ISPRS Journal of Photogrammetry and Remote Sensing 143 (2018) 191–204

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A multi-scale fully convolutional network for semantic labeling of 3D point clouds

Mohammed Yousefhussien [a,*], David J. Kelbe [b], Emmett J. Ientilucci [a], Carl Salvaggio [a]

[a] Rochester Institute of Technology, Chester F. Carlson Center for Imaging Science, Rochester, NY, USA
[b] Oak Ridge National Laboratory, Geographic Information Science and Technology Group, Oak Ridge, TN, USA

ARTICLE INFO

Article history:
Received 2 October 2017
Received in revised form 9 March 2018
Accepted 16 March 2018
Available online 16 May 2018

Keywords:
LiDAR
3D-labeling contest
Deep learning

ABSTRACT

When classifying point clouds, a large amount of time is devoted to the process of engineering a reliable set of features which are then passed to a classifier of choice. Generally, such features – usually derived from the 3D-covariance matrix – are computed using the surrounding neighborhood of points. While these features capture local information, the process is usually time-consuming and requires the application at multiple scales combined with contextual methods in order to adequately describe the diversity of objects within a scene. In this paper we present a novel 1D-fully convolutional network that consumes terrain-normalized points directly with the corresponding spectral data (if available) to generate point-wise labeling while implicitly learning contextual features in an end-to-end fashion. This unique approach allows us to operate on unordered point sets with varying densities, without relying on expensive hand-crafted features; thus reducing the time needed for testing by an order of magnitude over existing approaches. Our method uses only the 3D-coordinates and three corresponding spectral features for each point. Spectral features may either be extracted from 2D-georeferenced images, as shown here for Light Detection and Ranging (LiDAR) point clouds, or extracted directly for passive-derived point clouds, *i.e.* from multiple-view imagery. We train our network by splitting the data into square regions and use a pooling layer that respects the permutation-invariance of the input points. Evaluated using the ISPRS 3D Semantic Labeling Contest, our method scored second place with an overall accuracy of 81.6%. We ranked third place with a mean F1-score of 63.32%, surpassing the F1-score of the method with highest accuracy by 1.69%. In addition to labeling 3D-point clouds, we also show that our method can be easily extended to 2D-semantic segmentation tasks, with promising initial results.

1. Introduction

The past decade of computer/machine vision research and remote sensing hardware development has broadened the availability of 3D point cloud data through innovations in Light Detection and Ranging (LiDAR), Synthetic Aperture Radar (SAR), dense stereo- or multiview-photogrammetry and structure from motion (SfM). Despite the prevalence of 3D-point cloud data, automated interpretation and knowledge discovery from 3D-data remains challenging due to the irregular structure of raw point clouds. As such, exploitation has typically been limited to simple visualization and basic mensuration (Hackel et al., 2016). Or, some authors rasterized the point cloud onto a more tracTable 2.5D- Digital Surface Model (DSM) from which conventional image processing techniques are applied, *e.g.* (Hug and Wehr, 1997; Haala et al., 1998).

In order to generate exploitation-ready data products directly from the point cloud, semantic classification is desired. Similar to per-pixel image labeling, 3D-semantic labeling seeks to attribute a semantic classification label to each 3D-point. Classification labels, *e.g.* vegetation, building, road, etc., can subsequently be used to inform derivative processing efforts such as surface fitting (Xing et al., 2017), 3D modeling (Moussa and El-Sheimy, 2010), object detection (Jochem et al., 2009), and bare-earth extraction (Yunfei et al., 2008). However, the task of labeling every data point in the irregularly distributed point cloud captured by aerial platforms is challenging, especially in urban scenes with different object types and various scales ranging from very small spatial neighborhoods (power lines) to very large spatial neighborhoods (buildings). Moreover, point clouds are unstructured and unordered data with variable spatial densities. In order to scale the

* Corresponding author.
E-mail address: myhussien@gmail.com (M. Yousefhussien).

https://doi.org/10.1016/j.isprsjprs.2018.03.018

semantic classification task to meet the demands of emerging data volumes potentially at sub-meter resolution and global in coverage an efficient, streamlined, and robust model that directly operates on 3D point clouds is needed. The goal of this research is to introduce a flexible and simple multi-scale deep learning framework for direct semantic labeling of 3D aerial point clouds, thus eliminating the need for calculating costly, handcrafted features. The algorithm respects the permutation-invariance of input points and therefore avoids the need to transform the points to images or volumes.

2. Related work

Point cloud labeling algorithms can generally be grouped into two main categories. Section 2.1 describes "Direct Methods", which operate immediately on the point clouds themselves and do not change the 3D-nature of the data. Section 2.2 describes "Indirect Methods", which transform the input point cloud into an image or a volume as a preconditioning step to more traditional (raster-based) segmentation approaches. Considering the relative trade-offs of these techniques, Section 2.3 proposes a novel approach with 7 specific contributions for semantic classification of point clouds.

2.1. Direct methods

Direct methods assign semantic labels to each element in the point cloud based on a simple point-wise discriminative model operating on point features. Such features, known as "eigen-features", are derived from the covariance matrix of a local neighborhood and provide information on the local geometry of the sampled surface, *e.g.* planarity, sphericity, linearity (Lin et al., 2014a). To improve classification, contextual information can explicitly be incorporated into the model. For example, Blomley et al. (2016) used covariance features at multiple scales found using the eigenentropy-based scale selection method (Weinmann et al., 2014) and evaluated four different classifiers using the ISPRS 3D Semantic Labeling Contest.[1] Their best-performing model used a Linear Discriminant Analysis (LDA) classifier in conjunction with various local geometric features. However, scalability of this model was limited due to the dependence upon various handcrafted features and the need to experiment with various models that don't incorporate contextual features and require effort to tune.

Motivated by the frequent availability of coincident 3D data and optical imagery, Ramiya et al. (2014) proposed the use of point coordinates and spectral data directly, forming a per-point vector of (X, Y, Z, R, G, B) components. Labeling was achieved by filtering the scene into ground and non-ground points according to Axelsson (2000), then applying a 3D-region-growing segmentation to both sets to generate object proposals. Like Blomley et al. (2016), several geometric features were also derived, although specific details were not published. Without incorporating contextual features, each proposed segment was then classified according to the five classes from the ISPRS 3D Semantic Labeling Contest.

Alternatively, Mallet (2010) classified full-waveform LiDAR data using a point-wise multiclass support vector machine (SVM). And (Chehata et al., 2009) used random forests (RF) for feature detection and classification of urban scenes collected by airborne LiDAR. The reader is referred to Grilli et al. (2017) for a more complete review of discriminative classification models. While simple discriminative models are well-established, they are unable to consider interactions between 3D points.

To allow for spatial dependencies between object classes by considering labels of the local neighborhood, Niemeyer et al. (2014) proposed a contextual classification method based on Conditional Random Field (CRF). A linear and a random forest model were compared when used for both the unary and the pairwise potentials. By considering complex interactions between points, promising results were achieved, despite the added cost of computation speed: 3.4 min for testing using an RF model, and 81 min using the linear model. This computation time excludes the additional time needed to estimate the per-point, 131-dimensional feature vector prior to testing.

This contextual classification model was later extended to use a two-layer, hierarchical, high-order CRF, which incorporates spatial and semantic context (Niemeyer et al., 2016). The first layer operates on the point level, utilizing higher-order cliques and geometric features (Weinmann et al., 2014) to generate segments. The second layer operates on the generated segments, and therefore incorporates a larger spatial scale. Features included geometric- and intensity-based descriptors, in addition to distance and orientation to road features (Golovinskiy et al., 2009). By iteratively propagating context between layers, incorrect classifications can be revised at later stages; this resulted in good performance on a 2.25 million point dataset of Hannover, Germany. However, this method employed multiple algorithms, each designed separately, which would make simultaneously optimization challenging. Also, the use of computationally-intensive inference methods limits the run-time performance. In contrast to relying on multiple individually-trained components, an end-to-end learning mechanism is desired.

2.2. Indirect methods

Indirect methods – which mostly rely on deep learning – offer the potential to learn local and global features in a streamlined, end-to-end fashion (Yosinski et al., 2015). Driven by the reintroduction and improvement of Convolutional Neural Networks (CNNs) (LeCun et al., 1989; He et al., 2016), the availability of large-scale datasets (Deng et al., 2009), and the affordability of high-performance computing resources such as graphics processing units (GPUs), deep learning has enjoyed unprecedented popularity in recent years. This success in computer vision domains such as image labeling (Krizhevsky et al., 2012), object detection (Girshick et al., 2014), semantic segmentation (Badrinarayanan et al., 2017; Long et al., 2015), and target tracking (Wang and Yeung, 2013; Yousefhussien et al., 2016), has generated an interest in applying these frameworks for 3D classification.

However, the nonuniform and irregular nature of 3D-point clouds prevents a straightforward extension of 2D-CNNs, which were originally designed for raster imagery. Hence, initial deep learning approaches have relied on *transforming* the 3D data into more tracTable 2D images. For example, Su et al. (2015) rendered multiple synthetic "views" by placing a virtual camera around the 3D object. Rendered views were passed though replicas of the trained CNN, aggregated using a view-pooling layer, and then passed to another CNN to learn classification labels. Several other methods use the multiview approach with various modifications to the rendered views. For example, Bai et al. (2016) generated depth images as the 2D views, while other methods accumulated a unique signature from multiple view features. Still other methods projected the 3D information into 36 channels, modifying AlexNet (Krizhevsky et al., 2012) to handle such input. For further details, the reader is referred to (Savva et al., 2016).

