Toward a Deep Learning Approach for Automatic Semantic Segmentation of 3D Lidar Point Clouds in Urban Areas

Zouhair Ballouch, Rafika Hajji, and Mohamed Ettarid

Abstract

Semantic segmentation of Lidar data using Deep Learning (DL) is a fundamental step for a deep and rigorous understanding of large-scale urban areas. Indeed, the increasing development of Lidar technology in terms of accuracy and spatial resolution offers a best opportunity for delivering a reliable semantic segmentation in large-scale urban environments. Significant progress has been reported in this direction. However, the literature lacks a deep comparison of the existing methods and algorithms in terms of strengths and weakness. The aim of the present paper is therefore to propose an objective review about these methods by highlighting their strengths and limitations. We then propose a new approach based on the combination of Lidar data and other sources in conjunction with a Deep Learning technique whose objective is to automatically extract semantic information from airborne Lidar point clouds by enhancing both accuracy and semantic precision compared to the existing methods. We finally present the first results of our approach.

Keywords

Lidar • Deep learning • Semantic segmentation • Urban environment

Z. Ballouch (✉) · R. Hajji · M. Ettarid
College of Geomatic Sciences and Surveying Engineering, IAV Hassan II, Rabat, Morocco
e-mail: z.ballouch@iav.ac.ma

R. Hajji
e-mail: r.hajji@iav.ac.ma

M. Ettarid
e-mail: m.ettarid@iav.ac.ma

1 Introduction

Several challenges are facing contemporary cities such as urban sprawl, environmental degradation, climate change, etc. Understanding these issues and predicting their impact can only be achieved through a deep and rigorous analysis of the urban environment. In this context, 3D city models are today positioned as powerful tools to address several needs about urban planning and sustainable development. Monitoring of the dynamics of cities, urban space management, construction design, and environmental studies are some appealing examples where a 3D city model is needed [1]. To respond to several city challenges, 3D city models are intended to be semantically rich to meet the requirements of urban planning and monitoring. Currently, Lidar techniques are recognized as powerful tools for producing 3D city models by offering very accurate and dense 3D point clouds at a large scale.

Semantic segmentation is an essential step to automatically design a rich 3D city model from Lidar data. It consists of assigning a semantic label for each group of point clouds (or a group of pixels in the case of images) based on homogeneous criteria [2] (Fig. 1).

The segmentation of 3D Lidar point clouds has been widely investigated in the literature leading to several notable achievements. However, this is still an active research trend until the challenges about geometric and semantic accuracy as well as robustness and performance of the proposed methods are to be resolved.

Currently, there is a lot of interest in developing Deep Learning (DL) techniques for analyzing 3D spatial data. Thanks to their potential for processing huge amounts of data corresponding to large scale and complex urban areas with good performance in terms of accuracy and efficiency, DL methods revolutionizes the field of computer vision and are the state-of-the-art in object detection and semantic segmentation [4, 5].

According to the literature, several developments have been conducted in the field of segmentation of 3D Lidar

F. Barramou et al. (eds.), *Geospatial Intelligence*, Advances in Science, Technology & Innovation,
https://doi.org/10.1007/978-3-030-80458-9_6

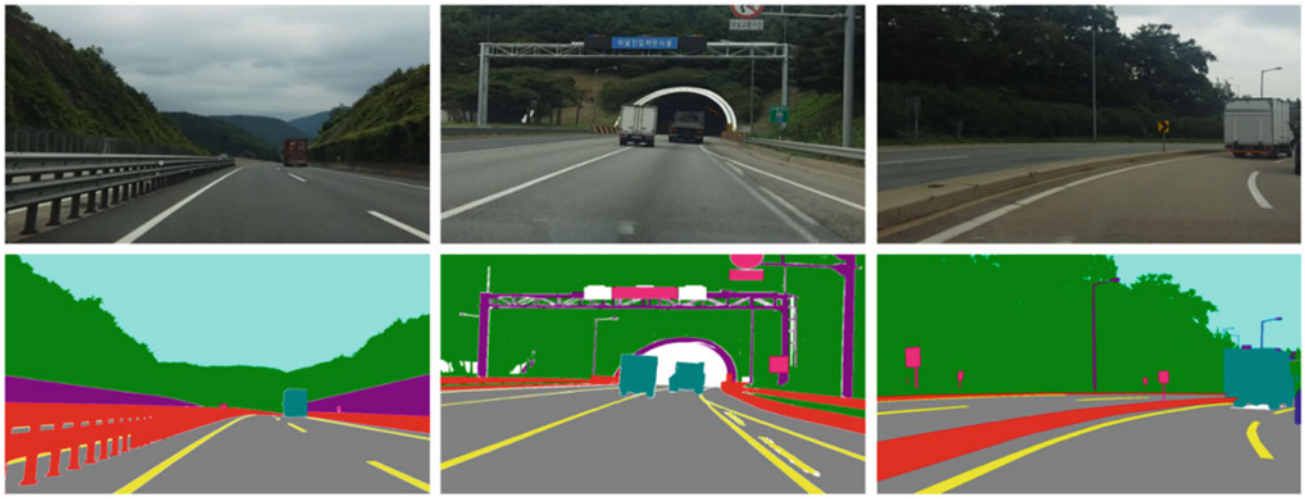

Fig. 1 3D semantic representation [3]

point clouds. The developed methods can be classified into three families. The first one is based on the raw point cloud, the second one is based on a product derived from the cloud, mainly Digital Surface Model (DSM), while the third one combines original point clouds and other data sources (aerial image, land map, etc.) [1].

