Article

3D Point Cloud Enriched Semantic Segmentation through Knowledge-Enhanced Deep Learning Technique

Zouhair BALLOUCH * , Rafika Hajji * , Abderrazzaq KHARROUBI * , Florent POUX * , Roland Billen *

Posted Date: 29 June 2023

doi: 10.20944/preprints202306.2022.v1

Keywords: smart fusion approach; enriched semantic segmentation; LiDAR point clouds; images data; data fusion; prior knowledge; deep learning; urban environment

Article

3D Point Cloud Enriched Semantic Segmentation through Knowledge-Enhanced Deep Learning Technique

Zouhair Ballouch [1,2,*], Rafika Hajji [1], Abderrazzaq Kharroubi [2], Florent Poux [2] and Roland Billen [2]

1 College of Geomatic Sciences and Surveying Engineering, IAV Hassan II, Rabat 6202, Morocco; r.hajji@iav.ac.ma
2 Geomatics Unit, Department of Geography, ULiège—Place du 20 Août, 4000 Liège, Belgium; fpoux@uliege.be (F.P.); akharroubi@uliege.be (A.K.); rbillen@uliege.be (R.B.)
* Correspondence: Zouhair.Ballouch@student.uliege.be; Tel.: +32-499391903

Abstract: Digital Twin Cities (DTCs) play a fundamental role in city planning and management. They allow three-dimensional modeling and simulation of cities. 3D semantic segmentation is the foundation for automatically creating enriched DTCs, as well as their updates. Past studies indicate that prior level fusion approaches demonstrate more promising precisions in 3D semantic segmentation compared to point level fusion, features level fusion, and decision level fusion families. In order to improve point cloud enriched semantic segmentation outcomes, this article proposes a new approach for 3D point cloud semantic segmentation through developing and benchmarking three prior level fusion scenarios. A reference approach based on point clouds and aerial images was proposed to compare it with the different developed scenarios. In each scenario, we inject a specific prior knowledge (geometric features,classified images, etc) and aerial images as attributes of point clouds into the neural network's learning pipeline. The objective is to find the one that integrates the most significant prior knowledge and enhances neural network knowledge more profoundly, which we have named the "smart fusion approach". The advanced Deep Learning algorithm "RandLaNet" was adopted to implement the different proposed scenarios and the reference approach, due to its excellent performance demonstrated in the literature. The introduction of some significant features associated with the label classes facilitated the learning process and improved the semantic segmentation results that can be achievable with the same neural network alone. Overall, our contribution provides a promising solution for addressing some challenges, in particular more accurate extraction of semantically rich objects from the urban fabric. An assessment of the semantic segmentation results obtained by the different scenarios is performed based on metrics computation and visual investigations. Finally,the smart fusion approach was derived based on the obtained qualitative and quantitative results.

Keywords: smart fusion approach; enriched semantic segmentation; LiDAR point clouds; images data; data fusion; prior knowledge; deep learning; urban environment

1. Introduction

Many cities around the world are building their Digital Twin Cities (DTCs) (Shahat et al., 2021). Semantic 3D city models are the foundation for developing these DTCs both for academic and industry research (Ruohomäki et al., 2018; White et al., 2021). In this context, semantic segmentation of LiDAR point clouds is a relevant input for building DTCs. It allows the semantic enrichment of 3D city models, their updates, and the performance of multiple spatial and thematic analyses for city management, urban planning, and decision-making.

The latest advancements in aerial mapping technology enable the capture of three-dimensional data at an extremely high level of spatial resolution. LiDAR (Light Detection And Ranging) technology stands out among other systems for its ability to quickly and precisely gather data with a high point density. However, there are some challenges in the acquisition phase and others in the processing phase, such as irregularity, rigid transformation, etc (Zhang et al., 2019). Before using the acquired data, several processing steps must be followed namely, pre-processing, registration, etc, in

doi:10.20944/preprints202306.2022.v1

order to obtain consistent data. The obtained data can be employed in a versatile manner across multiple domains, such as urban planning (Liu et al., 2022), outdoor navigation (Jeong et al., 2022), and urban environmental studies (Son et al., 2020).

The advancement of computer vision technology, along with the widespread utilization of Deep Learning (DL) methods, has resulted in the development of more robust and reliable 3D semantic segmentation techniques. Indeed, many DL algorithms have been developed recently for 3D semantic segmentation (Hu et al., 2020; Landrieu and Simonovsky, 2018; Qi et al., 2017). DL algorithms are proposed to handle complex tasks in various LiDAR applications. Among these algorithms, we can cite the Deep Neural Networks (DNNs) which have gained considerable popularity and attention due to their efficiency. The present focus is on developing new DL-based approaches to enhance the quality of semantic segmentation outcomes. Then, it is necessary to compare them with the existing approaches in order to derive the most suitable one for LiDAR point clouds processing.

Numerous studies have attested to the promising potential of point cloud and aerial image fusion for 3D semantic segmentation (Meyer et al., 2019; Poliyapram et al., 2019; Zhao et al., 2021). However, there are few initiatives proposed in the literature to enhance the DL technique knowledge for semantic segmentation by the injection of the prior knowledge into the learning pipeline (Ballouch et al., 2022b; Grilli et al., 2023). To the best of our knowledge, no study has evaluated all possible scenarios of injection of prior knowledge and aerial images as point clouds attributes to derive the approach that most improves the DL technique knowledge, as demonstrated in this paper. Specifically,we propose a novel approach that explicitly addresses the issue of the precise extraction of the maximum amount of urban objects (footpath, heigh vegetation, etc) by injecting point clouds, aerial images and prior knowledge (geometric features , classified images,…etc) to DL technique at prior level of the semantic segmentation pipeline. It's motivated by the fact that point cloud semantic segmentation can benefit from the use of aerial images and prior knowledge, especially in cases where the differentiation between detailed urban objects is challenging to detect. This study incorporates knowledge-based information and aerial images into the DL-based segmentation pipeline for 3D point clouds using the architecture "RandLaNet". By leveraging such knowledge, our approach can guide the semantic segmentation process towards the precise differentiation of several semantically rich classes. Our approach stems from the development and evaluation of three scenarios, each involving the fusion of point clouds, aerial images, and a specific type of prior knowledge. Based on the obtained results, we identify the scenario that will allow us to extract the maximum amount of semantic information present in the urban environment while providing satisfactory quality and quantity results. This scenario is derived as the "smart fusion approach". The potential of our research lies in the deployment of an automated urban object extraction pipeline with a high level of detail. Its primary aim is to meet the requirements of DTCs, specifically in creating the basic geometric structure of DTCs, enhancing their semantic enrichment, and ensuring regular updates. The following are the main contributions of this paper:

- Designing three possible prior level fusion scenarios of 3D semantic segmentation, each of which injects point clouds, aerial images, and a specific type of prior knowledge into the DL technique's learning pipeline;
- Evaluate the performance of each scenario developed in terms of enhancing DL technique knowledge;
- Highlighting the smart fusion approach for a precise extraction of maximum urban fabric detail;

The following paper is organized as described: Section 2 showcases the principal advancements made in fusion-based techniques for performing semantic segmentation on LiDAR point clouds. A detailed account of the fusion scenarios we developed is presented in Section 3. The experimental methodology and the obtained results are reported in Section 4. The discussion of our findings is reserved for Section 5. The paper concludes with a summary of our conclusions and proposed avenues for future research in Section 6.

2. Related Works

The increasing need for automated urban fabric extraction has resulted in 3D semantic segmentation of multi-sensor data becoming a rapidly growing and dynamic field of research. Although 3D urban semantic segmentation is indexed by 3D LiDAR data, other data sources (geometric features, classified images, etc) can also provide supplementary relevant information. The latter can compensate for the limits of 3D point clouds; such as the confusion between artificial and natural objects and the fact that point clouds are less suitable for delineating object contours. Promising results have been achieved in 3D semantic segmentation through the fusion of 3D point clouds with other data sources, as demonstrated by several studies in the literature (Ballouch et al., 2022a; Megahed et al., 2021; Weinmann and Weinmann, 2019). Furthermore, the addition of highly informative data is a major boost to semantic segmentation (Luo et al., 2016). At the same time, DL and advanced GPU (Graphics Processing Unit) techniques have significantly improved the performance of semantic segmentation processes. The DL revolution has demonstrated that many three-dimensional semantic segmentation challenges (the automation of traitements, their speed, the precision of results,..etc) are addressed by DL algorithms (PointNet++, SPGraph, etc). On the other hand, it is well known that more training labelled point clouds are required for learning models. Motivated by the high demand for training data, various datasets have been developed in recent years. The majority of them are freely available online. We can list , Toronto-3D (Tan et al., 2020), SensatUrban (Hu et al., 2021), Benchmark Dataset of Semantic Urban Meshes (SUM) (Gao et al., 2021), and Semantic3D (Hackel et al., 2017). The distributed datasets have boosted the scientific community, with an increasing variety of approaches being developed. They solved various challenges related to 3D semantic segmentation. Moreover, the created datasets are useful to inspire new methodologies and facilitate the evaluation of different approaches.

Despite the efforts made, 3D semantic segmentation remains a delicate and complex task due to the spectral and geometric similarity between different urban classes. Due to the remarkable performance achieved lastly by fusion approaches in semantic segmentation tasks, it would be interesting to advance in this research niche. The aim of this paper is to design a new fusion approach that is capable of semantically and automatically recognizing the objects. Our goal is to surpass the existing limitations of fusion approaches and automatically extract a maximum range of urban classes in an urban area with good accuracy.

Fusion-based approaches are applied by fusing data from different sensors at different fusion levels. As mentioned previously, four families of fusion-based approaches combine point clouds with other sources: 1) Prior-level fusion approaches 2) Point-level fusion approaches, 3) Feature-level fusion approaches, and 4) Decision-level fusion approaches.

2.1. Prior-level fusion approaches

One fusion approach involves fusing at the prior level, where classified images are assigned to the 3D point clouds, followed by applying a DL algorithm for the semantic segmentation of LiDAR data, as shown in Figure 1. This type of fusion approach has several advantages, including the ability to use semantic information directly from image classification, rather than depending on the original optical images. This leads to faster convergence and a lower loss function during both the training and testing phases. By integrating the results of optical image classification, the loss more quickly reaches a stable state and becomes smaller (Chen et al., 2021). However, the prior fusion methods suffer from the problems of non-overlapping regions and uncertainties (Oh and Kang, 2017).

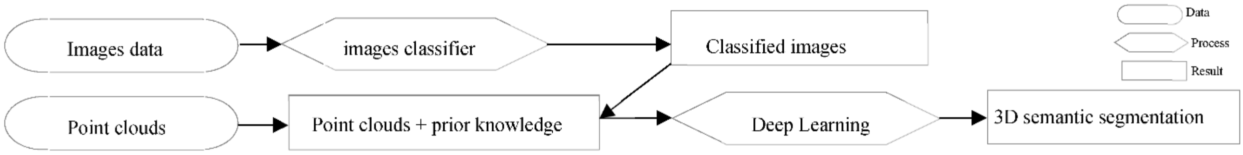


Figure 1. The general workflow of the prior-level fusion approaches.

There is a scarcity of prior-level fusion approach-based studies in the existing literature. Among them, (Chen et al., 2021) proposed a fusion approach of images and LiDAR point clouds for semantic segmentation. It provides a very low loss function in the training and testing steps. The proposed method was compared with point-level, feature-level, and decision-level fusion approaches. The evaluation using ISPRS dataset shows that the proposed method outperforms all other fusion methods with a good F1-score (82.79%). The research study of (Zhang et al., 2018) proposes a fusion methodology based on bidimensional images and 3D point clouds to segment complex urban environments. In this work, the prior knowledge obtained from bidimensional images is mapped to point clouds. Subsequently, the fine features of building objects are precisely and directly extracted from the point clouds based on mapping results. The assessment results show that the created model is adapted for high-resolution images and large-scale environments. Moreover, The used algorithm (FC-GHT) for refiningthe coarse segmentation allows to efficienctly extract the fine features of buildings. Finally, in a recent study by (Ballouch et al., 2022b), we presented a new fusion approach for semantic segmentation in urban areas, which operates at the prior level. This approach utilizes both aerial images and 3D point clouds, and employs an advanced DL architecture to improve semantic segmentation performance. In addition, we developed a solution to tackle the issue of inconsistent semantic labels found in both the image and LiDAR datasets. The results using the newly created dataset reveal that our approach outperforms the non-fusion process, achieving an Intersection over Union of 96% compared to 92%.

2.2. Point-level fusion approaches

Fusion approaches at the point level involve assigning spectral information from the optical image to each point, followed by utilizing a DL technique to perform semantic segmentation of the 3D point clouds with radiometric information, as illustrated in Figure 2. These approaches offer a number of advantages, mainly the good quality of results and the ease of use. However, despite their performance, this family has some disadvantages related to significant memory and computation time requirements. In addition, they require to have a simultaneous or minimal difference between the acquisition times of the two types of data (Ballouch et al., 2020).

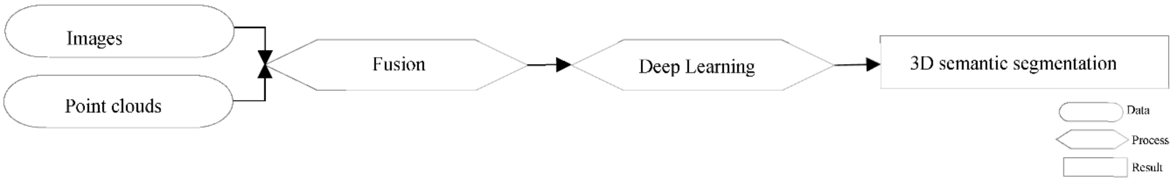


Figure 2. The general workflow of the point-level fusion approaches.

