ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023) 246–261

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Deep learning for filtering the ground from ALS point clouds: A dataset, evaluations and issues

Nannan Qin [a], Weikai Tan [b], Lingfei Ma [c], Dedong Zhang [d], Haiyan Guan [a,*], Jonathan Li [b,d,**]

[a] School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, 210044, JiangSu, China
[b] Department of Geography and Environmental Management, University of Waterloo, Waterloo, N2L 3G1, Ontario, Canada
[c] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, 102206, China
[d] Department of Systems Design Engineering, University of Waterloo, Waterloo, N2L 3G1, Ontario, Canada

ARTICLE INFO

Keywords:
Ground filtering
Deep learning
Point cloud dataset
Comparative evaluation

ABSTRACT

The capability of partially penetrating vegetation canopy and efficiently collecting high-precision point clouds over large areas makes airborne laser scanning (ALS) a valuable tool for various geospatial applications. However, automated ground filtering (GF), one fundamental and challenging step for most ALS applications, has remained a widely researched yet unsolved problem for decades. The recent breakthroughs in supervised deep learning (DL) techniques, which rely on sufficient and high-quality labeled datasets, provide a new solution to better solve this problem. Unfortunately, public 3D geospatial datasets are scarce, especially for those tailored for the landform-scale GF task. Moreover, whether advanced deep neural networks (DNNs) can be well-scaled to the problem of GF remains an open question. To comprehensively advance the development of effective DL-based GF pipelines, we first publish an ultra-large-scale GF dataset built upon open-access ALS point clouds of four different countries worldwide, which covers over 47 km^2 and nine different terrain scenes. Then, multiple attractive advantages of DL techniques in GF are evaluated through extensive experimental comparisons with traditional GF methods on the presented dataset. Furthermore, we reveal several issues faced by generalizing existing advanced 3D DNNs into GF tasks with a series of in-depth experimental analyses. Finally, some promising directions for future research are suggested in response to the identified challenges. Our dataset, named OpenGF, is available at https://github.com/Nathan-UW/OpenGF.

1. Introduction

With the fast development of 3D acquisition techniques, including inexpensive laser scanning, depth sensors, or advanced photogrammetric reconstruction pipelines, point clouds have become easier to capture and have driven a series of interesting studies in the fields of autonomous driving (Geiger et al., 2012), robots (Valada et al., 2017), and remote sensing (Xue et al., 2020; Bulatov et al., 2021). Owing to the ability of vegetation canopy penetration and efficiently collecting high-precision point clouds over a large area, airborne laser scanning (ALS) has been widely used in many large-scale geospatial applications. In forest monitoring, ALS data is used to estimate above-ground biomass or wildfire fuel consumption (Andersen et al., 2014; McCarley et al., 2020). In archaeology, ALS point clouds are important for the identification and monitoring of archaeological sites and landscapes which are hidden deeply underneath vegetation (Canuto et al., 2018; Doneus et al., 2020). In flood modeling, detailed elevation information of the ground surface for predicting flood-prone areas can be extracted from ALS data (Muhadi et al., 2020). In addition, ALS point clouds are also widely used in powerline corridor surveying (Ortega et al., 2019) and 3D urban scene understanding (Schmohl and Sörgel, 2019; Mao et al., 2022).

Since raw ALS data contains both ground points of the bare earth and non-ground points of land covers, one crucial and challenging pretreatment for most ALS applications is to discriminate ground points from non-ground points, often called ground filtering (GF). However, the significant topographic fluctuations and complex structures of objects result in notable differences within the class ground or non-ground. Meanwhile, highly similar geometric structures are frequently observed between the above two categories. The large intra-class differences and inter-class similarities bring great difficulties to the accurate GF of ALS point clouds. Furthermore, the ubiquitous outliers in point clouds make the problem more challenging. Although numerous GF methods were proposed in the last three decades, the problem of

* Corresponding author.
** Corresponding author at: Department of Geography and Environmental Management, University of Waterloo, Waterloo, N2L 3G1, Ontario, Canada.
E-mail addresses: guanhy.nj@nuist.edu.cn (H. Guan), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.isprsjprs.2023.06.005
Received 4 September 2022; Received in revised form 15 May 2023; Accepted 13 June 2023

extracting the ground from large-scale point clouds has not yet been completely solved.

More recently, a growing number of deep learning (DL) methods have been investigated for 3D scene parsing with a notable string of empirical successes (Qi et al., 2017b; Thomas et al., 2019; Hu et al., 2020; Zhang et al., 2020b; Zhu et al., 2021). The rapid advances in deep neural networks (DNNs), which are mainly benefited from abundant public datasets, has also attracted the attention of many studies (Hu and Yuan, 2016; Rizaldy et al., 2018; Gevaert et al., 2018; Jin et al., 2020; Zhang et al., 2020a; Nurunnabi et al., 2021) to resolve the limitations of traditional GF methods. Although these DL-based GF methods achieved impressive performance, the datasets they used are either unpublished or lack scene diversity. At the same time, the widely used public International Society for Photogrammetry and Remote Sensing (ISPRS) filtertest dataset (Sithole and Vosselman, 2004) contains too few labeled samples to release the power of supervised DL techniques. Thus, it is remarkably essential to publish a new high-quality GF dataset with wide coverage and diverse terrain scenes.

In addition, whether advanced deep neural networks (DNNs) can be well-scaled to the problem of GF remains an open question. On one hand, it is still unclear which problems of traditional GF methods are also faced by DL-based GF methods. For example, classic GF methods often make serious misclassifications in difficult local areas, resulting in time-consuming and labor-intensive manual refinements. Are the micro-topography errors made by the state-of-the-art DNNs few enough to eliminate post-processing? On the other hand, new issues brought by DL techniques to GF have not yet been in-depth explored. In contrast to the majority of 3D understanding tasks (e.g., semantic segmentation of rooms, streets, or even urban scenes), the research target of GF is the earth's surface which typically spans significantly wider areas. Can existing DNNs for small-scale 3D understanding effectively process large-scale topographic point clouds under limited GPU memory? Furthermore, existing supervised DL techniques typically tend to overfit a specific data distribution. Do the trained DL models for GF have the capability to be well applied to the unknown data? Extensive experimental comparisons and in-depth analyses would help answer these questions. Nevertheless, such evaluations were lacking in the field of GF.

To fully promote the development of advanced DL-based GF pipelines, we first built a public ultra-large-scale GF dataset, by taking advantage of worldwide open-access ALS point clouds with accurate ground labels. The dataset, named OpenGF, has a coverage of over 47 km^2 area and nine different terrain scenes. Meanwhile, over 542 million accurately labeled points make our dataset thousands of times larger than the ISPRS filtertest dataset. Then, extensive experimental comparisons of eight representative methods for GF were carried out on the proposed dataset, which highlights multiple strengths of DL techniques in GF. Furthermore, we revealed several key issues faced by generalizing existing 3D DNNs into GF tasks through a series of in-depth experimental analyses. Finally, some promising directions for future GF research were suggested in response to the identified challenges. A preliminary version of this work has been published in a conference proceedings (Qin et al., 2021). In this paper, we concentrate on conducting more detailed comparative evaluations and much deeper issue analyses, on the basis of supplementing new test data and methods. In short, the main contributions of our work include:

- Publishing a new ultra-large-scale ALS dataset dedicated to promoting the development of advanced DL-based GF pipelines, which contains both high-quality ground labels and diverse terrain types.
- Providing comparative evaluations of representative 3D DNNs and classic GF methods on OpenGF, which both highlights the strengths of DL techniques in GF and serves as a solid benchmark.
- Clarifying several key issues faced by generalizing existing DNNs to extract the ground from ALS point clouds, which provides some implications for future research.
- Presenting a detailed strategy for the rapid construction of landform-scale datasets with worldwide open-access data, which points to a new possibility to quickly build large-scale datasets.
- Converting traditional GF metrics into equivalent or similar modern classification metrics, in order to facilitate researchers in related fields.

The remainder of this paper is structured as follows: Related work is introduced in Section 2. Section 3 describes the detailed strategy to construct the OpenGF dataset, and Section 4 provides comparative evaluations of representative methods for GF on the proposed dataset. The issues faced by generalizing DL techniques into GF tasks are further clarified in Section 5. In addition, a comprehensive discussion and observed limitations are given in Section 6. Section 7 concludes the paper with outlooks.

2. Related work

2.1. Datasets for 3D geospatial labeling

The number of outdoor point cloud datasets has been increasing due to the fast-growing 3D acquisition techniques, and many of them have become publicly accessible. These datasets promote the development of 3D geospatial data intelligence with DL techniques.

In the light of different data collection techniques, existing geospatial datasets for 3D semantic segmentation can be roughly divided into three groups: **(1) Photogrammetric 3D datasets.** The point clouds/meshes of these datasets, such as Campus3D (Li et al., 2020), SensatUrban (Hu et al., 2022) and SUM-Helsinki (Gao et al., 2021), are produced using photogrammetry techniques. The photogrammetric 3D point clouds/meshes contain almost no ground points in areas under the vegetation canopy owing to the limitations of passive image acquisition, so such datasets have inherent shortcomings in training DL models for ground extraction in dense vegetation areas. **(2) TLS/MLS 3D datasets.** This kind of datasets is usually collected at the street level for roadway scene understanding, and some representative datasets include Semantic3D (Hackel et al., 2017), Paris-Lille-3D (Roynard et al., 2018), SemanticKITTI (Behley et al., 2019), and Toronto-3D (Tan et al., 2020). Due to the limited geographic coverage, such datasets are not suitable for large-scale ground extraction although they have a high point density with a large data volume. **(3) ALS/UAV-LS 3D datasets.** These datasets collected by airborne LiDAR sensors are commonly used for 3D urban classification and urban environmental perception, such as the ISPRS Vaihingen 3D (V3D) semantic labeling dataset (Niemeyer et al., 2014), DublinCity (Zolanvari et al., 2019), LASDU (Ye et al., 2020), DALES (Varney et al., 2020), and the Hessigheim 3D (H3D) dataset (Kölle et al., 2021). They concentrate on recognizing a variety of urban objects (e.g., grass, fence, cars, facades) rather than accurately extracting the ground. Besides, the coverage of typical non-urban scenes (e.g., mountains and forests) in this kind of dataset is very limited. All these factors make such datasets less suitable for GF studies.

To our best knowledge, there mainly exists two types of point cloud datasets for evaluating GF methods nowadays. The first kind is the well-known ISPRS filtertest dataset released before 2003. This dataset contains 15 reference samples with different terrain characteristics. Yet, each sample contains too few points to be used for training DL models. The latter is the experimental datasets used in recent publications on DL-based GF pipelines (Hu and Yuan, 2016; Jin et al., 2020). This kind of datasets typically contains massive training data, but it is either proprietary or contains insufficient terrain types.

The goal of the GF task is to robustly distinguish ground points from non-ground points in various terrain scenes, which requires a GF dataset containing both sufficient training data and different terrain types. To this end, OpenGF is published in this work to encourage the development of creative DL-based GF pipelines. Comprehensive specifications of the aforementioned representative datasets are compared in Table 1.

Table 1
Comprehensive specifications of representative geospatial datasets for 3D semantic labeling.

Dataset	Year	Coverage	Points	RGB	Collection	Application
ISPRS filtertest (Sithole and Vosselman, 2004)	–	1.1×10^6 m^2	0.4 M	No	ALS	Landform-scale
OpenGF (Ours)	2021	47.7×10^6 m^2	542.1 M	No	ALS	ground filtering
ISPRS V3D (Niemeyer et al., 2014)	2014	–	1.2 M	No	ALS	
DublinCity (Zolanvari et al., 2019)	2019	2×10^6 m^2	260 M	No	ALS	
LASDU (Ye et al., 2020)	2020	1.02×10^6 m^2	3.12 M	No	ALS	
DALES (Varney et al., 2020)	2020	10×10^6 m^2	505 M	No	ALS	Urban-scale
H3D (Kölle et al., 2021)	2021	1×10^5 m^2	73 M	Yes	UAV ALS	semantic segmentation
Campus3D (Li et al., 2020)	2020	1.6×10^6 m^2	937.1 M	Yes	UAV photogrammetry	
SensatUrban (Hu et al., 2022)	2020	7.6×10^6 m^2	2847 M	Yes	UAV photogrammetry	
SUM-Helsinki (Gao et al., 2021)	2021	4×10^6 m^2	19 M	Yes	Airplane photogrammetry	
Semantic3D (Hackel et al., 2017)	2017	–	4000 M	Yes	TLS	
Paris-Lille-3D (Roynard et al., 2018)	2018	1940 m	143 M	No	MLS	Street-scale
SemanticKITTI (Behley et al., 2019)	2019	39 200 m	4549 M	No	MLS	semantic segmentation
Toronto-3D (Tan et al., 2020)	2020	1000 m	78 M	Yes	MLS	

2.2. GF methods for ALS point clouds

The problem of GF has seen great progress over the last three decades, and numerous methods have been proposed. Generally, these approaches can be divided into two categories: classic ground filters and learning-based pipelines.

