CSF-Net: Color Spectrum Fusion Network for Semantic Labeling of Airborne Laser Scanning Point Cloud

Jihao Li, *Graduate Student Member, IEEE*, Wenkai Zhang, *Member, IEEE*, Wenhui Diao, *Member, IEEE*, Yingchao Feng, *Graduate Student Member, IEEE*, Xian Sun, *Senior Member, IEEE*, and Kun Fu, *Member, IEEE*

***Abstract*—Airborne laser scanning point cloud semantic labeling, which aims to identify the category of each point, plays a significant role in many applications, such as forest observing, powerline extraction, etc. Under the guidance of deep learning technology, the interpretation thought of point clouds has also greatly changed. However, owing to the irregular and unordered natures of point clouds, it is relatively difficult for classification model to distinguish some objects with similar geometry by single-modal data only. Fortunately, additional gain information, e.g., color spectrum which can be complementary to geometric information, is able to effectively promote the classification effect. Therefore, the design of fusion strategy is a critical part in model construction. In this article, aiming to capture more abstract semantic information for color spectrum data, we elaborate a color spectrum fusion (CSF) module. It can be flexibly integrated into a classification pipeline with just negligible parameters. Then, we expand data fusion thoughts for point clouds and color spectrum and investigate three possible fusion strategies. Accordingly, we develop three architectures to construct CSF-Nets. Ultimately, by taking a weighted cross entropy loss, we can train our CSF-Nets in an end-to-end manner. Experiments on two extensively used datasets: Vaihingen 3D and LASDU show that the presented three fusion approaches all can improve the performance, while the earlier fusion strategy performs the best. Besides, compared with other well-performed methods, CSF-Net is still able to achieve satisfactory performance on overall accuracy and mF_1-score indicator. This also validates the effectiveness of our multimodal fusion network.**

***Index Terms*—Airborne laser scanning (ALS) point cloud, color spectrum information, deep learning, multimodal fusion, semantic labeling.**

Manuscript received October 22, 2021; revised November 22, 2021; accepted December 2, 2021. Date of publication December 9, 2021; date of current version December 29, 2021. This work was supported by the National Natural Science Foundation of China under Grant 61725105. *(Corresponding author: Wenkai Zhang.)*

Jihao Li, Yingchao Feng, Xian Sun, and Kun Fu are with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, with the University of Chinese, Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: lijihao17@mails.ucas.ac.cn; fengyingchao17@mails.ucas.ac.cn; sunxian@mail.ie.ac.cn; kunfuiecas@gmail.com).

Wenkai Zhang and Wenhui Diao are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China.

Digital Object Identifier 10.1109/JSTARS.2021.3133602

I. Introduction

In THE past few decades, great progress has been achieved in the aspect of sensor technology, making it easier to access remote sensing data, like hyperspectral image [1], [2] and synthetic aperture radar image [3], [4]. In particular, with the increasing maturity of light detection and ranging (LiDAR) sensor technology, 3-D data have gradually attracted more and more attention, especially for airborne laser scanning (ALS) point cloud data which have been broadly applied in daily life and production, such as powerline extraction [5]–[7], city reconstruction [8]–[10], vegetation detection [11]–[13], and forest observing [14]–[16]. This prosperity also brings a growing requirement of automatic interpretation for point cloud data. It deeply fascinates a considerable number of scholars and experts to actively involve themselves in this study as well.

Point cloud semantic labeling, also referred to as point cloud classification, is intended to distinguish the category of each point, according to the semantic information of the scene. In recent years, deep learning technology significantly promotes the rapid development of 2-D remote sensing image interpretation [17], [18] due to its powerful capability of feature extraction and feature expression. Meanwhile, the thrive of it also gives rise to a substantial shift for point cloud data processing. Many traditional methods, such as support vector machine [19] and Bayesian discriminant classifiers [20], are gradually replaced by deep learning based methods with more superior performance, including PointNet [21], PointNet++ [22], PointConv [23], etc. The design thought has also been shifting from artificial feature extraction to deep network architecture construction. Commonly, ALS point clouds contain rich geometric information and are able to express a more detailed structural characteristic of different objects in complex scenes, compared to 2-D image data. Nevertheless, due to the irregular and unordered properties of ALS point clouds, deep learning networks just for single-modal data still face great challenges in large scene point cloud semantic labeling task. From Fig. 1, we can find that there occurs serious misclassification phenomenon in the joint part of roof and tree for the performance achieved only by point cloud data. Because the height of these two categories are relatively close and both of them appear similar in geometric structure (a tip-like shape), it is relatively difficult for classification model to identify them. Fortunately, several recent studies [24]–[28]