Several research teams have stated that the combination of Lidar data with other sources (aerial image, satellite image, etc.) is promising, thanks to the altimeter accuracy of the 3D point clouds and the planimetric continuity of the images [6]. This motivated us to conduct our research in this field where we propose to design a methodology based on the integration of Lidar data and other sources with the aim to enhance the quality of the semantic segmentation results for urban scenes.

In the next sections, we propose to give a global overview about the main developments in semantic segmentation by highlighting the strengths and the weakness of the developed approaches. Section 2 gives an overview about the main developed methods for automatic segmentation of Lidar point clouds. Then, Sect. 3 presents some DL approaches for semantic segmentation. The discussion of the main outcomes is the subject of Sect. 4. While Sect. 5 proposes the basic guidelines as well as the preliminary results of a new approach based on our investigations and the outcomes of the literature review. Finally, the paper ends with a conclusion.

2 Automatic Segmentation of 3D Point Clouds

Point cloud segmentation is an essential step for various applications. Besides clarifying the spatial relationships between point clouds and facilitating pattern recognition, the segmentation improves the quality of subsequent classifications. This process partitions a cloud of points into a set of segments characterized by spatial and/or geometric coherence. The definition of this coherence forms the critical part of the segmentation process [7]. Numerous segmentation approaches have been developed and applied to 3D Lidar data. In this section, we mainly focus on the general research methods that are widely used for the segmentation of 3D point clouds. Three families of approaches exist to perform a semantic segmentation of Lidar point clouds. The first one is based on the raw point cloud (Direct approaches). The second one is based on a product derived from the cloud (Derived Product Based Approaches). While the third one combines original point clouds and other data sources (aerial image, land map, etc.) (Hybrid approaches) [6].

2.1 Direct Approaches

Direct approaches are applied to 3D raw point clouds without any sampling method. Among the benefits of this family of approaches, we can cite the preservation of the original characteristics of data, including accuracy and topographical relationships. On the other hand, we raise some shortcomings and gaps that hinder the effectiveness and the relevance of this family of approaches, mainly the need of a too high computing time and a rather large memory.

In the literature, many studies have been based on direct approaches. Among the developed methods, Lee [8] has proposed a segmentation process based on 3D surface detection, specifically by using Lidar raw data directly without any prior interpolation. This method allows automatic division of the point cloud into two classes: ground

and buildings, considered as the main objects of an urban scene. The method of [9] proposes a cluster analysis of 3D airborne Lidar data by using a slope adaptive neighborhood system based on accuracy, point density, and distance between 3D point clouds in order to define the neighborhood between the measured points. According proximity and local continuity, points that are on the same surface are connected [9]. The method gives good results in extracting vertical walls and modeling objects with a precision of few centimeters. Lari [10] proposed a method for segmentation of planar patches using Lidar data. In this approach, the authors used an adaptive cylinder for establishing the neighborhood of each point by considering surface trend and density. This definition of neighborhood positively influences the calculation of segmentation attributes (vegetation, flat and gable roofs, walls …etc.). The approach demonstrates efficiency and reliability for both airborne Lidar and Mobile Mapping Systems data. Finally, a segmentation method applied to a mobile and airborne mapping system has been proposed by [11] where the main objective is to bypass the drawbacks of point-based classification techniques; its principle is based on grouping point clouds in regions with similar characteristics. The proposed methodology demonstrates a high potential in classification of both terrestrial and airborne Lidar data.

2.2 Derived Product Based Approaches

Since direct approaches require a very high processing time and large storage capacity, many researchers recommend the transformation of 3D data into 2D in order to have a regular form that is easy to manipulate. This is the principle of derived based approaches which are based on derived products from Lidar data such as DSM (Digital Surface Model) and 2D images. This family of approaches offers a wide range of advantages such as the ease of handling and the efficiency of data processing. However, these approaches require a 2D transformation of 3D data or voxels representations which result in a huge loss of geometric and radiometric information, and thus a loss of precision due to the resampling operation.

In the literature, there is a large number of approaches that have been developed for the segmentation of 3D point clouds from the regular data generated from the point clouds. Among these approaches, Yuan [12] proposed a new technique called “Pointseg” that allows a real time semantic segmentation of road objects based on spherical images where the structure of the proposed network is based on SqueezeNet [13] and SqueezeSeg [14]. The proposed network has three main functional layers: (1) fire layer, (2) squeeze reweighting layer, and (3) enlargement layer. The results show compatibility with robot applications by achieving competitive accuracy with 90 frames per second on a single GPU (Graphics Processing Unit) and high efficiency when tested with KITTI 3D objet detection dataset. Milioto [15] proposed a semantic segmentation approach called RangeNet++. This approach has been applied to Lidar data recorded by a rotating Lidar sensor in order to enable the autonomous vehicles to make the best decisions in a timely manner. The authors proposed a projection based 2D CNN (Convolutional Neural Network) processing of Lidar data and used a range image representation of each laser scan to perform the semantic inference. The results show that this method outperforms the state of the art both in runtime and accuracy. Moreover, a new approach for semantic labeling of unstructured 3D point clouds has been proposed in [16]. The authors proposed a framework that applies CNN on multiple 2D image views of the Lidar data based on two steps: (1) generation of two types of images: depth composite view and RGB view and (2) labeling each pair of bidimensional image views by means of CNNs. After that, they project back the semantized images. This approach showed good results when evaluated using a dataset called Semantic-8. Another method for Lidar data segmentation using voxel structure and graph-based clustering was proposed by [17]. The authors used a geometric method that not require any radiometric information. The process consists of three steps: (1) voxelisation of 3D point clouds, (2) calculation of geometric cues, and (3) the graph-based clustering. The method has demonstrated good results mainly for complex environment and non-planar areas, compared to several segmentation methods proposed in the literature. Riegler and Osman Ulusoy [18] proposed a method called “OctNet” as a novel tridimensional representation for point clouds labeling, which enables 3D CNN that are both high resolution and deep. The method was evaluated using RueMong2014 dataset [19] and achieved good results. Finally, another work has been proposed by [20] where the authors evaluated various bidimensional image models using four datasets which are DUT1, NC, DUT2, and KAIST. The results, compared to those of direct approaches, show that the use of bidimensional image models give an interesting improvement in computational efficiency with a little loss of precision. Furthermore, the authors concluded that 2D image models are better suited to real-time segmentation of outdoor areas.