Classic ground filters. Prior to the emergence of machine learning techniques, the GF problem was often formulated as artificially designed rules or optimization problems constrained by geometric priors, in which no supervised information is required. According to different algorithm concepts, they can be broadly categorized into five groups: (1) Morphology-based filters (Zhang et al., 2003; Pingel et al., 2013; Duan et al., 2019). These methods usually conduct a number of morphological operations (e.g., erosion and dilation) within a local window to filter non-ground points. (2) Surface-based filters (Axelsson, 2000; Evans and Hudak, 2007; Nie et al., 2017; Pfeifer et al., 2001; Elmqvist, 2002; Hu et al., 2015; Zhang et al., 2016). In this case, a reference surface is firstly constructed by discriminant functions, and then the ground points are determined by analyzing the spatial relationship between the point location and the corresponding buffer zone of the reference surface. (3) Slope-based filters (Vosselman, 2000; Sithole and Vosselman, 2001; Susaki, 2012). In these algorithms, the category is judged according to the slope between two points. The higher point will be labeled as the class *non-ground*, if the slope is greater than a certain threshold. (4) Segmentation-based filters (Sithole and Vosselman, 2005; Hingee et al., 2016; Beumier and Idrissa, 2016). In such approaches, the similar points are first segmented by applying segmentation methods, and then some rules (e.g., slope and elevation differences) are adopted to separate ground segments. (5) Statistic-based filters (Bartels et al., 2006; Özcan and Ünsalan, 2016). They usually assume that the elevation of ground points obeys a certain distribution while the existence of non-ground points may disturb this distribution. In addition, many hybrid methods (Zhang and Lin, 2013; Mongus et al., 2014; Su et al., 2015; Zhao et al., 2016) based on the above two or more typical filters have been proposed to improve the filtering performance. In general, classic ground filters perform well in relatively flat and low-vegetated areas, but they typically have defects to deal with complex landforms such as hybrid terrains, steep slopes, dense vegetation, and terrain discontinuities.

Learning-based pipelines. Different from the above classic filters, learning-based pipelines convert GF into a probability classification problem and rely on sufficient supervised samples. Early works (Lu et al., 2009; Jahromi et al., 2011; Ayazi and Saadat Seresht, 2019) focused on various traditional classifiers that can directly learn the discrimination rules from hand-crafted features. Nevertheless, their performance is generally limited by the low-level descriptiveness of hand-crafted features. In contrast, semantic information learned by DNNs can achieve superior performance in GF tasks. For example, Hu and Yuan (2016) first proposed a GF pipeline based on 2D DNNs and achieved impressive filtering performance. However, this approach is inefficient due to its point-wise classification strategy and redundant calculations in the conversion of points to images (Rizaldy et al., 2018). To alleviate this problem, improved GF pipelines based on 2D semantic segmentation networks were proposed (Rizaldy et al., 2018). However, this kind of pipelines will not be suitable for mountain areas with dense vegetation because of the considerable overlap between ground and non-ground points in the 2D projection map. To tackle this limitation, Schmohl and Sörgel (2019) adapted submanifold sparse convolutional networks (Graham et al., 2018) for the classification of voxelized ALS point clouds. However, the loss of 3D detail information is still unavoidable due to the voxelization process. Point-based DNNs were then introduced to further improve the accuracy and efficiency of GF in mountain areas (Zhang et al., 2020a; Jin et al., 2020). Such methods effectively reduced the time consumption and errors caused by converting point clouds into feature maps or voxels. However, due to the small-scale input that can be processed at one time, they are not suitable for urban areas with large buildings (Zhang et al., 2020a). To reduce the dependence on massive training data, Nurunnabi et al. (2021) introduced a feature-based DL architecture for ground extraction, but it needs a thorough understanding of the related hand-crafted features. In addition, DL-based digital terrain model (DTM) regressions were also explored (Luo et al., 2017; Amini Amirkolaee et al., 2022) to improve the quality of DTM generation in steep and dense forested areas. Overall, whether state-of-the-art DNNs can be well-scaled to the challenging GF problem remains an open question.

2.3. Ground extraction from photogrammetric data

Due to dependency on only 3D point coordinates, many GF methods designed for ALS point clouds are also applicable to photogrammetric point clouds (Serifoglu Yilmaz et al., 2018; Serifoglu Yilmaz and Gungor, 2018; Zeybek and Şanlıoğlu, 2019; Klápště et al., 2020). For example, Serifoglu Yilmaz et al. (2018) investigated the performances of seven widely used ground filters on UAV photogrammetric point clouds. Zeybek and Şanlıoğlu (2019) compared the performance of four different ground filters on UAV photogrammetry-based point clouds. To determine whether can the same performance be achieved by GF methods on ALS and photogrammetric data, Klápště et al. (2020) compared six classic ground filters on both types of point clouds in the same area. In addition, a number of raster-based methods (Perko et al., 2015; Mousa et al., 2017; Gevaert et al., 2018; Mousa et al., 2019; Duan et al., 2019; Hingee et al., 2019) were designed for DTM extraction from photogrammetry-based digital surface models (DSMs). In particular, Gevaert et al. (2018) proposed a DL framework for DTM extraction from photogrammetric DSMs, in which additional true-orthophotos were also used. However, the concept of DTMs and DSMs is generally senseless in overhanging landforms (Bulatov et al., 2021). For extremely steep or even overhanging 3D point clouds, Bulatov et al.

(2021) and Štroner et al. (2021) proposed novel filtering strategies recently. On the whole, there are relatively few ground extraction methods tailored for photogrammetric data, especially those based on DL techniques. This study is expected to provide a reference for more advanced ground extraction from photogrammetric data.

3. The openGF dataset

It is extremely difficult for one individual academic institution or research team to build large-scale datasets in a short time, especially high-quality GF datasets with wide spatial coverage and rich terrain types. Fortunately, lots of worldwide ALS data in different terrain scenes have become publicly available. In particular, some of them have been finely labeled with the category *ground*. These existing friendly open-access data provide us an excellent opportunity to quickly build an ultra-large-scale GF dataset, named OpenGF. The specific details of constructing OpenGF are as follows.

3.1. Preprocessing

3.1.1. Selection of point cloud blocks

Although ground annotations have been provided in most open-access ALS point clouds, a lot of inconsistencies are observed in the classification quality. To tackle this problem, only point clouds with high-quality ground annotations were chosen. Then, these data was further picked to cover four prime terrain types, including *Metropolis* (areas with flat ground and dense/large roofs), *Small City* (areas with flat/rugged ground and many middle-sized roofs), *Village* (areas with natural ground and scattered roofs), and *Mountain* (areas with sloped ground and dense/sparse forests). Eventually, several large-sized point clouds with accurate ground annotations in different prime terrain scenes were extracted from the open-access ALS data (Qin et al., 2021).

Similar to previous works (Varney et al., 2020; Hu et al., 2022), the extracted large-sized point clouds were further split into blocks in the size of 500 × 500 m^2. There is no overlapping areas between adjacent blocks, so that each block is distinctly independent. In order to keep a balanced scene distribution, 40 point cloud blocks were selected from each prime terrain scene. In particular, to ensure as much scene diversity as possible, we concentrated on choosing blocks belonging to different sub-scenes. Finally, the four prime terrain types were subdivided into nine fine-grained terrain types (i.e., sub-scenes $S1$–$S9$). Specifically, among the selected 160 point cloud blocks, 20 belong to $S1$ (metropolis with large roofs), 20 belong to $S2$ (metropolis with dense roofs), 10 belong to $S3$ (small city with flat ground), 20 belong to $S4$ (small city with local undulating ground), 10 belong to $S5$ (small city with rugged ground), 40 belong to $S6$ (village with scattered roofs), 10 belong to $S7$ (mountain with gentle slopes and dense vegetation), 10 belong to $S8$ (mountain with steep slopes and sparse vegetation), and 20 belong to $S9$ (mountain with steep slopes and dense vegetation).

3.1.2. Annotation of ground and non-ground points

Besides high-quality ground labels, the majority of selected point cloud blocks have also been labeled with many other detailed categories such as *vegetation*, *buildings*, *water*, *bridges*, *outliers*, *overlap*. Nevertheless, the GF task focuses only on distinguishing two categories (i.e., *ground* and *non-ground*), which means that the other categories have to be merged.

Owing to containing both ground and non-ground points, the surface points of water mixed with vegetation and the points labeled *overlap* were firstly removed directly. Then, we changed the labels of the remaining points to *non-ground*, *ground*, or *unclassified*. Table 2 shows the definitions of the final three categories. Note that, low and high outliers were relabeled as a separate category *unclassified*, so that researchers have the option to merge them into non-ground points or delete them directly. Eventually, the three consolidated categories were repeatedly checked for quality assurance. In fact, only a few hours were needed to complete the annotation of the selected point cloud blocks. Several examples of the annotation results are shown in Fig. 1.

Table 2
Class definitions of OpenGF.

Class number	Class name	Definition
0	*Unclassified*	Low/high outliers
1	*Non-ground*	Buildings, low/medium/high vegetation, bridges, cars, etc.
2	*Ground*	Bare terrain, clean water surfaces

3.2. Statistics of the training set

The training set is composed of the 160 point cloud blocks processed above (see Section 3.1.1), which are distributed in four countries worldwide (see the red stars in Fig. 2a). The different geographical locations of these samples enhance the data diversity of the training set.

In addition, the number of ground and non-ground points in each terrain scene was displayed in Fig. 2b. It can be observed that the number distribution of points in the two categories is extremely unbalanced in partial terrain scenes. For example, mountain areas with dense vegetation (i.e., $S7$ and $S9$) contain significantly more non-ground points than ground points, while the number of ground points in village areas ($S6$) is far more than that of non-ground points. Besides, the scene $S5$ includes significantly more points than the scene $S3$, despite the fact that the two scenes have the same spatial coverage (i.e., 10 blocks), which demonstrates the diversified point density of the training set.

To facilitate choosing the best trained models for testing, nine representative point clouds belonging to different terrain types were carefully selected as the validation set, which are shown in Fig. 3. The remaining point cloud blocks were used for training.

3.3. Challenges of the test set

To fully evaluate the performance of DL-based GF pipelines, we first chose three challenging point clouds in areas outside the coverage of the training set, and then relabeled them as the test set using the methods in Section 3.1.2. The details of the test set, consisting of Test I, Test II, and Test III, are as follows.

Test I is a large-sized point cloud spanning around 6.6 km^2 in hybrid terrain scenes. It contains around 26.2 million non-ground points and 20.6 million ground points. The point density of Test I is uneven due to the mixing of scenes. In Fig. 4a, it can be observed that Test I contains three prime terrain types, including *village*, *small city*, and *mountain*. Four typical local areas (named I-A, I-B, I-C, and I-D) belonging to different terrain types are shown in Fig. 4b. Through the profiles from points, it can be found that the fine separation of ground and non-ground points in micro-topography areas of Test I is very challenging.

Test II is a noise-contaminated point cloud spanning about 1.1 km^2 in metropolis areas. It contains over 3.1 million non-ground points and 3.1 million ground points. The calculated point density of Test II is approximately 6 pts./m^2 on average, although a larger average point density of 14 pts./m^2 is reported on the related websites. As shown in Fig. 5b, Test II includes objects in various sizes, such as grass, cars, and small/large buildings. In particular, the maximum length and width of one building exceeds 200m. Proper recognition of such a large object requires sufficient spatial context. Test II also contains a large number of low and high outliers, which brings great challenges to classic ground filters.

Test III is a point cloud spanning over 2.3 km^2 in complex mountain areas. It contains about 16.2 million non-ground points and 3.6 million ground points. The point density of Test III is uneven owing to the aggregative distribution of dense vegetation. Fig. 6 presents Test III and its four typical local areas (named III-A, III-B, III-C, and III-D). In Fig. 6b, it can be seen that (1) The ground points under dense forests

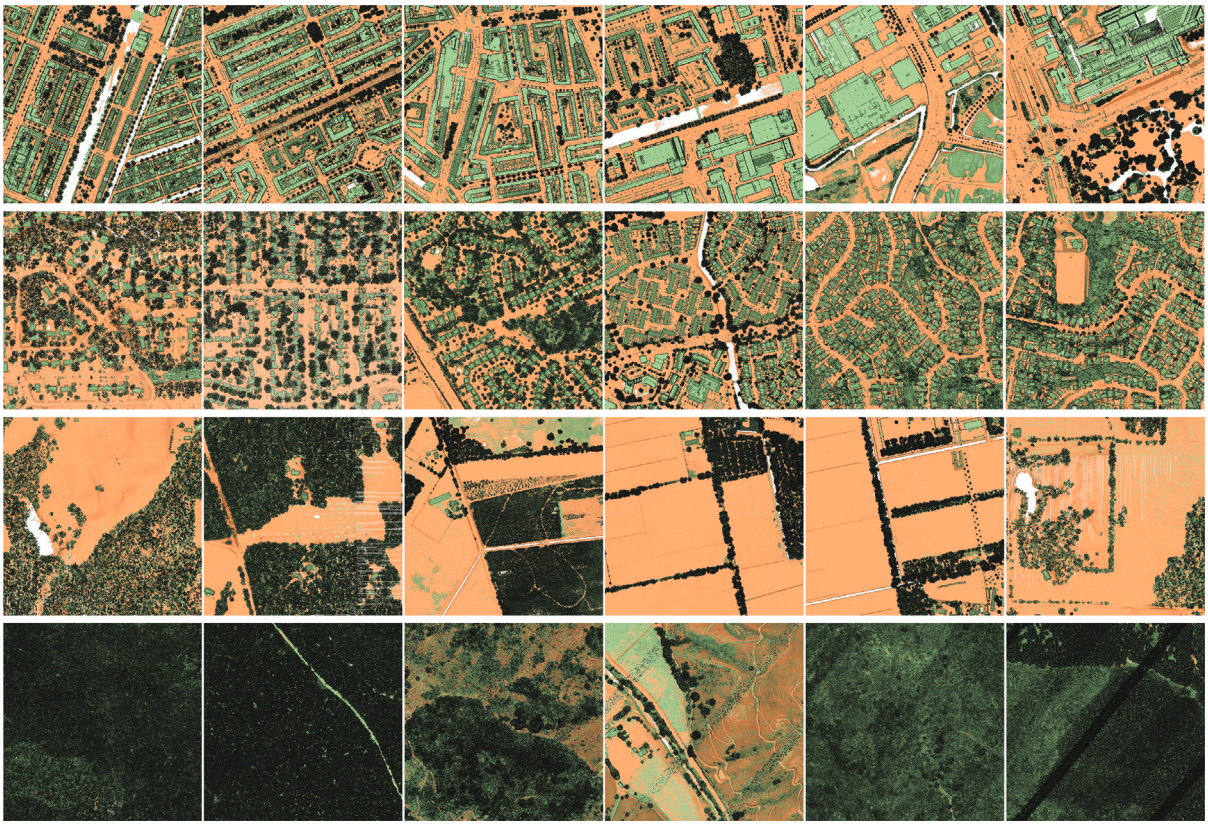

Fig. 1. Examples of the annotation results (top view). Ground and non-ground points are displayed in orange and green with shade effect, respectively. Note that, outliers are merged into non-ground points when displaying. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