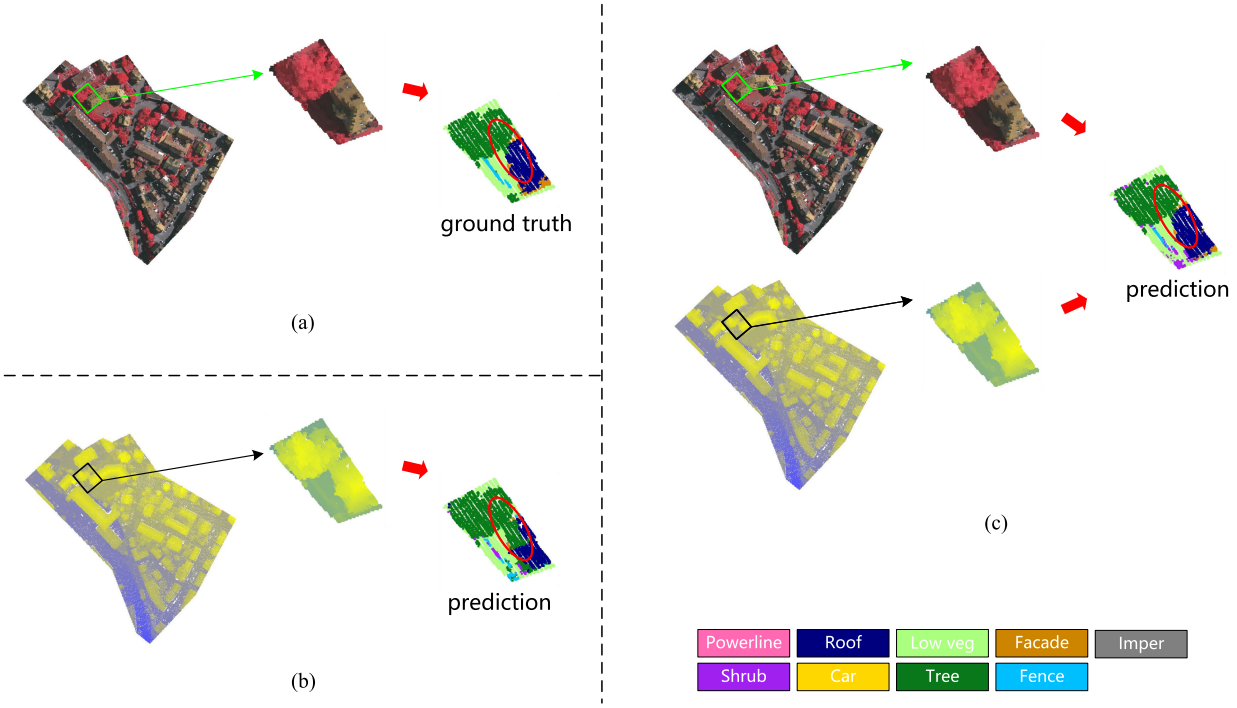

Fig. 1. Visualization results obtained by (b) single-modal data and (c) multimodal data, respectively. We can see that it is easier to distinguish roof and tree by the aid of color spectrum, compared to point cloud only. And the qualitative results obtained by point cloud and color spectrum data (c) are also obviously better than that of point cloud (b). Best viewed in color. (a) Color spectrum information. (b) Point cloud data. (c) Point cloud and color spectrum.

have convincingly implied that fusing information of other modal data is able to effectively boost the classification effect. Although roof and tree are about the same height, the colors of the two categories are quite different (the roof exhibits brown, while the tree presents red.). As shown in Fig. 1(c), in addition to the geometric information, color information is also introduced by fusing color spectrum data, which expands the discrimination of different objects. The visualization results have been significantly promoted. Therefore, the design of fusion strategy for point cloud and color spectrum data is an extremely critical part in neural network construction procedure.

In general, existing approaches for the fusion processing of 3-D point clouds and their color spectrum can be roughly divided into two types. The first one is projection fusion method [29]–[32]. It first projects 3-D point cloud data to 2-D space and then establishes the corresponding relationship between projection point cloud data and color spectrum. After that, it leverages powerful and effective 2-D convolutional neural networks (CNNs) to extract key features of fusion data and output the prediction results. Such a method fully exerts the effect of the existing technology; however, it cannot be ignored that the dimensionality reduction process necessarily leads to the information loss of raw 3-D data. The other is attribute attachment fusion method [23], [33]–[36]. This kind of practice finds the RGB color spectrum value of each point, and it takes the color values as attributes of the point cloud data. Then, a deep learning based semantic labeling model is carefully constructed to process the fused multimodal data. Such a method is relatively straightforward and easy to implement, but simply attaching colors to point clouds in the input layer of the network does not fully exploit the specificity and relationship of heterogeneous data. For color spectrum data, merely taking them as attributes of point clouds lacks abstract feature perception in high-dimensional space. Consequently, there is still much to do on how to effectively utilize color spectrum data and how to fuse the information of position modal data and color modal data.

