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2014; Zhao et al., 2016). Meanwhile, CNNs are also widely applied to the semantic segmentation of aerial imagery (Volpi and Tuia, 2017; Marmanis et al., 2016; Maggiori et al., 2017; Paisitkriangkrai et al., 2016; Chen et al., 2018a).



(a) True orthophoto (b) DSM (c) Labeling

Figure 1. Semantic segmentation of aerial imagery: given (a) the true orthophoto and (b) the corresponding DSM, the objective is to derive (c) a semantic labeling, whereby the classes of interest are given by *Impervious Surfaces* (white), *Building* (blue), *Low Vegetation* (cyan), *Tree* (green) and *Car* (yellow).

The adaptation of networks designed for image classification is the main approach to derive networks for semantic image segmentation (Long et al., 2015; Noh et al., 2015; Chen et al., 2014; Zhao et al., 2016). Thereby, several layers are involved which cause a reduction of resolution so that the derived labeling is very coarse. Consequently, a popular research topic is the question of how to successfully transfer the reduced resolution to the original resolution. The proposed methods can be categorized into

two groups, which are *fixed upsampling* and *learnable upsampling*. Bilinear interpolation is a method of fixed upsampling for recovering the resolution (Long et al., 2015; Chen et al., 2014; Chen et al., 2016). However, methods using learnable upsampling (Noh et al., 2015; Badrinarayanan et al., 2017; Zhao et al., 2016; Chen et al., 2018a) have become more popular, amongst which the deconvolution (or transposed convolution) (Noh et al., 2015) is widely used in semantic segmentation of aerial imagery (Volpi and Tuia, 2017; Marmanis et al., 2016). Alternatively, the shuffling operator provides another solution to recover the resolution (Shi et al., 2016; Chen et al., 2018a; Chen et al., 2018b).

In this paper, we focus on the use of a shuffling operator (Shi et al., 2016) to effectively upsample feature maps and take into account the benefit of using multi-scale predictions. On the one hand, features derived from deeper layers have stronger semantic information which are robust to translation, rotation and scale, but such features lack spatial information. In contrast, features derived from shallower layers contain more spatial information due to the higher resolution and are thus significant for localization (e.g. in terms of boundaries between objects). Accordingly, we address a multi-scale extension of the Shuffling Convolutional Neural Network (SCNN) (Chen et al., 2018a; Chen et al., 2018b). To achieve this, we fuse multi-scale features derived from the intermediate layers of the SCNN, which results in the Multi-scale Shuffling Convolutional Neural Network (MSCNN). In addition, we introduce additional losses to the fused features of the MSCNN and thus derive the Deeply-supervised Shuffling Convolutional Neural Network (DSCNN) as an MSCNN with deep supervision. These additional losses address the susceptibility of deep networks to the vanishing gradient problem by injecting additional errors into the intermediate layers of the deep network, resulting in a better convergence. Besides the presentation of the MSCNN and the DSCNN, we involve a variety of hand-crafted radiometric and geometric features extracted from the true orthophotos and the corresponding DSMs for the classification task. Based on a separate and combined consideration of these features, we explore the value of the different modalities for the classification task. For performance evaluation, we test our approaches on a benchmark dataset provided with the *ISPRS Benchmark on 2D Semantic Labeling*.

After briefly describing related work in Section 2, we explain our methodology for the semantic segmentation of aerial imagery based on multi-modal data in Section 3. Thereby, we focus on the extraction of hand-crafted features as the basis for classification and on the construction of three different types of deep networks given by the SCNN, the MSCNN and the DSCNN, respectively. To demonstrate the performance of these networks, we present the results achieved for a standard benchmark dataset in Section 4. A detailed analysis and discussion of the derived results is given in Section 5. Finally, we provide concluding remarks and suggestions for future work in Section 6.

2. RELATED WORK

Since the great success of the AlexNet (Krizhevsky et al., 2012) in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015), Convolutional Neural Networks (CNNs) have become the most popular method in computer vision regarding image classification, object detection and semantic segmentation. The Fully Convolutional Network (FCN) (Long et al., 2015) can be considered as the first approach to apply CNNs for semantic segmentation. Since then, a variety of approaches addressing this task has been presented (Noh et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2014; Chen et al., 2016; Zhao et al., 2016; Liu et al., 2015).

Most of the CNNs presented for semantic segmentation (Long et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2014; Chen et al., 2016) are adapted from popular networks for image classification such as the VGGNet (Simonyan and Zisserman, 2014) or the Residual Network (ResNet) (He et al., 2016a). As such networks for image classification contain several layers that cause a reduction of resolution, the outputs are of low resolution and therefore rather coarse. Consequently, such networks need to be adapted to fit in the task of semantic segmentation. In this regard, an intuitive solution is given by the removal of all the layers that cause resolution reduction, resulting in no-downsampling networks (Sherrah, 2016). However, such networks will suffer from computational overload and cost much more training time (Chen et al., 2018a; Sherrah, 2016). To reach a trade-off between computational efficiency and a reasonable resolution, DeepLab (Chen et al., 2016) keeps the first three layers that will cause resolution reduction and removes the remaining ones of such layers, resulting in an output stride of 8. To recover the resolution to the original size, DeepLab adopts bilinear interpolation. Instead, Shuffling Convolutional Neural Networks (SCNNs) (Chen et al., 2018a) replace the bilinear interpolation with the shuffling operator (Shi et al., 2016) to recover resolution. Besides, encoder-decoder architectures (Noh et al., 2015; Badrinarayanan et al., 2017) provide another solution to recover resolution, where the encoder part is responsible for encoding the input to a compressed representation with a low resolution and the decoder part adequately transfers the compressed representation to the original image size.