2.3 Hybrid Approaches

Despite the simplicity and the efficiency of Derived Product Based Approaches, several researchers argued that Lidar data need to be combined with other data sources (aerial photos, satellite images, etc.) to take benefits from the planimetric continuity of images and the altimetric precision

of 3D point clouds [6]. Several investigations in this field have shown promising results in terms of accuracy and quality of the segmentation. However, despite their performance, these approaches have many disadvantages related to memory requirements, difficulty of handling and implementation, and the need to have a minimum difference in the time of acquisition of the two types of data.

The first method has been proposed by [21] for automatic building detection from 3D point clouds and multispectral imagery. This method is capable of detecting different urban objects (industrial buildings, urban residential, etc.) of different shapes with very high precision. The authors of [22] applied a multi-filter CNN for semantic segmentation based on the combination of 3D point clouds and high-resolution optical images, and then they used a MRS (Multi-Resolution Ssegmentation) for delimiting the contours of objects. The results show that this approach improves the overall accuracy over other methods using Potsdam and Guangzhou datasets and is more suitable for the processing of objects with a regular shape such as cars and buildings. Furthermore, Xiu [23] proposed a new method to study the influence of integrating two types of data which are aerial images and 3D point clouds for semantic segmentation which shows an accuracy of 88%. Additionally, a new semantic segmentation study combining images and 3D point clouds has been proposed by [24] by adopting the DVLSHR (Deeplab-Vgg16 based Large-Scale and High-Resolution) model which is satisfactory for semantic segmentation of large-scale scenes when compared to other methods developed in the literature using CityScapes dataset. Another approach called SPLATNet was proposed in [25]. This approach has been tested with RueMonge2014 dataset [19] where an Intersection Over Union score was computed for all classes in order to evaluate the semantic segmentation results. The proposed approach scores well among the state-of-the-art algorithms for semantic segmentation. Recently, [26] proposed a new methodology for semantic segmentation which grasps bidimensional textural appearance and tridimensional structural characteristics in an integrated framework. The authors evaluated this approach using ScanNet Dataset [27]. The method has demonstrated good results compared to 3DMV (3D-Multi-View) and SplatNet (Sparse lattice Networks) approaches. Similarly, Li [28] designed a 3D real-time semantic map using 3D point clouds and images of road scenes. The method consists of using a CNN to segment 2D images acquired by a camera, and then the semantic segmentation results and the 3D point clouds are fused to generate a unified point cloud with an associated semantic information. The proposed technique is effective for several complex tasks including autonomous driving, robot navigation, etc.

2.4 Summary

3D Lidar data segmentation methods can be grouped into: Direct approaches, Derived Product Based Approaches, and Hybrid approaches. The direct approaches are the least used in the literature because they require a very large storage capacity and are very demanding in processing and computing time. Despite their limitations, their strengths lie in the preservation of the characteristics and the original topological relationships of the point cloud. Derived Product Based Approaches are the most dominant, simplest, and quickest approaches in the literature. However, the resampling operation applied to the point cloud causes a huge loss of information and so a loss of precision of the segmentation process. Finally, approaches combining 3D Lidar data and other sources allow improving the accuracy of the segmentation. However, these approaches do not accept large time differences between the acquisition of Lidar and images and require a very high storage capacity and a very important processing time (Table 1).

Actually, the development of DL methods offers a best opportunity to satisfy the need of computer vision field and demonstrates a high potential in semantic segmentation in terms of accuracy and efficiency. Their performance in segmentation process would enhance the quality of the results. The next section tries to give a brief overview of researches addressing DL in semantic segmentation.

3 Contribution of DL to Semantic Segmentation

Actually, DL methods revolutionize the field of computer vision and demonstrate good performance in semantic segmentation by solving a wide range of difficult problems in this field [29]. In this section, we examine some DL techniques used in semantic segmentation of Lidar data acquired in urban areas.

PointNet is a reference network which opened the way for the use of DL techniques for semantic segmentation of Lidar data [30]. Its performance, combined with its ease of implementation, makes it a perfect baseline for semantic segmentation of 3D point clouds. The core principle of PointNet is to implement the permutation invariance of the points in a cloud directly into the network. To evaluate its performance, the authors used the Stanford 3D dataset where data are annotated with 13 classes (floor, chair, table, etc.). PointNet has demonstrated satisfactory results compared to the literature. Similarly, Qi [31] proposed a hierarchical DL model called “PointNet++” in order to process a set of points that have been sampled in metric space in a hierarchical