Similar multiview approaches have also been applied to ground-based LiDAR point clouds. For example, Boulch et al. (2017) generated a mesh from the Semantic3D Large-scale Point Cloud Classification Benchmark (Hackel et al., 2017); this allowed for the generation of synthetic 2D views based on both RGB information and a 3-channel depth composite. A two-stream Seg-

[1] https://goo.gl/fSK6Fy.

Net (Badrinarayanan et al., 2017) network was then fused with residual correction (Audebert et al., 2016) to label corresponding pixels. 2D labels were then back-projected to the point cloud to generate 3D semantic classification labels. Likewise, Caltagirone et al. (2017) generated multiple overhead views, embedded with elevation and density features, to assist with road detection from LiDAR data. A Fully-Convolutional Network (FCN) (Long et al., 2015) was used for a single-scale binary semantic segmentation {road, not-road} based on training from the KITTI dataset (Geiger et al., 2013).

Despite their initial adoption, such multiview transformation approaches applied to point clouds lose information on the third spatial dimension through a projective rendering. Simultaneously, they introduce interpolation artifacts and void locations. Together, this complicates the process by unnecessarily rendering the data in 2D and forcing the network to ignore artificial regions caused by the voids. While this is less consequential for binary classification problems, multi-class problems require that each point be assigned to a separate class; this increases the complexity and may reduce the network's performance.

In light of these limitations of multiview transformation methods, other authors have taken a volumetric approach to handle points clouds using deep learning. For example, Li (2017) presented a method for vehicle detection in ground-based LiDAR point clouds. The input point cloud was voxelized and then appended with a fourth binary channel representing the *binary occupancy*, *i.e.* the presence or the absence of a point within each voxel. Using the KITTI dataset, a 3D-FCN was then trained and evaluated to produce two maps representing the objectness and bounding box scores. Similarly, Huang and You (2016) generated occupancy voxel grids based on LiDAR point cloud data, labeling each voxel according to the annotation of its center point. A 3D-CNN was then trained to label each voxel into one of seven classes; individual points were then labeled according to their parent voxel. Other authors have explored variations of voxelization methods including a binary occupancy grid, a density grid, and a hit grid. In VoxNet, Maturana and Scherer (2015) tested each voxelization model individually, to train 3D-CNNs with $32 \times 32 \times 32$ grid inputs. To handle multi-resolution inputs, they trained two separate networks, each receiving an occupancy grid with a different resolution.

Parallel development of both multiview and volumetric CNNs has resulted in an empirical performance gap. Qi et al. (2016) suggested that results could collectively be improved by merging these two paradigms. To address this, a hybrid volumetric CNN was proposed, which used long anisotropic kernels to project the 3D-volume into a 2D-representation. Outputs were processed using an image-based CNN adapted from the Network In Network (NIN) architecture (Lin et al., 2014b). To combine the multiview approach with proposed volumetric methods, the 3D-object was rotated to generate different 3D-orientations. Each individual orientation was processed individually by the same network to generate 2D-representations, which were then pooled together and passed to the image-based CNN.

Finally, Liu et al. (2017) took a different approach by combining image-like representations with a CRF. Instead of directly operating on the LiDAR data, they interpolated the DSM map as a separate channel. Using the imagery and the LiDAR data, two separate probability maps were generated. A pre-trained FCN was used to estimate the first probability map using optical imagery. Then, by handcrafting another set of features from both the spectral and the DSM map, a logistic regression was applied to generate a second set of probability maps. At the end of this two-stream process, the two probability maps were combined using high-order CRF to label every pixel into one of six categories.

2.3. Contribution

Although indirect methods introduced the application of deep learning for the semantic labeling task, they typically require a transformation of the input data, *i.e.* to views or volumes, in order to meet the ingest requirements of conventional image-based networks. Unfortunately, these transformations introduce computational overhead, add model complexity, and discard potentially relevant information. Likewise, direct methods have relied on the proliferation of various handcrafted features, in addition to contextual relationships, in order to meet increasing accuracy requirements with simple discriminative models. This added complexity has come at the cost of computational efficiency.

Meanwhile, the generation of 3D point cloud data has increased rapidly in recent years due to the availability of high-resolution optical satellite/airborne imagery and the explosion of modern stereo photogrammetry algorithms leveraging new computational resources. Such algorithms triangulate 3D-point coordinates directly from the optical imagery, and thus retain inherent spectral information; these attributes should be considered in the development of a successful model. In order to scale the semantic classification task to meet the demands of emerging data volumes – potentially at sub-meter resolution and global in coverage – an efficient, streamlined, and robust model that directly operates on 3D point clouds is needed.

It is in this context that we propose a simple, fully-convolutional network for direct semantic labeling of 3D point clouds with spectral information. Our proposed approach utilizes a modified version of PointNet (Qi et al., 2017), a deep network which operates directly on point clouds and so provides a flexible framework with large capacity and minimal overhead for efficient operation at scale. Moreover, it respects the permutation-invariance of input points and therefore avoids the need to transform the points to images or volumes. The network allows for summarizing the entire input point set with a feature vector that can be used for object classification, retrieval, and 2D space visualization as shown in (Qi et al., 2017). We will refer to this vector as a global feature for the input set. The use of a global descriptor of an entire set of input samples is widely used in the video understanding domain. For example, Ryoo et al. (2015) temporally pooled the per-frame features extracted by a CNN to generate a global descriptor for the entire video to recognize the activity within the video.

Inspired by the success of PointNet in applications such as object classification, part segmentation, and semantic labeling, we make the following contributions:

1. We extend PointNet to handle complex 3D data obtained from overhead remote sensing platforms using a multi-scale approach. Unlike CAD models, precisely-scanned 3D objects, or even indoor scenes, airborne point clouds exhibit unique characteristics, such as noise, occlusions, scene clutter, and terrain variation, which challenge the semantic classification task.
2. We present a deep learning algorithm with convolutional layers that consume unordered and unstructured point clouds directly, and therefore respects the pedigree of the input 3D data without modifying its representation and discarding information.
3. We eliminate the need for calculating costly handcrafted features and achieve near state-of-the-art results with just the three spatial coordinates and three corresponding spectral values for each point. At the same time, the overhead of adding additional features to our model is minimal compared to adding new channels or dimensions in 2D and volumetric cases.

4. We avoid the need to explicitly calculate contextual relationships (*e.g.* by using CRF) and instead use a simple layer that can learn a per-block global features during training.
5. Being fully convolutional, our network mitigates the issue of non-uniform point density, a common pitfall for stationary LiDAR platforms.
6. We show that test time is on the order of seconds, compared to minutes for existing techniques relying on handcrafted features or contextual methods, and operating on the same dataset.
7. Finally, we show how the network can easily be applied to handle the issue of multimodal fusion in 2D-semantic segmentation, with a simple modification to the data preparation step.

This paper is organized as follows: Section 3 will describe the network used and the adaptation of convolutions to 3D point clouds. Sections 4 will present the methodology used to evaluate our approach. Finally, Section 5 draws the conclusions and presents the future work.

3. Methodology

In this section we present a CNN-based deep learning method that is able to learn point-level and global features directly in an end-to-end fashion, rather than relying upon costly features or contextual processing layers. In Section 3.1, we describe how convolutional networks can be adapted to irregular point cloud data. Section 3.2 describes how batch normalization is used to precondition the outputs of the activation functions. Section 3.3 describes a pooling layer that is used to learn contextual features. Finally, Section 3.4 details the inference of semantic classification labels from the learned local and global features.

3.1. Adaptation of CNN to point clouds

CNN architectures consist of multiple, layered convolution operations, wherein at each layer, the set of filter weights is learned based on training data for a specific task. Recall that for a single 2D-convolution (Eq. (1)) a filter (h) in layer (ℓ) and channel (d) "slides" across the domain of the input signal, $x[u, v]$, accumulating and redistributing this signal into the output, $f^{(\ell,d)}[m, n]$.

$$f^{(\ell,d)}[m,n] = \sum_{u=-\infty}^{\infty}\sum_{v=-\infty}^{\infty} x[u,v]h^{(\ell,d)}[m-u,n-v] \tag{1}$$

This sliding process can be thought of as replicating the filter at every spatial location. Replicating the filter in this way allows for the extraction of features regardless of their position, and enables a linear system to be shift-invariant. Additionally, sharing the same set of weights at multiple locations increases the learning efficiency by reducing the number of parameters in the model. Based on the nature of the convolution, CNNs typically require highly regular data, *e.g.* images, which are organized based on a 2D-grid. Also, note that such a convolution (applied to gridded data, *e.g.* images), is not invariant to permutations of the input members *i.e.* pixels. In other words, the spatial distribution of pixels within a filter window is important to capture local features such as edges. Therefore, reordering the input points will result in meaningless output.

This introduces challenges for the application of CNNs to classify irregular, unstructured 3D point cloud data. Given an input set of N 3D-points $X = \{x_1, x_2, x_3, \ldots, x_N\}$, where every point represents a row in a 2D-array, the goal of point cloud classification is to assign every point x_i an object-level label from a set of predefined labels $Y = \{y_1, y_2, y_3, \ldots, y_C\}$, where C is the total number of classes. Since point clouds are not defined on regular grids, and convolutional layers require regular inputs, a modification to either the input or the network architecture is needed. In order to directly operate on point clouds and avoid transforming the data to a different representation (see Section 2.2), we follow (Qi et al., 2017) by adapting convolutional operations to point clouds.