There is a large number of point-level fusion processes that have been developed for 3D semantic segmentation. Among them, (Poliyapram et al., 2019) have introduced a DL architecture, PMNet, that performs point-wise fusion of optical images and point clouds while considering the permutation invariance properties of point clouds. By acquiring different data from the same location and fusing them, the proposed approach enhances the 3D semantic segmentation. Evaluation results demonstrate that the proposed approach outperforms other techniques such as observational-level fusion, non-fusion approach, and global feature-level fusion. In the same direction, the authors in (Ye et al., 2022) have developed a methodology based on CNN (Convolutional Neural Network) for 3D semantic segmentation, which integrates radiometric properties from image data. The developed method was evaluated using the SemanticKITTI dataset. It demonstrates an average accuracy that is 8.7% higher in several classes compared to another process that integrates image and point clouds. In addition, the runtime was lower. In a study by (Luo et al., 2016b), the contribution of fusing Compact Airborne Spectrographic Imager (CASI) hyperspectral and airborne LiDAR data for land cover semantic segmentation was investigated. The data were fused using Principal Components Analysis (PCA) and layer stacking techniques, and the Maximum Likelihood (ML) and Support Vector Machine (SVM) classifiers were employed for data classification. The study reported a significant improvement in overall accuracies, 9.1% and 19.6% respectively, compared to the results obtained from using LiDAR and CASI data alone. Additionally, the study found that the SVM classifier showed higher potential than the ML classifier in semantic segmentation obtained by the fusion of CASI and LiDAR data. Finally, (Yousefhussien et al., 2018) proposed a DL framework to classify 3D point clouds with spectral information, which processes the data with varying densities directly without any prior transformation. Moreover, the proposed methodology has attained results near to the literature although it uses only XYZ coordinates and their corresponding spectral information. In addition, the proposed approach has demonstrated promising results when semantically segmenting non-normalized points.

2.3. Feature-level fusion approaches

In feature-level fusion techniques, the features extracted from optical images and 3D point clouds are concatenated using neural networks, as illustrated in Figure 3. The concatenated features are then processed using a Multi-Layer Perceptron (MLP) to obtain the semantic segmentation results (Ballouch et al., 2022b). The methodologies based on fused features from images and LiDAR data demonstrate robust results (Bai et al., 2015; Mirzapour and Ghassemian, 2015). Besides, they reach greater precision compared with the techniques solely using either the radiometric information from images or geometrical information from LiDAR data (Dong et al., 2018). Moreover, feature-level fusion allows objective data compression. It guarantees a certain degree of precision and retains enough important information (Man et al., 2015). Nevertheless, some drawbacks should not be ignored. For instance, the orthophoto presents the wrapping phenomenon. Hence, the orthophoto's features (such as texture) may not reflect the true objects. Additionally, the LiDAR data (DSM) that depicts the z-value of the tallest surfaces, cannot provide information about occluded objects (such as low-rise buildings) (Dong et al., 2018).

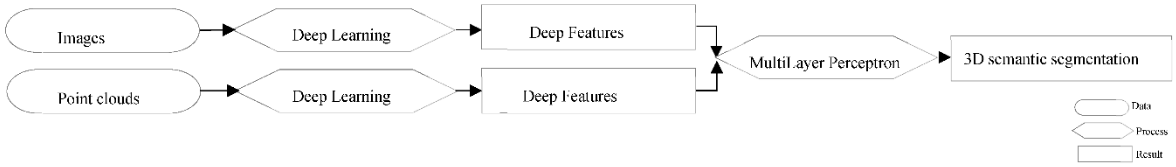


Figure 3. The general workflow of the feature-level fusion approaches.

The literature widely acknowledges that the semantic segmentation quality may benefit from convenient feature fusion. Among the methodologies developed in this sense, the authors (Mirzapour and Ghassemian, 2015) have utilized spectral information, texture, and shape features extracted from hyperspectral images to reduce classification errors. The proposed methodology is easy, i.e. only staking spectral and spatial features. This work concludes that is difficult to find a unique combination of the spatial and spectral features that give good results for all the datasets. Furthermore, it shows that the extraction of complementary features can yield good results even in a simple combinatorial process. Based on the obtained results, the study concluded also that integrating spatial information (shape features, texture, etc) in the process improves the semantic segmentation results. (Song et al., 2015) presented a learning-based super-resolution method that leverages the high spatial resolution of SPOT5 imagery and the wider swath width of Thematic Mapper / Enhanced Thematic Mapper images. The proposed technique involves using learned dictionary pairs to perform spectral and spatial fusion of these two types of data. The primary advantage of this method is that it enhances the spatial resolution of Thematic Mapper data, particularly in scenarios where there is no corresponding high-resolution SPOT5 data available. The semantic segmentation experiments were implemented using ground truth and fusion results. The objective is to demonstrate the advantages of the developed methodology for further semantic segmentation applications. Additionally, (Golipour et al., 2016) proposed a new methodology for integrating hierarchical statistical region merging segmentation of hyperspectral images with Markov Random Field (MRF) model. The objective is to provide a precise prior model in a Bayesian framework. The results obtained show that the proposed technique can create more homogeneous regions similar to MRF-based techniques. Besides, as segmentation-based techniques, it preserves precisely the class boundaries. The additional computational burden of the proposed hierarchical segmentation step is negligible considering its enhancement that provides the classification results. Finally, In a study conducted by (Bai et al., 2015), a feature fusion method was presented for classification tasks that utilized softmax regression. This method took into account the likelihood of an object sample belonging to different classes and incorporated object-to-class similarity information. The results of the experiment demonstrated that the proposed technique outperformed other baseline feature fusion techniques such as SVM and logistic regression. Specifically, the method exhibited a superior ability to measure feature similarity across multiple feature spaces, showcasing the effectiveness of the softmax regression-based approach.

2.4. Decision-level fusion approaches

Les approches de fusion au niveau décisionnel se distinguent des techniques de fusion au niveau des caractéristiques car elles utilisent un processus spécifique pour apprendre la couche de fusion finale en utilisant les résultats de la segmentation sémantique, comme le montre la Figure 4. The latter are generated by various single neural networks (Zhang and Chi, 2020). Decision-level fusion approaches combine the outputs of two classifiers that each work on LiDAR space or either pixel. Precisely, on the one hand, an algorithm processes the spectral information and produces a semantic segmentation of the images. On the other hand, the LiDAR data are also segmented semantically. Afterwards, the two types of results are fused using a fusion algorithm as the heuristic fusion rule (Chen et al., 2021). The decision-level fusion approaches have several advantages, such as the non-interference of the two-classification processes. This means that, the two processes are trained and validated independently. So, this type of fusion has good flexibility and low-complexity. Moreover, It can achieve a good performance; for the reason that each modality is employed to train a single DL technique. So, this allows us to learn the representation of independent features (Zhou et al., 2019). However, the fact that the approach draws on the prior decisions achieved by the two classifiers, can be affected by their shortcomings. Besides, while it can achieve a modest improvement, it is still limited by other fusion approaches' performance. Moreover, decision-level fusion approaches necessitate more memory due to the fact that the DL structure combines the features at a later stage. Additionally, these techniques require supplementary parameters for convolutional layers and other operations (Zhou et al., 2019).

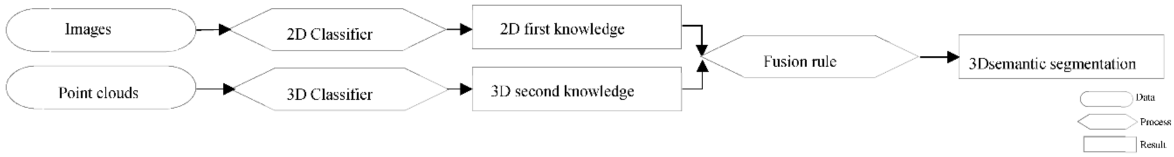


Figure 4. The general workflow of the decision-level fusion approaches.

The existing literature includes only a few decision-level fusion techniques. Among them, (Oh and Kang, 2017) developed a fusion approach for classification and object-detection. The semantic segmentation results from unary classifiers were fused using a CNN. Experimental results using the KITTI benchmark show that the proposed methodology acheives an average precision of 77.72%. Despite its performance, this approach presents some limitations, including the difficulty of generating the method in real time. In addition to the low accuracies obtained in the two classes “cyclists” and “pedestrians” due to the incomplete data obtained from the sensor. In the same direction, (Tabib Mahmoudi et al., 2015) have proposed a fusion approach of the object-based image analysis on multiviews very-high resolution imagery and DSM. The proposed methodology enhanced the precision and completeness of the object recognition results. The minimum values of improvements obtained in kappa and overall accuracy metrics are about 18 and 0.24 for the DMC benchmark, with 15 and 0.19 for the WorldView-2 benchmark respectively. Despite the significant results obtained, the correctness values of the approach results have not improved in all DMC benchmark classes. Finally, a study conducted by (Zhang et al., 2015) introduced a late fusion method that combines information from multiple modalities. The approach includes a pairwise CRF (Conditional Random Field) to enhance the spatial consistency of the structured prediction in a post-processing stage. The KITTI dataset was used for evaluation, and the findings indicate that the approach achieved an average class accuracy of 65.4% and a per-pixel accuracy of 89.3%.

2.5. Summary

Previous research has highlighted the effectiveness of semantic segmentation methods that leverage point clouds in combination with other data sources, such as satellite or aerial images, demonstrating high levels of accuracy and quality in the resulting visual outputs. In the literature, the commonly used fusion approaches of 3D LiDAR and image data can be categorized into four main types as mentioned before: prior-level, point-level, feature-level, and decision-level fusion approaches. The prior level approaches are the new fusion approaches in the literature. They have enhanced the accuracy of semantic segmentation results. Additionally, they demonstrated good performances in semantic segmentation, especially in terms of precision. The latter was improved by the direct use of semantic knowledge from classified images. Moreover, they demonstrated the low loss function in training and testing steps. Thus, the fact that this approach type integrates semantic information from images, the loss reaches a stable state faster and becomes smaller. However, these processes are a bit long. The point-level fusion approaches are the most dominant, quickest, and simplest fusion approaches in the literature. However, these processes are not able to classify complex urban scenes containing a diversity of urban objects. Especially, the geo-objects with geometric and radiometric similarity. The feature-level fusion approaches allow objective data compression. Consequently, they guarantee a certain degree of precision and retain enough important information. Nevertheless, the features extracted sometimes don't reflect the real objects. The decision-level fusion approaches are less complex and flexible. For the reason that the two semantic segmentation processes (one of the image and the other of the point clouds) do not interfere. Nonetheless, these approach types can be affected by errors in both processes. In addition, decision-level fusion approaches require more memory since the DL structure fuses feature later. Besides, the extra-parameters are needed for layers to perform convolution and other operations. The performance and limitations of each method are summarised in the following table (Table 1).

Table 1. Performances and limitations of the different fusion approaches.

Fusion approach	Performances	Limitations
Prior-level	-Direct use of semantic information from images -Fast convergence -Low loss function -High classification accuracy.	-Problems of non-overlapping regions and uncertainties -Bit long process
Point-level	-Fast drive -Easy handling -No prior information is required.	- High cost - Not able to classify diversified urban contexts - Relatively low classification accuracy
Feature-level	-Objective data compression -Retaining enough important information	-Training loss higher -Features may not reflect the real objects.
Decision-level	-Non-interference of the two semantic segmentation processes -Good flexibility -Low-complexity	-Impacted by the shortcomings of both classifiers. - Additional parameters for layers are required

	-Learning the representation of independent features is allowed	- More memory requirement

In this research, we are interested in the extraction of 3D urban objects with high geometric accuracy and good semantic richness. Based on the literature review, the prior fusion approaches have demonstrated good results compared to those of the others. This conclusion opens the way for further developments in this family. Therefore, we proposed and thoroughly evaluated three prior level fusion scenarios with the aim of deriving the smart fusion approach, which enhances the knowledge of the DL technique. To implement this, we injected an aerial image and a given knowledge K as point clouds attributts into the learning pipeline, where K comprises prior knowledge obtained from an aerial image or a point cloud, including classified images,geometric features,etc. Each time, we incorporate a type of knowledge into one of the scenarios to derive a semantic segmentation scenario that effectively integrates high-weighted, positively influential neural network knowledge.

3. Materials and Methods

3.1. Dataset

Our developed scenarios were evaluated using the SensatUrban dataset (Hu et al., 2021), which contains nearly three billion annotated points and was released at CVPR2021. The utilization of this dataset is justified by its high semantic richness compared to other existing airborne datasets. The 3D point clouds were obtained by UAV (Unmanned Aerial Vehicle) which follows a double-grid flight path. Three sites of Birmingham, Cambridge, and York cities were covered. The dataset covers about six square kilometers of an urban area with a diversity of urban objects. The SensatUrban dataset contains 13 different semantic classes, which are street furniture, traffic road, water, bike, footpath, car, rail, parking, bridge, wall, building, vegetation, and ground (Figure 5). The point clouds are manually labeled using the CloudCompare tool. Each point contains 6 attributes, which are X, Y, Z, and RGB information. The allocation of semantic categories to objects within the dataset is extremely imbalanced, with the bike and rail classes collectively accounting for just 0.025% of the overall points present in the dataset.

The SensatUrban dataset is freely available online at (https://github.com/QingyongHu/SensatUrban, accessed on 10 March 2023). However, it should be noted that the dataset's semantic labels for the testing data are not provided. Thus, to evaluate the proposed approach, the training data of SensatUrban was partitioned into new training and testing sets. The dataset utilized in the present study consisted of 34 blocks from various cities, all of which were labeled. In our experiments, a part of the training data (18 sets) was used to implement the first parts of the developed scenarios S1 ＆ S3 (Section 3). While the remaining part of the data (16 sets) was used to implement the second steps of scenarios S1 ＆ S3, the S2 and the reference approach (the main part of this work).