In this article, to investigate a more feasible and effective fusion approach and further boost the ALS point cloud semantic labeling performance, we propose a novel color spectrum fusion network (CSF-Net). In particular terms, aiming to obtain more abstract feature description and deeper feature perception for color spectrum information, a color spectrum fusion (CSF) branch is elaborated. This module is simple yet effective and is able to extract deep semantic information. In addition, it can be flexibly introduced in a symmetric encoder–decoder network architecture with just negligible extra parameters. Additionally, on the basis of the CSF module, we widen the fusion thoughts for different modal data and explore three possible fusion strategies. Accordingly, we develop three network architectures, i.e., fusion in encoder [CSF-Net (E)], fusion in decoder [CSF-Net (D)], and fusion in both encoder and decoder [CSF-Net (ED)]. The difference between previous attribute attachment fusion strategy and our method is briefly illustrated in Fig. 2. Ultimately, taking the weighted cross entropy (WCE) loss as the optimization objective, we train the designed CSF-Nets in an end-to-end way. Extensive experiments on Vaihingen 3D and LASDU indicate the feasibility and availability of our CSF-Nets. Furthermore, compared with other well-performed methods, our proposed approach shows the superiority as well.

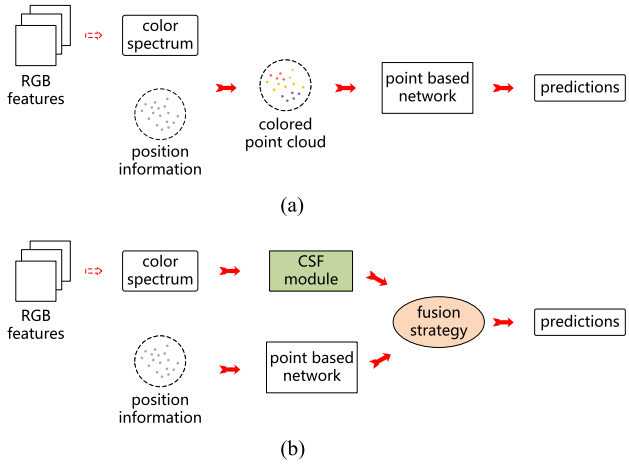

Fig. 2. Comparison between previous attribute attachment fusion method and the presented CSF-Net. We propose a CSF module to mine the information contained in color spectrum. Additionally, we also explore several fusion strategies to integrate different modal data. Best viewed in color.

The main contributions of this article are briefly summarized as follows.

1) This article reveals that it is relatively difficult for the classification model to distinguish some objects with similar geometry by point cloud modality only. However, additional gain information (e.g., color spectrum) can supplement this deficiency. Thus, we propose a CSF-Net framework to better fuse different modal data.
2) To obtain more abstract feature perception for color spectrum data, instead of just using low-level information, we carefully design a CSF module. It can be flexibly integrated into a classification pipeline with just negligible parameters. And the whole model is also able to be optimized in an end-to-end fashion.
3) Based on the CSF module, we expand previous fusion thoughts and explore in detail three possible fusion strategies. Moreover, we further construct three network architectures, i.e., fusion in encoder [CSF-Net (E)], fusion in decoder [CSF-Net (D)], and fusion in both encoder and decoder [CSF-Net (ED)].
4) Experimental results show that all CSF-Nets can obtain competitive performance on two challenging ALS datasets: Vaihingen 3D and LASDU, especially for CSF-Net (E) which reaches the highest mF_1. More importantly, our model also achieves excellent effect compared with various state-of-the-art (SOTA) methods.

The rest of this article is organized as follows. Section II introduces a brief review of some related researches about the fusion for point clouds and color spectrum data. In Section III, we provide a fully detailed description of the proposed CSF-Net. Section IV shows the experiments we conduct to confirm the effectiveness of our method. Moreover, the analysis of the experimental results is also presented. Then, we further compare with other excellent methods in Section V. Ultimately, Section VI gives a summary of the full thesis and puts forward some suggestions for future consideration.

II. Related Work

In order to better understand the proposed approach, in this section, we have a brief review of some recent research achievements with respect to the fusion of point cloud and color spectrum. Section II-A describes the projection-based fusion method which projects LiDAR data into 2-D space to realize the fusion of different modal data. Furthermore, attribute attachment fusion method that attaches color spectrum information to 3-D point clouds as the additional attributes is briefly introduced in Section II-B.