The complete architecture of our network is shown in Fig. 1. The input to the network is an $N \times M$ array of unordered data points, where N is the number of points, and M is the number of features for each point, *i.e.* spatial coordinates and/or spectral information. Shown in Fig. 1 is the simple case where the input data is the raw point cloud X, defined by its spatial coordinates (x, y, z) as columns of the array. The input could optionally be expanded to include any other features, such as spectral information. The first layer of the network applies a 1D-convolution – with a filter width equal to the width of the input vector – across the columns to capture the interactions between coordinates for each row (data point). The output of this layer is a single column where each value corresponds to an individual 3D-point within the set, *i.e.* an $(N \times M)$ input array is transformed to an $(N \times 1)$ output array. This layer operates on each point independently, an advantage which allows for the incorporation of point-wise operations such as scaling, rotation, translation, etc. Since such operations are differentiable, they can be included within the network as a layer and trained in an end-to-end fashion. This concept, first introduced in (Jaderberg et al., 2015), allows the network to automatically align the data into a canonical space and therefore makes it invariant to geometric transformations.

The subsequent convolutional layers perform feature transformation – such as dimensionality reduction or feature expansion – using (1×1) convolutions. We avoid fully-connected layers due to their expensive computational cost. Since the convolution process is a linear operation that performs a weighted sum of its inputs, a non-linear operation – known as the activation function – is needed in order to derive and learn more complex features. Three activation functions are frequently used in the literature: *sigmoid*, hyperbolic tangent (*tanh*), and rectified linear unit (*ReLU*). Sigmoid activation, $\sigma(x) = 1/(1 + e^{-x})$, reduces the input values to the range of [0,1]; this can be thought of as assigning a likelihood value to the input. Similarly, *tanh* activation, $\tanh(x) = 2\sigma(2x) - 1$, maps the input values to the range of $[-1, 1]$; this has the added advantage that the output is zero centered. Finally, *ReLU* activation, $f(x) = \max(0, x)$, applies a simple ramp function. It reduces the likelihood of vanishing gradient, greatly accelerates the convergence rate (Krizhevsky et al., 2012), and involves simpler mathematical operations; therefore it is the most common activation function used in deep convolutional networks. We implement the *ReLU* function in our network as follows:

$$f^{(1)} = \max(0, \langle h, x_i \rangle + b) \tag{2}$$

where the convolution in Eq. (2) is represented now by the dot product $\langle \cdot \rangle$, b is the bias, and $f^{(1)}$ is the first layer's output.

3.2. Batch normalization

Although the *ReLU* activation function has many advantages, it does not enforce a zero-centered distribution of activation values, which is a key factor to improve the gradient flow. One way to adjust this distribution is to change the weight initialization mechanism. Glorot and Bengio (2010) and He et al. (2015) showed that good initialization is the key factor for a successful, convergent network, however, control over the distribution of activations was handled indirectly. For improved control, Ioffe and Szegedy (2015) introduced Batch Normalization (BN) that directly operates on the activation values. An empirical mean and variance of the output (after the convolutional layer and before the non-linearity) are computed during training; these are then used to standardize the output values, *i.e.*

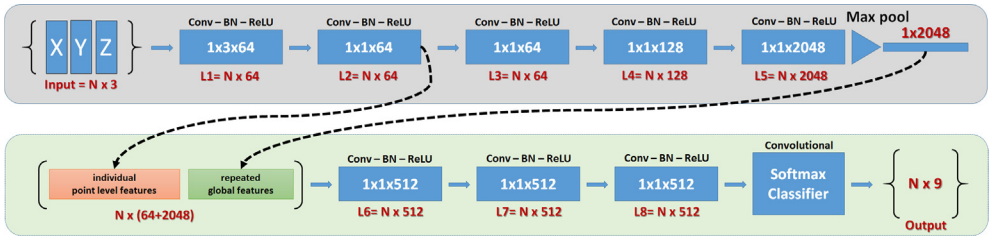

Fig. 1. The basic semantic labeling network takes as input an Nx3 set of points (a point "cloud"), and, in the first stage, passes it through a series of convolutional layers to learn local and global features. In the second stage, concatenated features are passed though (1×1) convolutional layers and then to a softmax classifier to perform semantic classification. Text in white indicates the filter size, while text in red indicates the layer's output shape. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$$\hat{s} = \frac{s - E[s]}{\sqrt{Var[s]}} \tag{3}$$

$$z = \gamma \cdot \hat{s} + \beta \tag{4}$$

where $s = \langle h, x_i \rangle$ is the output after the convolutional layer, and γ and β are the scale and shift parameters learned during training. Setting $\gamma = \sqrt{Var[s]}$, and $\beta = E[s]$ allows the network to recover the identity mapping. BN improves flow through the network, since all values are standardized, and simultaneously reduces the strong dependence on initialization. Also, since it allows for homogeneous distributions throughout the network, it enables higher learning rates, and acts as a form of regularization.

Incorporating these advantages, we initialize weights using the method of Glorot and Bengio (2010) and insert BN layers after every convolutional layer, as shown in Fig. 1. Note that the BN layer functions differently during testing and training. During testing, the mean and the variance are not computed. Instead, a single fixed value for the mean and the variance – found empirically during training using a running average – is used during testing. After integrating BN, the activation function (Eq. (2)) can now be written as follows:

$$f = \max(0, BN(h^T x + b)) \tag{5}$$

Evaluating Eq. (5) multiple times for different values of h allows each layer to capture various aspects of its input. This results in an output array of $(N \times K)$ dimensions, where K is the total number filters used. The size of K per-layer is shown in Fig. 1 (white text). Following the output with a series of convolutional layers, as shown in the upper part of Fig. 1, can be defined mathematically as a sequence of nested operations as follows:

$$f^{(\ell)} = f^{(\ell-1)}(f^{(\ell-2)}(\ldots f^{(1)}(x))) \tag{6}$$

where $f^{(\ell)}$ is defined as in Eq. (5), and ℓ is the layer index.

3.3. Contextual feature learning

During training, we desire a network that can learn both local features (at the point-level) and global features (at the block-level), which provide additional contextual information to support the classification stage. Here, we describe how global features can be extracted using a pooling layer, which simultaneously provides permutation-invariance, *i.e.* the order of the input points does not affect classification results. Contrary to 2D-images, point clouds are unstructured and unordered, and so an appropriate network should respect this original pedigree.

Three options are available: sorting the data as a preprocessing step, using a Recurrent Neural Network (RNN), or incorporating a permutation-agnostic function that aggregates the information from all points. In the first case, the optimal sorting rules are not obvious. In the second case, the point cloud must be considered as a sequential signal, thus requiring costly augmentation of the input with all possible permutations. While some architectures such as long short-term memory (LSTM) and gated recurrent unit (GRU) neural networks can deal with relatively long sequences, it becomes difficult to scale to millions of steps, which is a common size in point clouds.

Given these considerations, we incorporate a permutation-agnostic function as an additional layer. Specifically, pooling layers, commonly used to downsample 2D-images, work perfectly for this purpose. Instead of downsampling the point set, we pool values across all points to form a single global signature that represents the input point cloud. Such signature could be used directly to label the whole set. However, recall that for the task of semantic labeling, a label for each 3D-point is desired. Therefore, both local *and* global features are needed to describe the point and capture contextual information within the input set.

In this network, local features are obtained from the second convolutional layer $f^{(2)}$ with an output shape of $(N \times 64)$, *i.e.* this represents each 3D-point with a 64D-vector. On the other hand, a global signature for the point set is derived from the output of the fifth convolutional layer with dimensions of $(N \times 2048)$. This serves as input to the global feature extraction function, specifically, a *max-pooling* layer, which aggregates the features across all points and produce a signature with a shape of (1×2048) as follows:

$$g = \max_{row}(f^{(5)}) \tag{7}$$

where g is the global feature vector, $f^{(5)}$ is the output at the 5^{th} layer, and *row* indicates that the aggregation is applied vertically across the rows, *i.e.* points.

3.4. Inference

The global feature vector is concatenated with the point level features, yielding a per-point vector that contains both local and contextual information necessary for point-wise labeling. This concatenated feature vector is then passed to a series of feature transformation layers (1×1 convolutions) and finally to a softmax classifier as shows in the lower part of Fig. 1. We use the cross-entropy cost function to train the network. Cross-entropy is a special case of the general Kullback-Leibler (KL)-divergence D_{KL}, which measures how the ground-truth probability distribution p diverges from the output probability distribution q, *i.e.*:

$$\begin{aligned} D_{\mathrm{KL}}(p||q) &= \sum_i p(x_i) \cdot (\log p(x_i) - \log q(x_i)) \\ &= -\sum_i p(x_i) \log q(x_i) - \sum_i p(x_i) \log \tfrac{1}{p(x_i)} \\ &= H(p,q) - H(p) \end{aligned} \tag{8}$$

If the discrete distribution p is zero everywhere except a single location with maximum probability, the expression is reduced to the cross-entropy $H(p,q)$. In our case, p is the ground truth distribution represented by one-hot vectors encoding the label per-point, while q is the output of the softmax layer, representing the normalized class probabilities. Here, each individual class probability is calculated as follows:

$$P(y_i|f^{(\ell)}(x_i);W) = \frac{e^{f_{y_i}}}{\sum_c e^{f_c}} \quad (9)$$

where $f^{(\ell)}$ is the last convolutional layer and the input to the softmax layer, y_i is the current correct label for the input x_i, W is the weight matrix for the softmax classifier, and f_{y_i} is the unnormalized log probability of the output node indexed by y_i, and c is the class (output) index. Fig. 2 illustrates the methodological concepts presented in Section 3.3 ans Section 3.4.