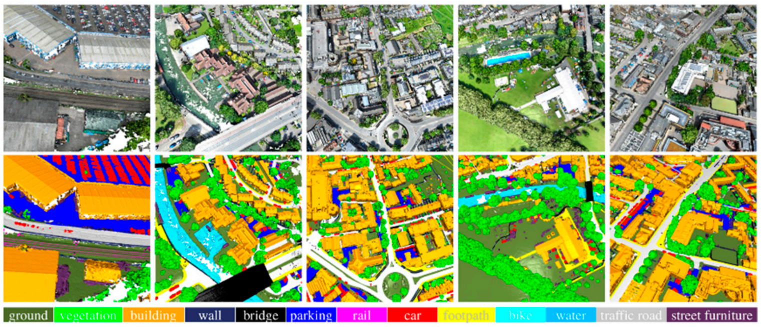


Figure 5. Classes of the SensatUrban dataset (Hu et al., 2021).

doi:10.20944/preprints202306.2022.v1

3.2. RandLaNet Deep Learning Algorithm

For this study, we employed the RandLaNet algorithm (Hu et al., 2020) which is currently regarded as a cutting-edge deep learning model for semantic segmentation. Specifically designed for large-scale point clouds, the algorithm is both lightweight and efficient (Guo et al., 2021), utilizing random point sampling as a learning method to achieve superior computational and memory efficiency. Furthermore, it does not require any preprocessing or postprocessing operations. Additionally, to preserve and capture geometric features, a local feature aggregation module was introduced.

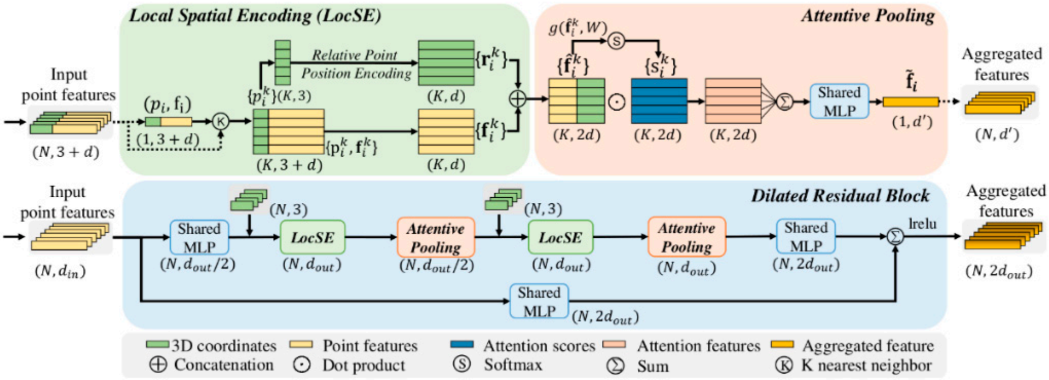


Figure 6. RandLaNet Algorithm(Hu et al., 2020).

Due to its high level of accuracy, the RandLaNet model has demonstrated exceptional efficiency in numerous applications involving airborne LiDAR systems. We can cite the classification of urban tissues, in which it reached higher results, as demonstrated by several recent research studies (Ballouch et al., 2022b; Hu et al., 2021). The performance of the RandLaNet algorithm has been evaluated on various datasets including, Semantic KITTI, Semantic 3D, S3DIS, etc. showing good qualitative and quantitative results (Hu et al., 2020).

The RandLaNet algorithm was trained six times. First, to run the first part of S1. Second to run its second part (Section 3) . The third is to run S2. The fourth is to run the first part of S3. Then, the fifth is to run its second part. The latter is to run the baseline approach. To ensure a fair comparison of the various methodologies, identical hyperparameters of the algorithm were utilized during implementation.

3.3. Methodology

Our study aims to evaluate three prior level fusion scenarios to determine which one leads to the maximum improvement of the semantic richness extracted from urban environment. The objective of this work is to find a smart fusion approach for the accurate extraction of semantically enriched 3D urban objects. Such extractions can be performed using airborne sensor data (Lidar and spectral) by employing DL techniques of artificial intelligence. The smart approach was derived based on a deep evaluation of the different developed scenarios. This research aims to provide new guidance to users and offer an operational methodology for the adoption of the derived approach.

The general work methodology for semantic segmentation consists of four processes, as depicted in Figure 7. The first fuses the classified images and aerial images with point clouds (S1). The second is based on geometrical features (extracted from point clouds), XYZ coordinates, and aerial images (S2). The third classifies urban space using classified geometrical information, aerial images, and point clouds (S3). The fourth process, known as the reference approach (RA), directly combines raw point clouds and images. The reference approach is utilized for the purpose of comparing it with the three proposed scenarios. Afterwards, the advanced DL algorithm “RandLaNet” was adopted to implement the different processes (developed scenarios + reference approach). An assessment of the results obtained by the different processes is performed based on metrics computation and visual investigations. Finally, the smart fusion approach was derived based on the obtained qualitative and quantitative results.

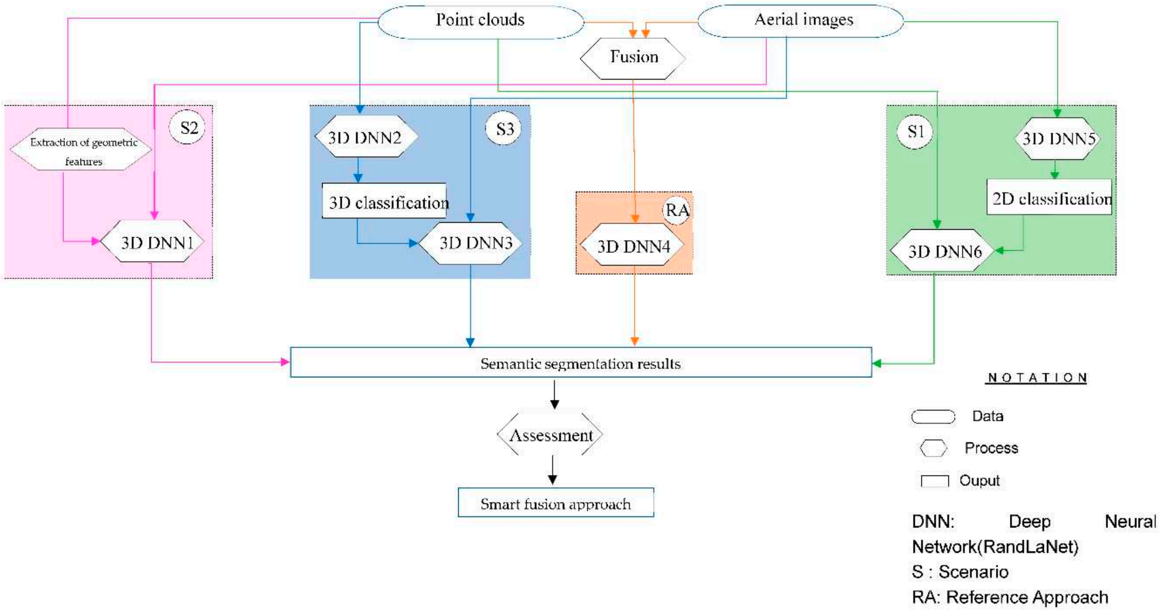


Figure 7. The general workflow.

For aerial image integration, we averaged the Red, Green and Blue bands to obtain a single column in the attribute table for each cloud. Our aim is to propose less costly scenarios with fewer attributes. While spectral data is present in all scenarios, we more rely on the attributes that differ between the developed scenarios, as our aim is to identify the most performing scenario rather than to maximize the results' accuracy.

3.3.1. Classified images and point clouds based scenario (S1)

The flowchart depicted in Figure 8 summarizes the first proposed scenario (S1), which is a prior level scenario utilizing 3D point clouds, aerial images, and classified images. The incorporation of aerial images into the point cloud has already been justified. However, the injection of classified images, spectral information as attributes of point clouds into the DL technique during its training is justified by several reasons. Firstly, integrating classified images brings a semantic dimension to the scenario. Classification images provides detailed information about different object categories present in the urban scene. This prior knowledge enhances the neural network's knowledge during the learning pipeline. Furthermore, it can be valuable in guiding object detection and semantic segmentation by reducing false negatives and false positives. By leveraging this semantic information, this scenario can achieve more consistent results in object identification. Furthermore, this scenario derives additional benefits from leveraging semantic knowledge directly obtained from image classification. This heightened reliability accelerates the convergence of this scenario, resulting in enhanced the precision of urban object extraction.

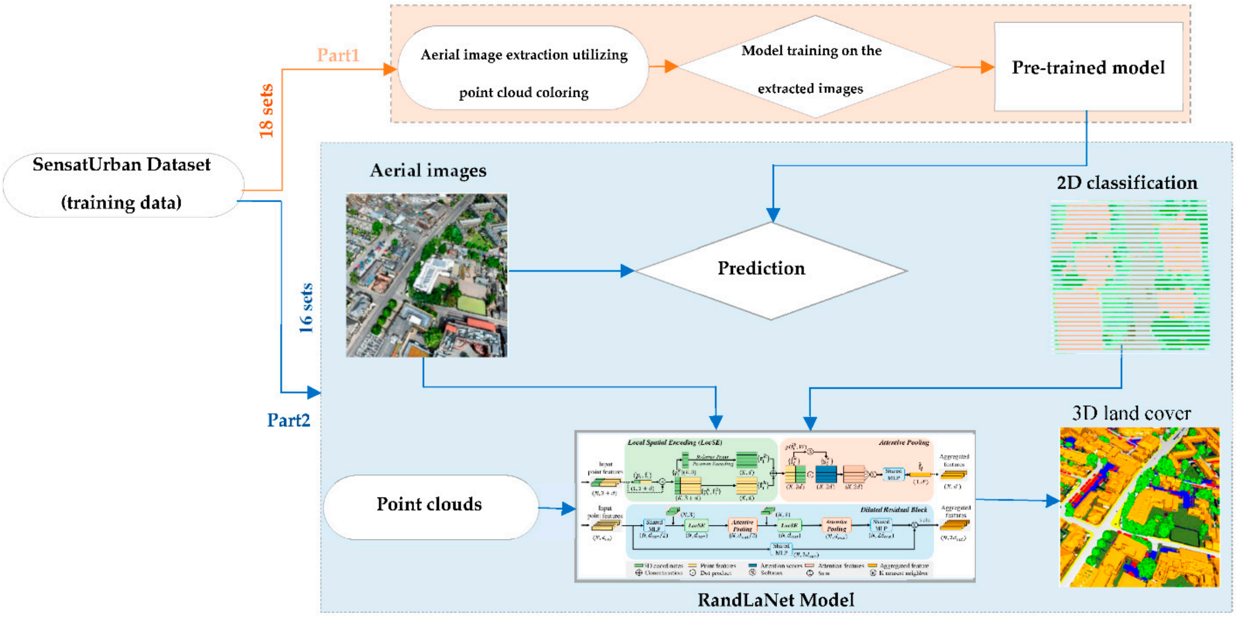


Figure 8. The first proposed scenario (S1).

To implement this scenario, we randomly divided the SensatUrban dataset into two parts: one containing 18 point clouds and the other containing 16. First, the aerial images used to colorize the point clouds were extracted from the dataset with 18 clouds. Then, the RandLaNet model was trained on these images, integrating them into RandLaNet along with their attributes :X, Y, Red, Green, Blue, and labels. In this work, we chose to use RandLaNet for 2D image classification instead of advanced 2D semantic segmentation networks, for several reasons. Primarily, the classes in our dataset differ from those in publicly accessible image datasets. Our goal in this work is twofold: to preserve the semantic richness of the SensatUrban dataset, i.e., its 13 classes, and to use the same classes to ensure a fair comparison with other scenarios developed in this work. After obtaining the trained model, we returned to the dataset containing 16 point clouds and extracted the images used to colorize these clouds in the same manner. We then classified them using the pre-trained model and merged them with the point clouds (XYZ coordinates) and aerial images. Thus, each point cloud contains the following attributes: XYZ coordinates, aerial image, classified image, and semantic label. Finally, these prepared point clouds were used to train the RandLaNet model. The fundamental hyperparameters of the original version of the algorithm have been preserved, and the algorithm was evaluated using the test data.

3.3.2. Geometric features, point clouds, and aerial images based scenario (S2)

Finding the best attributes to inject into the DL technique during training is a crucial element that can affect the accuracy of semantic segmentation. The idea of the second proposed scenario is to combine the geometric feature produced from point clouds, the original point clouds, and aerial images. The aim of this scenario is to examine the contribution of geometric properties to improving knowledge of the DL technique in semantic segmentation pipeline. As shown in Figure 9, S2 mainly contains two parts:

(A) Selection of the proper geometric features for the semantic segmentation based on the visual inspection;

(B) Injection of selected geometric features with aerial images as point clouds attributes to improve knowledge of the RandLaNet model;

doi:10.20944/preprints202306.2022.v1

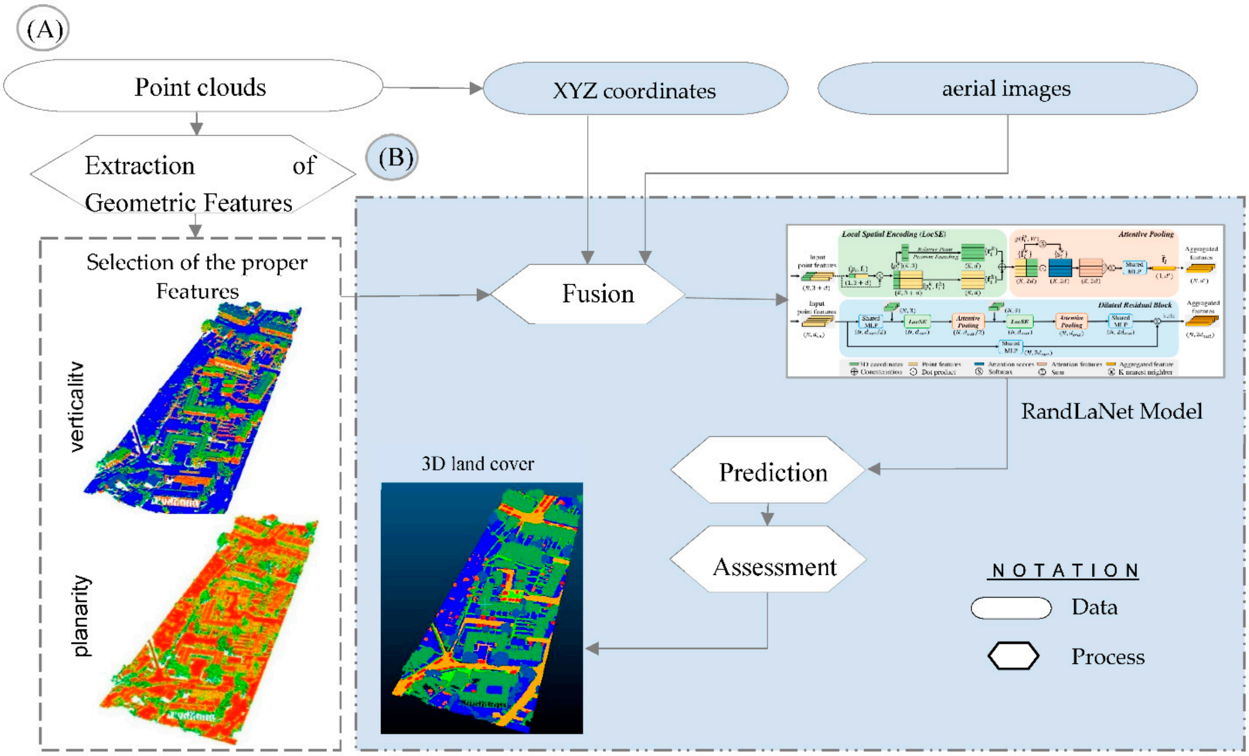


Figure 9. The second proposed scenario (S2).