Features derived from deeper layers have stronger semantic information but lack spatial information. In contrast, features derived from shallower layers contain more accurate spatial information due to the higher resolution, which in turn can be beneficial for a better localization. Therefore, an intuitive idea is borrowing spatial information from features with high resolution, which has actually been adopted by many approaches (Long et al., 2015; Noh et al., 2015; Badrinarayanan et al., 2017; Zhao et al., 2016; Ronneberger et al., 2015). For the DecovNet (Noh et al., 2015) and the SegNet (Badrinarayanan et al., 2017), the *unpooling* operator has been proposed to transfer the values in the low-resolution feature maps to the corresponding positions in the high-resolution feature maps. Alternatively, a more intuitive solution is given by fusing high-resolution features from shallower layers with low-resolution features from deeper layers progressively (Long et al., 2015; Ronneberger et al., 2015) or at once (Zhao et al., 2016). In this paper, we adopt the intuitive way of fusing multi-scale features progressively instead of using the *unpooling* operator.

Especially for very deep networks, the vanishing gradient problem represents a big challenge for the optimization of neural networks. More specifically, the gradient of the error function decreases when being backpropagated to previous layers during training and, if the gradient becomes too small, the respective weights of the network remain unchanged (He et al., 2016b). Though some concepts like batch normalization (Ioffe and Szegedy, 2015) and residual connection (He et al., 2016a) have been proposed to address this issue, it cannot be thoroughly solved. Deep supervision (Szegedy et al., 2015; Marmanis et al., 2016; Lee et al., 2014; Lin et al., 2016) provides an option for better training by appending additional losses in the intermediate layers. In this paper, we inject two additional losses in the intermediate layers in order to achieve deep supervision.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-1-29-2018 | © Authors 2018. CC BY 4.0 License.

Further approaches directly address the deep semantic segmentation of aerial imagery based on multi-modal data, e.g. by involving an encoder-decoder architecture (Volpi and Tuia, 2017) or adaptations of the VGGNet (Chen et al., 2018a) and the ResNet (Chen et al., 2018b). To aggregate multi-scale predictions within a deep network, a modification of the SegNet introduces a multi-kernel convolutional layer allowing for convolutions with several filter sizes (Audebert et al., 2016). Alternatively, a deep network in which spatial features are learned at multiple resolutions and a specific module which learns how to combine these features can be combined (Maggiori et al., 2017). Further strategies to fuse the multi-modal geospatial data within a deep learning framework have been presented in (Marmanis et al., 2016; Audebert et al., 2016; Audebert et al., 2018; Liu et al., 2017). An alternative strategy to better retain boundaries between objects in the classification results is to take into account semantically meaningful boundaries, e.g. by including an explicit object boundary detector in the SegNet encoder-decoder architecture or in FCN-type models (Marmanis et al., 2018). While all these approaches focus on the classification pipeline, only little attention has been paid to the input data itself. Few investigations involve very basic hand-crafted features given by the Normalized Difference Vegetation Index (NDVI) and the normalized Digital Surface Model (nDSM) (Gerke, 2014; Audebert et al., 2016; Liu et al., 2017). Other kinds of hand-crafted radiometric or geometric features which can be extracted from a local image neighborhood (Gerke and Xiao, 2014; Weinmann and Weinmann, 2018) have however only rarely been involved, although, in the context of classifying aerial imagery based on given true orthophotos and the corresponding DSMs, it has recently been demonstrated that the additional consideration of such hand-crafted radiometric and geometric features on a per-pixel basis may lead to improved classification results (Chen et al., 2018b). In this paper, we focus on a multi-scale extension of Shuffling Convolutional Neural Networks (Chen et al., 2018a; Chen et al., 2018b) involving deep supervision, and we thereby also involve a diversity of hand-crafted radiometric and geometric features extracted from the true orthophotos and their corresponding DSMs, respectively.

3. METHODOLOGY

In this section, we describe our methodology for the semantic interpretation of aerial imagery by exploiting data of several modalities. Thereby, we first focus on the extraction of hand-crafted radiometric and geometric features and the creation of feature maps (Section 3.1). Subsequently, we provide a detailed explanation of our proposed deep networks receiving the defined feature maps as input (Section 3.2). The result is a dense labeling, i.e. each pixel is assigned a respective semantic label.

3.1 Feature Extraction

Given an orthophoto and the corresponding DSM on a regular grid, we first extract a set of hand-crafted radiometric and geometric features for all points on the grid. The derived features may thus be stored in the form of a stack of feature maps (i.e. images containing the values of a respective feature on a per-pixel basis) which later serves as input to a deep network.

Radiometric Features: We take into account the three spectral bands used for defining the orthophoto, whereby we assume that a representation with respect to the reflectance in the near-infrared (NIR), red (R) and green (G) domains is given (Rottensteiner et al., 2012). Furthermore, we involve normalized colors as a simple example of color invariants with improved robustness with respect to changes in illumination (Gevers and Smeulders, 1999), yielding normalized near-infrared (nNIR), normalized red (nR) and normalized green (nG) values. Finally, we also consider the Normalized Difference Vegetation Index (NDVI) (Rouse, Jr. et al., 1973) as a strong indicator for vegetation and a slight variation represented by the Green Normalized Difference Vegetation Index (GNDVI) (Gitelson and Merzlyak, 1998) which is more sensitive to the chlorophyll concentration than the original NDVI.