4. Evaluation

This section describes the methodology used to evaluate the performance of our approach. In Section 4.1 we describe the datasets used. In Section 4.2 we describe preprocessing steps. In Section 4.3 we outline training parameters. In Section 4.4 we present our results on 3D-point clouds along with various experiments showing qualitative performance of our method when applied to different testing cases. In Section 4.5 we analyze the effects of the input feature selection. Finally, in Section 4.6 we show the extension of our network to handle the fusion of LiDAR and spectral data in 2D-semantic segmentation tasks.

4.1. Dataset

For this paper, we use data provided by both the ISPRS 2D[2] and 3D-Semantic Labeling Contest, as part of the urban classification and 3D-reconstruction benchmark. Both airborne LiDAR data and corresponding georeferenced IR-R-G imagery are provided from Vaihingen, Germany (Fig. 3). The georeferenced image of the whole area has a ground sampling distance of 8 cm and a size of 20250 × 21300 pixels. For the 3D contest, 9 classes have been defined, including *Powerline*, *Low vegetation*, *Impervious surfaces*, *Cars*, *Fence/Hedge*, *Roof*, *Facade*, *Shrub*, and *Tree*. The area is subdivided into two regions, for training and testing. Each region includes a text file that contains the LiDAR-derived (x, y, z) coordinates, backscattered intensity, and return count information, acquired using a Leica ALS50 system at a mean height of 500 m above ground. The point density is approximately 4 points/m^2 with a total of 753,859 training points, and 411,721 testing points. The test area is within the center of Vaihingen city, and is characterized by dense, complex buildings. It covers an area of 389 m × 419 m. The training area, on the other hand, is mostly residential, with detached houses and high rise buildings. It covers an area of 399 m × 421 m.

4.2. Preprocessing

Two preprocessing methods were employed to obtain our desired input. First, spectral information was attributed to each (x, y, z) triplet in the point cloud by applying a bilinear interpolation using the georeferenced IR-R-G imagery as shown in Fig. 3. Note that in the case of stereo-derived point clouds from optical imagery, this spectral information is inherently available, and would not need to be obtained separately. Next, we normalize the z values in the point cloud by subtracting a Digital Terrain Model (DTM), generated using LAStools,[3] in order to obtain height-above-ground. Then, to train our deep learning method from scratch, a sufficiently large amount of labeled data are required; however, a single and small training scene is provided. We solve this issue by subdividing our training and testing regions into smaller 3D-blocks. Such blocks are allowed to overlap (unlike PointNet), thus increasing the quantity of data available, and robustness by allowing overlapped points to be part of different blocks. Each point within the block is represented by a 9D-vector, containing the per-block centered coordinates (X, Y, Z), spectral data (IR, R, G), and normalized coordinates (x, y, z) to the full extent of the scene. Note that since our method is fully convolutional, the number of points per-block can vary during training and testing. This contribution resolves the typical challenge of working with point clouds of varying density. While we test using different densities, we sample fixed number of points per-block during training for debugging and batch training purposes.

To sample points from each block, we randomly choose 4096 points during training without replacement. If the number of points per-block is lower than the desired number of samples, random points within the block are repeated. However, if the number of points per-block is lower than 10 points, the block is ignored. To learn objects with different scales, *e.g.* building vs. car, one can train separate networks, with each network trained using a down-sampled version of the point cloud, as in Maturana and Scherer (2015). However, this is not practical, as it introduces an additional and unnecessary computational overhead. Instead, current deep learning approaches – *e.g.* a single network with high capacity – should be able handle multi-scale objects in an end-to-end fashion given appropriate inputs. To handle different resolutions, we generate blocks with different sizes and train our network using all scales simultaneously. Blocks of size 2 m × 2 m, 5 m × 5 m, and 10 m × 10 m work well in our case given the scale of the features of interest *e.g.* cars and roofs. Our final result during testing is the average of all three scales. While splitting the data into blocks with different sizes increases the number of training samples, robustness to noise and orientation could be further improved by augmenting the training data with modified versions of the original training data. We augment the training data by randomly rotating points around the z-axis (*i.e.* to adjust their geographic orientation), and jittering the coordinates. Jitter is added by applying an additive noise, sampled from a zero-mean normal distribution with $\sigma = 0.08$m for the x, y coordinates, and $\sigma = 0.04$m for the z coordinates. Next we clip the values to a maximum horizontal and vertical jitter of 30 cm and 15 cm respectively. The values were chosen empirically to add sufficient noise while preserving the relative differences between various objects. Rotation and jitter are applied before splitting the data into blocks. However, we also apply jitter during the per-block sampling process, in order to avoid duplicating points.

4.3. Training parameters

To asses and monitor the performance of our model during training, a validation set is needed. Instead of taking samples from the training set directly, we desire a validation set that is as different as possible to the training data. We address this by splitting the original training data before augmentation into a new training and validation subsets using stratified splits to preserve the distribution of classes in both sets. Class-distributions within each set are then balanced by repeating under-represented classes until they match the class with the highest number of members, resulting in a uniform distribution of classes. We then augment the new training data by applying the jitter and rotations as described in Section 4.2.

After splitting both the training and validation sets into blocks, we train using the augmented data only as input to the network

[2] https://goo.gl/mdvbwM.
[3] http://www.lastools.org/.

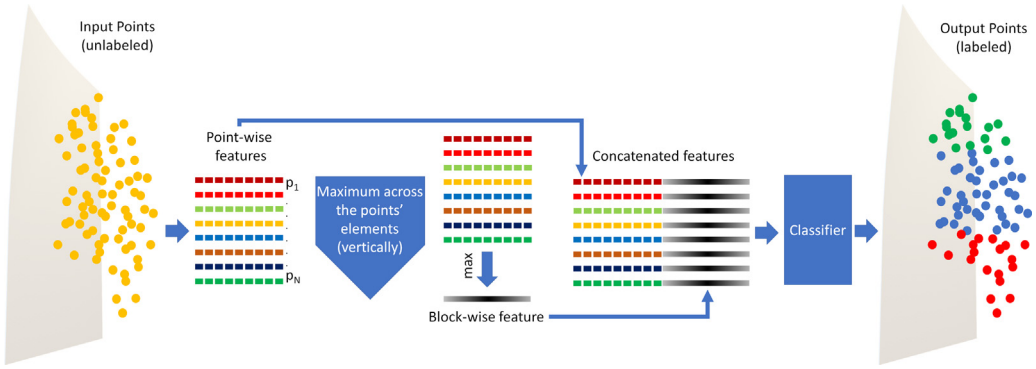

Fig. 2. A graphical illustration of the methodological concepts presented in Sections 3.3 and 3.4. The Point-wise features are the output of L2 in Fig. 1. For more details about the classifier, please see the bottom part of Fig. 1.

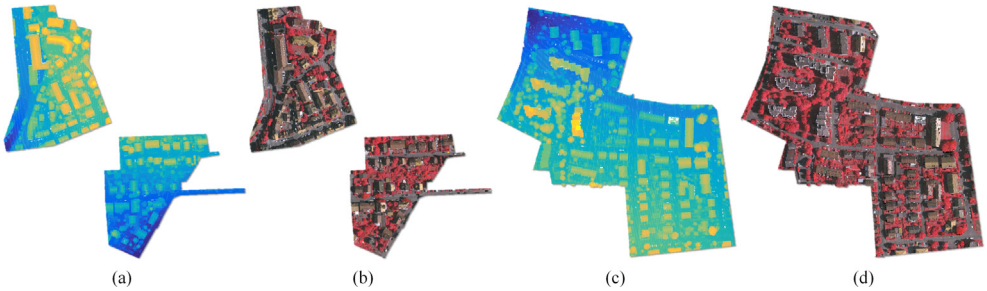

Fig. 3. From left to right: point cloud (a) color-coded by height, and (b) by spectral information (IR, R, G) for the test area; and point cloud (c) color-coded by height, and (d) by spectral information (IR, R, G) for the training data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

shown in Fig. 1. We use the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate 0.001, a momentum of 0.9 and a batch size 32. The learning rate (lr) is iteratively reduced based on the current number of epochs, according to:

$$lr_{new} = lr_{initial} \times \left(1.0 - \frac{epoch_{current}}{epoch_{total}}\right) \tag{10}$$

This proceeds for a total of 30 epochs, *i.e.* $epoch_{total} = 30$. We monitor the progress of the validation loss and save the weights if the loss improves. If the loss does not improve after 3 epochs, training is terminated and the weights with the best validation loss are used for testing. Training our network takes around 12–18 h to converge using a Tesla p40 GPU and Keras (Chollet, 2015) with the Tensorflow backend. The feed forward time during testing is 3.7 s for the full scene (~412k points). Fig. 4 shows the loss and overall accuracy progress during training and validation.