A) Selection of the proper geometric features

The use of geometric features can help elucidate the local geometry of point clouds and is now commonly employed in 3D point clouds processing. By extracting these properties at multiple scales instead of a single scale, the aim is to improve precision values. "Geometric features are calculated by the eigenvalues ($\lambda 1$, $\lambda 2$, $\lambda 3$) of the eigenvectors (v1,v2, v3) derived from the covariance matrix of any point p of the point cloud"(Atik et al., 2021):

$$cov(S) = \frac{1}{S}\sum_{p\epsilon S}(p - \bar{p})(p - \bar{p})^T$$

"where p is the centroid of the support S"(Atik et al., 2021). Several properties are calculated using eigenvalues: omnivariance, the sum of eigenvalues, eigenentropy, linearity, anisotropy, planarity, surface variation, verticality, and sphericity.

Geometric feature extraction is a crucial part of 3D semantic segmentation. Specifically, the exploitation of the extracted features in the semantic segmentation process improves the quality of results. Independently from the urban object to be semantically segmented and the data resolution, the geometric properties have a significant impact on the results. The geometric features have great importance by providing the DL structure with useful information about each urban class. Consequently, it helps the classifier to distinguish between different semantic classes. However, these geometric properties may mislead the semantic segmentation process. So, these errors should be considered during the analysis of results.

To select the proper geometric properties for the semantic segmentation, firstly, all geometric features were calculated (anisotropy, planarity, linearity…etc). After that, two proper ones were selected from the experimental results through visual inspection. Precisely, planarity and verticality were selected to integrate them as the attributes of point clouds thanks to their satisfactory visual quality. The geometric features having the least impact on the model training have been removed. The following are the geometric properties used in this study:

doi:10.20944/preprints202306.2022.v1

- Planarity is a characteristic that is obtained by fitting a plane to neighboring points and computing the average distance between those points and the plane (Özdemir and Remondino, 2019).
- Verticality: The angle between the xy-plane and the normal vector of each point is calculated using its 3D surface normal values (Özdemir and Remondino, 2019).

B) Data Training and Semantic Segmentation Based on RandLaNet

In addition to the selected geometric properties (planarity and verticality), aerial images and point clouds have also been used in this scenario as attributes for implementing the RandLaNet algorithm. To implement this scenario, we started with the preparation of the training data. As mentioned earlier, we divided our dataset into two parts. One of these parts contains 18 point clouds, while the other contains 16. In the case of this specific scenario, we worked only with the set that contains 16 point clouds. These are the same point clouds that are used to implement the second part of the other scenarios proposed in this work. The generation of training data is performed by calculating the geometric features (planarity and verticality) for each point cloud. The calculations were done using the Cloud Compare tool. Afterward, these geometric properties were merged with point clouds and aerial images. This data preparation methodology is applied to all the point clouds in the 16 datasets used.

The following illustrates the data preparation principle according to the formalities of the proposed scenario:

Point cloud 1 (X1 + Y1 + Z1 + planarity 1 + verticality 1+ average RGB1) (1)
Point cloud 2 (X2 + Y2 + Z2 + planarity 2 + verticality 2+ average RGB2) (2)
•
•
•
•
•
•
Point cloud N (Xn + Yn + Zn + planarity n + verticality n+ average RGBn) (N)

Afterwards, during the training step, certain parameters of the original RandLaNet version were adjusted including input tensor, data types, etc. However, the basic hyper-parameters (number of layers, learning rate,..etc) are kept the same. Finally, after training and validation of the RanLaNet algorithm, the pre-trained model was used to predict the labels of the test data.

3.3.3. Classified geometrical information, point clouds, and optical images based scenario (S3)

We intend to explore a novel scenario that also has not been previously examined in the literature. The proposed scenario (Figure 10) involves injecting classified geometric data (obtained solely based on XYZ coordinates), and radiometric information extracted from aerial images as attributes of point clouds into the DL technique's learning pipeline. Since the contours of the different objects in the point clouds are not precise, the use of point clouds only in semantic segmentation may be insufficient, due to the confusion between some semantic classes. To address this challenge, we decided to incorporate a classification of geometric information and aerial images as attributes of point clouds. This integration into DL technique's training would enableit to learn and enhance the delineation of 3D object contours more effectively. As a result, it becomes easier to differentiate between different objects. Furthermore, a low loss function in the training step is also expected. The desired objective is to improve model learning. Therefore, the confusion between classes could be reduced and the semantic segmentation results improved.

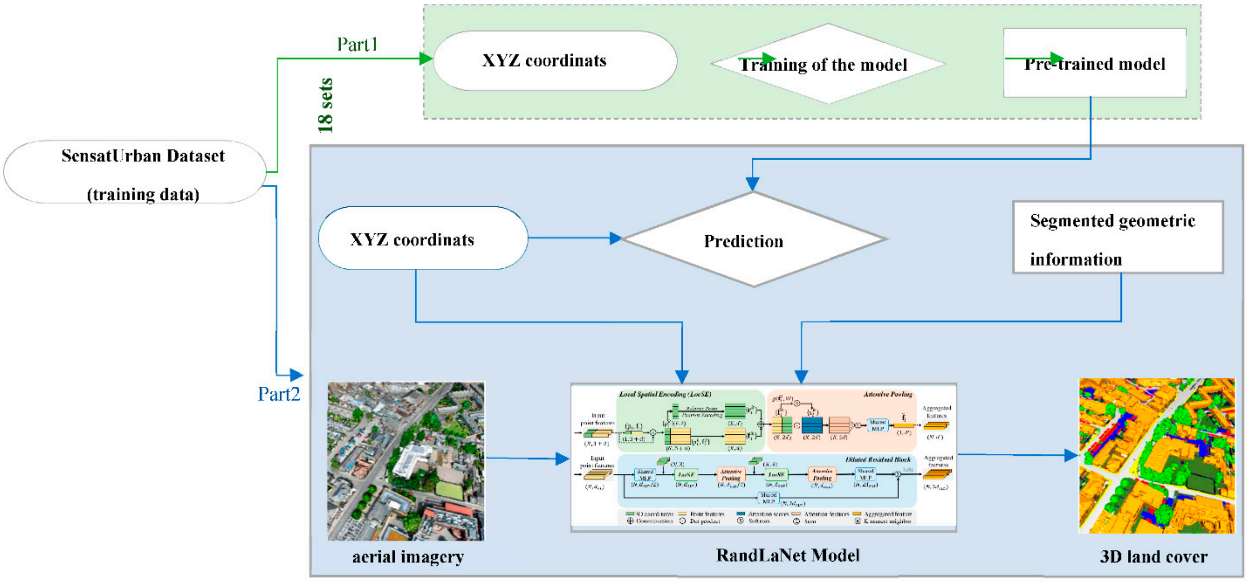


Figure 10. The third proposed scenario (S3).

For the implementation of S3, 18 sets of the SensatUrban dataset has been used to perform the first part of the scenario and 16 sets to perform its second part (see Figure 10). The proposed process includes two main steps. Firstly, 18 sets of point clouds that contain only the three attributes X, Y, and Z (additional attributes were eliminated) from the SensatUrban dataset are used to train the RandLaNet algorithm. After that, the pre-trained model is used to predict all point clouds (that contain also only XYZ coordinates) from the rest of the dataset (The part of the dataset that contains 16 clouds). The segmented geometric information obtained was considered as prior knowledge to obtain refined semantic segmentation results. Secondly, this prior knowledge is assigned to point clouds (XYZ+ aerial image) based on their coordinates. The same process of data preparation was followed to prepare all point clouds from 16 sets of the dataset. The merged data is then used to train the RandLA-Net model. The fundamental hyperparameters of the original version of the algorithm have been retained. However, the basic input tensor was modified into several channels as follows: XYZ coordinates, aerial image in addition to the classified geometric information. Finally, the pre-trained model is utilized to predict the test data, which is prepared in the same manner as the training data, in order to evaluate the model's performance.

3.3.4. Reference approach

The reference approach is a point-level fusion approach that directly combines geometric and radiometric information. It involves running the RandLaNet algorithm using the following attributes: XYZ coordinates and aerial images. The data used in the developed scenarios within the scope of this work differ from those of this reference approach (Hu et al., 2021). However, we compared them to better understand how the developed scenarios have improved the results of semantic segmentation compared to the reference approach that employs the most commonly used fusion method in the literature.

The reference approach includes two parts. The first one is the assignment of spectral information from images to point clouds. While the second one is the adoption of RandLaNet to classify the LiDAR point clouds with spectral information. Figure 11 summarizes the general process followed for the implementation of the reference approach.

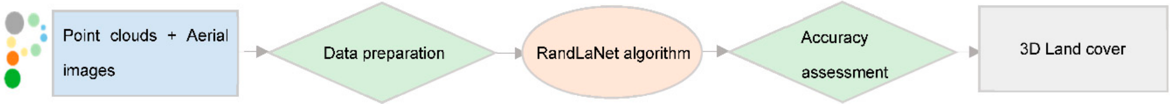


Figure 11. The general workflow of the reference approach.

To perform the RandLaNet technique, the same methodology of the existing approach (Hu et al., 2021) was followed with a slight difference. In our case, we used only 16 sets of the SensatUrban dataset to ensure a fair evaluation, similar to the developed scenarios. Additionally, we utilize the average of RGB instead of three separate columns containing the R, G, and B bands. That is, the basic input tensor was modified as follows: X, Y, Z, and Average RGB.

4. Experiments and Results Analysis

4.1. Implementation

The calculations for the study were carried out using Python programming language version 3.6, with Ubuntu version 20.04.3 as the operating system. Cloud Compare version 2.11.3 was used to calculate geometric properties and average RGB values, where geometric features were computed with a 0.4 m radius sphere representing support obtained with a radius of 4 m. Tensorflow-GPU v 1.14.0 was used as the code framework to implement the RandLaNet algorithm, with CUDA version 11.4 utilized to accelerate deep learning through parallel processing power of GPUs. All experiments were conducted on an NVIDIA GeForce RTX 3090, and a workstation with 256G RAM, a 3.70 GHz processor, and Windows 10 Pro for workstations OS (64-bit) was used for data processing. Furthermore, Scikit-learn, a free Python machine learning library, was employed to implement various processes, where optimized parameters remained unchanged throughout the study.

The RandLaNet algorithm is publicly available on GitHub at https://github.com/QingyongHu/RandLA-Net (accessed on March 10, 2023). The original version of the code was used to train and test the algorithm. For each scenario, the algorithm was adapted and trained six times using the provided data, and the hyperparameters were kept constant. The Adam optimization algorithm (Kingma and Ba, 2017) was used for training with an initial learning rate of 0.01, an initial noise parameter of 3.5, and a batch size of 4. The model was trained for 100 iterations, and all layers were included in the training. Every training process passes through two stages. The first is a forward pass, which deduces the prediction results and compares them with ground truth to generate a loss. While the second is a backward pass, in which the network weights are then updated by stochastic gradient descent. The obtained pre-trained networks were used for the prediction of the blocks selected to test all processes. Consequently, a semantic label was assigned for each cloud point.

In order to assess the efficacy of the suggested methodologies, we adopted five metrics: Precision, Recall, F1 score, Intersection over Union, and Confusion Matrix. Precision (1) gauges the percentage of points identified as positive in semantic segmentation. Recall (2) evaluates the proportion of true positives in relation to all actual positive instances. F1 score (3) represents the harmonic mean of precision and recall. Intersection over union (4) quantifies the extent of overlap between predicted and actual results. Evaluation of these metrics was conducted on Google Colaboratory.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}} \tag{2}$$

$$\text{F1} - \text{score} = \frac{2\,(\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}} \tag{3}$$

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{FP} + \mathrm{TP} + \mathrm{FN}} \tag{4}$$

"False positive (FP) refers to the number of points that are predicted positive but their actual label is negative. False negative (FN) refers to the number of points that are predicted negative and their actual label is positive" (Atik et al., 2021). A true positive (TP) is determined by the number of points that have identical semantic labels in both the ground truth and the predicted results.

The confusion matrix is also a commonly used metric for the quantitative evaluation of semantic segmentation results. "It is a good indicator of the performance of a semantic segmentation model by measuring the quality of its results. Each row corresponds to a real class; each column corresponds to an estimated class" (Ballouch et al., 2022b). Table 2 below summarizes the set of indicators that can be derived from the confusion matrix.