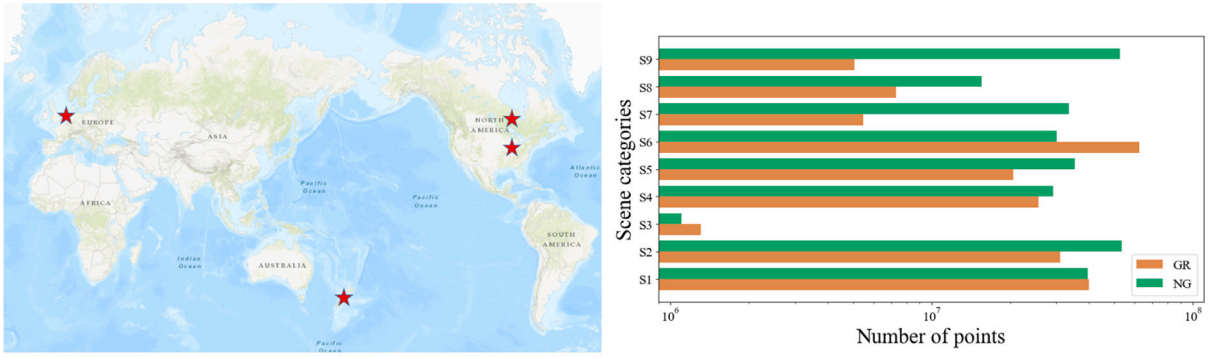

Fig. 2. Statistics of the training set. (a) Different geographical locations (the red stars) of the training samples. (b) The quantity of ground and non-ground points in nine different terrain scenes. GR and NG denote *ground* and *non-ground*, respectively.

are extremely sparse. (2) There are several large concave crack areas covered by vegetation. (3) Many small buildings are scattered on the terraced slopes.

Note that, our dataset only provides limited samples in typical terrain scenes, since its main objective is to facilitate the development of novel DL-based GF pipelines. If necessary, more training or test samples can be quickly produced according to the method provided above.

4. Comparative evaluations

Benefiting from the above constructed dataset, we can fairly compare the performance of classic ground filters and DL-based pipelines on a common dataset. We hope this evaluation would be highly complementary to the experimental comparison of filter algorithms conducted by ISPRS Working Group III/3 (Sithole and Vosselman, 2004).

4.1. Evaluated techniques

We carefully selected eight representative techniques for evaluation, including four classic ground filters and four state-of-the-art 3D DNNs. These methods cover the two mainstream categories as discussed in Section 2.2, and serve as solid baselines of our OpenGF benchmark.

4.1.1. Compared ground filters

PTD (Axelsson, 2000) is a progressive triangular irregular network (TIN) densification method. It first constructs a coarse TIN as the

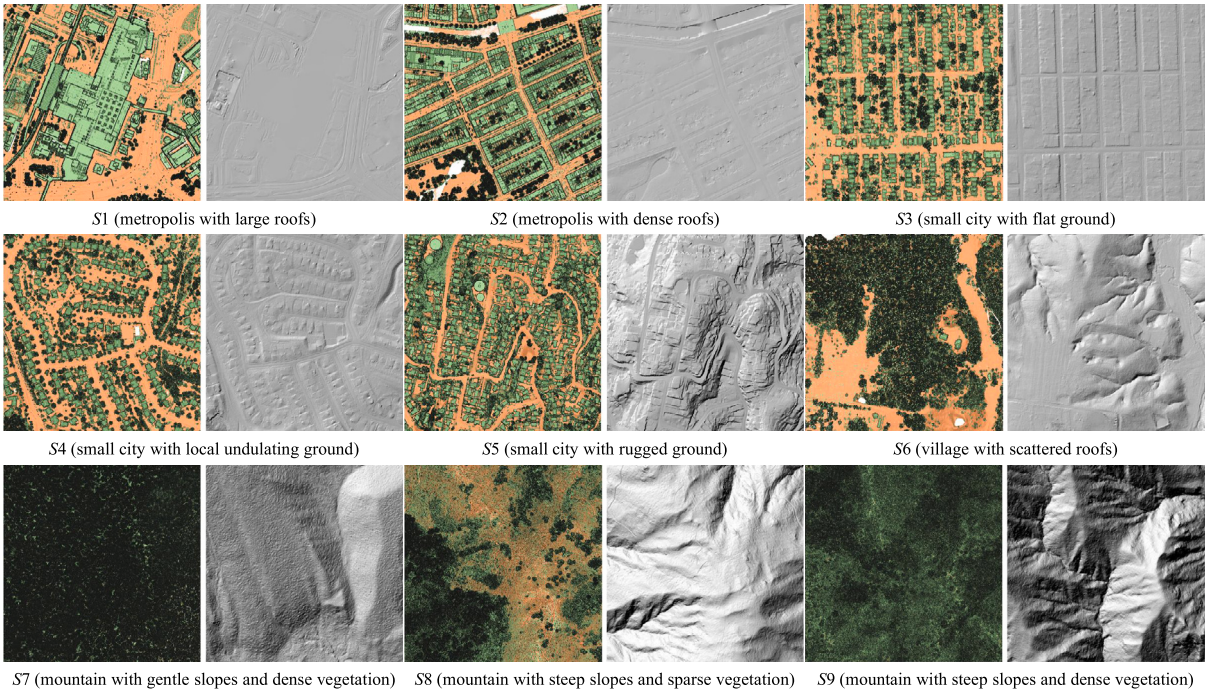

Fig. 3. Visualization of the validation set. Different blocks correspond to the nine defined terrain types. Each point cloud block is shown in two forms: the annotation result (*left*) and the reference DTM generated from the labeled ground points (*right*).

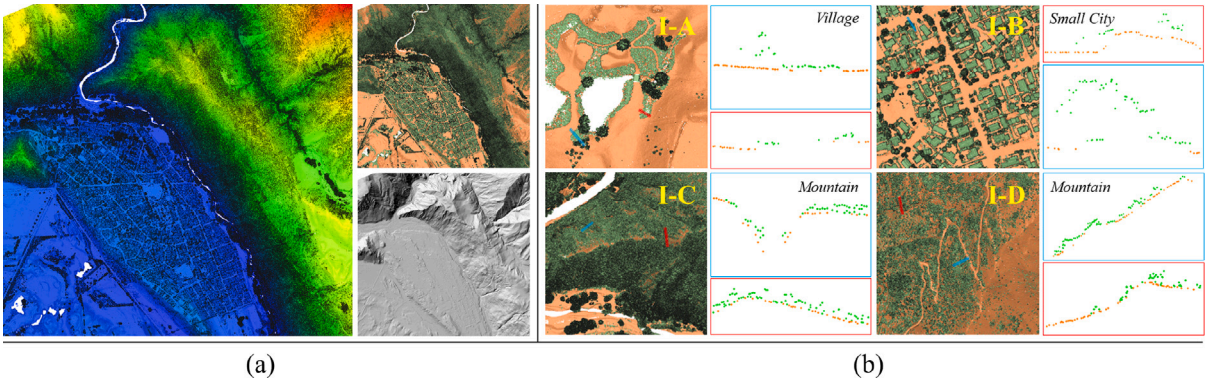

Fig. 4. An illustration of Test I. (a) The raw point cloud (*left*), the annotated result (*top right*), and the reference DTM generated from the labeled ground points (*bottom right*). (b) Examples of typical local areas. I-A belongs to *Village*, I-B belongs to *Small city*, I-C and I-D belong to *Mountain*. Two profiles are extracted from points in each local area, and shown in different boxes.

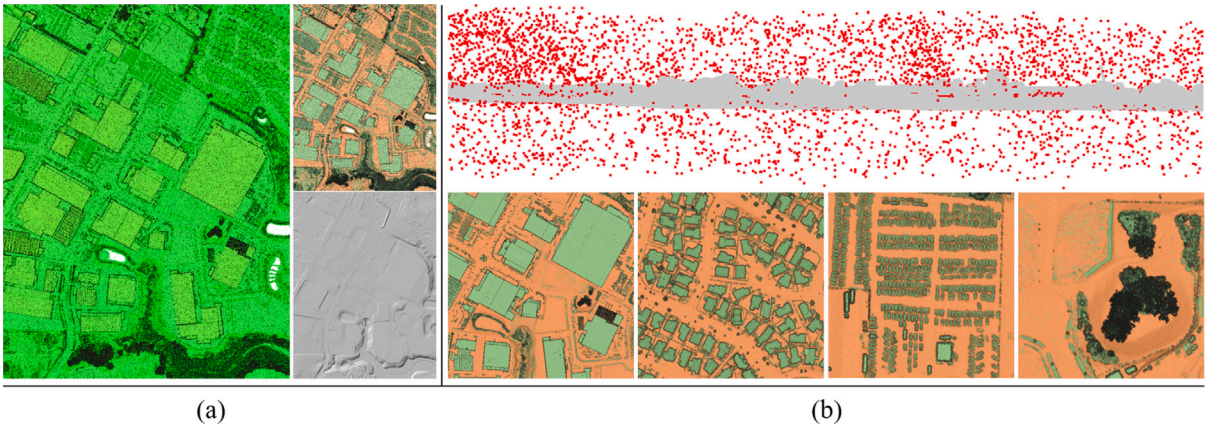

Fig. 5. An illustration of Test II. (a) The raw point cloud (*left*), the annotated result (*top right*), and the reference DTM generated from the labeled ground points (*bottom right*). (b) Dense outliers highlighted in red from the side view (*top*), and examples of objects in various sizes (*Bottom*). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

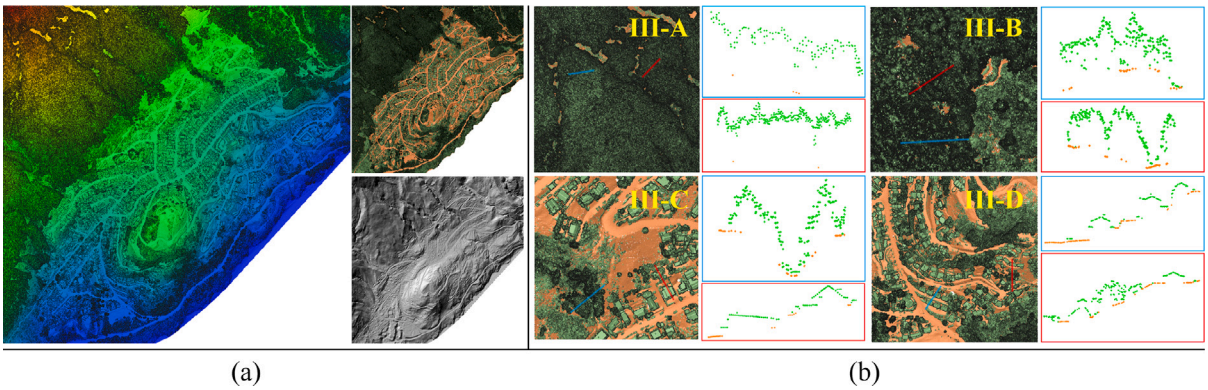

Fig. 6. An illustration of Test III. (a) The raw point cloud (*left*), the annotated result (*top right*), and the reference DTM generated from the labeled ground points (*bottom right*). (b) Examples of typical local areas. III-A and III-B are characterized by dense forests. III-C and III-D are characterized by terraced slopes. Two profiles are extracted from points in each local area, and shown in different boxes.

reference surface using the local lowest points, then classifies and adds more ground points to this TIN progressively.

PMF (Zhang et al., 2003) is a progressive morphological algorithm. It first defines a range of sliding windows with progressively changing sizes, and then separates ground points from non-ground points by the leverage of multiple morphological operations within these windows.