4.4. Classification results

Classification results are based on the data provided by the ISPRS 3D Semantic Labeling Contest (see Section 4.1). During testing, we split the data into blocks similar to the training stage, in order to recover objects at different scales. The block sizes, overlap, and number of points per-block are reported in Table 2. We forward pass the data to output a 9D-vector per-point indicating the probability of each point belonging to each of the nine categories. Since we use a fixed number of points for each block size, some points from the original data may remain unclassified due to sampling, while others may be duplicated due to repetition. Therefore, we interpolate the output results to classify the original set of points. We use nearest neighbor interpolation on each class probability separately, to generate a total of nine class probability maps. A label corresponding to the index of the highest probability from the nine maps is assigned to each point, indicating the classification label. Quantitative performance metrics are based on the ISPRS contest: Per-class accuracy, precision/correctness, recall/completeness, and F1-score, in addition to the overall accuracy. Although it is possible to submit results while excluding some classes, we chose to evaluate our method on all available classes. The confusion matrix in Table 1 shows resulting per-class accuracies. Fig. 5 shows the corresponding classification map, along with the error map provided by the contest organizers. As shown in Table 1, the proposed method performs well on *impervious surfaces* and *roof* classes. The worst performance is for the *fence/hedge* and *powerline* classes, which according to Table 1 is due to the confusion between closely-related classes. For example, *powerline* is mainly confused with *roof*, and *fence/hedge* is confused with *shrub*, both of which have similar topological and spectral characteristics. Likewise, the accuracy of *low vegetation* is affected by the presence of *impervious surfaces* and *shrubs*. *Shrub* appears to be causing most of the confusion. This is likely due to the fact that the spectral information is similar among the vegetation classes *low vegetation*, *shrub*, and *trees*. While height information may improve the results, the presence of classes with similar heights such as *car* and *fence/hedge* make this differentiation challenging.

To evaluate our performance against others, we compare our method to all submitted ISPRS contest results (both published and unpublished) at the time of paper submission. Since some submissions are unpublished we will review them briefly using the available information on the contest website[4]; We refer to the submitted methods according to names posted on the contest website.

[4] https://goo.gl/6iTj6W.

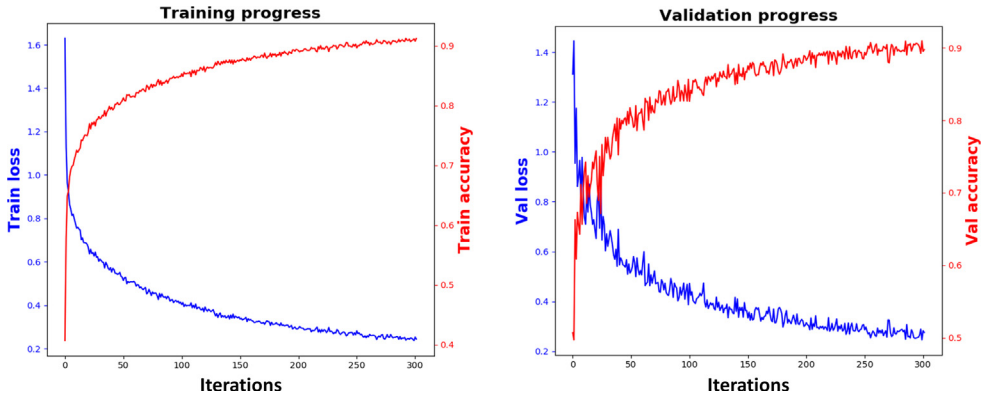

Fig. 4. The loss and overall accuracy progress for training (left) and validation (right). Every 10 iterations correspond to a single epoch.

Table 1
Confusion matrix showing the per-class accuracy using our deep learning framework. The overall accuracy (not shown) is **81.6**%.

Classes	power	low_veg	imp_surf	car	fence_hedge	roof	fac	shrub	tree
power	**29.8**	00.0	00.2	00.0	00.0	54.2	00.7	00.2	15.0
low_veg	00.0	**69.8**	10.5	00.5	00.2	00.5	00.6	16.1	01.9
imp_surf	00.0	05.2	**93.6**	00.2	00.0	00.2	00.1	00.7	00.0
car	00.0	05.0	01.2	**77.0**	00.2	02.6	00.8	12.9	00.3
fence_hedge	00.0	05.9	01.4	01.7	**10.4**	01.5	00.6	68.5	10.0
roof	00.1	00.5	00.4	00.0	00.0	**92.9**	02.8	02.3	00.9
fac	00.2	04.3	00.8	00.9	00.1	23.3	**47.4**	19.9	03.1
shrub	00.0	07.9	00.5	01.0	00.5	02.6	02.0	**73.4**	12.0
tree	00.0	00.8	00.0	00.2	00.1	01.2	01.3	17.1	**79.3**
Precision/Correctness	50.4	88.0	89.6	70.1	66.5	95.2	51.4	33.4	86.0
Recall/Completeness	29.8	69.8	93.6	77.0	10.4	92.9	47.4	73.4	79.3
F1 Score	37.5	77.9	91.5	73.4	18.0	94.0	49.3	45.9	82.5

Table 2
The block sizes and the corresponding overlap and number of points during testing.

Size	Overlap	# Points
2 m × 2 m	1 m	1024
5 m × 5 m	2 m	3072
10 m × 10 m	2 m	4096

Interested readers are encouraged to review the website for further details.

The **IIS_7**[5] method used spectral and geometrical features, combined with a segmentation based on supervoxels and color-based region growing. In contrast to **IIS_7**, we don't handcraft geometrical features or utilize the spectral information to segment the point cloud into similar coherent regions before classification. The **UM**[6] method used a One-vs-One classifier based on handcrafted features, including point attributes such number of returns, textural properties, and geometrical attributes. The **HM_1**[7] method utilized a CRF with RF classifier and contrast-sensitive Potts models, along with 2D-and 3D-geometrical features. The **WhuY3** deep learning method (Yang et al., 2017) used a window around each 3D point and divided such window into 128 × 128 cells. Five features (planarity, sphericity, intensity, variance of angles with respect to the normal, and height above ground) were estimated for each point in a cell. The whole cell was then transformed into a single pixel to for 128 × 128 image. A convolutional network was then used to classify the images. The **K_LDA** (Blomley et al., 2016) method used covariance features at multiple scales. Finally, the **LUH**[8] method used a two-layer hierarchical CRF that explicitly defines contextual relationships and utilizes voxel cloud connectivity segmentation, along with handcrafted features such as Fast Point Feature Histograms (FPFH).

Per-class accuracy, and overall accuracy (OA) for each submission, including ours, are shown in Table 3. As seen in Table 3, our method ranks second overall, sharing an overall accuracy of 81.6% with **LUH**. The method with the highest overall accuracy is **WhuY3**, achieving 82.3%. However, **WhuY3** achieved only one per-class highest accuracy score, as opposed to two for our method: the *car* and *shrub* classes. Likewise, the **IIS_7** method, which achieved the most (3) highest scores on a per-class basis, and use spectral data as well, only ranked eighth overall with an overall accuracy of 76.2%.

We also evaluated our results using the F1-score, which is generally considered to be a better comparative metric when an uneven class distribution exists and/or when the costs of false positives and false negatives are very different. Table 4 compares our method to others using the F1-score. Our method performed well across all classes except for the *fence/hedge* class. Other methods demonstrated similarly poor results on this same class. While **LUH** did score the highest on the *roof* and *tree* classes, their scores were only marginally better than ours, 0.2% and 0.6%, respectively. Their higher performance for the *fence/hedge* and *powerline* classes did, however, allow them to achieve the highest F1-score of 68.39%, with **HM_1** ranking second with a score of 66.39%. Our presented technique did ranked third with a score of 63.33%, surpassing **WhuY3** which scored the highest on overall accuracy (Table 3). Fig. 6 shows a qualitative visual comparison between the top 4 methods in Table 4.

The scores in both Tables should be thought of in context of the algorithm complexity. In other words, minimal accuracy gains may not be worth the added computational overhead required by other methods. The **LUH** method, for example, fares well in both overall,

[5] https://goo.gl/zepC6d.
[6] https://goo.gl/uzZFcs.
[7] https://goo.gl/d7AmXY.
[8] https://goo.gl/fGzrf3.

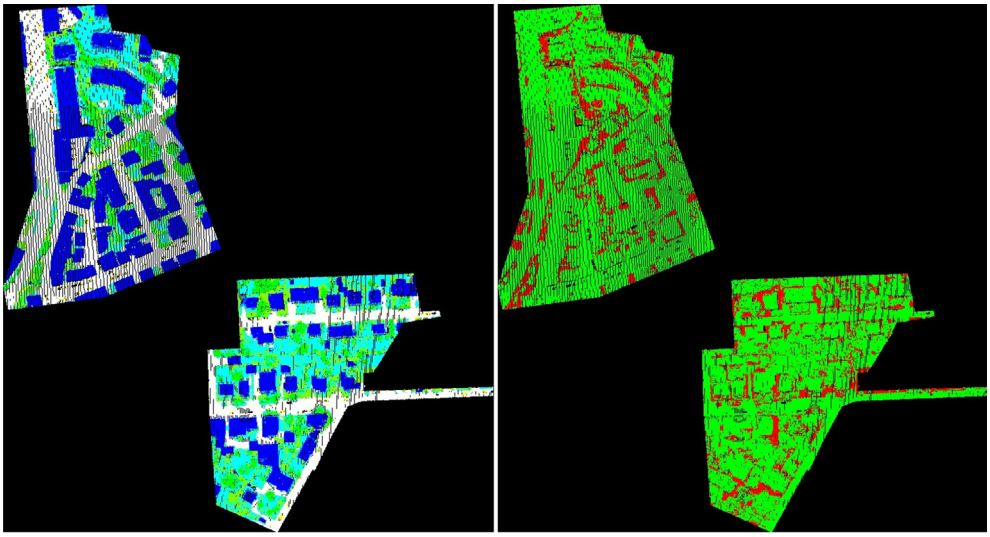

Fig. 5. The left image shows the classification map, while the right image shows the corresponding error map. The results were provided by the contest organizers.

and per-class, accuracy and F1-scores. However, this method uses two independent CRFs with handcrafted features, and segmentation methods that force points within a segment to share the same label. Likewise, **HM_1** uses a variety of contrast-sensitive Potts models, which may help preserve edges during segmentation, but adds NP-hard components to the problem (Boykov et al., 2001). As a result, while smoother results may be achieved for small data sets, this comes at the cost of slow run-time performance (see Section 2.1) and scalability limitations for massive data sets. In contrast, we utilize a series of simple 1D-convolutions that operate directly on the point cloud, without engineering additional features, or requiring structured representations such as segments. Instead, point-wise features (*e.g.* spatial coordinates and/or spectral information), and per-block contextual descriptors are learned in a straightforward, end-to-end fashion. Our average inference time is 3.7 s for a point cloud with N approximately equal to 412k.