Table 2. The confusion matrix principle.

		ground truth	
		-	+
Predicted class	-	*True Negatives*	*False Negatives*
	+	*False Positives*	*True Positives*

4.2. Results and Discussion

To show the semantic segmentation results of the developed scenarios more prominently, this section compares them using the RandLaNet algorithm. The performances of the proposed prior-level fusion scenarios were evaluated using the SensatUrban dataset. Moreover, the results have been compared with those of the reference approach. The study compared the efficiency of the developed scenarios by evaluating the main metrics namely, F1-score, Recall, Precision, Intersection over Union (IoU), and Confusion matrix. In addition, a qualitative results evaluation was made based on visual analysis of predicted (synthetic) and observed (actual) data. This section presents a detailed demonstration of the results, including a comparative analysis with the reference approach, as well as between the various developed scenarios.

4.2.1. Quantitative Assessments

In this subsection, we evaluate the scenarios S1, S2, and S3 with the reference approach using test set data. The comparisons are reported in Table 3. Since several scenarios were evaluated in this work, the same data splits were used for the RandLaNet algorithm' training, validation, and test to ensure a fair and consistent evaluation. Four urban scenes (4 test sets) were used to evaluate the pre-trained models and did not contribute to the training processes. We can see that all developed scenarios outperform the reference approach in all evaluation metrics. A comparison of the three developed scenarios shows that scenario S1 surpassed S2 and S3 scenarios in all indicators. The experimental results show that the first proposed fusion scenario (S1) delivers the best performance over other scenarios, which is manifested mainly in the higher IoU and highest precision in the semantic segmentation results. For example in scene 1, the Intersection over Union of S1, S2, S3, and the reference approach was 80%, 77%, 75%, and 63% respectively. Table 3 exhibits the semantic segmentation accuracies for the SensatUrban dataset as obtained using the different scenarios and the reference approach.

Table 3. Quantitative results of the proposed scenarios and the reference approach.

	Processes	F1-score	Recall	Precision	IoU
Scene 1	Reference approach	0.71	0.77	0.71	0.63
	S1	**0.87**	**0.87**	**0.88**	**0.80**
	S2	0.85	0.86	0.85	0.77
	S3	0.83	0.84	0.84	0.75

Scene 2	Reference approach	0.82	0.86	0.79	0.75
	S1	**0.93**	**0.92**	**0.94**	**0.88**
	S2	0.92	0.91	0.92	0.86
	S3	0.90	0.90	0.91	0.85
Scene 3	Reference approach	0.75	0.78	0.74	0.67
	S1	**0.86**	**0.85**	**0.88**	**0.79**
	S2	0.84	0.83	0.87	0.77
	S3	0.83	0.82	0.86	0.76
Scene 4	Reference approach	0.61	0.68	0.58	0.50
	S1	**0.80**	0.78	**0.84**	**0.68**
	S2	0.79	0.78	0.82	0.67
	S3	0.70	0.72	0.76	0.57

The proposed fusion scenarios in this paper show a significant improvement for all indicators compared to the reference approach. The first developed approach (S1) has obvious advantages but the difference between it and S2 is relatively small. From the results of each metric, we can see that the S1 scenario achieved 88/80%, 94/88%, 88/79%, and 84/68% semantic segmentation Precision/IoU in the four urban scenes. Compared to the reference approach, the S1 scenario which integrated classified images, RGB information, and point clouds, increases the semantic segmentation IoU of each scene by 17%, 13%, 12%, and 18%, respectively. Also, the S2 scenario integrated geometrical features, optical images, and point clouds; it increases the semantic segmentation IoU of each scene by 14%, 11%, 10%, and 17%, respectively. Additionally, the S3 scenario integrated classified geometrical information, optical images, and point clouds; it increases the semantic segmentation IoU of each scene by 12%, 10%, 9%, and 7%, respectively. The poor precision obtained by the reference approach could be explained by the lack of prior information from images or point clouds (geometric features), which could provide useful information related to urban space. Therefore, it is difficult to obtain accurately diversified objects' semantic segmentation by the direct fusion of LiDAR and optical image data. On the other hand, the S1 scenario has advantages over both the scenarios with geometric features (S2) and with classified geometrical information (S3). The results obtained by the S1 scenario indicated that the integration of prior knowledge from images (image classification) improves the 3D semantic segmentation. It improved the semantic segmentation precision to around 94% for example in scene 2. Additionally, with the help of prior knowledge from classified images in S1, we achieve about 17% increase in overall precision compared to the reference approach. The increased accuracies obtained by S1 can be attributed to complementary semantic information from the optical images. That is, the usage of classified images and spectral data together with the point clouds (XYZ coordinates) improved the accuracy. They are useful to facilitate the distinction of diverse objects. Therefore, we can make the satisfactory conclusion that the overall performance of S1 is a promising scenario by considering the different evaluation metrics.

The results obtained with different developed scenarios have been studied in detail by computing a percentage-based confusion matrix using ground truth data. “This percentage-based analysis provides an idea about the percentage of consistent and non-consistent points”(Ballouch et al., 2022b). The percentage-based confusion matrix obtained by all scenarios for scene 1 is depicted in Figure 12. The corresponding confusion matrices for the other urban scenes (2, 3, and 4) are available in the Appendix A.

It can be seen from confusion matrices that the developed scenarios have a real improvement compared to the reference approach. The improvement is more significant. This is because the direct fusion of images and point clouds is not sufficient for the semantic segmentation of complex urban scenes.

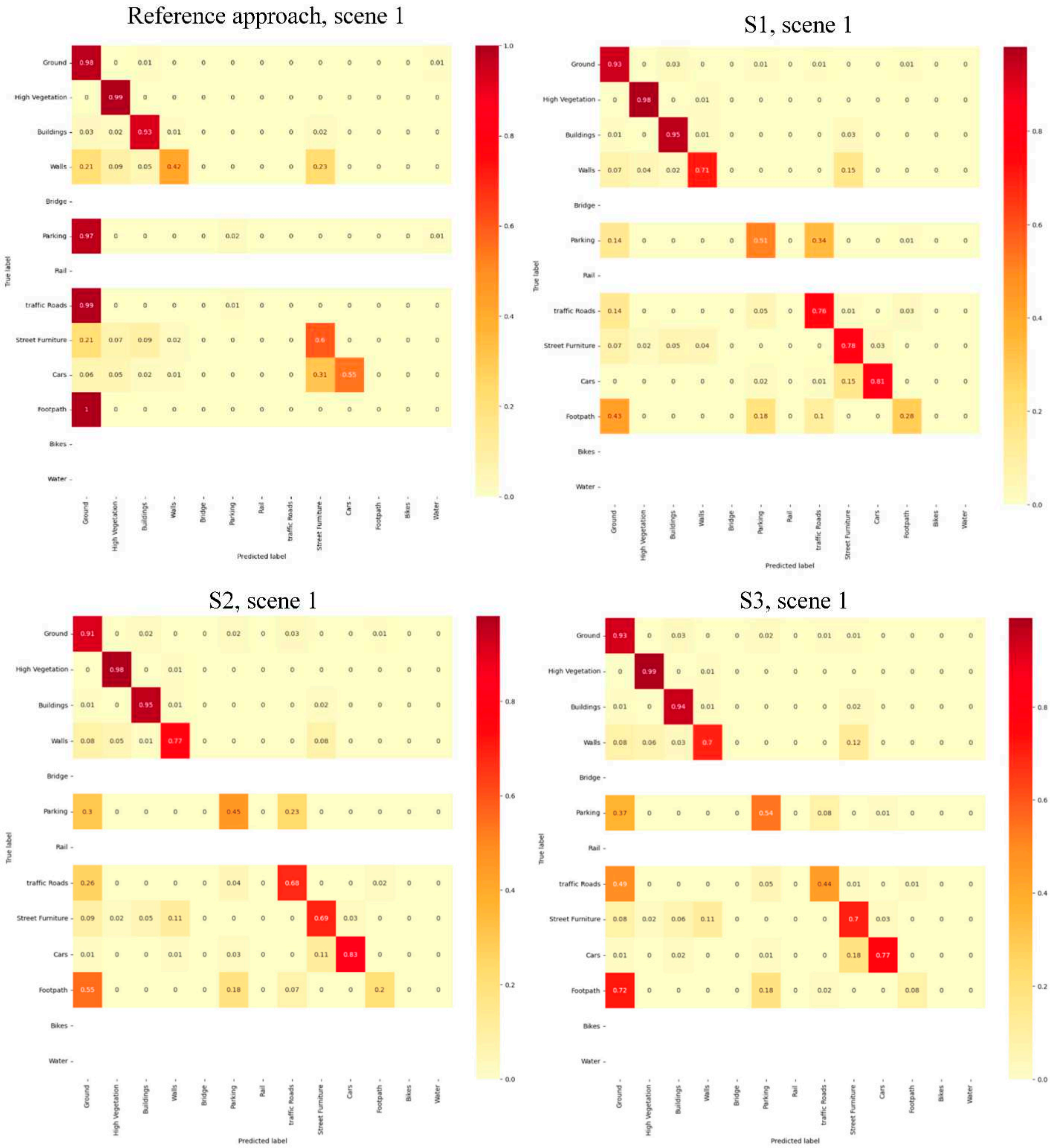


Figure 12. Normalized Confusion Matrix for Proposed Scenarios and the Reference Approach in Scene 1.

The obtained results allow general inferences to be made regarding the selection of a smart fusion approach for DL-based point clouds semantic segmentation. The following are the detailed results of each class or classes group independently:

Firstly, ground and high vegetation classes were successfully extracted in all scenes with all evaluated processus. It is due to their geometric and spectral characteristics which are easy to recognize. That is, they are easily distinguished from other classes. This means that only the point clouds and the optical images fused in the reference approach are sufficient to correctly segment the two classes.

Secondly, the building class is extracted accurately by the first proposed scenario (S1), but the difference between it and other developed scenarios is relatively small. However, despite its performance, a slight confusion was observed between this class and the street furniture object.

doi:10.20944/preprints202306.2022.v1

Thirdly, by observing the four scenes, we can see that the S1 scenario has a good performance on the point clouds scenes containing rail, traffic roads, street furniture, footpath, and parking objects. The five semantic classes were extracted precisely by this scenario, except for the footpath class, the precision of it was low. Besides, the percentage of consistent points obtained by it surpassed all other developed scenarios and the reference approach; The reference approach failed to label these classes. The two other scenarioes S2 and S3 have also segmented these classes (except the footpath class is not detected) but with lower precision compared to S1. Since these are the geoobjects that have approximately the same geometric features, the scenarioes based on geometric features (S2) or already classified geometric information (S3), failed to extract them precisely as scenario S1, although with the addition of optical images. For example in scene 4, The S1 scenario increases the percentage of consistent of each class by 12%(parking), 2% (rail), 7% (traffic roads),13% (street furniture), and 7% (footpath), respectively compared to the S2 scenario. The S1 scenario increases the percentage of consistency of each class by 2%(parking), 12% (Rail), 47% (traffic roads),8% (street furniture), and 9% (footpath), respectively compared to the S3 scenario. So, the positive effect of prior knowledge (image classification) from images on the percentage of consistency is seen more clearly. We can conclude that the percentage of consistent increases if classified images, and spectral information are used together with point clouds (scenario S1). That is, the semantic knowledge from optical images is very useful for separating these complex classes. However, Those semantic classes are often confused with others with similar characteristics. We can list the confusion between the parking class with the ground and traffic roads classes. And also, the confusion between the rail with street furniture and water objects. In addition to the confusion between the traffic roads class with ground and parking geoobjects , and also with bridge class in scene 4.

Fourthly, by observing the four scenes, we can see that the S2 scenario has a good performance on the point clouds scenes containing cars, walls, and bridge objects. The tree classes are successfully extracted by S2. The obtained result in these classes indicated that S2 generally performed better than the other scenarios. If we still take the example of scene 4, the S2 scenario increases the percentage of consistency of each class by 2%(cars), 14% (walls), 12% (bridge), respectively compared to the S1 scenario. Besides, it increases the percentage of consistency of each class by 5%(cars),4% (walls), and 62% (bridge), respectively compared to the S3 scenario. Additionally, the S2 scenario increases the percentage of consistency of each class by 22% (cars), and 42 % (walls), respectively compared to the reference approach. Hence, the bridge class has not been detected completely by the reference approach. So, the semantic segmentation of cars, walls, and bridge objects shows that the positive effect of geometric features on the percentage of consistency is seen clearly. The percentages of consistent points have increased in all urban scenes where optical images, point clouds, and 3D geometric features are used together. The results obtained show that the influence of adding suitable geometric features to point clouds is better addressed in these classes than in the others. Since these classes have different geometrical characteristics, the addition of planarity and verticality as point clouds attributes in S2 has facilitated their distinction of them. On the other hand, since these three classes have almost similar radiometric characteristics, S1 was not obtained the same accuracies as S2, but its results are generally good. Considering that the precisions obtained by the reference approach in cars and wall objects were moderately sufficient and insufficient in the bridge class, we can conclude that the addition of optical images to point clouds was not sufficient in the semantic segmentation of these three objects. However, Those semantic classes are often confused with others objects with similar characteristics. We can cite the confusion between the classes' cars and street furniture in scenes 1 and 4. In addition to the confusion between the class wall and street furniture. Thus, we notice a slight confusion between the wall object and the class buildings (scene 4) and ground (scene 1). Finally, we observe confusion between the class bridge and building in scene 4.

Consequently, it was concluded that the combination of geometric features, optical images, and point clouds (XYZ coordinates) is important for the semantic segmentation of cars, wall, and bridge classes. Moreover, The effective geometric features provide a real advantage and are very useful for the semantic segmentation of these irregular classes.

Fifthly, concerning the water class, the only scenario that can detect it precisely is the S1 scenario (see confusion matrix results). The water class is confused with the wall object in the S2 scenario. That is, the verticality and planarity selected are not enough to distinguish the water class. We can conclude that XYZ coordinates, optical images, and geometric properties are not useful in the semantic segmentation of water. In addition, it is confused with the ground object in the S3 scenario. On the other hand, the classified images and RGB information merged with the point clouds (case of the S1 scenario) showed a great performance in the semantic segmentation of this water class.