Geometric Features: Based on the DSM, we derive the normalized DSM (nDSM) describing the heights of objects above ground, which might be more informative than the DSM itself. For this purpose, we use the approach presented in (Gerke, 2014) which classifies pixels into ground and off-ground pixels using LAStools[1] and then adapts the height of each off-ground pixel by subtracting the height of the closest ground point. Furthermore, we consider geometric features in the form of local 3D shape features extracted from the DSM. Based on the spatial 3D coordinates corresponding to a local 3×3 image neighborhood, we efficiently derive the 3D structure tensor (Weinmann and Weinmann, 2018) and normalize its three eigenvalues by their sum. The normalized eigenvalues, in turn, are then used to calculate the features of linearity (L), planarity (P), sphericity (S), omnivariance (O), anisotropy (A), eigenentropy (E) and change of curvature (E) (West et al., 2004; Pauly et al., 2003) which have been involved in a variety of investigations for 3D scene analysis (Demantké et al., 2011; Weinmann, 2016; Hackel et al., 2016).

3.2 Supervised Classification

Once feature maps have been extracted, we use them as input to a deep network. Relying on the idea of a Shuffling Convolutional Neural Network (SCNN, presented in Section 3.2.1), we introduce the fusion of features of different scales as common strategy to address the localization/recognition trade-off (Maggiori et al., 2017). This results in a Multi-scale Shuffling Convolution Neural Network (MSCNN, presented in Section 3.2.2). Furthermore, we present an extension in the form of an MSCNN with deep supervision (DSCNN, presented in Section 3.2.3) which allows an improved classification due to the use of additional losses.

3.2.1 SCNN: In theory, any network designed for image classification can be adapted to a Shuffling Convolutional Neural Network (SCNN) for dense semantic image segmentation. The original SCNN (Chen et al., 2018a) is adapted from the VGGNet (Simonyan and Zisserman, 2014), while the Residual Shuffling Convolutional Neural Network (RSCNN) (Chen et al., 2018b) is adapted from the ResNet (He et al., 2016a). Thereby, the adaptation consists in involving a shuffling operator as an efficient operator to realize the upscaling of feature maps without introducing additional parameters. The concept of a shuffling operator has originally been introduced for super-resolution (Shi et al., 2016) and also been used for the semantic segmentation of aerial imagery (Chen et al., 2018a; Chen et al., 2018b). More specifically, the upscaling of feature maps is achieved by combining feature maps in a periodic shuffling manner to increase the resolution, which forces the network to learn upscaling. For example, if we need to double the resolution of the feature map, we can combine four feature maps as illustrated in Figure 2. The only hyper-parameter of a shuffling operator is the upscaling rate u. Generally, the process of constructing an SCNN can be split into four steps as described in the following:

[1] http://rapidlasso.com/lastools/

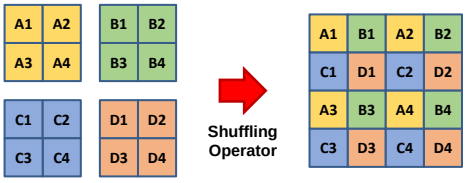


Figure 2. Basic concept of the shuffling operator: it converts $c \times u^2$ feature maps of size $H \times W$ into c feature maps of size $(H \times u) \times (W \times u)$. Here: $H = 2, W = 2, u = 2, c = 1$.

Backbone Network Extraction: Given a CNN for image classification such as the VGGNet (Simonyan and Zisserman, 2014) or the ResNet (He et al., 2016a), we will get the corresponding backbone network after removing the last pooling layer and its subsequent layers. In this paper, we use the respective adaptation of the ResNet-101 as backbone network. This is motivated by the fact that the use of standard networks such as the VGGNet (Simonyan and Zisserman, 2014) with many layers allows learning complex non-linear relationships, yet the performance of such networks tends to decrease when adding further layers via simply stacking convolutional layers due to the vanishing gradient problem. To effectively address this issue, we use the ResNet (He et al., 2016a) which is motivated by the idea that optimizing the residual mapping is easier than optimizing the original mapping. The additional gain in computational efficiency allows to form deep networks with more than 100 convolutional layers.

Resolution Adjustment: Pooling layers or convolutional layers with a stride larger than 1 will cause a reduction of resolution. We refer to such layers as Resolution Reduction Layers (RRL). Based on the backbone network, we keep the first three RRLs and change the strides of the remaining RRLs to 1. Though networks without RRLs (Sherrah, 2016) have been proposed, such networks always suffer from a severe computational overload (Sherrah, 2016; Chen et al., 2018a). Therefore, to have a trade-off between computational efficiency and a reasonable resolution, keeping three or four RRLs is an acceptable choice (Chen et al., 2014; Chen et al., 2017b; Chen et al., 2018a).

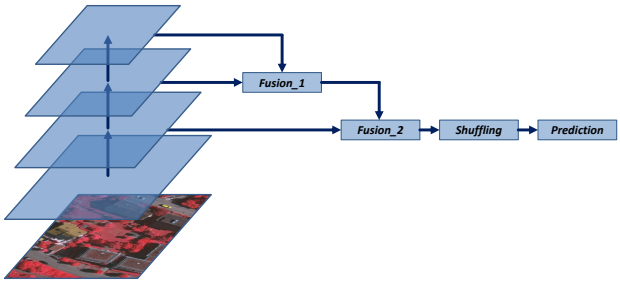
Field-of-View Enhancement: RRLs have been shown to not only be beneficial to learning features robust to the translation, but also to increase the field-of-view of filters (Chen et al., 2016). To compensate for the decrease of the field-of-view after removing some RRLs, atrous convolution (Chen et al., 2014; Chen et al., 2016) is introduced. Once the stride of an RRL is set to 1, the strides of its subsequent convolution layers are doubled.