Table 1 Advantages and disadvantages of the different segmentation approaches

Approach	Advantages	Disadvantages
Direct approaches	– Preserve the original topological relationships of point cloud	– Expensive – Few developed programs
Derived product based approaches	– Easy and fast drive – Requires few parameters	– Loss of information and accuracy due to re-sampling – False data caused by resampling step – Errors accumulation
Hybrid approaches	– Accurate – Efficient	– Expensive – Require a minimum difference in time of acquisition of the two types of data

manner. To test this approach, four datasets have been used, namely, ModelNet40, MNIST, SHREC15, and ScanNet. The results show that the proposed approach is more suitable to process point sets robustly and efficiently compared to other existing methods. Besides, this methodology introduced hierarchical feature learning and captures spatial features at different scales which is important in case of objects of different sizes. Another semantic segmentation approach named SegCloud was proposed in [32]. The proposed approach combines the advantages of trilinear interpolation, neural networks, and FC-CRF (Fully Connected Conditional Random Fields). The authors used the trilinear interpolation to transform voxels predictions to raw 3D points, then the FC-CRF allows overall consistency, and fine semantic segmentation. The authors evaluated the performance of the proposed algorithm using four multi-scale datasets about indoor or outdoor scenes (NYU V2, S3DIS, KITTI, and Semantic3D). The results show that CRF allows a significant improvement of the network and a high ability to extract the contours of objects in a very clear way. Moreover, a novel fully CNN approach for semantic segmentation of images named SegNet has been developed by [33]. It consists of an encoder-decoder structure based on the convolution layers of the VGG-16 algorithm. The architecture of SegNet is symmetrical and allows precise positioning of abstract features with good spatial locations. CamVid dataset has been used to evaluate the performance of the proposed method. This dataset is divided into two sets: the first contains 367 images used for training the model while the second contains 233 images used for performance evaluation. The results show that this algorithm gives good results and achieves very high scores in the case of semantic segmentation of road environments. Furthermore, Landrieu and Simonovsky [34] proposed a new Lidar approach applicable for large 3D Lidar data where the main objective is to divide the point clouds into simple forms. The process is based on three main steps: (1) a new concept called a superpoint graph to encode the relationships between object parts by edge attributes is proposed, (2) a neural network is used for the representation of each simple shape, and (3) two public datasets (S3DIS and Semantic3D) are used to improve the average of mIOU (mean Intersection Over Union). In addition, Qi [35] proposed a 3D object detection approach based on collaboration between Haugh Voting and point set network called VoteNet. It is a geometric method that does not require any radiometric information but shows clear improvements over hybrid methods. Additionally, Yang [36] proposed a new large-scale urban semantic segmentation framework by integrating multiple aggregation levels (point-segment-object) of features and contextual features for road facilities recognition from 3D Lidar data. This study achieved very satisfactory results with an object recognition accuracy of more than 90%. Finally, Hu [37] developed a new neural network architecture called "RandLA-Net" that directly uses 3D Lidar data based on point sampling in a random manner. In order to reduce the point density, to avoid loss of information caused by the resampling step, the authors proposed a new local feature aggregation module. Compared to the literature, the proposed approach demonstrates a good performance in terms of precision, calculation time and is not demanding a fairly large memory.

4 Discussion

Today, 3D city models allow better understanding of urban spaces which is crucial for optimal management of cities. They are capable to meet several needs related to simulation and decision making processes. However, most of 3D city models lack rich semantics about urban knowledge and are far to respond to several challenges about smart and sustainable cities.

In computer vision, semantic segmentation is defined as the assignment of a class to each coherent region of an image [2] or 3D point clouds. Many recent studies have shown the effectiveness of DL in this context [30, 34–37]. The first experiments of approaches dedicated to semantic segmentation of 3D point clouds began by the use of conventional image processing programs by transforming the 3D Lidar data into regular shapes (for example, series of images) as in the case of the approach proposed by [16] that requires a

transformation of 3D point clouds to 2D images. Other DL techniques are based on the transformation of the Lidar data into a grid of voxels that have a regular form as the case of the SegCloud method that was proposed by [32]. These regular representations do not really allow a clear writing of the particular organization of Lidar data which limits the performance of this type of approach [34]. Besides, the voxel representation does not take into account the small details of 3D forms.

Several research teams have proposed a range of dedicated approaches directly analyzing Lidar data. Among these approaches, the PointNet approach, proposed by [30], operates at the point level, which allows a very fine segmentation. This method is adapted to 3D point clouds acquired in indoor scenes, but it requires a necessary adaptation or additional training to be adapted to large datasets [32]. Similarly, the PointNet++ method is applied to the raw point clouds [25] without any sampling operation, which saves the initial information [35]. This method has demonstrated better performance in semantic segmentation and object classification [35]. However, it shows some limitations, namely, large computation and memory cost [38, 39]. Furthermore, this approach is not able to aggregate the scene context around the object centers due to more clutter and inclusion of neighboring elements [35], and also lacks a relevant specification of the spatial connectivity between points [25]. We note that "PointNet" and "PointNet++" have not been tested on data acquired by a large scale airborne mapping system that contains more complicated urban geographic features [23]. Recently, several approaches have been developed for processing of large-scale 3D point clouds. In this context, we find the SPG method that allows the preprocessing of 3D Lidar data as super-graphs in order to subsequently apply a neural network to assign a semantic label for each group of points [15]. The main advantage of this approach is its ability to handle large point clouds simultaneously by cutting point clouds into simple shapes that are easier to classify than points, but despite the low number of network parameters, this approach is high demanding in terms of time of processing required by super-graph construction and geometric partitioning [15]. We can state that most of the existing semantic segmentation approaches require a variety of blocks partitioning steps, pre/post–processing as well as the construction of graphs. In contrary, the "RandLA-Net" approach is able to directly process large scale 3D Lidar data in a single pass with high efficiency (1 million points in a single pass) without any pre-processing or post-processing steps compared to the existing methods [39].