A. Projection Fusion Method

Generally, point clouds are obtained by a LiDAR system. They are in 3-D data space, while RGB color information is commonly stored in the form of 2-D images. Different from other multimodal fusion methods, such as [27], [28], [37], and [38], that integrate hyperspectral image and elevation information without alignment preprocessing, there may not exist obvious corresponding relationship between points and pixels. Hence, aiming to align these two types of data, some academics attempt to utilize projection technique to transfer point cloud data to 2-D domain. Caltagirone *et al.* [29] project point clouds onto camera image plane and upsample the obtained projection images. Then, the authors fuse multimodal information through a fully convolutional network (FCN) [39] to conduct road extraction task. FCN with cross fusion pattern reaches the SOTA in KITTI benchmark [40], [41] at that time. Besides, SqueezeSeg [30] is a great lightweight and high real-time method. It adopts spherical projection technique to achieve LiDAR data conversion. Next, proven 2-D CNN technology can be directly applied to reduced-dimension data and spectrum data, and then the model can extract critical features from these CNN-friendly data. Also, a conditional random field (CRF), following the output of CNN, can further optimize the segmentation results. After SqueezeSeg achieved outstanding results, Wu *et al.* [31] did not stop the steps for exploration and proposed SqueezeSegV2. It inserts a novel context aggregation module (CAM) between two FireModules [42] and leverages focal loss [43] to modulate the attention of the model. After taking several measures, there exists a significant improvement in the aspect of segmentation accuracy for SqueezeSegV2, compared to original SqueezeSeg model. Subsequently, based on SqueezeSeg and SqueezeSegV2, Krispel *et al.* [32] elaborate a simple but effective approach, called FuseSeg. It fully utilizes RGB/LiDAR calibration to establish the corresponding relationship between RGB spectrum data and projection data. After that, FuseSeg fuses the high-dimensional abstract features of the two modal data output by two independent network branches. Experimental results show that this method not only reaches a relatively good segmentation performance but also possesses a fast processing speed.

B. Attribute Attachment Fusion Method

Although projection fusion method is able to appropriately leverage powerful 2-D CNN, it inevitably leads to the information loss of 3-D point cloud data. Meanwhile, with the advent of

PointNet [21], a pioneering work in 3-D point cloud processing on the basis of deep learning technology, it has gradually become a popular practice to apply deep neural network directly to point clouds. Regarding the fusion of point cloud and spectrum data, Yousefhussien *et al.* [33] cleverly extract spectrum features corresponding to 3-D point clouds from 2-D geo-referenced optical remote sensing imagery and fuse them in the point clouds. Then, they extend classical PointNet architecture and employ a multiscale method to cope with complex fusion data. Full and reliable experiments verify the superiority and feasibility of this method. Moreover, Wang *et al.* [34] develop a graph-based model, dubbed dynamic graph CNN (DGCNN). It attaches RGB spectrum data to point clouds as the additional attributes and exploit k-nearest-neighbor (k-NN) graph to model the relationship between unstructured data. Then, a novel edge convolution (EdgeConv) operator is presented to learn point cloud features. Wu *et al.* [23] also fuse colors and point clouds directly and propose PointConv. This approach can flexibly and dramatically learn a convolution kernel through several multilayer perceptrons (MLPs) to handle the fusion data with a relatively high computational efficiency. Extensive experiments conducted by authors demonstrate the effectiveness and superior performance of these two approaches. In addition, different from the aforementioned methods, Su *et al.* [35] present a SParse LATtice Network which projects both multiview images and the point clouds into a lattice space and draws support from the sparse bilateral convolution layer [44], [45] to directly process fused data. Jaritz *et al.* [36] introduce a multiview PointNet. This method first selects several RGB images from critical 2-D perspectives. Next, it aggregates them into point clouds by using a 2-D–3-D feature lifting network. Then, the authors leverage a point cloud based pipeline to process the aggregated features and produce predicted labels. Validated by experiments, these methods compare favorably against some top-performing methods in semantic labeling task as well.

Attribute attachment fusion method is relatively simple and effective; yet, we argue that there is still much room for further investigation in the fusion of different modal data. Motivated by this, we propose a novel model for the fusion of point clouds and color spectrum data and explore the impact of different fusion strategies on semantic labeling performance. A series of experiments verify the availability and superiority of the proposed method.

III. Methodology

In this section, we introduce the proposed CSF-Net in detail. Section III-A gives a specific description of the CSF module. In Section III-B, we expand three possible fusion strategies. Ultimately, the whole CSF-Net pipeline is expounded in Section III-C.

A. Color Spectrum Fusion Module

Aiming to additionally exploit relatively deep semantic information of color spectrum data and further enhance the performance of point cloud semantic labeling task, we introduce a CSF module. The structure of the presented CSF module is illustrated in detail in Fig. 3. This module is able to aggregate color spectrum features from high-dimensional space. Furthermore, it can be flexibly attached to the whole pipeline with negligible extra model parameters and the network architecture can still be optimized in an end-to-end way. The concrete integration measure for the proposed CSF module will be provided in detail in the next subsection.