Constructing SCNNs: On top of the modified backbone network, we append one convolution layer to generate a reasonable number of feature maps for the shuffling operator. This is followed by a shuffling operator and a softmax operator. An intuitive description can be seen in Figure 3(a). For simplicity, the shuffling operator in Figure 3(a) includes the convolution layer for creating the correct number of feature maps.

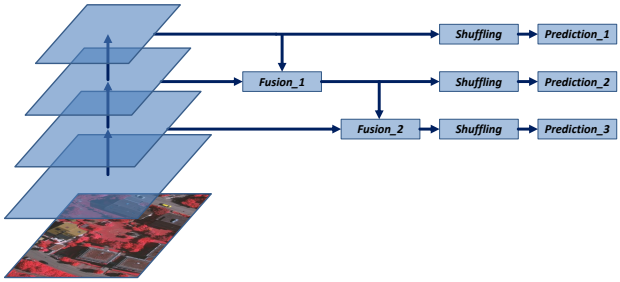
3.2.2 Multi-scale SCNN: By fusing feature maps of different scales, we construct a Multi-scale Shuffling Convolutional Neural Network (MSCNN). Deeper features with lower resolution are assumed to provide semantic information which are robust to variations in translation, rotation and scale. In contrast, features with higher resolution are assumed to contain more spatial information for better localization.

Focusing on computational efficiency, we keep the first four RRLs instead of only three RRLs. The feature maps with strides

(a) Original SCNN



(b) Multi-scale SCNN



(c) Multi-scale SCNN with deep supervision

Figure 3. The used networks: the prediction of the SCNN is based on the feature maps with an output stride of 8, while the MSCNN fuses outputs with strides of $\{16, 8, 4\}$ and the DSCNN

of $\{16, 8, 4\}$ are fused as shown in Figure 3(b). Correspondingly, the shuffling operator and the softmax operator are applied on the finest fused feature maps. The details of the used fusion module are presented in Figure 4. Thereby, the kernel of the convolution layers represented by *Convolution_1*, *Convolution_2* and *Convolution_4* is $(1, 1)$ and the kernel of *Convolution_3* is $(3, 3)$. We adopt deconvolution with a kernel of $(4, 4)$, a stride of $(2, 2)$ and a padding of $(1, 1)$ to realize upsampling. The number of filters for all these convolution/deconvolution layers is 512 and 256 for *Fusion_1* and *Fusion_2*, respectively.

3.2.3 Multi-scale SCNN with Deep Supervision: Deep supervision (Lee et al., 2014) is a technique used to achieve a better training (Marmanis et al., 2016; Szegedy et al., 2015), especially for very deep networks. Based on the MSCNN, we insert two additional losses in the form of cross-entropy losses into the intermediate layers as shown in Figure 3(c) and thus construct our Deeply-supervised Shuffling Convolutional Neural Network (DSCNN). Compared with the original SCNN, though additional fusion modules and additional prediction modules are introduced, the DSCNN is computationally more efficient than the original SCNN as four RRLs are kept instead of three. Also, these additional modules introduce some additional parameters for training. When e.g. using the adaptation of the ResNet-101 as backbone network, the number of trainable parameters of the RSCNN-101, MSCNN-101 and DSCNN-101 are shown in Table 1.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-1-29-2018 | © Authors 2018. CC BY 4.0 License.

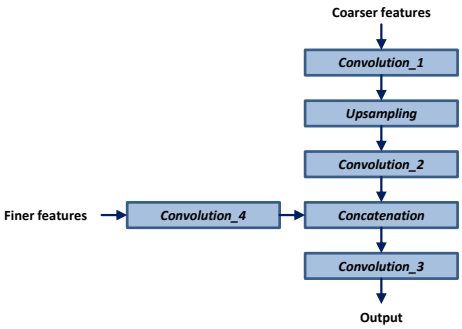


Figure 4. The fusion module: we adopt a deconvolution layer with a kernel of $(4, 4)$, a stride of $(2, 2)$ and a padding of $(1, 1)$ to realize upsampling.

Network	# Parameters	t_{train}
RSCNN-101	43.2M	15.3 h
MSCNN-101	55.9M	7.0 h
DSCNN-101	59.2M	7.8 h

Table 1. Number of parameters and time t_{train} required for training when using the same hyper-parameters for all networks.

4. EXPERIMENTAL RESULTS

In this section, we first provide a brief description of the dataset used in our experiments (Section 4.1). Subsequently, we explain implementation details as well as experimental configurations before presenting the derived results (Section 4.2).

4.1 Dataset

To evaluate the performance of our methods, we use the *Vaihingen Dataset* (Cramer, 2010; Rottensteiner et al., 2012). This dataset was acquired over a small village with many detached buildings and small multi-story buildings. It contains 33 tiles of different sizes and the spatial resolution is specified with 9 cm. For each tile, a very high-resolution true orthophoto (with three channels corresponding to the near-infrared, red and green domains) and the corresponding DSM derived via dense image matching techniques are provided. In addition, a reference labeling with respect to six semantic classes represented by *Impervious Surfaces*, *Building*, *Low Vegetation*, *Tree*, *Car* and *Clutter/Background* is given for 16 of the 33 tiles on a per-pixel basis. For extensive tests, we split the set of 16 labeled tiles into two subsets (Volpi and Tuia, 2017). One subset comprises the tiles with IDs $\in \{1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37\}$ and is used for training. The other subset comprises the tiles with IDs $\in \{11, 15, 28, 30, 34\}$ and is used for evaluation. Please note that the definition of training and test data thus differs from the one used for the *ISPRS Benchmark on 2D Semantic Labeling*.