Therefore, we can conclude that the use of prior knowledge from optical images provides an advantage in the semantic segmentation of the water class. However, It is difficult to detect water objects with the attributes integrated into scenarios S1 and S2.

Finally, The bike class is not detected in the case of all scenarios. It is due to the very low percentage of bike samples.

4.2.2. Qualitative Assessments

In addition to the quantitative evaluation, a qualitative analysis was also carried out by visualizing the 3D point clouds to check the semantic segmentation results in detail for the test data set. Figure 13 demonstrates the visual comparison of the predicted results obtained by the proposed scenarios and their corresponding ground truths. To show the semantic segmentation effect of all scenarios more intuitively, Figures 13 and 14 shows the semantic segmentation effect of different scenarios on different point clouds scenes. It can be seen from the figures that the semantic segmentation result of the S1 scenario is closest to the true value effect. Besides, its results are more accurate and coherent compared to the other scenarios. Additionally, the semantic segmentation results show that the 3D urban scene was better semantically segmented with S1, where all classes were extracted precisely with clear boundaries.

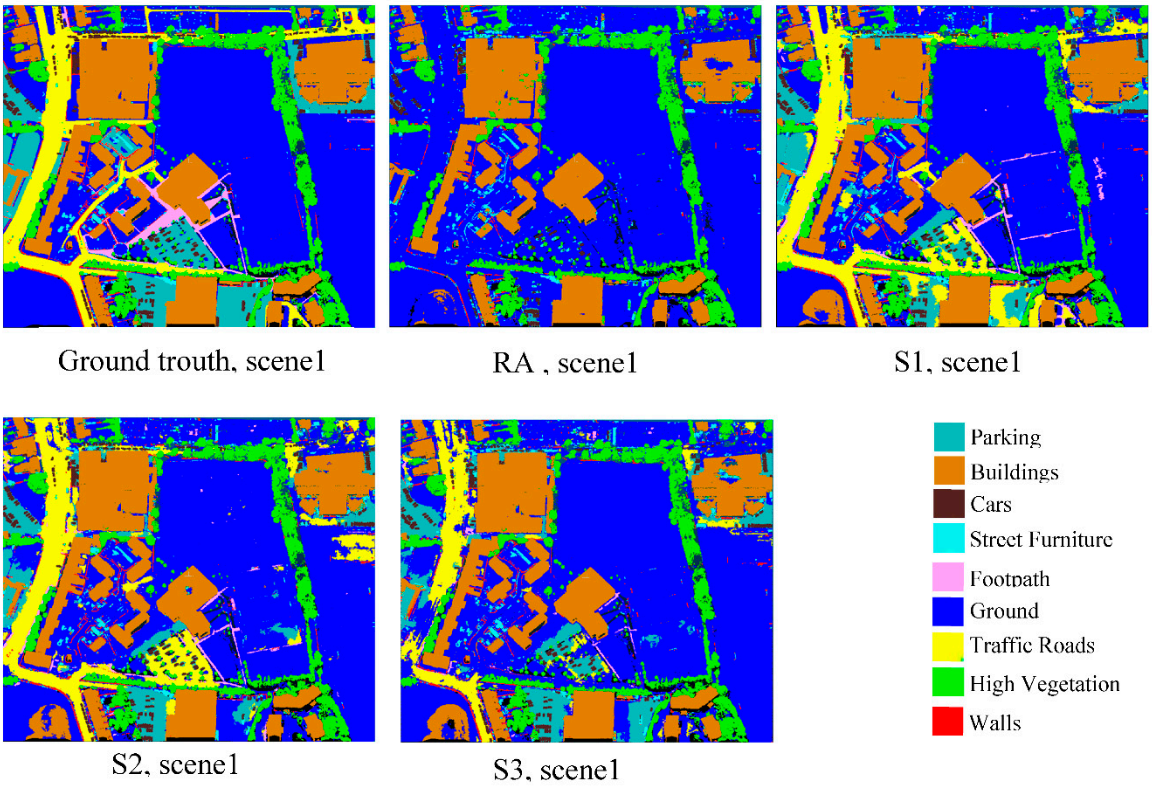


Figure 13. 3D semantic segmentation results of the reference and the three developed scenarios. Ground truth is also displayed.

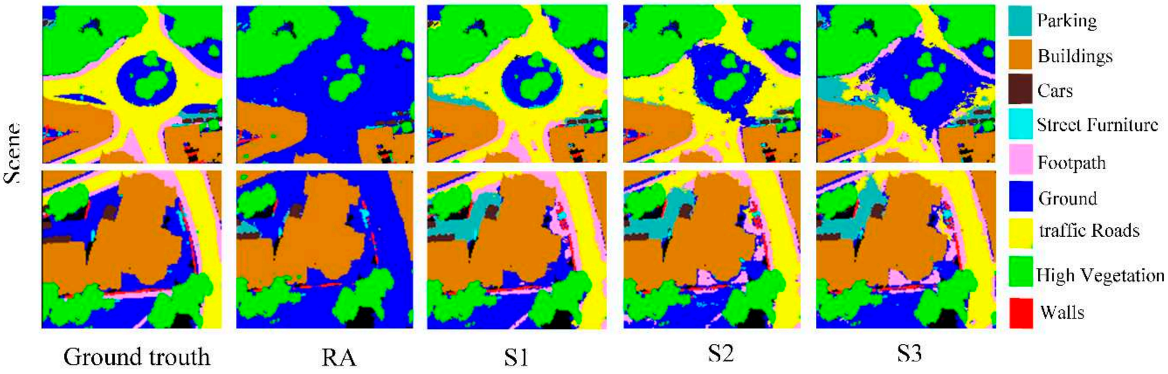


Figure 14. Selected regions from 3D semantic segmentation maps of the all evaluated processus.

The qualitative results of each class are further explained in the following paragraphs:

The semantic segmentation results indicate that the ground and high vegetation classes have been effectively semantically segmented in general by all the processes. Visual investigations show that the ground class is confused with buildings objects in all scenarios. This confusion is caused by their similar spectral information, but the overall semantic segmentation is generally good.

Scene 4 (see confusion matrix results in the Appendix A) is a perfect patch that demonstrates mixed semantic classes with a diversity of urban objects; we are interested here in the rail, traffic roads, street furniture, and parking classes. We can observe that these classes are hard to be recognized by the reference approach. It failed to label these semantic classes. Furthermore, as observed in quantitative results, S1 shows better performance in these classes by producing very few miss-segmented points compared to others scenarios. Thanks to the prior knowledge from optical images in scenario S1, errors of it were lower than those delivered by other scenarios for these semantic classes. In the case of S2, S3, and the reference approach, several parking class points have been miss-segmented as ground. It is due to the similarity of their geometric and radiometric properties. Moreover, the three scenarios have miss-classified the traffic roads as a ground class. The street furniture shares a similar color to the building and wall in this dataset: in fact, as shown in Figure 13, part of the street furniture is labeled as a building in semantic segmentation results of S2, S3, and reference approach. Finally, the rail object had not detected by the reference approach. Besides, S2 and S3 have miss-classified it as water and street furniture (see confusion matrix results in the Appendix A).

Concerning the building class, the visual evaluation has shown that different developed scenarios have extracted correctly this object compared to the reference approach. Generally, the developed scenarios have improved the semantic segmentation results of buildings. In the case of the reference scenario, we observe a slight confusion between the building class and those of ground and high vegetation. Besides, errors of S1 were slightly lower than those delivered by S2 and S3 scenarios for the building class.

Visually, we can observe in Figures 13 and 14 that the footpath object is hard to be recognized. S1, S2, and reference scenario failed to label this kind of footpath correctly. While S1 achieves an acceptable performance on it (scene 2).

Concerning the cars, wall, and bridge objects, thanks to the suitable geometric features calculated from point clouds in S2, errors of it were lower than those delivered by the other scenarios. The results indicated that the bridge class is labeled as buildings with the reference approach. Additionally, a part of this class is labeled as buildings in the segmentation results of S1 and S3. Moreover, as shown in Figure 13, various car class points have been miss-segmented as street furniture, especially in scene 4 (see confusion matrix results). Additionally, the wall is confused with several classes mainly street furniture and building geoobjects.

According to visual comparison, the semantic segmentation based on classified images (S1), geometric features (S2), and classified geometric information (S3) demonstrated a real complementary effect compared to the reference approach. The visual results indicated that the developed scenario S1 generally performed better than S2 and S3. Particularly, the S2 scenario improves the semantic segmentation results of some classes (wall, cars, bridge) more than the other scenarios.

4.3. Discussion

Three-Dimensional semantic segmentation is a crucial task for producing DTCs for city planning and management. In this context, automatic extraction of the maximum amount of semantic information with good accuracy is still a challenging research topic. In this paper, new approach based on the fusion of point clouds and other sources has been developed and evaluated to meet this challenge.

This work develops three prior-level fusion scenarios based on DL for 3D semantic segmentation. To summarize the performance of different developed scenarios, the results were compared to a reference approach, using both qualitative and quantitative assessments. The quantitative assessment is performed on five main metrics, namely, F1-score, Recall, Precision, IoU, and confusion matrix. On the other hand, the qualitative assessment is performed based on visual analysis of observed and predicted semantic segmentation. To ensure a fair evaluation, all scenarios (reference approach, S1, S2, and S3) used the same training, validation, and test data. From the comparison in Table 3, generally, the semantic segmentation effect of developed scenarios is significantly improved compared to the reference approach. Precisely, it can be seen that S1 shows significant improvements for most indicators compared to S2, S3, and the reference approach in all scenes. Compared to the reference approach, the increase in IoU was 17% for the S1, 14% for the S2, and 12% for the S3 (scene 1, Table 3). The positive effect of prior knowledge from optical images (integrated into S1) on accuracy is seen more clearly. This demonstrates that prior knowledge from classified images has an improved effect on enhancing knowledge of the utilized DL technique. It is because the classified images help the DL technique at learning time to distinguish correctly the different classes, mainly those that present a geometric and radiometric similarity. Afterwards, the fastest convergence was offered. Therefore, feature selection in S1 improves evaluation metrics in the semantic segmentation performed with RandLaNet. On the other hand, the S2 scenario has also given satisfactory results, which are in some cases close to the S1 scenario, for example in scene 4. It achieved higher overall accuracy in some classes than that of S1 and S3. This means that the usage of suitable geometric features and optical images together with point clouds improves the semantic segmentation results, mainly for the classes that have different geometric characteristics. For example, the calculation of verticality has highlighted the objects that follow a straight line which facilitates their semantic segmentation in a precise manner.

To evaluate each semantic class independently, the confusion matrices were calculated in the four scenes. By observing their results, it can be seen that the developed fusion scenarios achieved the best semantic segmentation compared to the reference approach; That is, the prior-level fusion scenarios designed in this article are better than the existing approach. Despite the good results of the reference approach obtained in some classes such as ground, it fails to label completely some others namely bridge, traffic roads, and footpath classes. Besides, its results in parking classes are not acceptable. Thus, it's quite difficult to detect these objects using only point clouds and aerial images.

As a first conclusion of this work, we point out that the direct fusion of point clouds and aerial images is not sufficient for the semantic segmentation of complex scenes with a diversity of objects. Compared to the reference approach, S2, and S3, we can see that S1 has the best performance on the point clouds scenes containing rail, traffic roads, street furniture, footpath, and parking objects. Despite the choice of the most appropriate geometric properties in S2 and the addition of classified geometric information in S3, these two scenarios did not succeed in obtaining the high accuracies as those obtained by S1. The attributes selected in these scenarios are not enough to distinguish these types of terrains more. It can be due to the geometric similarity of these classes. However, in the case of the S1 scenario, the fusion of the classified images and spectral information with the point clouds has improved the semantic segmentation results of these classes. The confusion matrices calculated in four scenes have confirmed this situation. We can conclude here that the preliminary results of image classification have guided the model to know these different classes and distinguish them precisely. On the other hand, the second scenario S2 performed well in the cars, wall, and bridge objects in all scenes. It demonstrated the best precisions compared to S1, S3, and the reference approach. The low accuracy obtained by S1 to those obtained by S2 can be due to the similarity of the radiometric information of these geoobjects. Nevertheless, the description of local geometric properties by selected geometric features has facilitated the distinction of these three classes in scenario S2. We can conclude here that the two geometric properties "planarity and verticality" have primordial contributions to the semantic segmentation of cars, walls, and bridge objects.

Figure 13 shows the ground truth and results of the four fusion scenarios. The developed scenarios achieved acceptable visual quality compared to the reference approach. Globally, the qualitative results show that the three developed fusion scenarios had fewer semantic segmentation errors compared with the reference. Precisely, the first scenario S1 has demonstrated results that are fine and very close to the ground truth. Visually, they exceed the results obtained by S2, S3, and reference processes for the majority of classes (traffic roads, street furniture….etc). As well, the visual comparison shows that semantic segmentation errors of it were lower than those delivered by other scenarios for the majority of objects. The results obtained by the evaluation metrics confirmed this situation. As mentioned before, these classes do not differ geometrically. Therefore, it is not possible to distinguish them using only classified geometric information (S3) or 3D geometric features (S2). In particular, the three geoobjects, namely wall, cars, and bridge were extracted precisely with the second scenario S2 (see Figure 13 and confusion matrix results). The latter improves the visual quality of these classes compared to S1, S3, and reference processes.