4.5. Effect of individual features

Our 1D convolutional network has flexible input feature requirements, and can consume directly (i) spatial coordinates only, *i.e.* a 3D point cloud, (ii) spatial coordinates and spectral information, or (iii) spectral information only. Moreover, 3D spatial coordinates (x, y, z) may optionally be normalized to remove the effects of terrain, providing (x, y, height-above-ground). Finally, models can be trained at different scales by adjusting the block size. In this section we provide two experiments to analyze the impact of feature selection and digital terrain model on model performance for various use-cases.

In the first experiment, we investigate the effect of the input feature selection, *i.e.* spatial and/or spectral information, by training our model based on three sets of input data. The first model is trained using 3D coordinates only. The second model is trained using spectral information only. The third model is trained using both 3D-coordinates and spectral information (IR-R-G) for each point; this is the best-performing network, which was evaluated in more detail in Section 4.4. Results are shown in Fig. 7 for multiple scales, *i.e.* block sizes of 2 m × 2 m, 5 m × 5 m, and 10 m × 10 m, and, in column 4, the average result across all scales.

When using the 3D-coordinates only (first row), the larger scale performs better than the smaller scale. On the other hand, when using the spectral information only (second row), the smaller scale performs better than the larger scale. This result is interesting since it shows the effect of global features at multiple scales. For example, when using the spectral data, smaller scales will generally include similar points as opposed to larger scales. Therefore, the global features will accurately describe the group (block) of points at a smaller scale. In contrast, the global features will not sufficiently describe structures when using only the 3D-coordinates at a smaller scale. In this case, a larger scale is needed to capture structural information that can help distinguish between 3D objects.

This suggests that combining both features could improve the results. The result submitted to the ISPRS 3D Semantic Labeling Contest (bottom right corner) used the average of 3 block sizes, trained on both (x, y, z) point coordinates and spectral information. The red circle highlights how using all features helped to correctly classify a troublesome *low vegetation* region. Likewise, the red box highlights how the spectral data, in the absence of 3D coordinates, tends to classify all vegetation as *tree*, while the 3D-coordinates, in the absence of spectral data tends to confuse *impervious surfaces* with *low vegetation*. See Fig. 5 for reference regrading correctly classified regions in our submission.

In the second experiment, we investigated the effect of normalizing the z-coordinates to height-above-ground, based on the DTM model. This was done both in the absence (Fig. 8) and presence (Fig. 9) of spectral information, which may not always be available. Fig. 8 shows that in the absence of spectral information, normalizing the 3D coordinates to height-above-ground provides a cleaner and less-fragmented classification at all scales, including the average.

As shown in Fig. 9, the best classification results are achieved when spectral information is available, and after obtaining height-above-ground using a DTM. Close scrutiny of the regions marked in red shows that terrain-normalized input points improves the classification, especially for parts of roofs. However, in general, when spectral information is available, the results without using a DTM are still reasonable. This is exciting, as it opens the door to more streamlined point cloud exploitation workflows that can directly ingest 3D data in its original form. Furthermore, this may enable new techniques for generating, at scale, more precise DTMs based on the spectral content inherent in stereo-derived point clouds (Tapper, 2016).

Table 3
The per-class accuracy for each method and the corresponding Overall Accuracy (OA).

Methods	power	low_veg	imp_surf	car	fence_hedge	roof	fac	shrub	tree	OA
Ours	29.8	69.8	93.6	**77.0**	10.4	92.9	47.4	**73.4**	79.3	**81.6**
IIS_7	40.8	49.9	**96.5**	46.7	**39.5**	**96.2**	–	52.0	68.8	76.2
UM	33.3	**79.5**	90.3	32.5	02.9	90.5	43.7	43.3	**85.2**	**80.8**
HM_1	**82.8**	65.9	**94.2**	67.1	25.2	91.5	49.0	**62.7**	**82.6**	80.5
WhuY3	0.247	**81.8**	91.9	**69.3**	14.7	**95.4**	40.9	38.2	78.5	**82.3**
K_LDA	**89.3**	12.4	47.6	28.9	20.4	80.7	**51.3**	38.4	72.8	50.2
LUH	53.2	72.7	90.4	63.3	**25.9**	91.3	**60.9**	**73.4**	79.1	**81.6**

Table 4
The F1-scores per-class for each method and the corresponding average value.

Methods	power	low_veg	imp_surf	car	fence_hedge	roof	fac	shrub	tree	Avg. F1
Ours	37.5	77.9	**91.5**	**73.4**	18.0	**94.0**	49.3	45.9	**82.5**	**63.33**
IIS_7	54.4	65.2	85.0	57.9	28.9	90.9	–	39.5	75.6	55.27
UM	46.1	**79.0**	89.1	47.7	05.2	92.0	52.7	40.9	77.9	58.96
HM_1	**69.8**	73.8	**91.5**	58.2	**29.9**	91.6	**54.7**	**47.8**	80.2	**66.39**
WhuY3	37.1	**81.4**	90.1	63.4	23.9	93.4	47.5	39.9	78.0	61.63
K_LDA	05.9	20.1	61.0	30.1	16.0	60.7	42.8	32.5	64.2	37.03
LUH	**59.6**	77.5	**91.1**	**73.1**	**34.0**	**94.2**	**56.3**	**46.6**	**83.1**	**68.39**

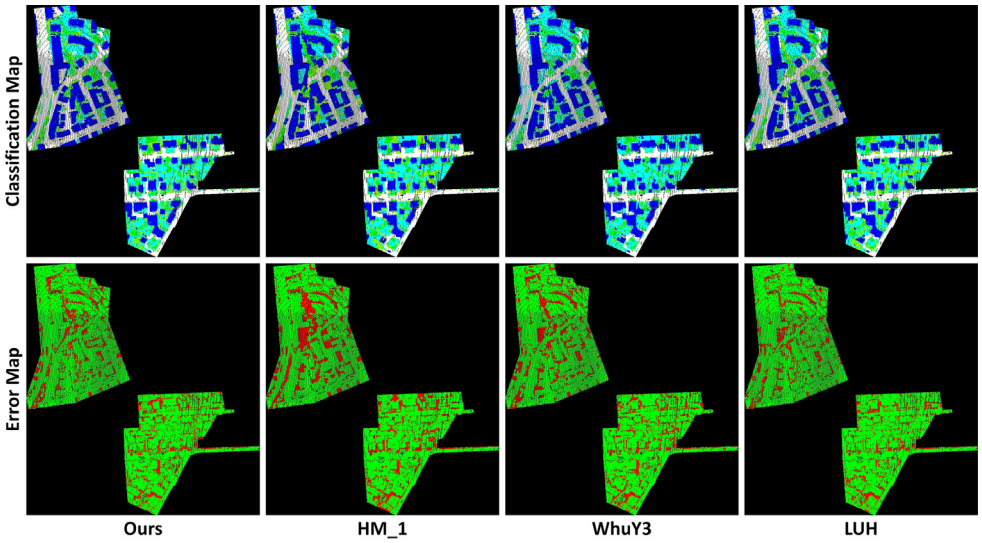

Fig. 6. A qualitative comparison showing the output classification maps (top row) and the error maps (bottom row) for the top 4 methods (columns) in Table 4.

4.6. Extension to 2D semantic labeling

A primary contributions of our method is the compact, 1D CNN architecture. This complements the input data characteristics of point clouds in ways that traditional 2D CNNs do not. In other words, deep learning methods that rely on mapping features to 2D image-like maps, *i.e.* DSM or density maps (Su et al., 2015; Caltagirone et al., 2017; Liu et al., 2017) don't take into account the increased complexity when adding new features. In such a case, adding a new feature involves concatenating the data with a new (full) 2D-channel, subsequently requiring an additional set of 2D-filters and greatly increasing the amount memory required. On the other hand, our method operates directly on the point cloud, representing each point with a 1D-vector. Adding a new feature only requires appending each per-point vector with a single value; this increases the width of the 1D-convolutions by only one element in the first layer and does not modify the rest of the network. This advantage provides an elegant solution to the 3D multi-sensor fusion problem, allowing us to readily combine complementary information about the scene from different modalities for semantic segmentation.