MCC (Evans and Hudak, 2007) is an iterative multi-scale classification method. It first adopts a multi-scale strategy in the process of surface interpolation and a progressive curvature tolerance that takes into account slope interaction with the points, then iteratively classifies ground and non-ground points according to surface curvature thresholds across multiple scales.

CSF (Zhang et al., 2016) is an easy-to-use approach based on cloth simulation. It first inverts a point cloud and covers the inverted data with a rigid cloth, then approximates the bare earth's surface according to interactions between the rigid cloth nodes and the points in corresponding areas, and finally discriminates ground points from non-ground points based on the generated reference surface.

4.1.2. Selected 3D DNNs

PointNet++ (Qi et al., 2017b) is an improved version of PointNet (Qi et al., 2017a), the pioneering work of point-based DL architectures which can directly consume raw 3D points without voxelization or projection operations. To overcome the limitation of characterizing local structures of PointNet, local geometrical features are extracted in a hierarchical way in the pipeline of PointNet++.

KPConv (Thomas et al., 2019) is a DL-based pipeline with kernel point convolution as the core operation, which can learn spatial correlation from point clouds directly. By the leverage of powerful rigid or deformable kernel points, its semantic segmentation architectures have achieved state-of-the-art classification performance on multiple 3D benchmarks.

RandLA-Net (Hu et al., 2020) is a DL-based pipeline focusing on efficient interpretation of large-scale point clouds. By taking advantage of the efficient random sampling and the lightweight local spatial encoding, this work has achieved leading efficiency performance on several point cloud datasets.

SCF-Net (Fan et al., 2021) is a DL-based pipeline with a learnable module consisting of a local polar representation block, a dual-distance attentive pooling block, and a global contextual feature block. By effectively learning spatial contextual features from large-scale point clouds, it has achieved state-of-the-art performance on two point cloud benchmarks.

4.2. Implementations and configurations

Ground filters. The implementations of the PTD algorithm in TerraScan software, PMF algorithm in PDAL library, MCC algorithm in MCC-LiDAR software, and CSF algorithm in CloudCompare software were directly adopted. We carefully chose the fine-tuned parameters of the four filters for each test data, through the trial-and-error strategy. The key parameters are specified in Table 3.

3D DNNs. The official or other reliable open-source implementations of the four selected DNNs were adopted in this work (Qin et al., 2021). Since none of these DL-based pipelines was tailored for the GF task, some minor modifications were introduced to make them adapt better to the OpenGF dataset. In the data preprocessing stage, we regarded outliers as non-ground points to participate in training by converting label 0 into label 1. Besides, to avoid numerical overflow in the process of model calculation, the minimum coordinates of each point cloud block were offset to the origin of coordinates. The grid downsampling (GD) size (i.e., the grid size for downsampling) of input point clouds was set to 1.0 m for the best classification performance through multiple attempts. We also adjusted the hyper-parameters of the DNNs according to the downsampled point density. In the training stage, we set the batch sizes of all DNNs as 4 or 2 due to the limited GPU memory. In the test stage, we used the trained models with the best overall accuracy (OA) on the validation set.

All experiments were conducted, under Ubuntu 18.04/Windows 10 system, on a workstation PC equipped with an NVIDIA RTX 2080Ti GPU and an Intel Core i9-9900K CPU @3.60 GHz, 32 GB RAM.

4.3. Evaluation metrics

Earlier works evaluated their GF methods with the metrics provided in the ISPRS filter test (Sithole and Vosselman, 2004), including $Type\ I$ (the proportion of ground points misclassified as the category *non-ground*), $Type\ II$ (the proportion of non-ground points misclassified as the category *ground*), and $Total$ errors (the proportion of all misclassified points). To facilitate more researchers in related fields to participate in GF studies, these conventional metrics are converted into similar or equivalent semantic segmentation metrics in this paper. Specifically, class-wise intersection-over-union (IoU) and OA are introduced. The adapted IoUs are computed as follows:

$$IoU_1 = \frac{TP_1}{TP_1 + FP_1 + FP_2},\ IoU_2 = \frac{TP_2}{TP_2 + FP_2 + FP_1} \tag{1}$$

where IoU_1 and IoU_2 refer to the IoU of *non-ground* (class 1) and *ground* (class 2), respectively. TP_1, FP_1, TP_2, and FP_2 denote, respectively, the number of correctly identified non-ground points, the

Table 3
Fine-tuned key parameters of the compared ground filters. Mb: "Maximum building size", Ia: "Iteration angle", Id: "Iteration distance", Cs: "Cell size", Mw: "Maximum window size", Sd: "Spacing for scale domain", Ct: "Curvature threshold", Cr: "Cloth resolution", Mi: "Maximum iterations".

	Test I	Test II	Test III
PTD	Mb, Ia, Id = 30 m, 40°, 1 m	Mb, Ia, Id = 210 m, 50°, 1.2 m	Mb, Ia, Id = 60 m, 30°, 0.4 m
PMF	Cs, Mw = 0.7 m, 20 m	Cs, Mw = 0.7 m, 200 m	Cs, Mw = 1 m, 30 m
MCC	Sd, Ct = 1 m, 0.3	Sd, Ct = 14 m, 0.3	Sd, Ct = 1 m, 0.3
CSF	Cr, Mi = 0.5 m, 800	Cr, Mi = 1.2 m, 500	Cr, Mi = 0.5 m, 500

Table 4
Evaluation of the selected eight representative methods on Test I.

	OA (%)	$RMSE$ (m)	IoU_1 (%)	IoU_2 (%)
PTD	94.82	**0.37**	91.10	89.00
PMF	90.63	0.51	85.22	79.62
MCC	**96.29**	0.42	**93.63**	**91.86**
CSF	93.07	0.95	88.17	85.64
PointNet++	97.58	0.25	95.75	94.68
KPConv	**97.79**	**0.20**	**96.10**	**95.17**
RandLA-Net	96.29	0.29	93.74	91.65
SCF-Net	95.75	0.28	92.90	90.43

number of non-ground points wrongly identified as the category *ground*, the number of correctly identified ground points, and the number of ground points wrongly identified as the category *non-ground*.

Furthermore, the filtering performance is also evaluated based on the quality of the extracted bare-earth surface. For convenience, following many previous GF studies (Hu and Yuan, 2016; Duan et al., 2019; Amini Amirkolaee et al., 2022), the root mean square error ($RMSE$) between the produced and reference DTMs is selected as another essential evaluation metric, which is computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(P_i - R_i)^2}{N}} \tag{2}$$

where N is the number of valid pixels in the raster DTM. P_i and R_i represent the elevation value of the ith valid pixel in the produced DTM and the reference DTM, respectively.

4.4. Performance comparison

The performance of the selected eight representative methods was compared qualitatively and quantitatively on the test set of OpenGF (i.e., Test I, Test II, and Test III), which highlights multiple attractive advantages of existing DL techniques in the problem of GF.

4.4.1. Evaluation on Test I

Test I covers multiple typical terrain types (see Fig. 4), which makes it a perfect case for evaluating the flexibility of classic ground filters and DL-based pipelines in hybrid terrain scenes. Table 4 lists the quantitative filtering results of the evaluated eight methods on Test I. With the help of highlighted errors, we also qualitatively compared the filtering results in Fig. 7.

In Table 4 and Fig. 7, it can be seen that: (1) In hybrid terrain scenes, DL-based pipelines have an obvious advantage over classic ground filters in terms of OA and $RMSE$. Only the worst OA (95.75%) achieved by DL-based pipelines is slightly lower than the best OA (96.29%) obtained by classic ground filters. (2) Among DL-based pipelines, KPConv achieves the best classification accuracy. Its $RMSE$ also surpasses those of other 3D DNNs. (3) Among classic ground filters, MCC achieves the best classification accuracy, but it is inferior to PTD in terms of $RMSE$. (4) The performance of classic ground filters may vary greatly in different terrain areas. In contrast, DL-based pipelines have more adaptive performance in hybrid landforms (see Fig. 7).

In addition, according to the IoUs achieved by different methods (see Table 4), we can further find the following: (1) DL-based pipelines have more balanced IoU_1 and IoU_2 than classic ground filters. A balance of IoU_1 and IoU_2 will most probably produce a high OA and a low $RMSE$. (2) The performance of DL-based pipelines is generally more stable than those of classic ground filters in hybrid terrain scenes. (3) Notably, although PointNet++ is an early work, its performance on Test I is comparable to that of the more recent and advanced KPConv. The full potential of PointNet++ has yet to be explored.

Highlighted Strengths of DL: DL-based pipelines are more flexible than classic ground filters in hybrid terrain scenes.

4.4.2. Evaluation on Test II

A great quantity of high and low outliers exist in Test II (see Fig. 5b), which provides us a great opportunity to test the effect of dense outliers on filtering performance. For comparative analyses, the original Test II was converted into Test II (w outliers) by treating outliers as non-ground points, and Test II (w/o outliers) by directly removing outliers. The filtering results of Test II are compared in Table 5 and Fig. 8.

According to the quantitative and qualitative comparisons, it can be seen that: (1) The adverse effect of outliers on the performance of DL-based pipelines is much smaller than that of classic ground filters (see Table 5), despite that denoising prior to filtering may reduce this gap. (2) For classic ground filters, the influence of outliers on IoU_2 is much greater than that on IoU_1, which is mainly caused by the low outliers in Test II (w outliers). (3) On Test II (w/o outliers), PTD achieves a much better $RMSE$ than all other methods, although its classification accuracy is slightly inferior to RandLA-Net.

Highlighted Strengths of DL: The sensitivity of DL-based pipelines to dense outliers is much lower than those of classic ground filters.

4.4.3. Evaluation on Test III

Test III is characterized by extremely sparse ground points under dense vegetation and terraced slopes with sharply changing elevations, which provides us a good chance to compare the robustness of different representative GF methods in complex mountain areas. Table 6 lists the quantitative filtering results of the evaluated eight methods on Test III.

In Table 6, it can be seen that: (1) In complex mountain areas, DL-based pipelines outperform classic ground filters in terms of classification accuracy, but they have similar $RMSE$ values. (2) Among DL-based pipelines, KPConv achieves the best performance in terms of all evaluation metrics. (3) Among the classic ground filters, PMF obtains the best $RMSE$ although it has the worst classification accuracy. On the contrary, PTD achieves the best classification accuracy, but it has the worst $RMSE$.

Moreover, the highlighted misclassifications can be intuitively observed in Fig. 9. In general, the performance of all the methods is quite well in simple terrain scenes (e.g., areas with flat ground/gentle slopes). Complex landforms, such as areas with dense vegetation, buildings on terraced slopes, and the discontinuous ground, appear to be the greatest challenge. According to the qualitative results in Fig. 9, the reason for the high $RMSE$ of PTD maybe due to the wrong removal of key terrain points in ridge areas.

Highlighted Strengths of DL: DL-based pipelines are relatively more robust than classic ground filters in complex mountain areas.

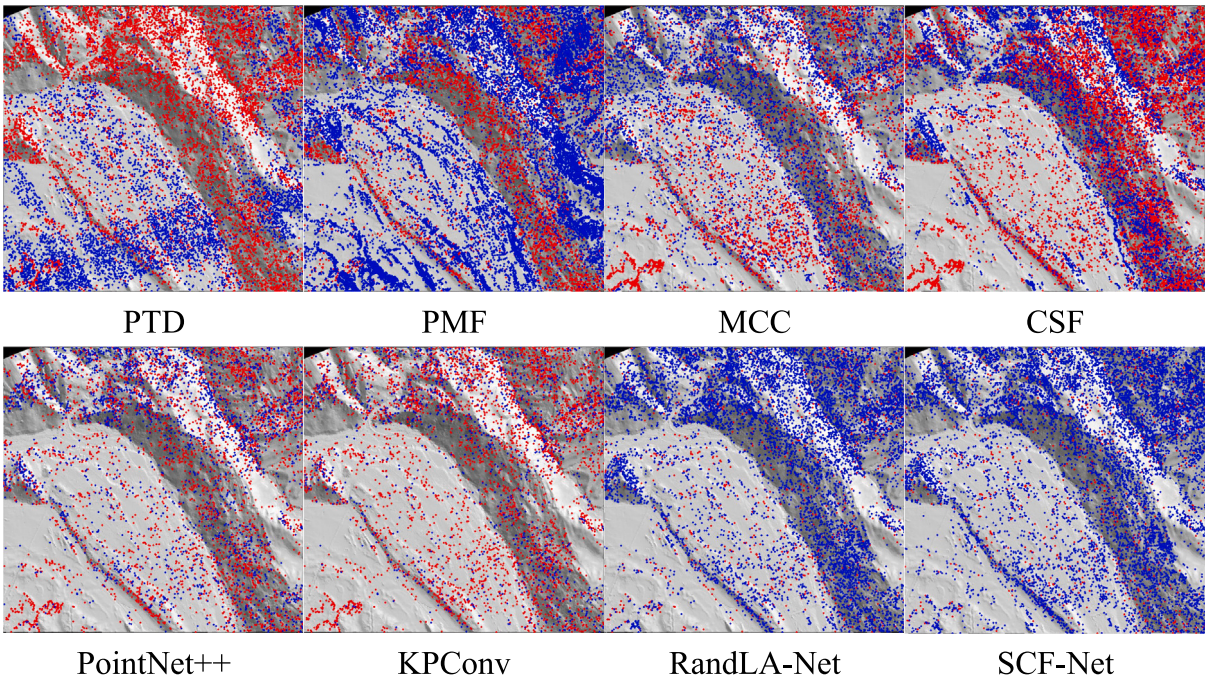

Fig. 7. Qualitative performance of different methods on Test I. The misclassified ground (blue) and non-ground (red) points are overlayed on the extracted bare ground surface. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Evaluation of the selected eight representative methods on Test II with/without outliers.