After analyzing both qualitative and quantitative results, it was found that the reference approach was unable to fully detect certain classes, indicating that the direct merging of point clouds and optical images makes it challenging to semantically segment complex urban environments. Furthermore, the results revealed that scenario S1 outperformed S2 and S3 in all indicators and demonstrated a significant improvement, producing more accurate predictions closer to the ground truth. Besides, it demonstrated a good balance between the accuracy and feasibility of the semantic segmentation according to the purposes and requirements of this work. Additionally, the S1 scenario allows for the utilization of classified images from various sources, including drone and satellite images, and can be processed by different neural networks of image classification, making it a practical option. The S1 scenario is also not highly data-intensive, as satisfactory results were obtained by training the model with only a portion of the dataset, which reduces the financial resources and hardware required since it relies solely on aerial images and point clouds. However, this scenario could be bit long , and classification errors in the images could negatively impact the 3D semantic segmentation results. Although S1 has several advantages, the difference between S1 and S2 is relatively small. Specifically, S2 is better suited for semantically segmenting walls, cars, and bridge geoobjects and outperforms S1 and S3 in these classes according to qualitative and quantitative results. Additionally, S2 is easier to handle than the other scenarios and does not require any pre-segmentation step. Nevertheless, this scenario is only applicable to geometrically distinct classes, and the challenge of similar features of the main geoobjects persists. Additionally, S2 necessitates the selection of features that have a positive impact on semantic segmentation. In regards to the S3 scenario, it is better suited for geometrically distinct geoobjects. The uniqueness of this scenario lies in its direct use of semantic knowledge from geometric information, which enhances the distinction of such objects. However, a pre-segmentation step is required, which makes the process can be bit long. Moreover, the accuracy of its 3D semantic segmentation is relatively low, and classification errors in geometric information could have a negative impact on semantic segmentation outcomes. In conclusion, considering the good qualitative and quantitative results in all classes and its superior performance compared to other scenarios, the S1 scenario is the smart fusion approach for semantic segmentation of point clouds acquired on a large scale. The selection of the S1 scenario is motivated by the accuracy of the outcomes and the excellent visual quality of the predictions. Additionally, we suggest considering the S2 scenario due to its high performance in certain semantic classes and its ease of handling.

Finally, it should be noted that this research work presents certain limitations including the usage of only 16 sets of the SensatUrban dataset which may not be sufficient to achieve the maximum accuracies of differents scenarios. In addition, the developed fusion scenarios should be tested on other datasets that contain other semantic classes. As a perspective, we suggest investigating the derived smart fusion approach in various urban contexts by choosing other urban objects and by considering also other datasets types, especially, the terrestrial point clouds. The goal is to evaluate the precisions and errors of the selected smart fusion approach when confronted with other urban environments.

5. Conclusion

This article presents a new prior level fusion approach for semantic segmentation based on a deep evaluation of three scenarios, which involve injecting aerial images and prior knowledge as point clouds attributes into the RandLaNet technique's learning pipeline. The three proposed scenarios were evaluated based on their qualitative and quantitative results to identify the one that successfully extract precisely the maximum level of urban fabric details. The derived scenario was named the "smart fusion approach". The purpose of injecting prior knowledge and spectral information from the aerial images is to enhance the original network knowledge, thereby facilitating the semantic segmentation of semantically rich urban objects. In each scenario, we incorporate aerial images and a specific type of prior knowledge as attributes to the point clouds. This process aims to determine the scenario that effectively integrates features with a greater positive impact on improving knowledge of the RandLaNet technique (smart fusion approach). The first one (S1) assigns the classified image and aerial images to the corresponding point clouds. The second one (S2) fuses the suitable geometric features and aerial images with the corresponding point clouds. The third one (S3) fuses the classified geometric information and aerial images with the corresponding point clouds. While the reference approach combines directly aerial images and points clouds. A detailed evaluation of the proposed scenarios and reference approach was conducted using the SensatUrban dataset, which exhibits a wide variety of semantic classes. Hence, The same data splits were used for the RandLaNet algorithm' training, validation, and test to ensure a fair and consistent evaluation. Four test sets of data (4 urban scenes) were used to evaluate the pre-trained models and did not contribute to the training process. A deep comparison shows that scenario S1 surpassed the reference approach, S2, and S3 scenarios in all indicators. It shows a good improvement for all evaluation metrics compared to others. However, in some cases, the S2 scenario has shown results close to those of S1. Compared to the reference approach, the S1 increases the semantic segmentation IoU of each scene by 17%, 13%, 12%, and 18%, respectively, also, the S2 increase the semantic segmentation IoU of each scene by 14%, 11%, 10%, and 17%, respectively, additionally, the S3 increase the semantic segmentation IoU of each scene by 12%, 10%, 9%, and 7%, respectively. The qualitative and quantitative analysis of each class shows that the reference approach failed to detect several classes. Therefore, it's difficult to classify complex large-scale urban by directly fusing point clouds and aerial images. Moreover, based on the visual analysis, S1 predictions seem more precise and closer to the ground truth. In addition, it demonstrated a good balance between semantic segmentation accuracy and efficiency according to the goals and requests of this work. In specific, the second proposed scenario (S2) is more adequate for the semantic segmentation of some geoobjects. It outperformed the qualitative and quantitative achievements of S1 and S3 in them. Furthermore, the manipulation of this scenario is very simple compared to the others. Finally, since the S1 scenario exhibits good scores in all classes and its performances surpass the other scenarios, we can conclude that the S1 scenario is the smart fusion approach for the semantic segmentation of large-scale point clouds. The preference for the S1 scenario is therefore motivated by its results' accuracy and the quality of its visual predictions. We also recommended the S2 scenario because of its high performance in some semantic classes and its simplicity of processing. The experiments show that the derived smart fusion approach can improve the knowledge of the RandLa-Net technique. It allows good metrics in particular for classes that are difficult to detect using original DL architecture without prior knowledge. Additionally, it succeeds in reducing the confusion between different semantic classes. Furthermore, the smart fusion approach can potentially be adapted for any 3D semantic segmentation DL techniques. So, we suggest investigating the semantic segmentation smart fusion approach in other complex urban environments to evaluate its efficiency and limits in different urban contexts.

Appendix A:

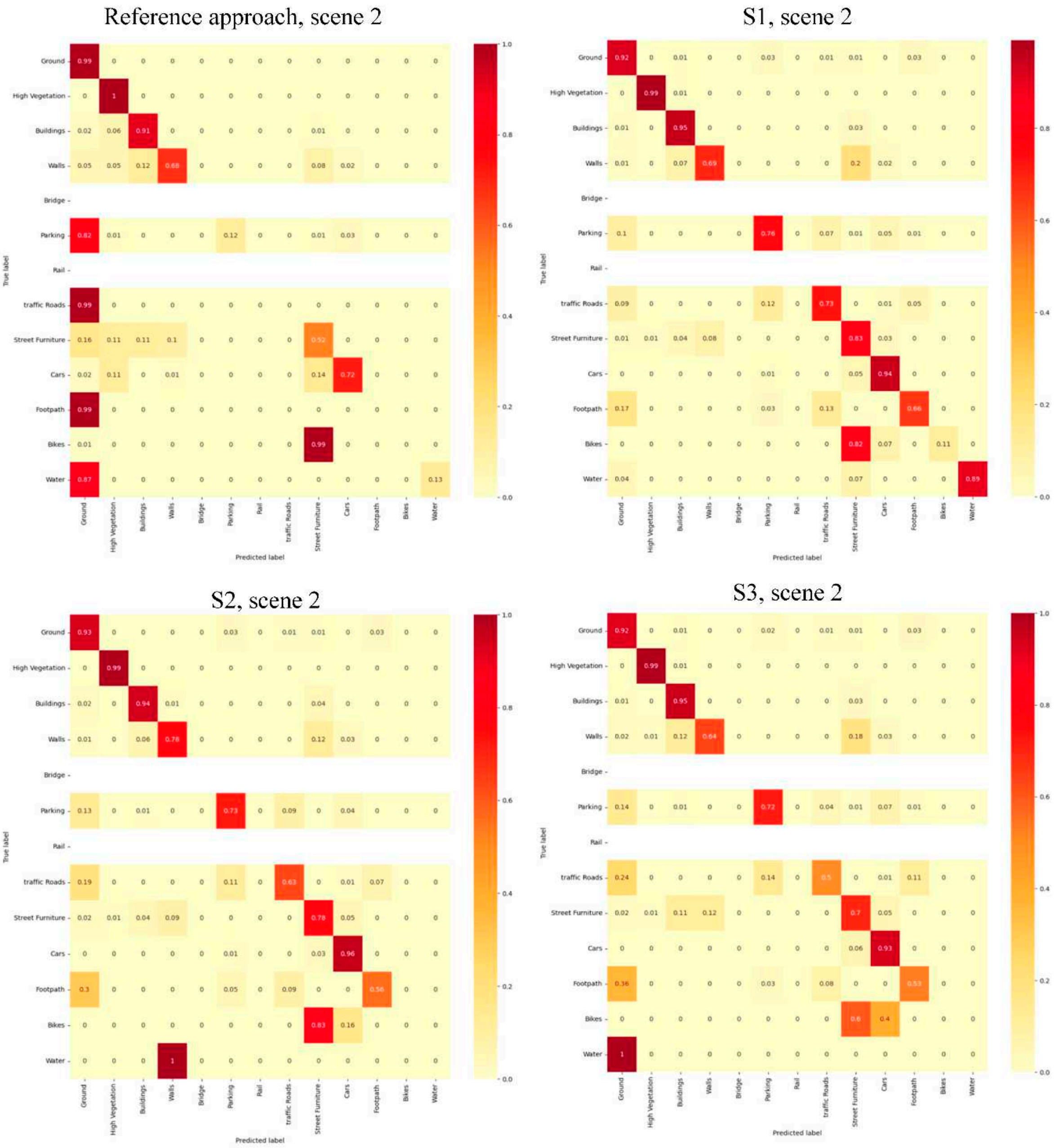
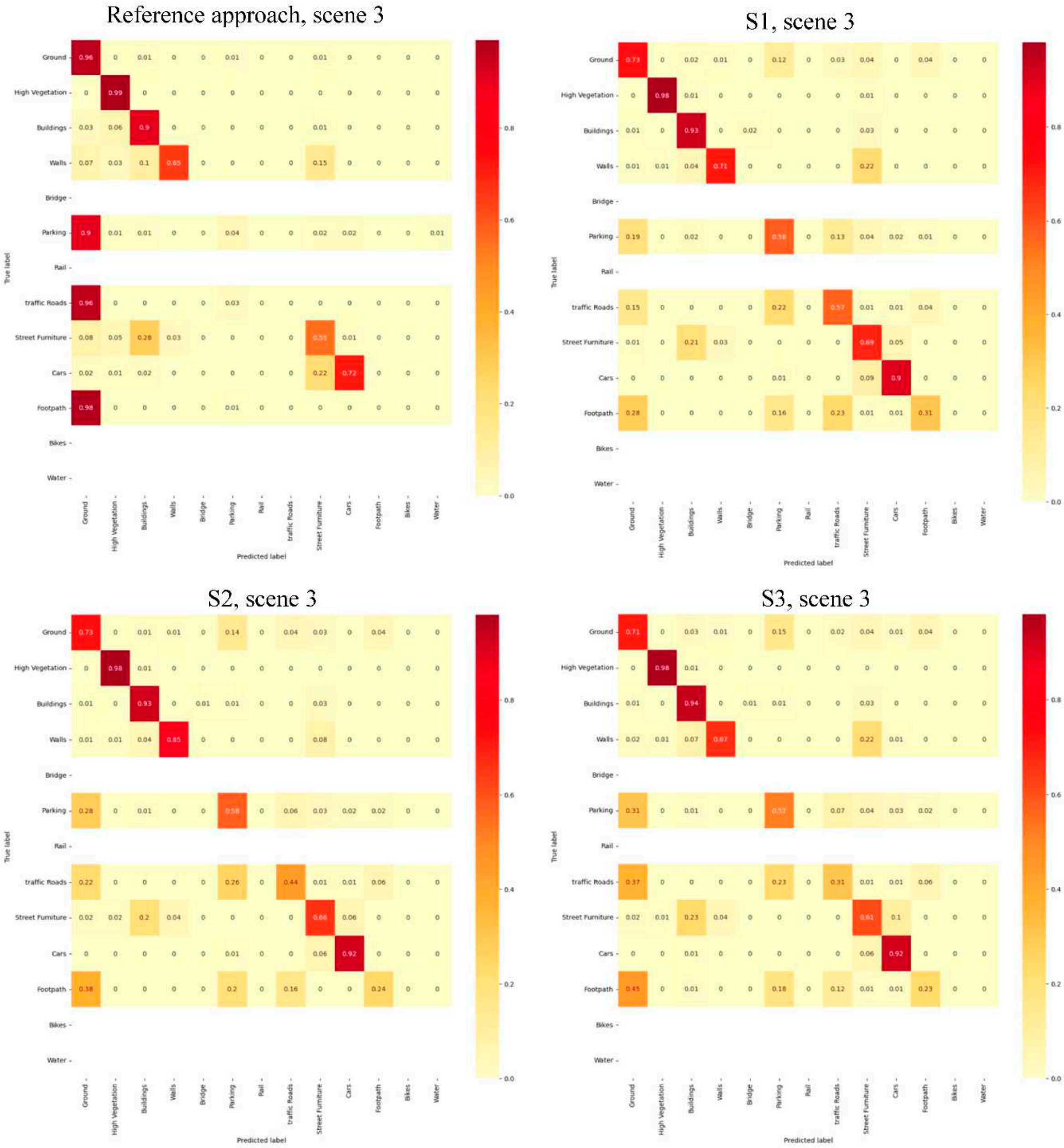


Figure A1. Normalized Confusion matrix of evaluated semantic segmentation approaches over the scene 2.



Reference approach, scene 3

S1, scene 3

S2, scene 3

S3, scene 3

Figure A2. Normalized Confusion matrix of evaluated semantic segmentation approaches over the scene 3.