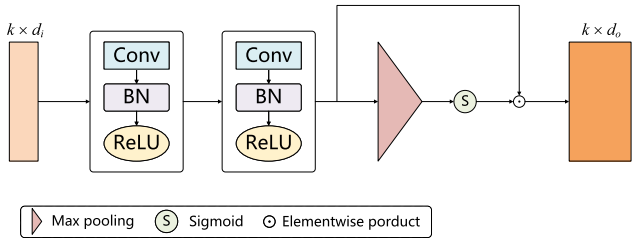

Fig. 3. Structure of the CSF module. It integrates the local information and global expression of color spectrum data. d_i and d_o are the input dimension and output dimension of CSF module, respectively. Best viewed in color.

The workflow of the CSF module is as follows. First, CSF takes the color spectrum information as input. Here, we let $\mathbf{x} = \{x_1, x_2, \ldots, x_k\} \subset \mathbb{R}^{d_i}$ to represent the color spectrum data. Then, two consecutive calculation units map $\mathbf{x}$ into a high-dimensional abstract space. The calculation unit consists of a convolution layer, a batch normalization (BN) layer and a rectified linear unit (ReLU) layer. We express the result at this time as $\mathbf{f} \subset \mathbb{R}^d$, which can be written as

$$\mathbf{f} = \mathcal{T}(\mathbf{x}) \tag{1}$$

where $\mathcal{T}$ stands for the mapping process for the color spectrum data. After that, in order to obtain a global expression of auxiliary color spectrum features, a max pooling operation with a nonlinear activation function *Sigmoid* is applied to the generated $\mathbf{f}$. Meanwhile, $\mathbf{f}$ propagates along the other branch unmodified and it is integrated with the global information through an elementwise multiplication or known as Hadamard product. Note that these values are broadcasted during the forward propagation. In summary, we can formulate the calculation procedure of the proposed CSF module as

$$\mathbf{h} = \mathbf{f} \odot \sigma(\mathcal{M}(\mathbf{f})) \tag{2}$$

where σ and $\mathcal{M}$ separately denote *Sigmoid* function and max pooling operation.

B. Fusion Strategy

Based on the CSF module, we further explore fusion strategies for point cloud and abstract semantic information of color spectrum. Previous deep learning based researches commonly extend additional attributes for point clouds and fuse color spectrum information into the point clouds at the input layer of the neural network, as illustrated in Fig. 4(a). The data of different modal are concatenated together and then fed into a symmetric encoder–decoder architecture. The details of this architecture will be presented in Section III-C. This fusion pattern is relatively uncomplicated and its implementation is not difficult as well. Whereas, inspired by [46] that broadens the