Finally, semantic segmentation is an active research trend which aims to reach robust methods to extract semantics from dense point clouds or images. The construction of these models from Lidar data requires designing new approaches capable of extracting the maximum amount of semantic information about a large-scale urban environment with high accuracy and efficiency. Our research tries to respond to this challenge by proposing an innovative hybrid approach which aims to enhance the quality of semantic segmentation of airborne Lidar point clouds.

5 Our Approach

The literature review about DL techniques that address semantic segmentation of Lidar point clouds shows that this is clearly a field that requires further research in order to improve the accuracy and the performance of the segmentation process. This has motivated us to conduct research in this field in order to propose an innovative approach for semantic segmentation of airborne Lidar data based on a hybrid solution. In this section, we expose the first guidelines and preliminary results of our proposed research in this context.

5.1 Methodology

We propose to design a DL approach based on the combination of 3D airborne Lidar data and aerial images for semantic segmentation of airborne Lidar point clouds corresponding to large-scale urban environments. Our methodology is expected to give better results in terms of precision and robustness to recognize 3D objects of urban scenes and associate them a rich semantic. Figure 2 summarizes the general workflow of our approach.

Our approach relies on the combination of the geometry of Lidar data and the spectral information of images. It is based on the use of raw data in order to retain the original characteristics and topological relationships of 3D point clouds. The first step consists of applying semantic segmentation to drone images which results will be integrated with Lidar data in order to refine the quality of the segmentation process (part 2). The test of the performance and the reliability of the proposed approach will be performed through several large-scale datasets. In the next section, we present and analyze the preliminary results related to the first step of the wokflow (Part1).

5.2 Preliminary Segmentation

Semantic segmentation from drone images is a first step of the general workflow. The results will be then integrated with Lidar point clouds to enhance the segmentation process. High spatial–resolution of data acquired by drones makes it possible to discriminate the different urban objects and associate them a semantic label. In this context, several DL

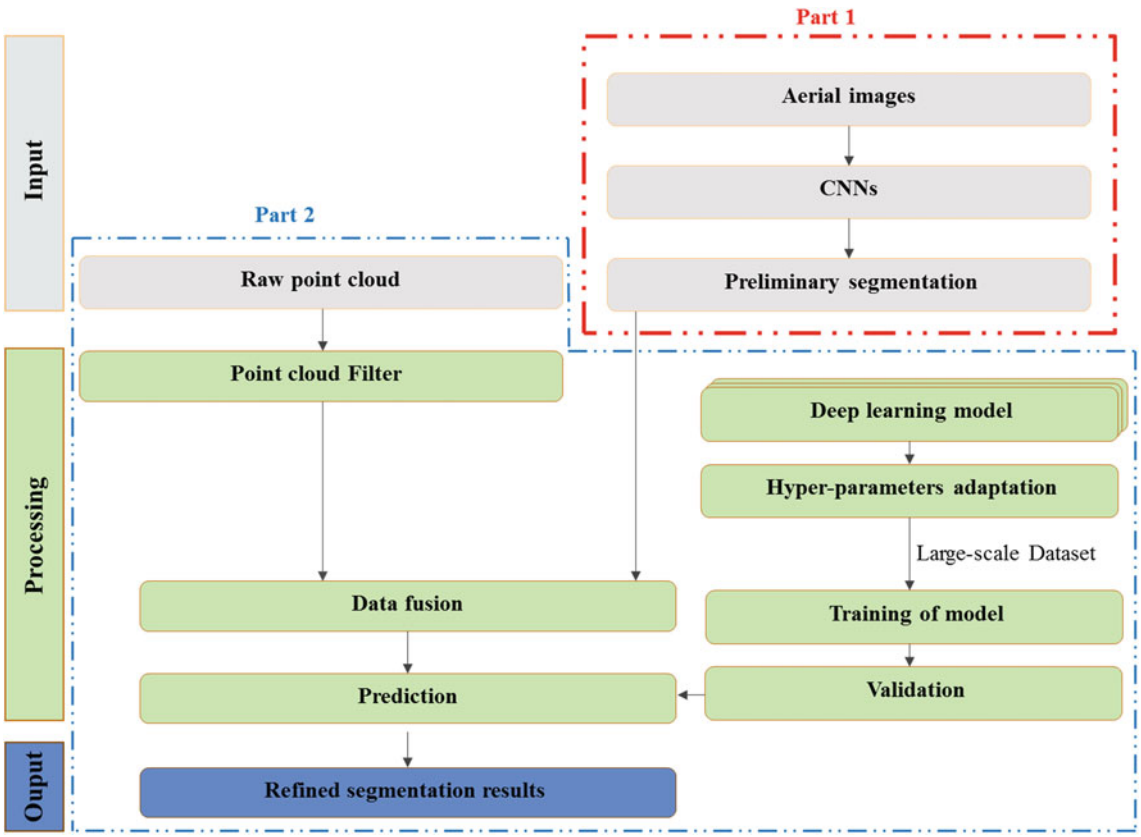

Fig. 2 The general workflow of our approach

techniques applied to drone images have been proposed in the literature [40–43]. To our knowledge, there is no literature review about the evaluation of the existing techniques. This is why we had to conduct several tests to evaluate different models (Unet, Vgg_Unet, Resnet50_Unet, Segnet, Vgg_Segnet, and Resnet50_Segnet) in terms of precision and calculation time in order to choose the most suitable one for semantic segmentation of drone images.