	Test II (w Outliers)				Test II (w/o Outliers)			
	OA (%)	$RMSE$ (m)	IoU_1 (%)	IoU_2 (%)	OA (%)	$RMSE$ (m)	IoU_1 (%)	IoU_2 (%)
PTD	50.32	90.37	50.32	0.00	**93.30**	**0.12**	**87.64**	**87.24**
PMF	50.87	66.76	50.59	1.15	86.56	1.11	78.50	73.61
MCC	51.84	**11.24**	51.05	3.25	84.44	1.63	75.39	70.27
CSF	**68.36**	12.24	**59.94**	**39.90**	89.34	1.89	81.08	80.38
PointNet++	86.40	4.79	75.46	76.63	87.38	4.89	75.19	79.63
KPConv	91.65	2.71	84.46	84.71	91.09	3.87	82.44	84.67
RandLA-Net	**94.28**	**1.84**	**89.29**	**89.05**	**94.96**	**1.20**	**90.38**	**90.42**
SCF-Net	90.43	2.58	83.08	81.95	90.91	2.88	83.35	83.32

Table 6
Evaluation of the selected eight representative methods on Test III.

	OA (%)	$RMSE$ (m)	IoU_1 (%)	IoU_2 (%)
PTD	**97.55**	3.24	**97.05**	**87.16**
PMF	94.93	**0.99**	94.09	73.67
MCC	96.97	1.29	96.40	84.12
CSF	95.35	1.77	94.42	78.29
PointNet++	98.12	3.64	97.72	90.24
KPConv	**98.31**	**0.79**	**97.94**	**91.28**
RandLA-Net	97.60	1.72	97.14	87.08
SCF-Net	97.23	2.94	96.70	85.18

Table 7
Approximate running time (minute) of six representative methods on the test set. Notably, the time consumed by fine-tuning of parameters of classic ground filters and the data preprocessing in DL-based pipelines are both not listed.

	Test I	Test II (w/o Outliers)	Test III
PTD	2.8	**0.2**	0.8
PMF	**0.7**	1.5	**0.4**
CSF	14.6	0.7	7.8
KPConv	2.3	**0.2**	1.3
RandLA-Net	**1.9**	0.3	**1.2**
SCF-Net	2.1	0.3	1.2

4.4.4. Computational efficiency

Computational efficiency is another important assessment indicator for GF techniques. Considering the running time is also decided by the coding quality, well-engineered software or codes were adopted (see Section 4.2) for the evaluation of time complexity. Note that, in order to evaluate as objectively as possible, MCC and PointNet++ were excluded because their open-source implementations are too time-consuming to be representative. Besides, the training time of DL-based pipelines is not reported here since the implementations and training strategies of different 3D DNNs are not directly comparable.

The running time of six representative methods are shown in Table 7. It can be seen that: (1) The inference time of DL-based pipelines increases approximately linearly with the increase of the data size, while that of classic ground filters is highly influenced by the algorithm strategy and parameter settings. (2) As the amount of points increases, the running time of CSF becomes much longer than those of KPConv and RandLA-Net. (3) Although the amount of data of Test I is about 8 times that of Test II (w/o outliers), the time spent by PMF on Test II (w/o outliers) is almost 2 times that of Test I. This is mainly because that PMF needs more iterative processing to filter out the large buildings in Test II (w/o outliers).

Highlighted Strengths of DL: The variation of the inference time of DL-based pipelines is more stable than that of classic ground filters on different amount of data.

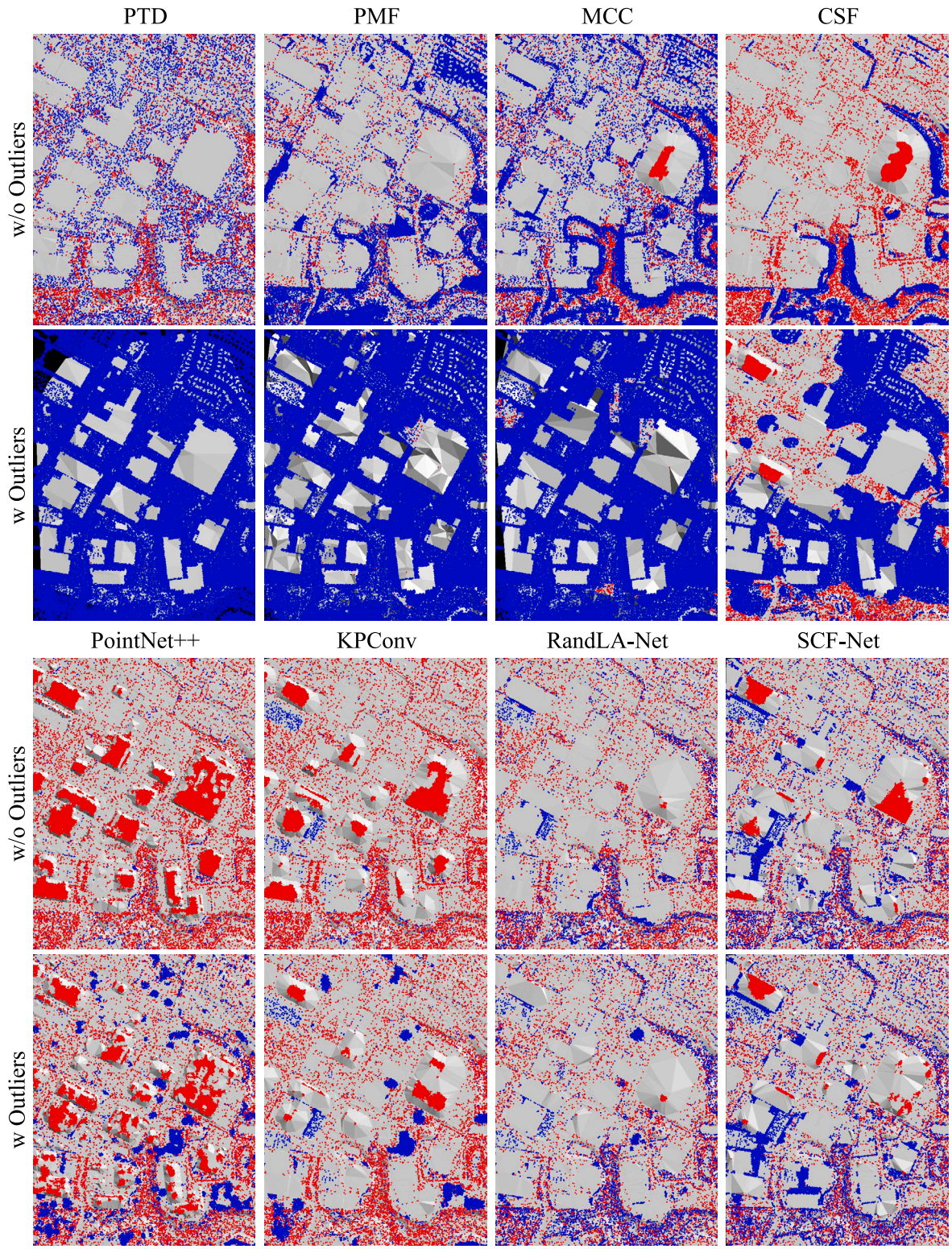

Fig. 8. Qualitative performance of different methods on Test II with/without outliers. The misclassified ground (blue) and non-ground (red) points are overlayed on the extracted bare ground surface. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. Research issues

As clarifying the remaining challenges is also helpful to stimulate future research, we further point out several issues faced by generalizing existing DL techniques into GF tasks with reference to in-depth experimental analyses.

5.1. Unsatisfied micro-topography errors

From a practical point of view, correcting a small amount of micro-topography errors in complex terrain scenes manually often spends much more time than fixing a large number of apparent misclassifications in simple terrain scenes. It remains an open question whether the

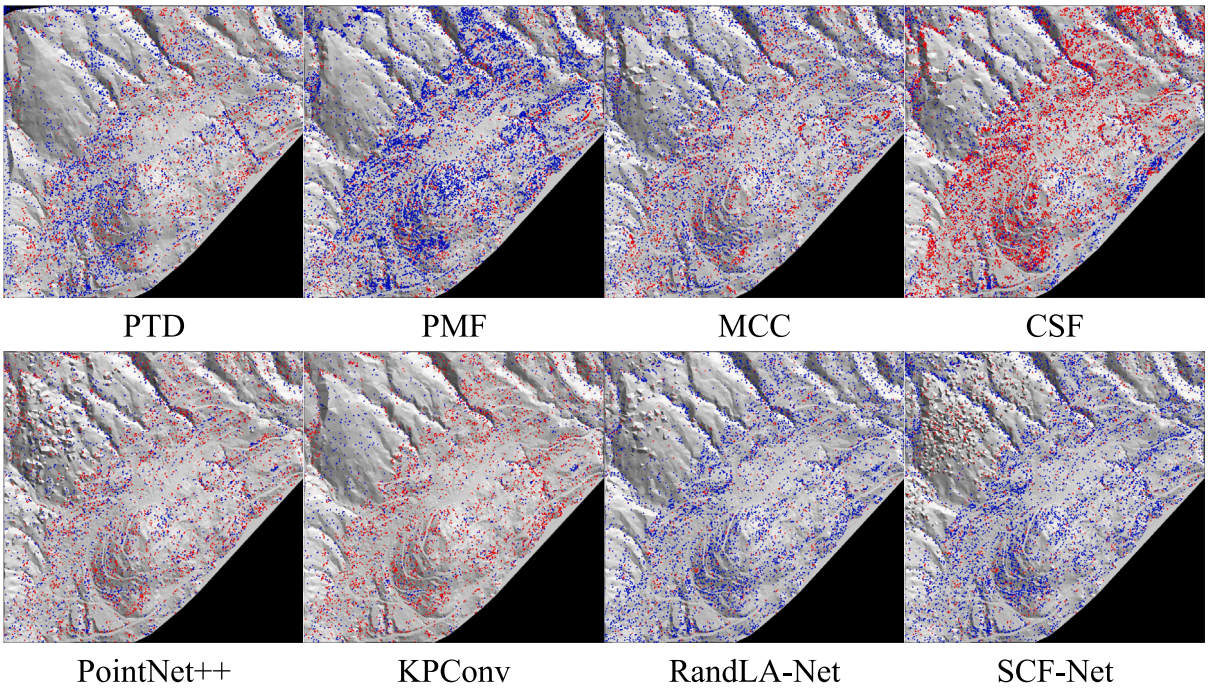

Fig. 9. Qualitative performance of different methods on Test III. The misclassified ground (blue) and non-ground (red) points are overlayed on the extracted bare ground surface. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Quantitative results of the selected eight methods in four challenging local areas (I-A, I-C, III-A, and III-D) of the test set.

	IoU_1 (%)				IoU_2 (%)				OA (%)				$RMSE$ (m)			
	I-A	I-C	III-A	III-D	I-A	I-C	III-A	III-D	I-A	I-C	III-A	III-D	I-A	I-C	III-A	III-D
PTD	45.84	91.64	**99.08**	**93.52**	84.60	78.68	**74.61**	**87.44**	86.38	93.61	**99.10**	**95.53**	0.19	0.61	5.83	**0.26**
PMF	46.82	89.84	98.38	87.62	80.00	70.91	55.10	74.68	83.00	91.86	98.41	90.93	**0.16**	0.80	**1.64**	0.80
MCC	50.75	**93.25**	98.75	93.29	90.88	**78.71**	66.06	86.91	91.66	**94.60**	98.78	95.36	0.19	**0.59**	2.97	0.40
CSF	**51.78**	88.13	98.06	88.20	**91.07**	65.39	56.95	80.21	**91.85**	90.30	98.10	92.02	0.18	2.79	4.67	0.81
PointNet++	74.84	94.40	99.13	95.46	95.14	83.30	77.96	91.50	95.75	95.62	99.16	96.95	0.18	0.47	9.32	0.67
KPConv	75.08	**95.07**	**99.41**	**95.74**	95.19	**85.73**	**84.37**	**92.16**	95.80	**96.20**	**99.43**	**97.16**	0.16	**0.37**	**1.90**	0.67
RandLA-Net	**88.62**	91.58	99.21	93.98	**97.56**	71.92	78.14	87.89	**97.95**	93.07	99.24	95.81	**0.13**	0.98	5.04	**0.36**
SCF-Net	88.15	91.64	98.74	93.00	97.43	72.28	66.50	85.88	97.84	93.14	98.77	95.09	0.13	0.59	7.22	1.00

micro-topography errors created by state-of-the-art 3D DNNs are few enough to be satisfactory.

To determine the impact of these micro-topography errors on the quality of the extracted bare ground surface, we conducted further experimental analyses in some challenging local areas of the test set. Table 8 lists the quantitative results achieved by the selected eight methods (see Section 4) in four challenging local areas (i.e., I-A, I-C, III-A, and III-D). Fig. 10 highlights the errors occurring in the extracted bare ground with different colors.