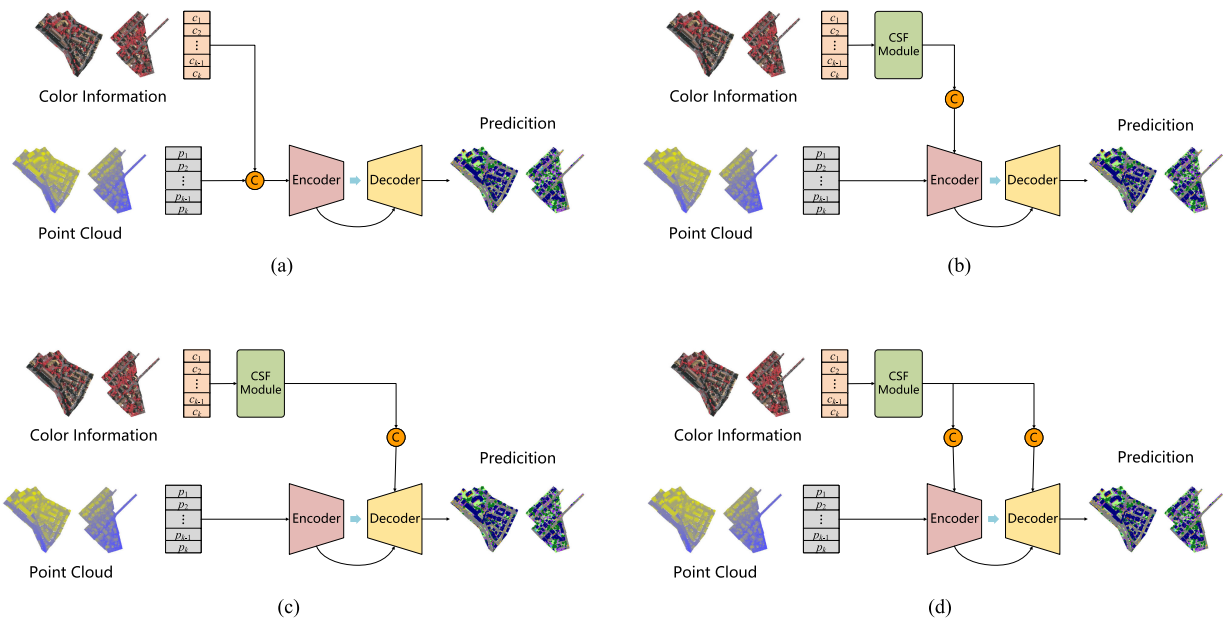

Fig. 4. Four fusion patterns of color spectrum information. CSF module is the abbreviation of color spectrum fusion module. All circular icons in this figure with letter C indicate concatenation. (a) Color information and point cloud are directly fused in the input layer. (b) Color information is first fed into the CSF module and then fused with the point cloud features in the encoder part. (c) Color information is first fed into the CSF module and then fused with the point cloud features in the decoder part. (d) Color information is first fed into the CSF module and then fused with the point cloud features in both the encoder part and decoder part. Best viewed in color. (a) Fusion in the input layer. (b) Fusion in encoder. (c) Fusion in decoder. (d) Fusion in both encoder and decoder.

fusion thinking for digital surface model and high-resolution remote sensing imagery, we also hope to expand more fusion strategies, as shown in Fig. 4(b)–(d), to further investigate the impact of different fusion approaches on point cloud semantic labeling performance. Here, we name the network exhibited in Fig. 4(b)–(d) as CSF-Net (E), CSF-Net (D), and CSF-Net (ED), respectively.

In CSF-Net (E), we fuse color features into the encoder part. Different from the approach of simple combination in the model input, the symmetric encoder–decoder network and the elaborated CSF module separately extract the features of point clouds and color spectrum data. The way that two branches work in parallel decouples the coupling of features between point clouds and color spectrum, making the deep network capture the critical and superior information of different modal data more effectively. Particularly, for color spectrum, CSF module provides more abstract semantic features, instead of low-level color attributes only. Based on the strategy of CSF-Net (E), we continue to explore other possible fusion approaches. We fuse the point cloud features and color spectrum features generated by CSF module in up-sampling procedure, as described in Fig. 4(c). Furthermore, combining the aforementioned two approaches, we also conduct fusion strategy in both the encoder and decoder parts, which is depicted in Fig. 4(d).

In order to obtain the relatively well-behaved network architectures, we compare different fusion methods in Section IV. Experimental results reveal that all CSF-Nets can achieve promising performance, whereas CSF-Net (E) reaches a more superior level. The relevant analysis for these models will be discussed in Section IV-E.

C. Whole CSF-Net Architecture

The whole CSF-Net architecture pipeline is illustrated in Fig. 5. It can be divided into two parts: encoder and decoder. Note that, due to the limitation of graphics processing unit (GPU) memory, the input is just a block of point clouds, and not the whole scene. Specific division principle will be explicated in Section IV-B. We set up four down-samplings in encoder, which can generate more sparse and abstract features. Correspondingly, we also utilize four up-samplings to restore the original point clouds. The fusion of color spectrum modal data is at the fusion point. Fig. 5 only indicates the possible locations where fusion points are placed in the network. Nevertheless, for a certain architecture, like CSF-Net (E) or CSF-Net (D), it may not contain all these fusion points at the same time. Here, for the purpose of keeping alignment of these two modal data, we also apply the same sampling, which is synchronized with the point clouds, to color spectrum data.

In the encoder part, the quantity of points after each down-sampling operation is set to 2048, 512, 128, and 32, respectively. We select the sampled points according to efficient farthest point sampling [22], which iteratively picks the one with the largest Euclidean distance from the remaining point cloud set. Then, the sampled points and their features are grouped through a k-NN algorithm to extract local information. Here, we gradually enlarge the search radius with the deepening of the neural network, namely, 2, 4, 8, and 16, on account of the increasingly sparse data points. Moreover, the number of nearest neighbors is all set to 32 in the whole pipeline. After that, a three-layer MLP maps the point cloud features into a high-dimensional space to obtain the expressions that are more conducive to neural network