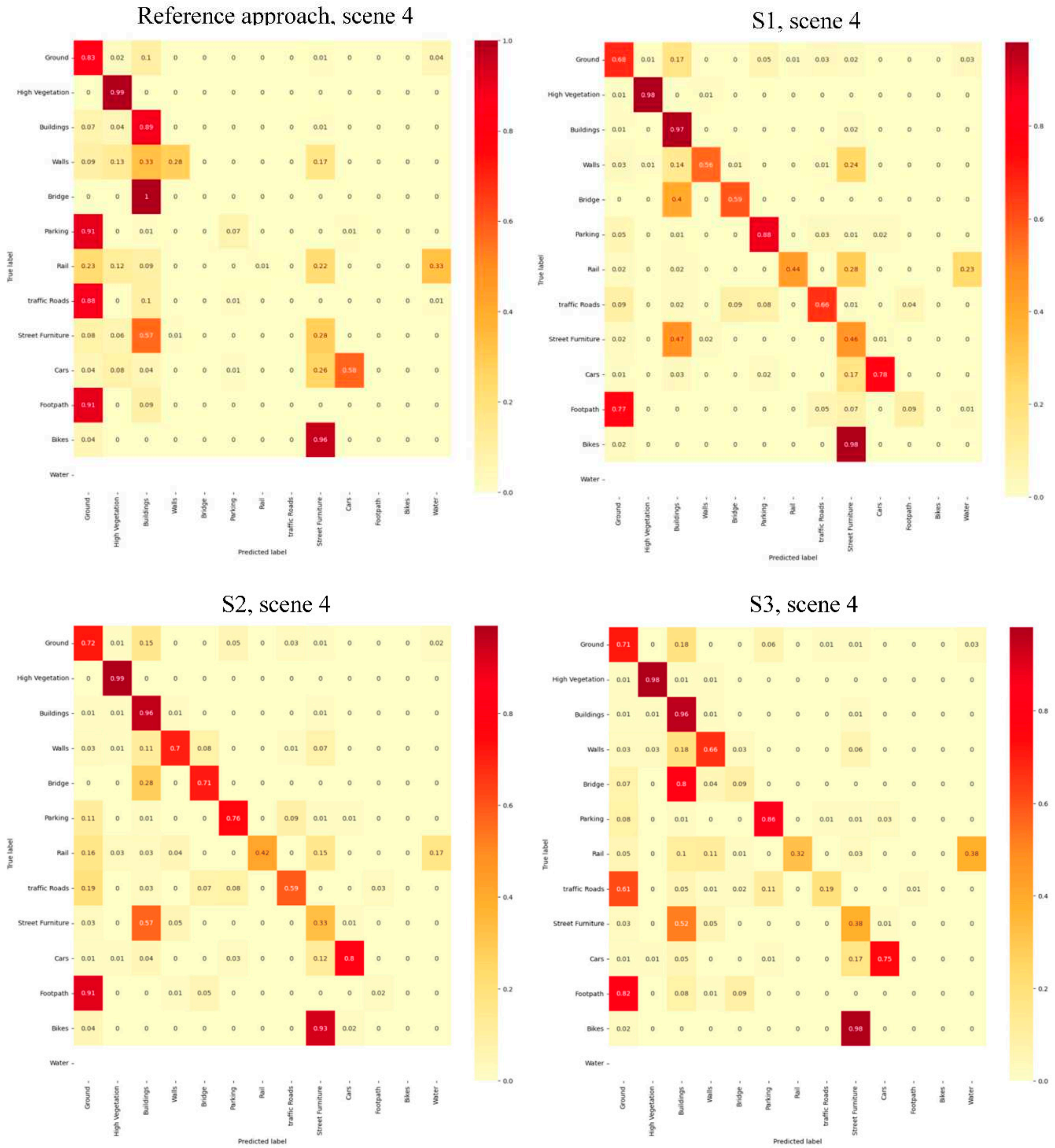


Figure A3. Normalized Confusion matrix of evaluated semantic segmentation approaches over the scene 4.

References

1. Atik, M.E., Duran, Z., Seker, D.Z., 2021. Machine Learning-Based Supervised Classification of Point Clouds Using Multiscale Geometric Features. ISPRS International Journal of Geo-Information 10, 187. https://doi.org/10.3390/ijgi10030187.
2. Bai, X., Liu, C., Ren, P., Zhou, J., Zhao, H., Su, Y., 2015. Object Classification via Feature Fusion Based Marginalized Kernels. IEEE Geoscience and Remote Sensing Letters 12, 8–12. https://doi.org/10.1109/LGRS.2014.2322953
3. Ballouch, Z., Hajji, R., Ettarid, M., 2022a. Toward a Deep Learning Approach for Automatic Semantic Segmentation of 3D Lidar Point Clouds in Urban Areas, in: Barramou, F., El Brirchi, E.H., Mansouri, K., Dehbi, Y. (Eds.), Geospatial Intelligence: Applications and Future Trends. Springer International Publishing, Cham, pp. 67–77. https://doi.org/10.1007/978-3-030-80458-9_6
4. Ballouch, Z., Hajji, R., Ettarid, M., 2020. The contribution of Deep Learning to the semantic segmentation of 3D point-clouds in urban areas, in: 2020 IEEE International Conference of Moroccan Geomatics

(Morgeo). Presented at the 2020 IEEE International conference of Moroccan Geomatics (Morgeo), pp. 1–6. https://doi.org/10.1109/Morgeo49228.2020.9121898

5. Ballouch, Z., Hajji, R., Poux, F., Kharroubi, A., Billen, R., 2022b. A Prior Level Fusion Approach for the Semantic Segmentation of 3D Point Clouds Using Deep Learning. Remote Sensing 14, 3415. https://doi.org/10.3390/rs14143415
6. Chen, Y., Liu, X., Xiao, Y., Zhao, Q., Wan, S., 2021. Three-Dimensional Urban Land Cover Classification by Prior-Level Fusion of LiDAR Point Cloud and Optical Imagery. Remote Sensing 13, 4928. https://doi.org/10.3390/rs13234928
7. Dong, Y., Zhang, L., Cui, X., Ai, H., Xu, B., 2018. Extraction of Buildings from Multiple-View Aerial Images Using a Feature-Level-Fusion Strategy. Remote Sensing 10, 1947. https://doi.org/10.3390/rs10121947
8. Gao, W., Nan, L., Boom, B., Ledoux, H., 2021. SUM: A benchmark dataset of Semantic Urban Meshes. ISPRS Journal of Photogrammetry and Remote Sensing 179, 108–120. https://doi.org/10.1016/j.isprsjprs.2021.07.008
9. Golipour, M., Ghassemian, H., Mirzapour, F., 2016. Integrating Hierarchical Segmentation Maps With MRF Prior for Classification of Hyperspectral Images in a Bayesian Framework. IEEE Transactions on Geoscience and Remote Sensing 54, 805–816. https://doi.org/10.1109/TGRS.2015.2466657
10. Grilli, E., Daniele, A., Bassier, M., Remondino, F., Serafini, L., 2023. Knowledge Enhanced Neural Networks for Point Cloud Semantic Segmentation. Remote Sensing 15, 2590. https://doi.org/10.3390/rs15102590
11. Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. Semantic3D.net: A new Large-scale Point Cloud Classification Benchmark. arXiv:1704.03847 [cs].
12. Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2021. Towards Semantic Segmentation of Urban-Scale 3D Point Clouds: A Dataset, Benchmarks and Challenges. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4977–4987.
13. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, pp. 11105–11114. https://doi.org/10.1109/CVPR42600.2020.01112
14. Jeong, J., Song, H., Park, J., Resende, P., Bradaï, B., Jo, K., 2022. Fast and Lite Point Cloud Semantic Segmentation for Autonomous Driving Utilizing LiDAR Synthetic Training Data. IEEE Access 10, 78899–78909. https://doi.org/10.1109/ACCESS.2022.3184803
15. Landrieu, L., Simonovsky, M., 2018. Large-Scale Point Cloud Semantic Segmentation With Superpoint Graphs. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4558–4567.
16. Liu, C., Zeng, D., Akbar, A., Wu, H., Jia, S., Xu, Z., Yue, H., 2022. Context-Aware Network for Semantic Segmentation Toward Large-Scale Point Clouds in Urban Environments. IEEE Transactions on Geoscience and Remote Sensing 60, 1–15. https://doi.org/10.1109/TGRS.2022.3182776
17. Luo, S., Wang, C., Xi, X., Zeng, H., Li, D., Xia, S., Wang, P., 2016. Fusion of Airborne Discrete-Return LiDAR and Hyperspectral Data for Land Cover Classification. Remote Sensing 8, 3. https://doi.org/10.3390/rs8010003
18. Man, Q., Dong, P., Guo, H., 2015. Pixel- and feature-level fusion of hyperspectral and lidar data for urban land-use classification. International Journal of Remote Sensing 36, 1618–1644. https://doi.org/10.1080/01431161.2015.1015657
19. Megahed, Y., Shaker, A., Yan, W.Y., 2021. Fusion of Airborne LiDAR Point Clouds and Aerial Images for Heterogeneous Land-Use Urban Mapping. Remote Sensing 13, 814. https://doi.org/10.3390/rs13040814
20. Meyer, G.P., Charland, J., Hegde, D., Laddha, A., Vallespi-Gonzalez, C., 2019. Sensor Fusion for Joint 3D Object Detection and Semantic Segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Long Beach, CA, USA, pp. 1230–1237. https://doi.org/10.1109/CVPRW.2019.00162
21. Mirzapour, F., Ghassemian, H., 2015. Improving hyperspectral image classification by combining spectral, texture, and shape features. International Journal of Remote Sensing 36, 1070–1096. https://doi.org/10.1080/01431161.2015.1007251

22. Oh, S.-I., Kang, H.-B., 2017. Object Detection and Classification by Decision-Level Fusion for Intelligent Vehicle Systems. Sensors 17, 207. https://doi.org/10.3390/s17010207
23. Özdemir, E., Remondino, F., 2019. CLASSIFICATION OF AERIAL POINT CLOUDS WITH DEEP LEARNING. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-2/W13, 103–110. https://doi.org/10.5194/isprs-archives-XLII-2-W13-103-2019
24. Poliyapram, V., Wang, W., Nakamura, R., 2019. A Point-Wise LiDAR and Image Multimodal Fusion Network (PMNet) for Aerial Point Cloud 3D Semantic Segmentation. Remote Sensing 11, 2961. https://doi.org/10.3390/rs11242961
25. Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.
26. Ruohomäki, T., Airaksinen, E., Huuska, P., Kesäniemi, O., Martikka, M., Suomisto, J., 2018. Smart City Platform Enabling Digital Twin, in: 2018 International Conference on Intelligent Systems (IS). Presented at the 2018 International Conference on Intelligent Systems (IS), pp. 155–161. https://doi.org/10.1109/IS.2018.8710517
27. Shahat, E., Hyun, C.T., Yeom, C., 2021. City Digital Twin Potentials: A Review and Research Agenda. Sustainability 13, 3386. https://doi.org/10.3390/su13063386
28. Son, S.W., Kim, D.W., Sung, W.G., Yu, J.J., 2020. Integrating UAV and TLS Approaches for Environmental Management: A Case Study of a Waste Stockpile Area. Remote Sensing 12, 1615. https://doi.org/10.3390/rs12101615
29. Song, H., Huang, B., Liu, Q., Zhang, K., 2015. Improving the Spatial Resolution of Landsat TM/ETM+ Through Fusion With SPOT5 Images via Learning-Based Super-Resolution. IEEE Transactions on Geoscience and Remote Sensing 53, 1195–1204. https://doi.org/10.1109/TGRS.2014.2335818
30. Tabib Mahmoudi, F., Samadzadegan, F., Reinartz, P., 2015. Object Recognition Based on the Context Aware Decision-Level Fusion in Multiviews Imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8, 12–22. https://doi.org/10.1109/JSTARS.2014.2362103
31. Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K., Li, J., 2020. Toronto-3D: A Large-Scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 202–203.
32. Weinmann, M., Weinmann, M., 2019. FUSION OF HYPERSPECTRAL, MULTISPECTRAL, COLOR AND 3D POINT CLOUD INFORMATION FOR THE SEMANTIC INTERPRETATION OF URBAN ENVIRONMENTS. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. XLII-2/W13, 1899–1906. https://doi.org/10.5194/isprs-archives-XLII-2-W13-1899-2019
33. White, G., Zink, A., Codecá, L., Clarke, S., 2021. A digital twin smart city for citizen feedback. Cities 110, 103064. https://doi.org/10.1016/j.cities.2020.103064
34. Ye, C., Pan, H., Yu, X., Gao, H., 2022. A spatially enhanced network with camera-lidar fusion for 3D semantic segmentation. Neurocomputing 484, 59–66. https://doi.org/10.1016/j.neucom.2020.12.135
35. Yousefhussien, M., Kelbe, D.J., Ientilucci, E.J., Salvaggio, C., 2018. A multi-scale fully convolutional network for semantic labeling of 3D point clouds. ISPRS Journal of Photogrammetry and Remote Sensing 143, 191–204. https://doi.org/10.1016/j.isprsjprs.2018.03.018
36. Zhang, J., Zhao, X., Chen, Z., Lu, Z., 2019. A Review of Deep Learning-Based Semantic Segmentation for Point Cloud. IEEE Access 7, 179118–179133. https://doi.org/10.1109/ACCESS.2019.2958671
37. Zhang, R., Li, G., Li, M., Wang, L., 2018. Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning. ISPRS Journal of Photogrammetry and Remote Sensing 143, 85–96. https://doi.org/10.1016/j.isprsjprs.2018.04.022
38. Zhang, Y., Chi, M., 2020. Mask-R-FCN: A Deep Fusion Network for Semantic Segmentation. IEEE Access 8, 155753–155765. https://doi.org/10.1109/ACCESS.2020.3012701
39. Zhao, L., Zhou, H., Zhu, X., Song, X., Li, H., Tao, W., 2021. LIF-Seg: LiDAR and Camera Image Fusion for 3D LiDAR Semantic Segmentation. arXiv:2108.07511 [cs].
40. Zhou, T., Ruan, S., Canu, S., 2019. A review: Deep learning for medical image segmentation using multi-modality fusion. Array 3–4, 100004. https://doi.org/10.1016/j.array.2019.100004.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.