According to Table 8, it can be seen that: (1) In terms of IoUs, the state-of-the-art 3D DNNs have a similar performance trend with classic ground filters. (2) In local areas dominated by ground points (e.g., the local area I-A in Fig. 4b), the classification accuracy of the category *ground* (i.e., IoU_2) is significantly higher than that of the category *non-ground* (see Table 8). Instead, the classification accuracy of the category *non-ground* (i.e., IoU_1) is significantly higher than that of the category *ground* (see Table 8) in local areas with lots of non-ground points (e.g., the local area III-A in Fig. 6b). In other words, the category with more points will have a higher classification accuracy, which can be attributed to the tendency of the evaluated methods to maximize the OA.

In Table 8 and Fig. 10, we can further see that: (1) Both classic ground filters and DL-based pipelines have made relatively serious errors in challenging local areas. (2) Although an OA of over 98% can be achieved by all the evaluated methods in the local area III-A of Test III, the corresponding $RMSE$ achieved by these methods is disappointing (see Table 8). This means that a high classification accuracy does not necessarily lead to a high-quality bare ground surface. The quality of bare ground surface depends more on where misclassifications occur in the terrain scene (see Fig. 10). (3) All the evaluated DNNs failed to correctly identify small objects close to the bare ground, although the classification accuracy exceeds that of classic ground filters. Meanwhile, like classic ground filters, many valuable ground points are wrongly classified (see Fig. 10) by DNNs. This indicates that existing DL-based pipelines cannot suppress the occurrence of a small amount of micro-topography errors.

5.2. Failures to recognize large objects

Large point cloud blocks are usually further partitioned into patches before feeding into existing DNNs owing to the limited GPU memory. However, the determination of the patch size is usually a thorny problem. Small-sized patches will break the geometrical structure of large objects, while large-sized patches will result in an unbearable GPU memory consumption. To ensure sufficient spatial coverage and affordable GPU memory at the same time, a compromise is to feed a downsampled large-sized input to the existing DNNs. Yet, an excessively large GD size will inevitably lose the details of the original data and decrease the overall classification accuracy.

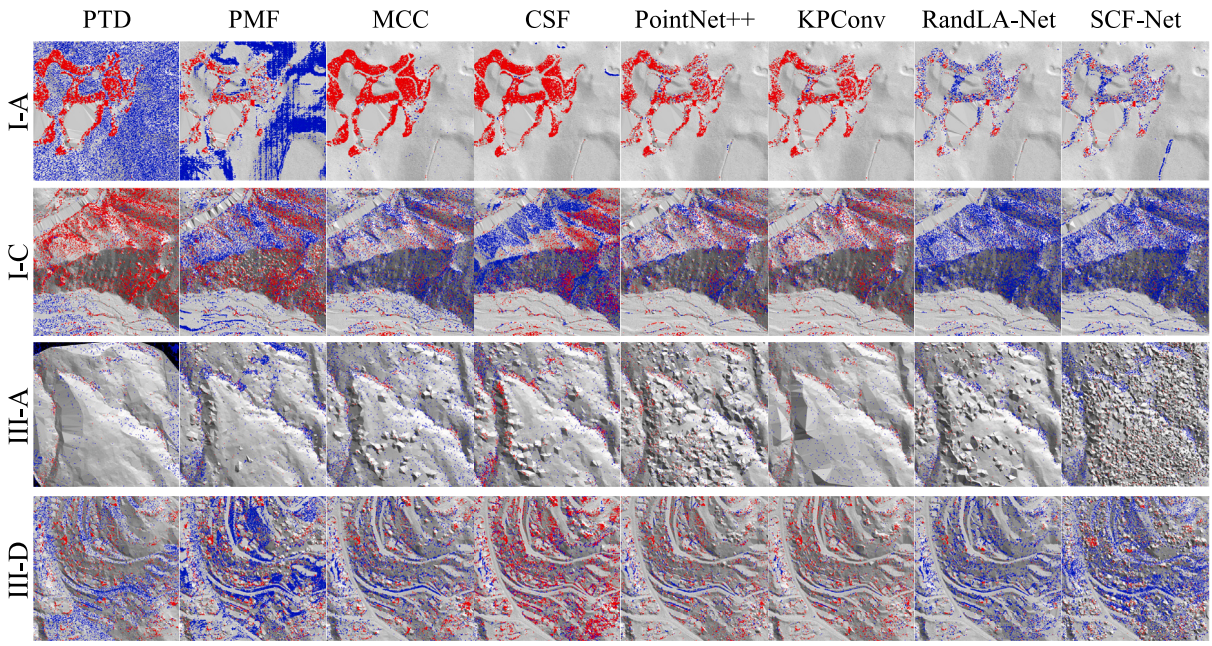

Fig. 10. Qualitative performance of the selected eight methods in four challenging local areas (i.e., I-A, I-C, III-A, and III-D) of the test set. The misclassified ground (blue) and non-ground (red) points are overlayed on the extracted bare ground surface. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 9
Quantitative results achieved by the evaluated 3D DNNs on Test II (w outliers) under different input configurations.

	Input configuration			Performance			Efficiency
	GD size (m)	Constant-volume input (m)	Constant-number input	OA (%)	IoU_1 (%)	IoU_2 (%)	Inference time (s)
PointNet++	0.5	35 (cube length)	8192	82.10	66.54	72.22	–
	1.0	50 (cube length)	8192	**86.40**	75.46	**76.63**	–
	1.5	140 (cube length)	8192	85.82	**75.91**	74.37	–
KPConv	0.5	25 (radius of sphere)	–	89.24	79.67	81.40	58
	1.0	50 (radius of sphere)	–	**91.65**	**84.46**	**84.71**	20
	1.5	75 (radius of sphere)	–	89.86	80.17	82.81	**12**
RandLA-Net	0.5	–	65535	91.89	85.01	84.99	39
	1.0	–	65 535	**94.28**	**89.29**	**89.05**	21
	1.5	–	65 535	91.28	84.39	83.51	**19**
SCF-Net	0.5	–	65 536	85.19	76.50	71.40	50
	1.0	–	65 536	**90.43**	**83.08**	**81.95**	19
	1.5	–	65 536	88.38	79.23	79.12	**17**

Accordingly, we further conducted comparative experiments to find a trade-off among the GD size, input coverage, performance as well as efficiency under the limited GPU memory. Table 9 reports the quantitative results achieved by the evaluated 3D DNNs on Test II (w outliers) under different input configurations. In addition, the qualitative performance achieved by RandLA-Net with different input configurations is shown in Fig. 11.

In general, it can be found from the quantitative and visualization results that: (1) The evaluated 3D DNNs show similar responses to the change in the GD size. As the GD size increases, the performance of the four DNNs increases first and then decreases. This result is not unexpected. When the GD size is small (e.g., 0.5 m in Table 9), the details of the original data can be well preserved, but the small coverage of inputs may lead to the incorrect recognition of many large objects (e.g., Fig. 11a). Instead, if the GD size is large (e.g., 1.5 m in Table 9), the input of DNNs will have a large coverage at the cost of losing the details of the original data, which may result in more misclassifications in local areas (e.g., Fig. 11c).(2) When using an appropriate GD size (e.g., 1.0 m in Section 4.2), all the four DNNs achieve their best performance on Test II (w outliers) at a relatively fast inference speed. However, lots of points belonging to large buildings are still misclassified by most DNNs (see Fig. 8). (3) The performance of RandLA-Net surpasses those of other DNNs by a large margin on Test II (w outliers) (see Table 9), which may attribute to that RandLA-Net has the capability of processing large-scale inputs in a single pass (Hu et al., 2020). (4) RandLA-Net fails to correctly identify large buildings (see Fig. 11a) when using a small grid size (e.g., 0.5 m in Table 9) for downsampling. This demonstrates that large-scale spatial context is a key factor for the accurate identification of large objects.

5.3. Insufficient generalization capability

Since DNNs are usually trained on given samples and tend to overfit a specific data distribution, it is meaningful to explore whether DL models trained on the OpenGF dataset can be well applied to other GF datasets (e.g., the ISPRS filtertest dataset).

In this paper, four trained DL models with the best performance on the OpenGF dataset (i.e., the PointNet++, KPConv, RandLA-Net, and SCF-Net in Section 4) were used directly to test the 15 reference study sites of the ISPRS filtertest dataset. Note that, the two datasets differ in many aspects such as the point density and geographical regions, and thus may follow a different data distribution.

The quantitative results (OA, IoU_1, and IoU_2) on the ISPRS filtertest dataset achieved by the four trained DL models are respectively

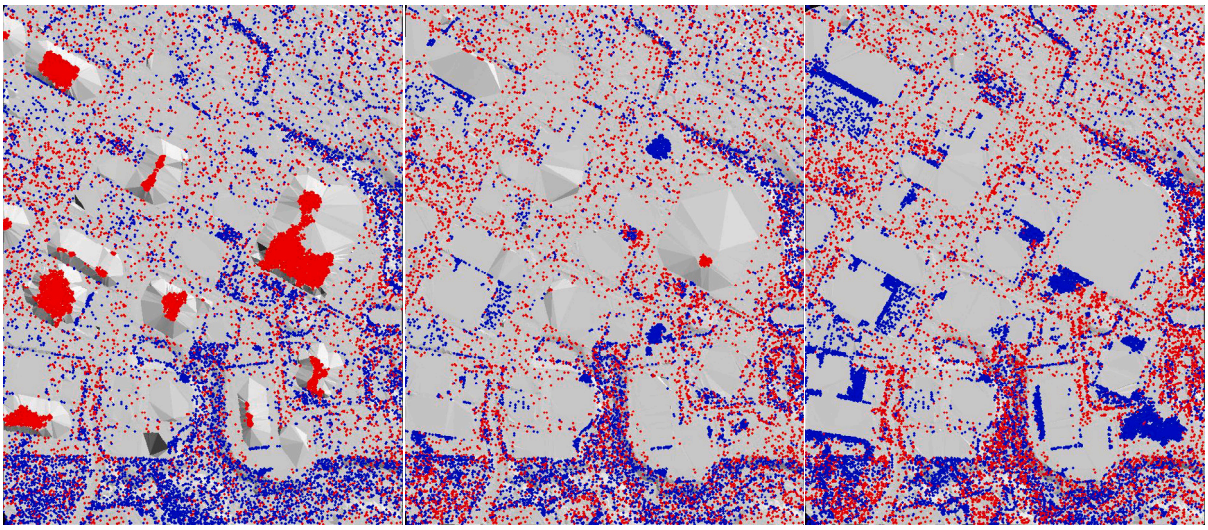

Fig. 11. Qualitative performance achieved by RandLA-Net on Test II (w outliers) with GD size of (a) 0.5 m, (b) 1.0 m, and (c) 1.5 m. The misclassified ground (blue) and non-ground (red) points are overlayed on the extracted bare ground surface. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 10
Cross-dataset generalization performance (OA, %) of four DL models trained on OpenGF.

	Reference samples of ISPRS filtertest															Avg.
	11	12	21	22	23	24	31	41	42	51	52	53	54	61	71	
PointNet++	82.20	95.02	95.96	90.61	86.29	90.90	97.56	83.51	98.01	88.87	62.33	63.07	82.99	86.16	77.60	85.40
KPConv	80.00	95.68	97.82	89.30	87.81	90.03	98.60	84.98	96.47	95.12	88.62	84.11	92.95	96.74	96.87	**91.67**
RandLA-Net	80.11	94.08	95.64	87.05	82.45	90.35	97.62	82.80	91.53	95.38	94.06	77.10	92.43	97.25	97.21	90.34
SCF-Net	80.05	91.46	94.07	85.49	80.39	87.41	96.56	73.61	86.86	96.51	92.55	72.17	93.63	89.17	54.38	84.95
Elmqvist	77.60	91.82	91.47	91.07	87.72	86.17	94.66	91.24	96.32	78.69	42.05	51.55	78.74	64.13	65.78	80.23
Sohn	79.51	91.61	91.20	92.46	90.16	86.67	93.61	88.73	98.22	90.69	87.96	79.81	94.32	97.01	97.80	90.14
Axelsson	89.24	96.75	95.75	96.37	96.00	95.58	95.22	86.09	98.38	97.28	96.93	91.09	96.77	97.92	98.37	**94.96**
Pfeifer	82.65	95.50	97.43	93.29	91.78	91.36	98.20	89.25	97.36	96.29	80.36	87.40	94.53	93.09	91.15	91.98
Brovelli	63.04	83.72	90.70	77.72	72.20	63.94	87.08	82.97	93.62	77.19	54.44	47.19	76.11	78.32	65.02	74.22
Roggero	79.20	93.39	90.16	76.22	76.80	76.75	97.86	87.79	95.70	96.99	90.22	82.71	95.04	81.01	94.89	87.65
Wack	75.98	93.39	95.45	92.49	89.03	88.47	97.79	90.99	96.46	88.55	76.17	72.76	92.37	86.53	83.03	87.96
Sithole	76.75	89.79	92.24	79.14	77.29	74.72	96.85	76.33	96.15	92.98	72.47	62.93	93.67	78.37	78.17	82.52

Note that, the OAs achieved by various ground filters, including Elmqvist (Elmqvist et al., 2001), Sohn (Sohn and Dowman, 2002), Axelsson (Axelsson, 2000), Pfeifer (Pfeifer et al., 2001), Brovelli (Brovelli et al., 2002), Roggero (Roggero, 2001), Wack (Wack and Wimmer, 2002), and Sithole (Sithole and Vosselman, 2001), are converted from the corresponding $Total$ errors published in Sithole and Vosselman (2003).