Additionally, we show that this 1D architecture can be easily extended to the traditional task of 2D-image labeling. To handle 2D images, a preprocessing step simply restructures the data from a raster representation to a 2D-array where each row represents a 3D-point vector. The spatial position of each point corresponds to the 2D-pixel coordinates, while the height value corresponds to the digital counts in the DSM image. The spectral information of each point is derived from the corresponding image's digital numbers. We then train our model as described previously, excluding the data augmentation or multi-scale stages, due to the high resolution nature of the images which provided sufficient training samples.

We qualitatively evaluated our method against two relevant submissions from the Potsdam 2D-Semantic Labeling Contest.[9] The first method, **RIT_L7** (Liu et al., 2017) used a CRF model with an FCN and a logistic regression classifier to fuse the normalized DSM and the spectral information, scoring an overall accuracy of 88.4%. The **RIT_L7** method was chosen for comparison since it uses

[9] https://goo.gl/rN3Cge.

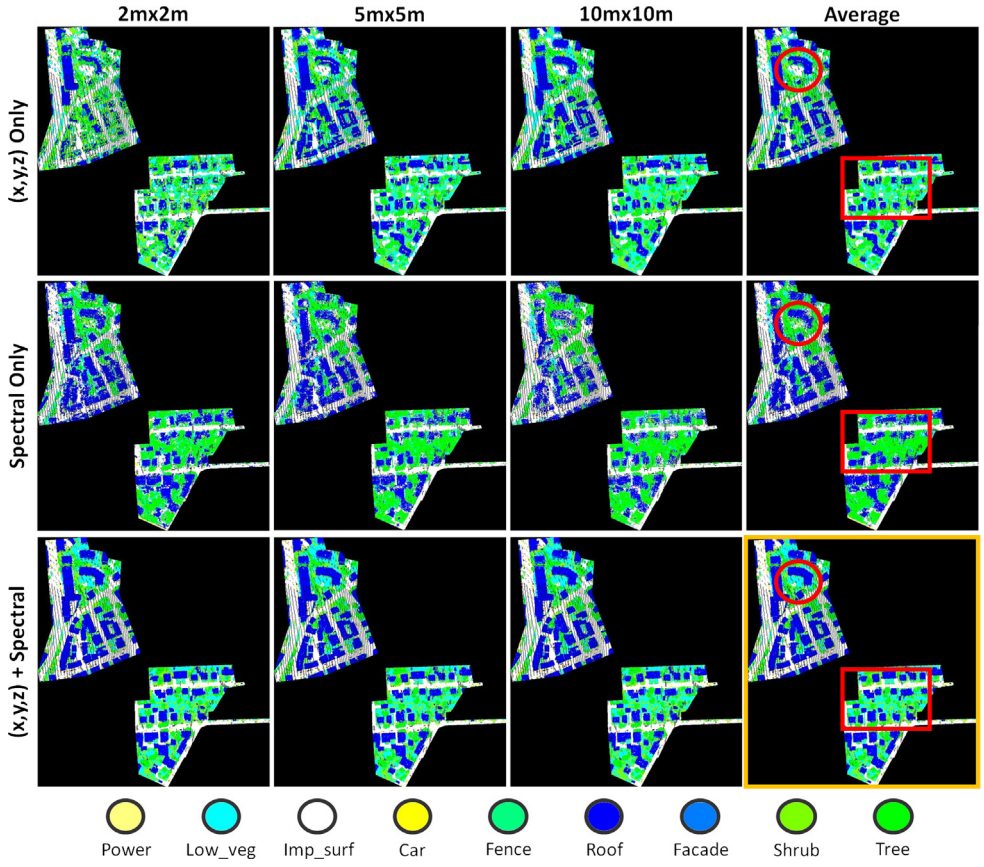

Fig. 7. Results matrix showing the effect of training the network using different input features (rows), and at different block sizes (columns). The submitted result to the ISPRS 3D Semantic Labeling Contest is shown in the bottom right. Red markers indicate regions of comparison. Class color keys are shown at the bottom. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

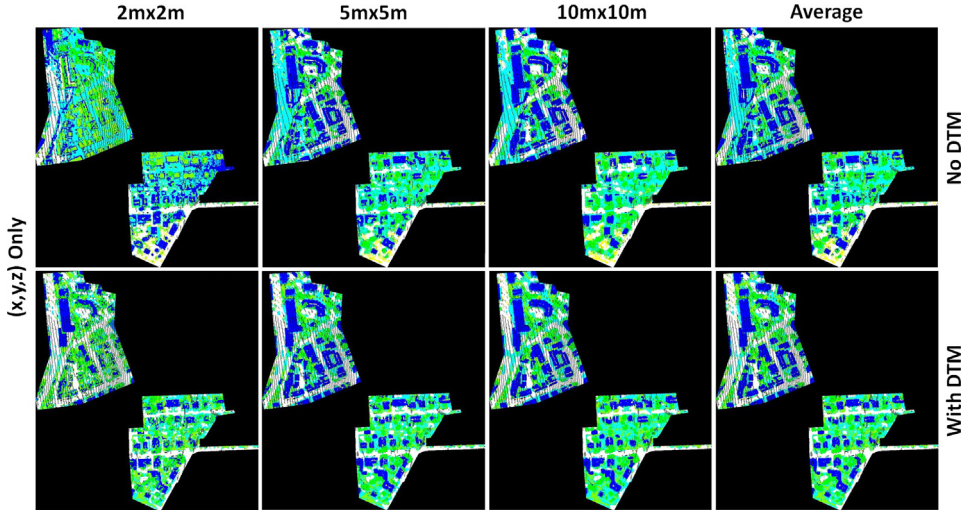

Fig. 8. A comparison between training with (bottom) and without (top) a digital terrain model using only 3D-coordinates for block sizes.

CRF to explicitly utilize contextual information. The second method (unpublished), **CASIA2**, fine-tuned Resnet101 (He et al., 2016) and used only the spectral data, scoring an overall accuracy of 91.1%. The **CASIA2** method was chosen for comparison since it relies on a very large and computationally intensive Network, in contrast to our very compact network. For example, our network has about 1.9 M parameters, while Resnet101 has about 44.5 M parameters.

Fig. 10 compares our output semantic labeling results against the **RIT_L7** and **CASIA2** methods, for two example 2D images. The circled areas in the upper image show that our method was able to correctly (see the corresponding spectral image) classify challenging low-vegetation regions, which were incorrectly classified as clutter by **RIT_L7**, and had residuals of the building class in **CASIA2**. Likewise, the circled regions in the lower image shows that both our method and the **RIT_L7** method produced good

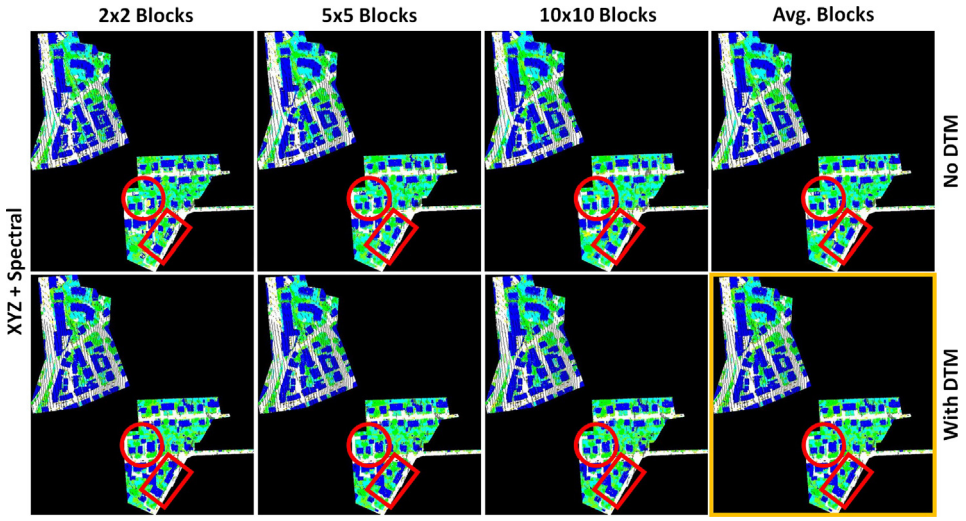

Fig. 9. A comparison between training with and without a digital terrain model using 3D-coordinates and spectral data. Regions marked with red highlight differences.