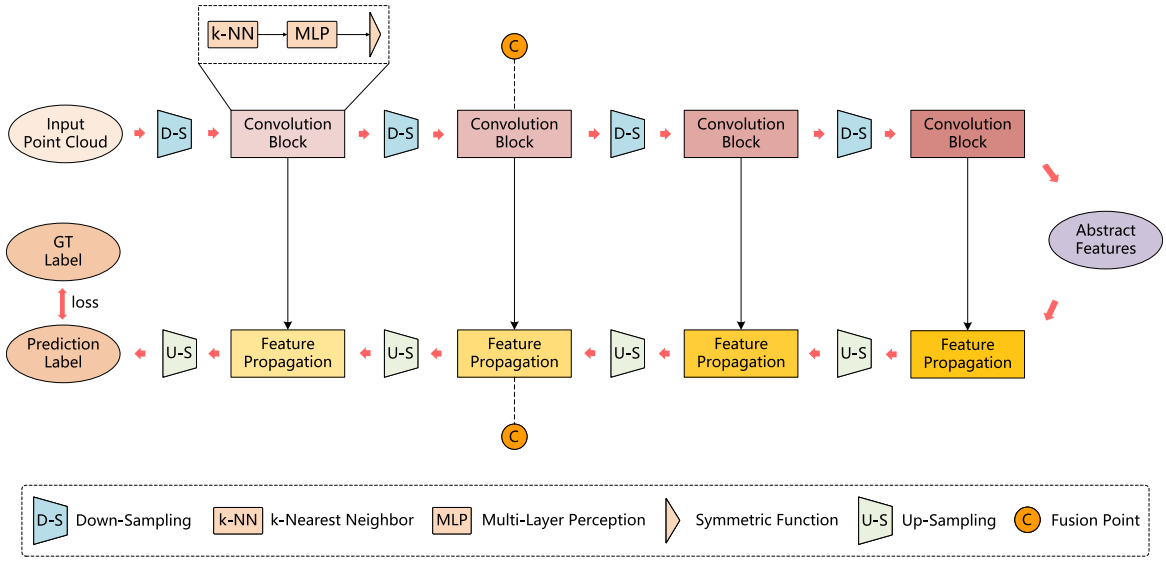

Fig. 5. Overview of the whole CSF-Net architecture. The upper level is the encoder part, while the lower level is the decoder part. Best viewed in color.

perception. With regard to symmetric function, we adopt max pooling which is akin to [21] and [22].

For the decoder part, we successively leverage a weighted interpolation method to hierarchically reconstruct the input point clouds. This algorithm first searches m ($m = 3$, in our network) neighbor(s) of each data point in Euclidean space. Then, it represents the interpolated feature of the up-sampling point as the inverse Euclidean distance weighted mean of the m point(s). In the up-sampling stage, a skip connection that integrates the feature of convolution block and feature propagation at the same height is employed to thicken features and remedy the information lost during the sampling process as well. This technique has been clearly proved that it can accelerate convergence and improve the performance in 2-D and 3-D semantic segmentation tasks [22], [47]–[49]. Considering the serious class imbalance problem, we attach a WCE objective function to the output layer of the neural network to minimize the margin between the prediction values and ground truth labels. WCE loss can be calculated as

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\omega_j * y_j(x_i) * \log p_j(x_i) \tag{3}$$

where N and C denote the total quantity of samples and annotated categories in a point cloud dataset, respectively. ω_j stands for the weight of the jth class. Relatively speaking, the category with a small number of points is assigned a larger weight value. In addition, $y_j(x_i)$ and $p_j(x_i)$ separately indicate the jth element of ground truth (that is, one-hot vector) and predicted vector for the ith sample x_i.

IV. Experiments

In this section, we conduct extensive experiments to confirm the effectiveness of our proposed CSF-Nets for point cloud semantic labeling task. Experimental datasets and the corresponding data preprocessing procedure are respectively introduced in Sections IV-A and IV-B. Then, we present the evaluation metrics utilized in our experiments in Section IV-C. Training parameters and other setups are stated in Section IV-D. Finally, we list the experimental results and provide a careful analysis of them in Section IV-E.

A. Datasets

The experiments are performed on two extensively used point cloud datasets, i.e., Vaihingen 3D [50], [51] and LASDU [52], [53]. A brief introduction about these two datasets are given below.

1) Vaihingen 3D: Vaihingen 3D[1] is a very classical ALS point cloud dataset. It is provided by International Society for Photogrammetry and Remote Sensing (ISPRS) and used to be a benchmark for 3-D Semantic Labeling Challenge. The point cloud data is collected in a small village of Vaihingen, in Germany, by the use of a Leica ALS50 system with a 45° field of view. All points are annotated into nine semantic categories, namely, *Powerline*, *Low vegetation*, *Impervious surfaces*, *Car*, *Fence/Hedge*, *Roof*, *Facade*, *Shrub*, and *Tree*. The annotated Vaihingen 3D can be seen in Fig. 6. This dataset has an extremely unbalanced category distribution (as recorded in Table I), which makes it quite challenging. In addition, point clouds of Vaihingen 3D are also very sparse in space, merely about 8 points/m^2. According to the statistics, there are 753 876 points in the training scene and 411 722 points in the test scene. Each point totally contains six attributes, that is, *X*–*Y*–*Z* values in space, intensity, return number, and number of returns. The acquisition of RGB color spectrum information will be depicted in Section IV-B.