4.2 Experiments and Results

Our implementations are based on the MXNet deep learning framework (Chen et al., 2015) and tested on a high-performance computer (Dual Intel Xeon E5-2609, 2.4 GHz, 256 GB RAM) equipped with a NVIDIA TITAN X GPU with 12 GB memory. The network parameters are initialized with the method introduced in (He et al., 2015). As loss function, we adopt the cross-entropy loss which is summed over all the pixels in a batch of 8 patches. To optimize this objective function, we use the standard Stochastic Gradient Descent (SGD) with a momentum of 0.9. Each patch fed into the network is cropped randomly and temporarily as proposed in (Chen et al., 2017a) and then normalized by the subtraction of the mean value and a subsequent division by the standard deviation of the patch. We employ "poly" learning (Chen et al., 2016), whereby the learning rate is multiplied by $(1 - \frac{N_{\text{iter}}}{N_{\text{max\_iter}}})^p$ with N_{iter} and $N_{\text{max\_iter}}$ denoting the number of iterations and the maximum number of iterations (and with $p = 0.9$ in our experiments). During training, we first use cropped patches of 224×224 pixels for 60k iterations with an initial learning rate of 0.007, and we then fine-tune the network using cropped patches of 448×448 pixels for further 15k iterations with an initial learning rate of 0.001. As evaluation metrics, we consider the Overall Accuracy (OA) and the mean Intersection-over-Union (mIoU). To reason about the performance for single classes, we additionally consider the classwise F_1-scores.

Flip+Rotate	MSA	OA	mIoU
		83.89	57.76
✓		84.29	58.60
✓	✓	84.65	58.48

Table 2. The contribution of data augmentation. In this experiment, only the reflectance values in the near-infrared (NIR), red (R) and green (G) bands are used as features.

Inference Strategy: When making inference for the tiles in the evaluation set, two different strategies may be applied. The first strategy relies on resizing the tiles to suitable sizes in order to meet the size requirement of the networks. After making the prediction, the results are resized to the original sizes of the tiles. In this way, the shapes of objects are always distorted before the resized tiles are fed into networks. Therefore, we denote this strategy as "warp". The second strategy is based on making predictions for the cropped patches from the tiles as the Field-of-View Enhancement (FoVE) (Chen et al., 2018a; Chen et al., 2017a). This involves two hyper-parameters given by the size of the cropped patches and the step size.

Data Augmentation: To address the problem of overfitting, we apply data augmentation by randomly scaling the input images from 0.5 to 2.0 (known as **M**ulti-**S**cale **A**ugmentation, **MSA**), horizontally flipping and rotating by 90°, 180° or 270°, respectively. Based on the RSCNN-101, we investigate the importance of data augmentation for which numerical results are provided in Table 2. In this experiment, we only make use of features based on the reflectance values in the near-infrared (NIR), red (R) and green (G) bands. To get the prediction for each tile, we crop patches of 992×992 pixels with a step size of 812 pixels from the tile and make a patch-wise prediction following (Chen et al., 2018a). As can be seen in the table, the overall accuracy increases from 83.89% to 84.65% when involving data augmentation.

Multi-scale Feature Fusion and Deep Supervison: To explore the contribution of multi-scale fusion and deep supervision, we train the RSCNN-101, the MSCNN-101 and the DSCNN-101, respectively, and evaluate their performance on the validation set. The numerical results are provided in Table 3 and corresponding visualizations for Patch 30 are provided in Figure 5. Here, we only make use of features based on the reflectance values in the near-infrared (NIR), red (R) and green (G) bands. When making predictions, we feed a whole tile into the respective network after being resized to a suitable size. As can be seen in this table, the OA increases from 84.37% to 85.71% and the mIoU increases from 58.31% to 60.64% after fusing multi-scale features and adopting deep supervision.

Multi-modal Data: To explore the influence of additionally using hand-crafted features to define the input to the deep network, we focus on a separate and combined consideration of radiometric and geometric information (cf. Section 3.1) as input to the

DSCNN-101. The achieved classification results when relying on the *Prediction_3* are provided in Table 4. In this table, we list the results derived for the use of both inference strategies given by FoVE and "warp", respectively. For Patch 30 of the dataset, a visualization of the ground truth and the semantic labelings derived with the DSCNN-101(3) based on different sets of involved features is given in Figure 6.

5. DISCUSSION

Inference Strategy: Table 4 reveals that the inference strategy of "warp" is not as stable as FoVE. Especially when the input contains the DSM, the performance of "warp" is much worse than that of FoVE (more than 1% drop with respect to OA). In other cases, there is no big difference between these two strategies.

Data Augmentation: Data augmentation is an efficient way to address the problem of overfitting. As can be seen in Table 2, when training with horizontal flip and rotation, the OA increases by 0.40% and the mIoU increases by 0.84%. Adding MSA, the OA is further increased by 0.36% while the mIoU shows a slight drop by 0.12%. This drop is due to the unbalanced convergences between classes as the F_1-score of the class *Car* drops from 74.39% to 71.37% when adding MSA.