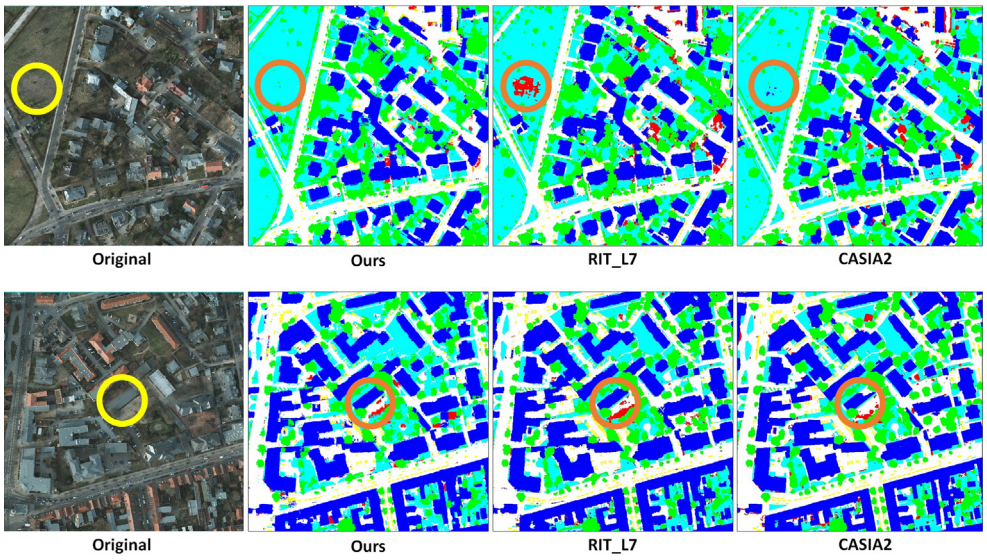

Fig. 10. A qualitative comparison between our method to two submissions that uses 2D-deep networks. The yellow circles shows the original regions of interest, while the brown circles mark the corresponding regions in the classification maps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results when classifying the center building. The **CASIA2** method missed the top right corner of the building; this is likely due to a lack of consideration for height information. Including height information as another channel in a very large model such as Resnet101 is not a trivial task due to three-channel design. Also, even if height information were included as a fourth channel, finetuning would not be possible and training would be infeasible given the limited number of images provided in the contest. This highlights the advantages of our compact model: adding another feature is as simple as adding a single value per-point. And, given the relatively few number of parameters, a simple data augmentation is sufficient to train our network.

5. Conclusions

In this paper we present a deep learning framework to semantically label 3D point clouds with spectral information. Our compact, 1D fully convolutional architecture directly consumes unstructured and unordered point clouds without relying on costly engineered features or 2D image transformations. We achieved near state-of-the-art results using only the 3D-coordinates and their corresponding spectral information. By design, our network can consume regions with varying densities, and is able to learn point-wise and block-wise features in an end-to-end fashion. Furthermore, our model is flexible, and can be readily extended to 2D semantic segmentation. Also, our experiments showed promising results when classifying unnormalized points. Given the compact, end-to-end framework, and fast testing time, our model has the potential to scale to much larger datasets, including those derived from optical satellite imagery. Future work includes addressing general drawbacks such as reducing the time and the memory needed for preparing the input data blocks, as well as improving the method to learn neighborhood features along with the point-wise and block-wise features. Also, extending the model to operate on optically derived point clouds, improve the performance with respect to unnormalized points, and investigate more fine-grained classes.

Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science.

References

Audebert, N., Saux, B.L., Lefèvre, S., 21–23 November 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Proceedings of the Asian Conference on Computer Vision (ACCV). Taipei, Taiwan.

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. In: ISPRS International Archives of Photogrammetry and Remote Sensing, vol. XXXIII-Part B4/1. pp. 111–118.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bai, S., Bai, X., Zhou, Z., Zhang, Z., Jan Latecki, L., 2016. GIFT: a real-time and scalable 3D shape search engine. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5023–5032.

Blomley, R., Jutzi, B., Weinmann, M., September 2016. 3D semantic labeling of ALS point clouds by exploiting multi-scale, multi-type neighborhoods for feature extraction. In: Proceedings of the International Conference on Geographic Object-Based Image Analysis (GEOBIA). Enschede, The Netherlands, pp. 1–8.

Boulch, A., Saux, B.L., Audebert, N., April 2017. Unstructured point cloud semantic labeling using deep segmentation networks. In: Proceedings of the Eurographics Workshop on 3D Object Retrieval, vol. 2. pp. 17–24.

Boykov, Y., Veksler, O., Zabih, R., 2001. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23 (11), 1222–1239.

Caltagirone, L., Scheidegger, S., Svensson, L., Wahde, M., June 2017. Fast LIDAR-based road detection using fully convolutional neural networks. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). pp. 1019–1024.

Chehata, N., Guo, L., Mallet, C., 2009. Airborne lidar feature selection for urban classification using random forests. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVIII-Part 3.

Chollet, F., et al., 2015. Keras. <https://github.com/fchollet/keras>.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2013. Vision meets robotics: the KITTI dataset. Int. J. Robot. Res. 32 (11), 1231–1237.

Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 580–587.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, vol. 9. pp. 249–256.

Golovinskiy, A., Kim, V.G., Funkhouser, T., September 2009. Shape-based recognition of 3D point clouds in urban environments. In: Proceedings of the 12th International Conference on Computer Vision (ICCV). pp. 2154–2161.

Grilli, E., Menna, F., Remondino, F., 2017. A review of point clouds segmentation and classification algorithms. In: ISPRS International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLII-2/W3. Nafplio, Greece, pp. 339–344.

Haala, N., Brenner, C., Anders, K.-H., 1998. 3D urban GIS from laser altimeter and 2D map data. In: ISPRS International Archives of Photogrammetry, Remote Sensing & Spatial Information Sciences, vol. 32. pp. 339–346.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. SEMANTIC3D.NET: a new large-scale point cloud classification benchmark. In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-1-W1. pp. 91–98.

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

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

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.

Huang, J., You, S., 2016. Point cloud labeling using 3D convolutional neural network. In: Proceedings of the 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 2670–2675.

Hug, C., Wehr, A., 1997. Detecting and identifying topographic objects in imaging laser altimeter data. In: ISPRS International Archives of Photogrammetry and Remote Sensing, vol. 32, Part 3–4W2. pp. 19–26.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37. pp. 448–456.

Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems 28. pp. 2017–2025.

Jochem, A., Höfle, B., Hollaus, M., Rutzinger, M., September 2009. Object detection in airborne LIDAR data for improved solar radiation modeling in urban areas. In: International Archives of Photogrammetry, Remote Sensing, and Spatial Information Sciences, vol. 38, Part 3/W8. Paris, France, pp. 1–6.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: International Conference on Learning Representations.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25. pp. 1097–1105.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1, 541–551.

Li, B., 2017. 3D fully convolutional network for vehicle detection in point cloud. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS).

Lin, C.-H., Chen, J.-Y., Su, P.-L., Chen, C.-H., 2014a. Eigen-feature analysis of weighted covariance matrices for LiDAR point cloud classification. ISPRS J. Photogramm. Remote Sens. 94, 70–79.

Lin, M., Chen, Q., Yan, S., 2014b. Network in network. In: International Conference on Learning Representations (ICLR).

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

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

Mallet, C., 2010. Analysis of Full-waveform Lidar Data for Urban Area Mapping (Ph. D. thesis). Télécom ParisTech.

Maturana, D., Scherer, S., 2015. VoxDet: a 3D convolutional neural network for real-time object recognition. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 922–928.

Moussa, A., El-Sheimy, N., September 2010. Automatic classification and 3D modeling of lidar data. In: Proceedings of the ISPRS Commission III symposium, vol. 38. ISPRS, Saint-Mand, France, pp. 155–159.

Niemeyer, J., Rottensteiner, F., Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. ISPRS J. Photogramm. Remote Sens. 87, 152–165.

Niemeyer, J., Rottensteiner, F., Soergel, U., Heipke, C., July 2016. Hierarchical higher order CRF for the classification of airborne LIDAR point clouds in urban areas. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLI-B3. Czech Republic, pp. 655–662.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Qi, C.R., Su, H., Nießner, M., Dai, A., Yan, M., Guibas, L.J., 2016. Volumetric and multi-view CNNs for object classification on 3D data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5648–5656.

Ramiya, A., Nidamanuri, R.R., Krishnan, R., December 2014. Semantic labelling of urban point cloud data. In: International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XL-8. Hyderbad, India, pp. 907–911.

Ryoo, M.S., Rothrock, B., Matthies, L.H., 2015. Pooled motion features for first-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 896–904.

Savva, M., Yu, F., Su, H., Aono, M., Chen, B., Cohen-Or, D., Deng, W., Su, H., Bai, S., Bai, X., et al., 2016. SHREC16 track: large-scale 3d shape retrieval from ShapeNet core55. In: Proceedings of the Eurographics Workshop on 3D Object Retrieval. pp. 89–98.

Su, H., Maji, S., Kalogerakis, E., Learned-Miller, E., 2015. Multi-view convolutional neural networks for 3D shape recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 945–953.

Tapper, G., 2016. Extraction of DTM from Satellite Images Using Neural Networks (Ph.D. thesis). Linkoping University.

Wang, N., Yeung, D.-Y., 2013. Learning a deep compact image representation for visual tracking. In: Advances in Neural Information Processing Systems, vol. 26. pp. 809–817.

Weinmann, M., Jutzi, B., Mallet, C., Aug. 2014. Semantic 3D scene interpretation: a framework combining optimal neighborhood size selection with relevant features. In: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. II-3. pp. 181–188.

Xing, S., Li, P., Xu, Q., Wang, D., Li, P., Sep. 2017. Surface fitting filtering of LIDAR point cloud with waveform information. In: ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. IV-2. pp. 179–184.

Yang, Z., Jiang, W., Xu, B., Zhu, Q., Jiang, S., Huang, W., 2017. A convolutional neural network-based 3D semantic labeling method for ALS point clouds. Remote Sens. 9 (936).

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H., 2015. Understanding neural networks through deep visualization. Available from: <1506.06579>.

Yousefhussien, M.A., Browning, N.A., Kanan, C., 2016. Online tracking using saliency. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1–10.

Yunfei, B., Guoping, L., Chunxiang, C., Hao, Z., Qisheng, H., Linyan, B., Chaoyi, C., 2008. Classification of lidar point cloud and generation of DTM from lidar height and intensity data in forested area. In: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXVII-7. pp. 313–318.