Table 11
Cross-dataset generalization performance (IoU_1, %) of four DL models trained on OpenGF.

	Reference samples of ISPRS filtertest															Avg.
	11	12	21	22	23	24	31	41	42	51	52	53	54	61	71	
PointNet++	69.93	90.39	83.34	75.63	75.99	74.00	94.77	72.13	97.22	63.37	19.96	9.17	74.74	18.98	31.29	63.39
KPConv	67.72	91.73	90.71	74.04	79.00	72.38	97.00	76.83	95.23	80.67	45.45	19.04	88.11	50.30	77.14	**73.69**
RandLA-Net	67.81	89.00	83.10	70.31	72.48	73.07	95.01	73.96	89.27	79.15	58.20	11.39	86.07	52.72	76.43	71.86
SCF-Net	67.18	84.00	78.45	68.03	70.07	67.64	92.85	62.92	84.31	84.17	54.45	12.00	88.60	23.78	19.69	63.88

reported in Tables 10, 11, and 12. In addition, Fig. 12 shows the qualitative comparison of cross-dataset generalization results achieved by the four trained DL models on two representative ISPRS samples.

It can be seen that: (1) The cross-dataset generalization performance of the four trained DL models (see Tables 10, 11, and 12) shows different degrees of degradation, compared with their intra-dataset performance (see Table 6). (2) In common urban areas (e.g., the sample 31 of ISPRS filtertest), the four trained DL models can still obtain relatively good performance, while in some special landforms (e.g., the sample 53 of ISPRS filtertest) the four trained DL models make serious errors (see Table 10 and Fig. 12). (3) KPConv achieves the best generalization performance among the evaluated 3D DNNs with an average OA of 91.67%, although the accuracy is still lower than that achieved by some classic ground filters (see Table 10). This demonstrates that the generalization capability of existing DL techniques is insufficient in the GF task.

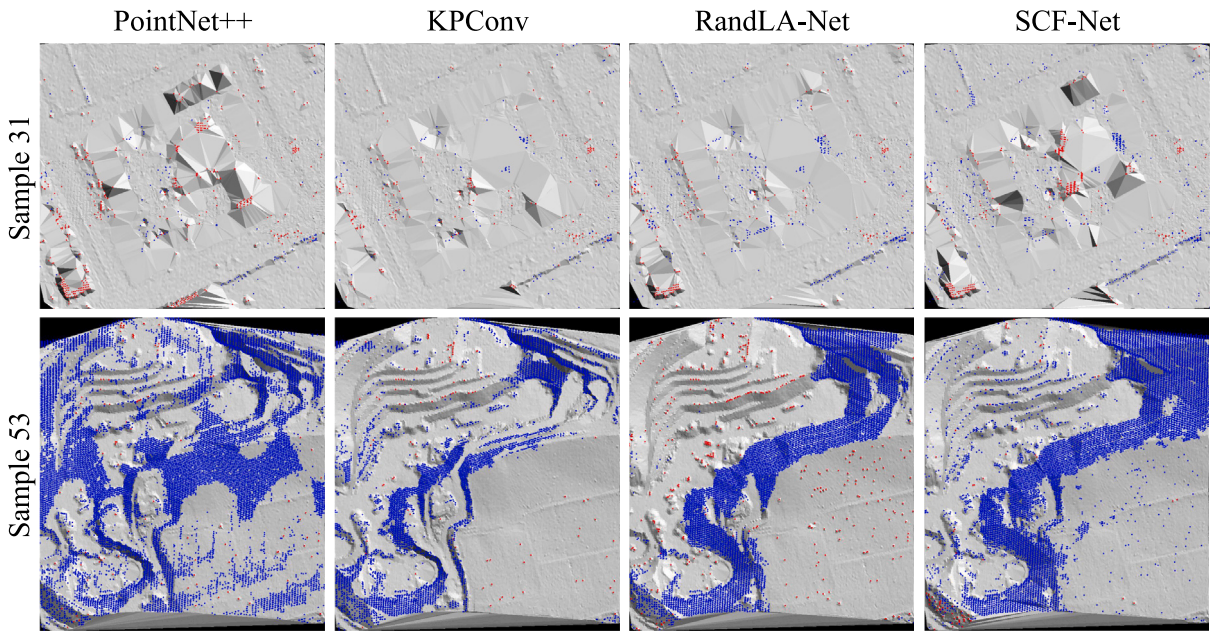

Fig. 12. Qualitative cross-dataset generalization results achieved by four DL models trained on OpenGF. The misclassified ground (blue) and non-ground (red) points are overlayed on the extracted bare ground surface. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 12
Cross-dataset generalization performance (IoU_2, %) of four DL models trained on OpenGF.

	Reference samples of ISPRS filtertest															Avg.
	11	12	21	22	23	24	31	41	42	51	52	53	54	61	71	
PointNet++	69.64	90.65	94.93	86.74	75.79	87.71	95.62	71.24	93.43	86.21	58.43	61.64	65.76	85.70	75.06	79.90
KPConv	65.55	91.70	97.24	84.61	77.49	86.50	97.43	70.08	88.02	93.87	87.43	83.50	85.22	96.63	96.51	**86.79**
RandLA-Net	65.76	88.64	94.45	81.33	67.35	86.93	95.65	66.37	71.25	94.39	93.52	76.41	85.76	97.16	96.94	84.13
SCF-Net	66.28	84.52	92.43	79.01	63.75	82.92	93.77	52.20	55.36	95.71	91.82	71.07	87.40	88.79	48.63	76.91

6. Discussion and limitations

6.1. Comprehensive discussion

Based on the above extensive experimental comparisons and analyses, the advantages, challenges, and directions of DL in GF can be summarized as following.

Advantages. In simple terrain scenes, both DL-based pipelines and classic ground filters can achieve good performance. DL-based pipelines outperform classic ground filters mainly in the following aspects: (1) In hybrid terrain scenes, DL-based pipelines have an obvious advantage over classic ground filters in terms of OA and $RMSE$. The performance of classic ground filters may vary greatly in different terrain areas. In contrast, DL-based pipelines have more adaptive performance in hybrid landforms. (2) Compared with classic ground filters, the adverse effect of outliers on the performance of DL-based pipelines is much smaller. The reason for this result may be that classic ground filters typically assume the local lowest points as the ground, while DL-based pipelines do not rely on any prior assumptions when classifying points. (3) In complex mountain areas, DL-based pipelines outperform classic ground filters in terms of classification accuracy, although both of them have poor performance in terms of $RMSE$. (4) The running time of classic ground filters is highly influenced by the algorithm strategy and parameter settings. The variation of the inference time of DL-based pipelines is more stable than that of classic ground filters on different amounts of data.

Challenges. Similar to classic ground filters, state-of-the-art 3D DNNs often lose key ground points or retain low object points in difficult local areas, indicating that DL-based pipelines also cannot suppress the occurrence of a small amount of micro-topography errors. In addition, adapting advanced DL techniques for GF brings some new challenges: (1) Due to the limited GPU memory, most of current 3D DNNs can only handle small-sized inputs while achieving their best overall performance, causing the misclassification of large buildings in the filtering results. This is not surprising considering that large-scale spatial context is essential for the accurate classification of large objects. (2) Although state-of-the-art 3D DNNs have multiple advantages in the GF task, they are usually trained on given samples and tend to overfit a specific data distribution. Owing to the domain gaps between different datasets, the cross-dataset generalization performance of the trained DL models for GF shows different degrees of degradation, compared with their intra-dataset performance.

Directions. In response to the above challenges, some promising research directions are suggested as follows: (1) To improve the quality of the extracted bare ground surface in difficult areas, it is considerably encouraged to develop effective strategies for micro-topography error suppression. (2) To accurately recognize large objects, it is urgently needed to design large-scale spatial context embedding mechanisms. (3) To improve the generalization capability of DL-based pipelines for GF, it is necessary to introduce advanced transfer learning techniques.

6.2. Current limitations

Although this work has considered many aspects of GF, there are still some limitations to address: (1) Since ALS point clouds are typically collected from high altitudes and have relatively low density, they have problems describing nearly vertical terrains (Štroner et al., 2021). Inevitably, our ALS point cloud dataset lacks extremely steep even overhanging point clouds. From this point of view, the proposed OpenGF is a little 2.5D-biased. Meanwhile, since the elevation of overhanging

point clouds cannot be represented as a continuous function of plane coordinates, the widely used metric $RMSE$ would not make much sense in this situation (Bulatov et al., 2021). In the future, more datasets and advanced GF techniques are needed to explore, especially in some important small-scale applications (e.g., civil engineering and natural disaster monitoring). (2) Thanks to the proposed OpenGF dataset, extensive investigations on the GF of ALS point clouds can be conducted here. Nevertheless, due to a lack of available large-scale GF datasets from other sources (e.g., UAV photogrammetry and MLS), the applicability of DL-based GF methods to other kinds of point clouds (e.g., photogrammetric point clouds and MLS point clouds) was not in-depth explored in this paper. In the future, it is encouraged to conduct a comprehensive survey on ground extraction from different types of 3D point clouds.

7. Conclusion and outlook

This paper first introduced an ultra-large-scale ALS dataset tailored for the GF task, which covers about 47.7 km^2 and contains nine different terrain types from four countries worldwide. Then, extensive comparative evaluations of eight representative methods, including four state-of-the-art 3D DNNs and four classic ground filters, were carried out on the proposed dataset, which highlights multiple strengths of DL techniques in GF. Furthermore, several key issues faced by generalizing existing DNNs into GF tasks were revealed with reference to in-depth analyses. Finally, some promising directions for future research were suggested in response to the identified challenges.

Through a series of comparative experiments, it can be concluded that: (1) The OpenGF dataset has the capability of effectively training advanced DL models for GF. (2) Compared with classic ground filters, DL-based GF pipelines have advantages in many aspects, such as flexibility in mixed terrain scenes, sensitivity to dense outliers, robustness in complex landforms, and stability in computational efficiency. (3) The issues faced by generalizing advanced DL techniques into GF tasks mainly lie in unsatisfied micro-topography errors, failures to recognize large objects, and insufficient generalization. (4) Promising directions for developing more advanced DL-based GF pipelines include micro-topography error suppression, large-scale spatial context embedding, and advanced transfer learning.

In the future, it is valuable to explore more types of GF datasets and advanced GF techniques, especially in some important small-scale applications. In addition, a comprehensive survey on the GF of multi-source point clouds will be useful for both academic and industrial communities. We hope this work can serve as a reference for these interesting directions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China [grant numbers 42001400, 42101451]. The authors would like to thank the reviewers for the constructive comments.

References

Amini Amirkolaee, H., Arefi, H., Ahmadlou, M., Raikwar, V., 2022. DTM extraction from DSM using a multi-scale DTM fusion strategy based on deep learning. Remote Sens. Environ. 274, 113014.

Andersen, H.-E., Reutebuch, S.E., McGaughey, R.J., d'Oliveira, M.V., Keller, M., 2014. Monitoring selective logging in western Amazonia with repeat lidar flights. Remote Sens. Environ. 151, 157–165.

Axelsson, P., 2000. DEM generation from laser scanner data using adaptive TIN models. Int. Arch. Photogramm. Remote Sens. XXXIII, 110–117.

Ayazi, S.M., Saadat Seresht, M., 2019. Comparison of traditional and machine learning base methods for ground point cloud labeling. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII-4/W18, 141–145.

Bartels, M., Wei, H., Mason, D.C., 2006. DTM generation from LiDAR data using skewness balancing. In: Proc. Int. Conf. on Pattern Recog., Vol. 1. pp. 566–569.

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In: Proc. IEEE Int. Conf. Comput. Vis.. pp. 9296–9306.

Beumier, C., Idrissa, M., 2016. Digital terrain models derived from digital surface model uniform regions in urban areas. Int. J. Remote Sens. 37 (15), 3477–3493.

Brovelli, M., Cannata, M., Longoni, U., 2002. Managing and processing LIDAR data within GRASS. In: Proc. of the Open Source GIS - GRASS Users Conference.

Bulatov, D., Stütz, D., Hacker, J., Weinmann, M., 2021. Classification of airborne 3D point clouds regarding separation of vegetation in complex environments. Appl. Opt. 60 (22), F6–F20.

Canuto, M.A., Estrada-Belli, F., Garrison, T.G., Houston, S.D., Acuña, M.J., Kováč, M., Marken, D., Nondédéo, P., Auld-Thomas, L., Castanet, C., et al., 2018. Ancient lowland Maya complexity as revealed by airborne laser scanning of northern Guatemala. Science 361 (6409), eaau0137.

Doneus, M., Mandlburger, G., Doneus, N., 2020. Archaeological ground point filtering of airborne laser scan derived point-clouds in a difficult mediterranean environment. J. Comput. Appl. Archaeol. 3 (1), 92–108.

Duan, L., Desbrun, M., Giraud, A., Trastour, F., Laurore, L., 2019. Large-scale DTM generation from satellite data. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops. pp. 1442–1450.

Elmqvist, M., 2002. Ground surface estimation from airborne laser scanner data using active shape models. Int. Arch. Photogramm. Remote Sens. XXXIV, 114–118.