Multi-scale Feature Fusion and Deep Supervison: Both the fusion of multi-scale features and deep supervision contribute to the improvement in performance. As can be seen in Table 3, fusing multi-scale features results in an improvement of 0.65% in OA and 1.21% in mIoU, respectively. Additionally considering deep supervision leads to a further improvement of 0.69% in OA and 1.12% in mIoU when considering the prediction based on DSCNN-101(3). Therefore, compared with the original RSCNN-101, our proposed network yields an improvement of 1.37% in OA and 2.23% in mIoU. This improvement comes from two aspects. On the one hand, an improvement arises from the fusion of multi-scale features, where high-level features provide more semantic information and low-level features provide more spatial information. Through the fusion modules, the network can take advantage of both high-level features and low-level features. On the other hand, an improvement arises from the deep supervision which provides a better optimization of the network. Though there exist skip connections in the residual modules, the vanishing gradient problem cannot be eliminated thoroughly. With deep supervision, it can be alleviated to some extent and put more semantic constraints on the intermediate layers. Furthermore, the derived results reveal that predictions based on finer features outperform predictions based on coarser features. As can be seen in Table 3, DSCNN-101(3) outperforms DSCNN-101(2) which in turn outperforms DSCNN-101(1) in OA. The same conclusion can be obtained through observing the second row of Figure 5. Meanwhile, taking the average of the score maps of *Prediction_1*, *Prediction_2* and *Prediction_3* in the DSCNN-101 does not contribute to an improvement in performance. As can be seen in Table 3, the OA achieved with the DSCNN-101(E) is only 0.01% higher than the OA achieved with the DSCNN-101(3) while the mIoU drops by 0.08%. Therefore, in the experiments of exploring the contribution of multi-modal data for classification, we adopt the results of DSCNN-101(3). Visualizations of the predictions of different networks are shown in Figure 5. The figures reveal the gridding effect (i.e. the result is not very smooth and contains many scatters in grids) in the prediction of RSCNN-101. The gridding effect is alleviated to a large extent with the fusion of multi-scale features. Adding deeper supervision, the result becomes much smoother.

Multi-modal Data: Height features such as the DSM and the nDSM can provide complementary information to the spectral features (near-infrared, red and green). As can be seen in the first block of Table 4, the combinations of spectral features and height features yield the best results. The highest OA (86.31%) is reached with the combination of NIR, R, G and nDSM. The highest mIoU (61.33%) is reached with the combination of NIR, R, G, DSM and nDSM. Similarly, as can be seen in Figure 6, the height information provides vital information for identifying some specific objects, e.g., the building at the top left corner. Interestingly, there is no numerical improvement when using additional hand-crafted features (cf. Section 3.1) in comparison to the standard RSCNN where an improvement can be observed (Chen et al., 2018b). One reason might be that these features can be learned by this deep network. However, in Figure 6, it can be observed for some objects that the additional use of hand-crafted features contributes to the classification, while they also introduce some noise which may result in incorrect classification. Therefore, we believe that the consideration of hand-crafted radiometric and geometric features is still significant.

6. CONCLUSIONS

In this paper, we have proposed the Deeply-supervised Shuffling Convolutional Neural Network (DSCNN) for semantic image segmentation. The DSCNN extends the standard SCNN by fusing multi-scale features and introducing deep supervision. The results derived for a benchmark dataset reveal that our proposed network outperforms the baseline network, where both the multi-scale fusion and the additional losses contribute to the improvement. By fusing multi-scale features, the proposed network effectively addresses the gridding effect and produces much smoother results than the original network. Via feeding different combinations of the multi-modal data and derived hand-crafted features, we have investigated the value of the data of both modalities and the derived features. The derived results reveal that using all radiometric and geometric features does not achieve the best result. However, via analyzing the visualization of the results, we find that the results derived from using all features are smoother and the predictions of some objects are improved. However, it also introduces some misclassifications. Hence, we conclude that the effective use of hand-crafted features remains a challenge to be addressed in future work.

ACKNOWLEDGEMENTS

This work is supported by the foundation of China Scholarship Council under Grant 201704910608. The *Vaihingen Dataset* was provided by the *German Society for Photogrammetry, Remote Sensing and Geoinformation* (*DGPF*) (Cramer, 2010): http://www.ifp.uni-stuttgart.de/dgpf/DKEP-Allg.html.

REFERENCES

Audebert, N., Le Saux, B. and Lefèvre, S., 2016. Semantic segmentation of Earth observation data using multimodal and multi-scale deep networks. In: *Proc. 13th Asian Conference on Computer Vision (ACCV)*, Taipei, Taiwan, Vol. I, pp. 180–196.

Audebert, N., Le Saux, B. and Lefèvre, S., 2018. Beyond RGB: very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing* 140, pp. 20–32.

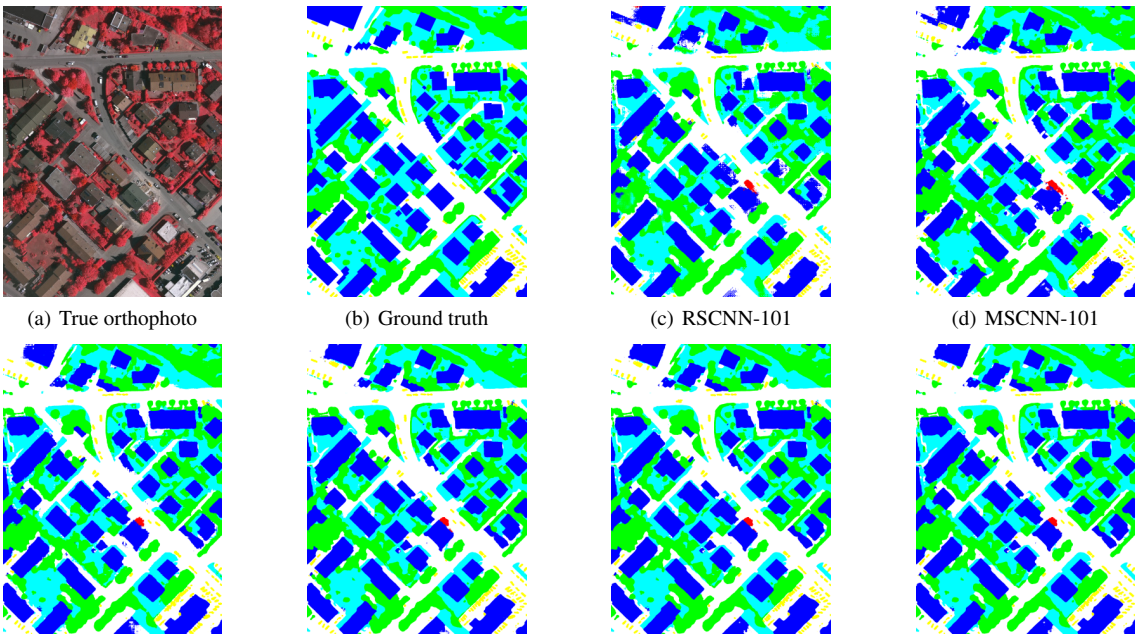
Badrinarayanan, V., Kendall, A. and Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(12), pp. 2481–2495.