5.2.1 Data

The case study consists of 400 large-scale drone images with a high resolution of 6000 * 4000 px (24Mpx) and an altitude of 5–30 m above ground which are available for free download (https://dronedataset.icg.tugraz.at). The images are annotated with 20 classes: tree, gras, other vegetation, dirt, gravel, rocks, water, paved area, pool, person, dog, car, bicycle, roof, wall, fence, fence-pole, window, door, and obstacle. Some examples of the dataset are shown in Fig. 3.

Another data is used for the evaluation of the process. It is relative to an urban zone of the city of Nador (Morocco), where the images was acquired with a ground resolution at 100 m flight height of 3.5 cm and resolution of 12 MegaPixel.

5.2.2 Results

For the implementation of the DL models used in this study, we used the Keras library and Google Colaboratory as a cloud computing server. Google Colaboratory is a free Google tool that allows performing computational simulations with support of Python and some other libraries. For conducting the tests, 80% of the dataset is used for training the model while 20% serves as testing data. In this section, we present the results about the evaluation of both accuracy and time of calculation of the segmentation process applied to the selected models.

Accuracy assessment

The semantic segmentation realized according the tested models is evaluated through two parameters: (1) accuracy and (2) frequency weighted IU (f.w.IU). Accuracy metric is the ratio of the number of correct predictions to the total number of input samples. While the frequency weighted IU defines the variations on region intersection over union (IU) used in target detection [44]. These metrics are obtained using the equations below:

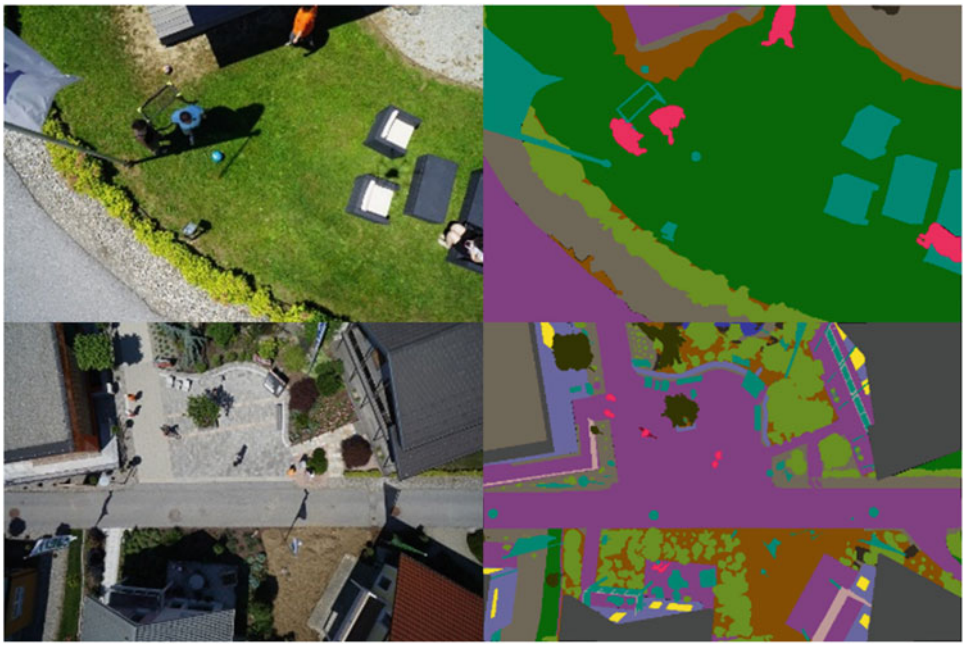

Fig. 3 Examples of classified Drone images from the dataset

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

$$\text{f.w. IU} = \Sigma i\left(\Sigma jUij \cdot \Sigma k\Sigma jUij\right)\frac{Uii}{\Sigma jUij + \Sigma jUji - Uii}[44]$$

where k represents the number of classes. The symbol u_{ij} corresponds to the number of samples belonging to category i in ground truth and are classified in class j in segmentation results [44].

The evaluation results according accuracy and frequency weighted IU are reported in Table 2 and Table 3 respectively.

Even though we conducted the tests with a limited number of epochs, we reached good results for both accuracy and frequency weighted IU with the different models. According the results, we can say that the "Resnet50_Unet" model outperforms the others both in accuracy and frequency weighted IU metrics.

Training duration

Besides the accuracy of the segmentation, we also evaluated the efficiency of the tested models in terms of processing time. The results are reported in Table 4.

According to the statistics in Table 4, we can state that the processing time is relatively negligible and all models require almost the same computation time with a bit difference of the Vgg-Unet model which requires slightly more time than the other ones.

The preliminary tests were necessary to test the performance of the selected models. According the results, the "Resnet50_Unet" has been elected as the most suitable model for semantic segmentation of drone images to be adopted in our approach. This model has been applied to the case study about the urban area in Morocco. The corresponding semantic segmentation results are shown in Fig. 4 and validated by comparison to the field reality.

6 Summary

In the previous section, we presented the first results of the general workflow of our approach. It consists of semantic segmentation of drone images as a first step of the process. The general objective is to integrate the preliminary results of the image segmentation process with Lidar data in order to enhance the quality of the segmentation in terms of accuracy and performance. We performed a series of experiments to compare the capabilities of the different DL techniques for semantic segmentation of urban objects using Drone images.