Elmqvist, M., Jungert, E., Lantz, F., Persson, A., Soderman, U., 2001. Terrain modelling and analysis using laser scanner data. Int. Arch. Photogramm. Remote Sens. XXXIV-3/W4, 219–226.

Evans, J.S., Hudak, A.T., 2007. A multiscale curvature algorithm for classifying discrete return LiDAR in forested environments. IEEE Trans. Geosci. Remote Sens. 45 (4), 1029–1038.

Fan, S., Dong, Q., Zhu, F., Lv, Y., Ye, P., Wang, F.-Y., 2021. SCF-Net: Learning spatial contextual features for large-scale point cloud segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog.. pp. 14504–14513.

Gao, W., Nan, L., Boom, B., Ledoux, H., 2021. SUM: A benchmark dataset of semantic urban meshes. ISPRS J. Photogramm. Remote Sens. 179, 108–120.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog.. pp. 3354–3361.

Gevaert, C., Persello, C., Nex, F., Vosselman, G., 2018. A deep learning approach to DTM extraction from imagery using rule-based training labels. ISPRS J. Photogramm. Remote Sens. 142, 106–123.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3D semantic segmentation with submanifold sparse convolutional networks. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog.. pp. 9224–9232.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. Semantic3D.Net: A new large-scale point cloud classification benchmark. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. IV-1/W1, 91–98.

Hingee, K.L., Caccetta, P., Caccetta, L., 2019. Modelling discontinuous terrain from DSMs using segment labelling, outlier removal and thin-plate splines. ISPRS J. Photogramm. Remote Sens. 155, 159–171.

Hingee, K., Caccetta, P., Caccetta, L., Wu, X., Devereaux, D., 2016. Digital terrain from a two-step segmentation and outlier-based algorithm. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLI-B3, 233–239.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2022. SensatUrban: Learning semantics from urban-scale photogrammetric point clouds. Int. J. Comput. Vis. 130, 316–343.

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: Proc. IEEE Conf. Comput. Vis. Pattern Recog.. pp. 11105–11114.

Hu, X., Ye, L., Pang, S., Shan, J., 2015. Semi-global filtering of airborne LiDAR data for fast extraction of digital terrain models. Remote Sens. 7 (8), 10996–11015.

Hu, X., Yuan, Y., 2016. Deep-learning-based classification for DTM extraction from ALS point cloud. Remote Sens. 8 (9), 730.

Jahromi, A.B., Zoej, M.J.V., Mohammadzadeh, A., Sadeghian, S., 2011. A novel filtering algorithm for bare-earth extraction from airborne laser scanning data using an artificial neural network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4 (4), 836–843.

Jin, S., Su, Y., Zhao, X., Hu, T., Guo, Q., 2020. A point-based fully convolutional neural network for airborne LiDAR ground point filtering in forested environments. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3958–3974.

Klápště, P., Fogl, M., Barták, V., Gdulová, K., Urban, R., Moudrý, V., 2020. Sensitivity analysis of parameters and contrasting performance of ground filtering algorithms with UAV photogrammetry-based and LiDAR point clouds. Int. J. Digit. Earth 13 (12), 1672–1694.

Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J.D., Ledoux, H., 2021. The Hessigheim 3D (H3D) benchmark on semantic segmentation of high-resolution 3D point clouds and textured meshes from UAV LiDAR and multi-view-stereo. ISPRS Open J. Photogramm. Remote Sens. 1, 100001.

Li, X., Li, C., Tong, Z., Lim, A., Yuan, J., Wu, Y., Tang, J., Huang, R., 2020. Campus3D: A photogrammetry point cloud benchmark for hierarchical understanding of outdoor scene. In: Proc. ACM Int. Conf. Multimedia. pp. 238–246.

Lu, W.-L., Murphy, K.P., Little, J.J., Sheffer, A., Fu, H., 2009. A hybrid conditional random field for estimating the underlying ground surface from airborne LiDAR data. IEEE Trans. Geosci. Remote Sens. 47 (8), 2913–2922.

Luo, Y., Ma, H., Zhou, L., 2017. DEM retrieval from airborne LiDAR point clouds in mountain areas via deep neural networks. IEEE Geosci. Remote Sens. Lett. 14 (10), 1770–1774.

Mao, Y., Chen, K., Diao, W., Sun, X., Lu, X., Fu, K., Weinmann, M., 2022. Beyond single receptive field: A receptive field fusion-and-stratification network for airborne laser scanning point cloud classification. ISPRS J. Photogramm. Remote Sens. 188, 45–61.

McCarley, T.R., Hudak, A.T., Sparks, A.M., Vaillant, N.M., Meddens, A.J., Trader, L., Mauro, F., Kreitler, J., Boschetti, L., 2020. Estimating wildfire fuel consumption with multitemporal airborne laser scanning data and demonstrating linkage with MODIS-derived fire radiative energy. Remote Sens. Environ. 251, 112114.

Mongus, D., Lukač, N., Žalik, B., 2014. Ground and building extraction from LiDAR data based on differential morphological profiles and locally fitted surfaces. ISPRS J. Photogramm. Remote Sens. 93, 145–156.

Mousa, Y., Helmholz, P., Belton, D., 2017. New DTM extraction approach from airborne images derived DSM. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII-1/W1, 75–82.

Mousa, Y.A., Helmholz, P., Belton, D., Bulatov, D., 2019. Building detection and regularisation using DSM and imagery information. Photogramm. Rec. 34 (165), 85–107.

Muhadi, N.A., Abdullah, A.F., Bejo, S.K., Mahadi, M.R., Mijic, A., 2020. The use of LiDAR-derived DEM in flood applications: A review. Remote Sens. 12 (14), 2308.

Nie, S., Wang, C., Dong, P., Xi, X., Luo, S., Qin, H., 2017. A revised progressive TIN densification for filtering airborne LiDAR data. Measurement 104, 70–77.

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

Nurunnabi, A., Teferle, F.N., Li, J., Lindenbergh, R.C., Hunegnaw, A., 2021. An efficient deep learning approach for ground point filtering in aerial laser scanning point clouds. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLIII-B1-2021, 31–38.

Ortega, S., Trujillo, A., Santana, J.M., Suárez, J.P., Santana, J., 2019. Characterization and modeling of power line corridor elements from LiDAR point clouds. ISPRS J. Photogramm. Remote Sens. 152, 24–33.

Özcan, A.H., Ünsalan, C., 2016. LiDAR data filtering and DTM generation using empirical mode decomposition. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (1), 360–371.

Perko, R., Raggam, H., Gutjahr, K., Schardt, M., 2015. Advanced DTM generation from very high resolution satellite stereo images. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. II-3/W4, 165–172.

Pfeifer, N., Stadler, P., Briese, C., 2001. Derivation of digital terrain models in the SCOP++ environment. In: Proc. of OEEPE Workshop on Airborne Laserscanning and Interferometric SAR for Detailed Digital Terrain Models, Vol. 3612. http://hdl.handle.net/20.500.12708/43018 (Accessed 14 May 2023).

Pingel, T.J., Clarke, K.C., McBride, W.A., 2013. An improved simple morphological filter for the terrain classification of airborne LIDAR data. ISPRS J. Photogramm. Remote Sens. 77, 21–30.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog.. pp. 77–85.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: Proc. Adv. Neural Inf. Process. Syst., Vol. 30. pp. 5105–5114.

Qin, N., Tan, W., Ma, L., Zhang, D., Li, J., 2021. OpenGF: An ultra-large-scale ground filtering dataset built upon open ALS point clouds around the world. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops. pp. 1082–1091.

Rizaldy, A., Persello, C., Gevaert, C.M., Oude Elberink, S.J., 2018. Fully convolutional networks for ground classification from LiDAR point clouds. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-2, 231–238.

Roggero, M., 2001. Airborne Laser Scanning: Clustering in raw data. Int. Arch. Photogramm. Remote Sens. XXXIV–3/W4, 227–232.

Roynard, X., Deschaud, J.-E., Goulette, F., 2018. Paris-Lille-3D: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. Int. J. Robot. Res. 37 (6), 545–557.

Schmohl, S., Sörgel, U., 2019. Submanifold sparse convolutional networks for semantic segmentation of large-scale ALS point clouds. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-2/W5, 77–84.

Serifoglu Yilmaz, C., Gungor, O., 2018. Comparison of the performances of ground filtering algorithms and DTM generation from a UAV-based point cloud. Geocarto Int. 33, 522–537.

Serifoglu Yilmaz, C., Yilmaz, V., Gungor, O., 2018. Investigating the performances of commercial and non-commercial software for ground filtering of UAV-based point clouds. Int. J. Remote Sens. 39, 5016–5042.

Sithole, G., Vosselman, G., 2001. Filtering of laser altimetry data using a slope adaptive filter. Int. Arch. Photogramm. Remote Sens. XXXIV-3/W4, 203–210.

Sithole, G., Vosselman, G., 2003. Report: ISPRS comparison of filters. https://www.itc.nl/isprs/wgIII-3/filtertest/report/ (Accessed 14 May 2023).

Sithole, G., Vosselman, G., 2004. Experimental comparison of filter algorithms for bare-Earth extraction from airborne laser scanning point clouds. ISPRS J. Photogramm. Remote Sens. 59, 85–101.

Sithole, G., Vosselman, G., 2005. Filtering of airborne laser scanner data based on segmented point clouds. In: ISPRS WG III/3, III/4, V/3 Workshop "Laserscanning 2005". pp. 66–71.

Sohn, G., Dowman, I., 2002. Terrain surface reconstruction by the use of tetrahedron model with the MDL criterion. Int. Arch. Photogramm. Remote Sens. XXXIV, 336–344.

Štroner, M., Urban, R., Lidmila, M., Kolář, V., Křemen, T., 2021. Vegetation filtering of a steep rugged terrain: The performance of standard algorithms and a newly proposed workflow on an example of a railway ledge. Remote Sens. 13, 3050.

Su, W., Sun, Z., Zhong, R., Huang, J., Li, M., Zhu, J., Zhang, K., Wu, H., Zhu, D., 2015. A new hierarchical moving curve-fitting algorithm for filtering lidar data for automatic DTM generation. Int. J. Remote Sens. 36 (14), 3616–3635.

Susaki, J., 2012. Adaptive slope filtering of airborne LiDAR data in urban areas for digital terrain model (DTM) generation. Remote Sens. 4 (6), 1804–1819.

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. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops. pp. 797–806.

Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L.J., 2019. KPConv: Flexible and deformable convolution for point clouds. In: Proc. IEEE Int. Conf. Comput. Vis.. pp. 6410–6419.

Valada, A., Vertens, J., Dhall, A., Burgard, W., 2017. AdapNet: Adaptive semantic segmentation in adverse environmental conditions. In: IEEE Int. Conf. Robot. Autom.. pp. 4644–4651.

Varney, N., Asari, V., Graehling, Q., 2020. DALES: A large-scale aerial LiDAR data set for semantic segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops. pp. 717–726.

Vosselman, G., 2000. Slope based filtering of laser altimetry data. Int. Arch. Photogramm. Remote Sens. XXXIII, 935–942.

Wack, R., Wimmer, A., 2002. Digital terrain models from airborne laser scanner data – A grid based approach. Int. Arch. Photogramm. Remote Sens. XXXIV (3/B), 293–296.

Xue, F., Lu, W., Chen, Z., Webster, C.J., 2020. From LiDAR point cloud towards digital twin city: Clustering city objects based on Gestalt principles. ISPRS J. Photogramm. Remote Sens. 167, 418–431.

Ye, Z., Xu, Y., Huang, R., Tong, X., Li, X., Liu, X., Luan, K., Hoegner, L., Stilla, U., 2020. LASDU: A large-scale aerial LiDAR dataset for semantic labeling in Dense Urban Areas. ISPRS Int. J. Geo-Inf. 9 (7), 450.

Zeybek, M., Şanlıoğlu, İ., 2019. Point cloud filtering on UAV based point cloud. Measurement 133, 99–111.

Zhang, K., Chen, S.-C., Whitman, D., Shyu, M.-L., Yan, J., Zhang, C., 2003. A progressive morphological filter for removing nonground measurements from airborne LIDAR data. IEEE Trans. Geosci. Remote Sens. 41 (4), 872–882.

Zhang, J., Hu, X., Dai, H., Qu, S., 2020a. DEM extraction from ALS point clouds in forest areas via graph convolution network. Remote Sens. 12 (1), 178.

Zhang, J., Lin, X., 2013. Filtering airborne LiDAR data by embedding smoothness-constrained segmentation in progressive TIN densification. ISPRS J. Photogramm. Remote Sens. 81, 44–59.

Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X., Yan, G., 2016. An easy-to-use airborne LiDAR data filtering method based on cloth simulation. Remote Sens. 8 (6), 501.

Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., Foroosh, H., 2020b. PolarNet: An improved grid representation for online LiDAR point clouds semantic segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog.. pp. 9598–9607.

Zhao, X., Guo, Q., Su, Y., Xue, B., 2016. Improved progressive TIN densification filtering algorithm for airborne LiDAR data in forested areas. ISPRS J. Photogramm. Remote Sens. 117, 79–91.

Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D., 2021. Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recog.. pp. 9934–9943.

Zolanvari, S.M.I., Ruano, S., Rana, A., Cummins, A., da Silva, R.E., Rahbar, M., Smolic, A., 2019. DublinCity: Annotated LiDAR point cloud and its applications. In: Proc. British Machine Vis. Conf..