Networks	F_1(IS)	F_1(B)	F_1(LV)	F_1(T)	F_1(C)	OA	mIoU
RSCNN-101	87.05	91.93	74.92	83.51	71.90	84.37	58.31
MSCNN-101	87.45	92.55	74.96	84.51	75.13	85.02	59.52
DSCNN-101(1)	87.76	93.07	75.97	84.64	72.60	85.38	59.47
DSCNN-101(2)	88.14	93.46	76.21	84.67	75.72	85.66	60.68
DSCNN-101(3)	88.20	93.55	76.26	84.66	76.57	85.71	60.64
DSCNN-101(E)	88.18	93.45	76.31	84.75	76.23	85.72	60.56

Table 3. The contribution of multi-scale fusion and deep supervision when solely using the reflectance information in the near-infrared (NIR), red (R) and green (G) domains. The expressions DSCNN-101(1), DSCNN-101(2) and DSCNN-101(3) denote the results of *Prediction_1*, *Prediction_2* and *Prediction_3* in Figure 3(c), respectively. The expression DSCNN-101(E) means that the final predictions are based on taking the average of the score maps of *Prediction_1*, *Prediction_2* and *Prediction_3* in Figure 3(c).

Input	FoVE	Warp	F_1(IS)	F_1(B)	F_1(LV)	F_1(T)	F_1(C)	OA	mIoU
NIR-R-G	✓		87.77	93.42	75.96	84.99	74.40	85.61	60.04
NIR-R-G-DSM	✓		88.07	94.14	76.75	85.01	77.57	86.00	61.20
NIR-R-G-nDSM	✓		88.59	94.00	77.21	85.23	75.82	86.31	61.06
NIR-R-G-DSM-nDSM	✓		88.36	93.81	77.97	85.09	77.93	86.11	61.33
NIR-R-G-nDSM-L-P-S	✓		88.34	93.92	76.53	85.36	77.16	86.15	61.15
NIR-R-G-NDVI-nDSM-L-P-S	✓		88.77	94.04	76.68	84.93	74.48	86.20	61.08
Radiometry & Geometry	✓		88.66	93.78	76.18	84.61	75.28	85.89	60.52
NIR-R-G		✓	88.20	93.55	76.26	84.66	76.57	85.71	60.64
NIR-R-G-DSM		✓	87.48	93.41	75.32	83.69	76.79	84.93	60.01
NIR-R-G-nDSM		✓	88.50	93.84	77.22	85.09	76.12	86.23	61.03
NIR-R-G-DSM-nDSM		✓	87.61	93.21	74.83	84.04	75.22	84.95	59.63
NIR-R-G-nDSM-LPS		✓	88.42	94.11	76.88	85.22	76.99	86.23	61.23
NIR-R-G-NDVI-nDSM-L-P-S		✓	88.76	93.85	76.94	84.91	76.94	86.18	61.18
Radiometry & Geometry		✓	88.93	93.81	76.69	84.80	75.09	86.13	60.72

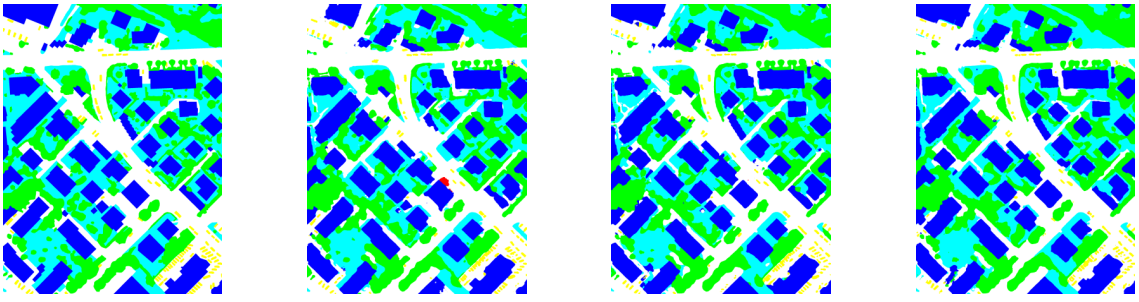
Table 4. The influence of different combinations of features and of different inference strategies on the result of the DSCNN-101(3).



(a) True orthophoto (b) Ground truth (c) RSCNN-101 (d) MSCNN-101

(e) DSCNN-101(1) (f) DSCNN-101(2) (g) DSCNN-101(3) (h) DSCNN-101(E)

Figure 5. Visualization of the true orthophoto of Tile 30, the corresponding ground truth and the results for semantic segmentation when using the different deep networks defined in Section 3.2 and the “warp” inference strategy (*Impervious Surfaces*: white; *Building*: blue; *Low Vegetation*: cyan; *Tree*: green; *Car*: yellow; *Clutter/Background*: red).