Table 2 Comparison of accuracy between the DL models

	Unet	Vgg_Unet	Resnet50_Unet	Segnet	Vgg_Segnet	Resnet50_Segnet
Accuracy	0.71	0.76	0.85	0.72	0.7215	0.82

Table 3 Comparison of frequency-weighted IU between the DL models

	Unet	Vgg_unet	Resnet50_Unet	Segnet	Vgg_segnet	Resnet50_Segnet
Frequency_Weighted_IU	0.56	0.63	0.76	0.58	0.56	0.72

Table 4 The required time for the segmentation process

	Unet	Vgg-Unet	Resnet50-Unet	Segnet	Vgg-Segnet	Resnet50_Segnet
Epochs	1310 s	1403 s	1225 s	1217 s	1208 s	1281 s
Epoch 1	1269 s	1385 s	1202 s	1198 s	1160 s	1222 s
Epoch 2	1243 s	1366 s	1219 s	1178 s	1159 s	1175 s
Epoch 3	1248 s	1319 s	1229 s	1154 s	1161 s	1171 s
Epoch 4	1205 s	1287 s	1209 s	1152 s	1163 s	1172 s
Total time (s)	6275	6760	6084	5899	5851	6021
Total time (m)	105	113	101	98	97	100

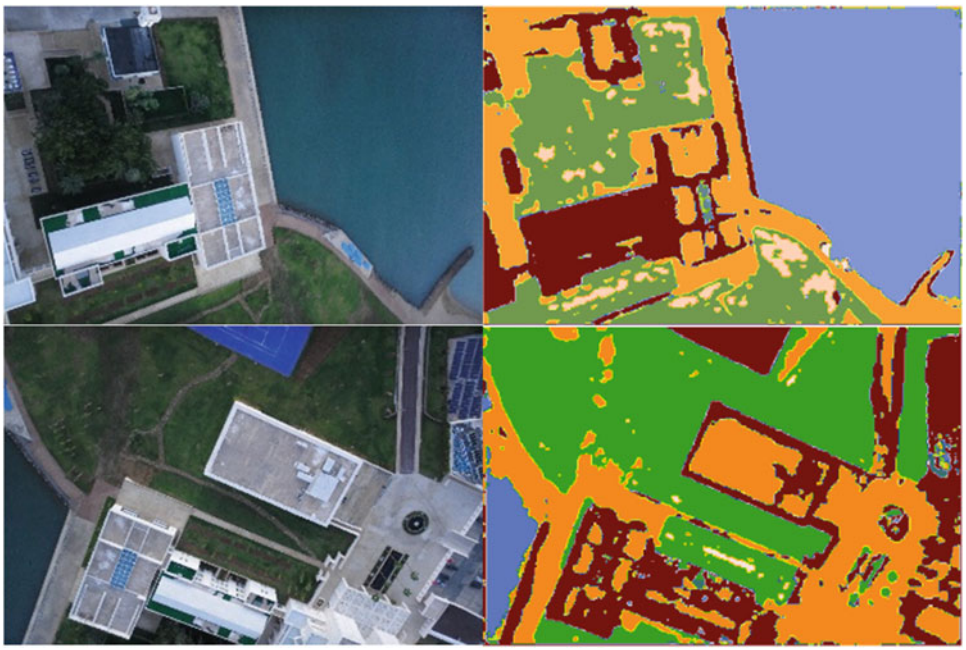

Fig. 4 Examples of semantic segmentation results

The results show that that all tested models give good results in terms of accuracy and frequency weighted IU. However, the Resnet50_Unet model scores well in both parameters. Hence, it has been selected as the most suitable one for semantic segmentation of drone images among the others. We should note that the quality of the results can be further improved by using a powerful dataset with more training data and by augmenting the number of epochs.

Finally, for a better evaluation of the performance of different DL models, we propose to use other types of datasets, as well as to apply the models to other images acquired in other different urban contexts.

7 Conclusion

In this paper, we have proposed a literature review about semantic segmentation methods of 3D Lidar point clouds based on DL. Several DL models have been presented and analyzed by highlighting their advantages and their limitations. We then presented the first guidelines about our proposed methodology which aims at developing a DL approach based on integrating 3D Lidar point clouds and aerial images for semantic segmentation in a large-scale urban environment. We aim to improve the object recognition accuracy and the efficiency of the existing methods.

As a first step of our approach, we investigated the performance of some DL models in terms of accuracy and performance for semantic segmentation of drone images by conducting several tests. In the next steps, our method will be tested on several datasets to confirm the reliability and the performance of the proposed approach.