(a) Ground truth (b) NIR-R-G (c) NIR-R-G-nDSM (d) All defined features

Figure 6. Visualization of the ground truth of Tile 30 and the classification results achieved with the DSCNN-101(3) when using different subsets of the features defined in Section 3.1 and the “warp” inference strategy, and the same color encoding as in Figure 5.

Chen, K., Fu, K., Gao, X., Yan, M., Sun, X. and Zhang, H., 2017a. Building extraction from remote sensing images with deep learning in a supervised manner. In: *Proc. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, TX, USA, pp. 1672–1675.

Chen, K., Fu, K., Yan, M., Gao, X., Sun, X. and Wei, X., 2018a. Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters* 15(2), pp. 173–177.

Chen, K., Weinmann, M., Gao, X., Yan, M., Hinz, S., Jutzi, B. and Weinmann, M., 2018b. Residual shuffling convolutional neural networks for deep semantic image segmentation using multi-modal data. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Riva del Garda, Italy, Vol. IV-2, pp. 65–72.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L., 2014. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A. L., 2016. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv preprint arXiv:1606.00915*.

Chen, L.-C., Papandreou, G., Schroff, F. and Adam, H., 2017b. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C. and Zhang, Z., 2015. MXNet: a flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*.

Cramer, M., 2010. The DGPF-test on digital airborne camera evaluation – Overview and test design. *PFG Photogrammetrie – Fernerkundung – Geoinformation* 2/2010, pp. 73–82.

Demantké, J., Mallet, C., David, N. and Vallet, B., 2011. Dimensionality based scale selection in 3D lidar point clouds. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Calgary, Canada, Vol. XXXVIII-5/W12, pp. 97–102.

Gerke, M., 2014. Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen). Technical report, ITC, University of Twente.

Gerke, M. and Xiao, J., 2014. Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification. *ISPRS Journal of Photogrammetry and Remote Sensing* 87, pp. 78–92.

Gevers, T. and Smeulders, A. W. M., 1999. Color based object recognition. *Pattern Recognition* 32(3), pp. 453–464.

Gitelson, A. A. and Merzlyak, M. N., 1998. Remote sensing of chlorophyll concentration in higher plant leaves. *Advances in Space Research* 22(5), pp. 689–692.

Hackel, T., Wegner, J. D. and Schindler, K., 2016. Fast semantic segmentation of 3D point clouds with strongly varying density. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Prague, Czech Republic, Vol. III-3, pp. 177–184.

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1026–1034.

He, K., Zhang, X., Ren, S. and Sun, J., 2016a. Deep residual learning for image recognition. In: *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770–778.

He, K., Zhang, X., Ren, S. and Sun, J., 2016b. Identity mappings in deep residual networks. In: *Proc. European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, pp. 630–645.

Ioffe, S. and Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proc. 32nd International Conference on Machine Learning (ICML)*, Lille, France, pp. 448–456.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. In: *Proc. 25th International Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, Vol. I, pp. 1097–1105.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z. and Tu, Z., 2014. Deeply-supervised nets. *arXiv preprint arXiv:1409.5185*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., 2016. Feature pyramid networks for object detection. *arXiv preprint arXiv:1612.03144*.

Liu, W., Rabinovich, A. and Berg, A. C., 2015. ParseNet: looking wider to see better. *arXiv preprint arXiv:1506.04579*.

Liu, Y., Piramanayagam, S., Monteiro, S. T. and Saber, E., 2017. Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order CRFs. In: *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, pp. 1561–1570.

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 3431–3440.

Maggiori, E., Tarabalka, Y., Charpiat, G. and Alliez, P., 2017. High-resolution image classification with convolutional networks. In: *Proc. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Fort Worth, TX, USA, pp. 5157–5160.

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M. and Stilla, U., 2018. Classification with an edge: improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 135, pp. 158–172.

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M. and Stilla, U., 2016. Semantic segmentation of aerial images with an ensemble of CNNs. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Prague, Czech Republic, Vol. III-3, pp. 473–480.

Noh, H., Hong, S. and Han, B., 2015. Learning deconvolution network for semantic segmentation. In: *Proc. 2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1520–1528.

Paisitkriangkrai, S., Sherrah, J., Janney, P. and van den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9(7), pp. 2868–2881.

Pauly, M., Keiser, R. and Gross, M., 2003. Multi-scale feature extraction on point-sampled surfaces. *Computer Graphics Forum* 22(3), pp. 81–89.

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*.

Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S. and Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Melbourne, Australia, Vol. I-3, pp. 293–298.

Rouse, Jr., J. W., Haas, R. H., Schell, J. A. and Deering, D. W., 1973. Monitoring vegetation systems in the Great Plains with ERTS. In: *Proc. 3rd Earth Resources Technology Satellite-1 Symposium (ERTS)*, Washington, D.C., USA, Vol. I, pp. 309–317.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3), pp. 211–252.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D. and Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 1874–1883.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1–9.

Volpi, M. and Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing* 55(2), pp. 881–893.

Weinmann, M., 2016. *Reconstruction and analysis of 3D scenes – From irregularly distributed 3D points to object classes*. Springer, Cham, Switzerland.

Weinmann, M. and Weinmann, M., 2018. Geospatial computer vision based on multi-modal data – How valuable is shape information for the extraction of semantic information? *Remote Sensing* 10(1), pp. 2:1–2:20.

West, K. F., Webb, B. N., Lersch, J. R., Pothier, S., Triscari, J. M. and Iverson, A. E., 2004. Context-driven automated target detection in 3-D data. *Proceedings of SPIE* 5426, pp. 133–143.

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J., 2016. Pyramid scene parsing network. *arXiv preprint arXiv:1612.01105*.