References

1. A. Bellakaout, Extraction automatique des batiments, végétation et voirie à partir des données Lidar 3D. Thèse de docteur de l'institut agronomique et vétérinaire Hassan II, Maroc (2016)
2. L. Haifeng, Unsupervised scene adaptation for semantic segmentation of urban mobile laser scanning point clouds. ISPRS J. Photogramm. Remote. Sens. **169**, 253–267 (2020)
3. B. Kim, Highway driving dataset for semantic video segmentation. School of Electrical Engineering Korea Advanced Institute of Science and Technology (KAIST), South Korea (2016)
4. J. Castillo-Navarro, Réseaux de neurones semi-supervisés pour la segmentation sémantique en télédétection. Colloque GRETSI sur le Traitement du Signal et des Images, Lille, France. hal-02343961 (2019)
5. A. Garcia-Garcia, A review on deep learning techniques applied to semantic segmentation. arXiv:1704.06857v1 [cs.CV] (2017)
6. M. Awrangjeb, Automatic detection of residential buildings using LIDAR data and multispectral imagery. ISPRS J. Photogram. Remote Sens. **65**, 457–467 (2010)
7. J. Ravaglia, Segmentation de nuages de points par octrees et analyse en composantes principales. GTMG 2014, Mar 2014, Lyon, France. hal-01376473 (2014)
8. I. Lee, Perceptual organization of 3D surface points, photogrammetric computer vision. ISPRS Comm. III. Graz, Austria. **XXXIV** part 3A/B. ISSN 1682-1750 (2002)
9. S. Filin, Segmentation of airborne laser scanning data using a slope adaptive neighborhood. ISPRS J. Photogramm. Remote Sens. **60**, 71–80 (2006). https://doi.org/10.1016/j.isprsjprs.2005.10.005 (2005)
10. Z. Lari, An adaptive approach for segmentation of 3D laser point cloud, in *ISPRS Workshop Laser Scanning*, Calgary, Canada (2011)
11. Z. Lari, A. Habib, Segmentation-based classification of laser scanning data, in *ASPRS 2012 Annual Conference Sacramento*, California, 19–23 Mar 2012
12. W. Yuan, PointSeg: real-time semantic segmentation based on 3D LiDAR point cloud. arXiv:1807.06288v8 [cs.CV] (2018)
13. F.N. Iandola, Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. CoRR abs/1602.07360 (2016)
14. B. Wu, Squeezeseg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3d lidar point cloud. CoRR abs/1710.07368 (2017)
15. A. Milioto, RangeNet++: fast and accurate LiDAR semantic segmentation. German Research Foundation under Germany's Excellence Strategy, EXC-2070 - 390732324 (PhenoRob) as well as grant number BE 5996/1–1, and by NVIDIA Corporation (2019)
16. A. Boulch, SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. Comput. Graph. (2017)
17. Y. Xu, Voxel- and graph-based point cloud segmentation of 3d scenes using perceptual grouping laws. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. **IV-1/W1** (2017)
18. G. Riegler, A. Osman Ulusoy, Octnet: learning deep 3d representations at high resolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 3 (2017)
19. H. Riemenschneider, A. Bódis-Szomorú, Learning where to classify in multi-view semantic segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2014)
20. Y. Liu, Comparison of 2D image models in segmentation performance for 3D laser point clouds. Neurocomputing (2017)
21. M. Awrangjeb, Automatic detection of residential buildings using LIDAR data and multispectral imagery. ISPRS J. Photogramm. Remote Sens. **65**, 457–467 (2010)
22. Y. Sun, Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. ISPRS J. Photogramm. Remote Sens. (2018)
23. H. Xiu, 3D semantic segmentation for high-resolution aerial survey derived point clouds using deep learning (Demonstration), in *Information Systems (SIGSPATIAL'18)*, 6–9 Nov 2018, Seattle, WA, USA, ed. by F. Banaei-Kashani, E. Hoel (ACM, New York, NY, USA, 2018)
24. R. Zhanga, Fusion of images and point clouds for the semantic segmentation of large scale 3D scenes based on deep learning. ISPRS J. Photogramm. Remote Sens. (2018)
25. H. Su, V. Jampani, Splatnet: sparse lattice networks for point cloud processing, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 2530–2539
26. H.-Y. Chiang, A unified point-based framework for 3D segmentation, in *International Conference on 3D Vision (3DV)* (2019)
27. A. Dai, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in *Proceedings of CVPR 2017* (2017)
28. J. Li, Building and optimization of 3D semantic map based on Lidar and camera fusion. Neurocomputing
29. Y. Li, Deep learning for remote sensing image classification: a survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, vol. 8, p. 1264 (2018)
30. C.R. Qi, Pointnet: deep learning on point sets for 3d classification and segmentation. CoRR abs/1612.00593 (2016)
31. C.R. Qi, PointNet++: deep hierarchical feature learning on point sets in a metric space. arXiv:1706.02413v1 [cs.CV] (2017)
32. L.P. Tchapmi, Segcloud: semantic segmentation of 3d point clouds, in *International Conference on 3D Vision (3DV)* (2017), pp. 537–547
33. B. Vijay, SegNet: a deep convolutional encoder- decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 2481–2495 (2017)
34. L. Landrieu, M. Simonovsky, Large-scale point cloud semantic segmentation with superpoint graphs, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 4558–4567
35. C.R. Qi, Deep Hough voting for 3D object detection in point clouds. arXiv:1904.09664v2 [cs.CV] (2019)
36. B. Yang, Computing multiple aggregation levels and contextual features for road facilities recognition using mobile laser scanning data. ISPRS J. Photogramm. Remote Sens. **126**, 180–194 (2017)
37. Q. Hu, RandLA-Net: efficient semantic segmentation of large-scale point clouds. arXiv:1911.11236v3 [cs.CV] (2020)
38. Z. Yang, Std: sparse-to-dense 3d object detector for point cloud, in *The IEEE International Conference on Computer Vision (ICCV)* (2019)
39. Y. Cui, Deep learning for image and point cloud fusion in autonomous driving: a review. arXiv:2004.05224v2 [cs.CV] (2020)
40. A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv print. 14 p (2014)
41. O. Ronneberger, P. Fischer, U-Net: convolutional networks for biomedical biomedical image segmentation, in International

Conference on Medical Image Computing and Computer-Assisted Intervention (2015), pp. 234–241
42. K. He, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778
43. A. Chaurasia, Linknet: exploiting encoder representations for efficient semantic segmentation. arXiv preprint 1707.03718 (2017)
44. M.H. Wu, ECNet: efficient convolutional networks for side scan sonar image segmentation. Sensors **19**(9), 2019 (2009). https://doi.org/10.3390/s19092009