2) LASDU: LASDU[2] is published by Technical University of Munich (TUM) and Tongji University in 2020. All the point cloud data are collected from a Leica ALS70 system. The study

[1][Online]. Available: https://www2.isprs.org/commissions/comm2/wg4/benchmark/3d-semantic-labeling/

[2][Online]. Available: https://github.com/Yusheng-Xu/LASDU-Semantic-Labeling-Benchmark

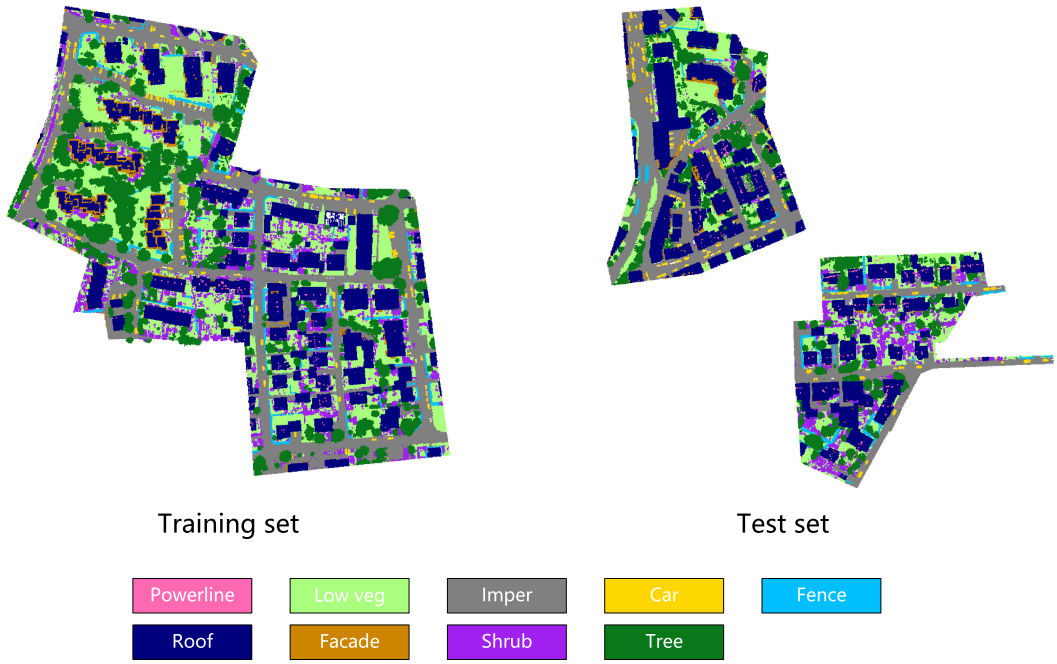

Fig. 6. Vaihingen 3D dataset. Left is the training set that is utilized for model training, while right is the test set that is for model evaluation. The label of each point in Vaihingen 3D has been rendered as the legend shown at the bottom. Best viewed in color.

TABLE I
PROPORTION OF DIFFERENT CATEGORIES IN VAIHINGEN 3D DATASET

Category	Training set	Test set
Powerline	546 / 0.07%	600 / 0.15%
Low vegetation	180,850 / 23.99%	98,690 / 23.97%
Impervious surfaces	193,723 / 25.70%	101,986 / 24.77%
Car	4,614 / 0.61%	3,708 / 0.90%
Fence/Hedge	12,070 / 1.60%	7,422 / 1.80%
Roof	152,045 / 20.17%	109,048 / 26.49%
Facade	27,250 / 3.61%	11,224 / 2.73%
Shrub	47,605 / 6.31%	24,818 / 6.03%
Tree	135,173 / 17.93%	54,226 / 13.17%
total	753,876	411,722

TABLE II
PROPORTION OF DIFFERENT CATEGORIES IN LASDU DATASET

Category	Training set	Test set
Ground	704,425 / 41.56%	637,257 / 45.98%
Building	508,479 / 30.00%	395,109 / 28.51%
Tree	204,775 / 12.08%	108,466 / 7.83%
Low vegetation	210,495 / 12.42%	192,051 / 13.86%
Artifact	66,738 / 3.94%	53,061 / 3.83%
total	1,694,912	1,385,944

region of LASDU is in a valley of Heihe River Basin, in the Northwest of China, covering an area of about 1.02 km^2. This dataset has been divided into four parts by publisher in advance, i.e., Sections I–IV. Each respectively consists of 0.77, 0.59, 1.10, and 0.62 million colored point clouds. These points are divided into five different categories, including *Ground*, *Building*, *Tree*, *Low vegetation*, and *Artifact*. Each region in LASDU with manual annotation is exhibited in Fig. 7. Akin to Vaihingen 3D dataset, there also exists a serious class-imbalanced problem in LASDU (as shown in Table II), which brings great difficulty to segmentation algorithm. Besides, the density of these point clouds is only about 3–4 points/m^2. The relatively sparse space distribution makes the interpretation more challenging. In line with [52], we train the proposed model on Sections II and III while validate its performance on Sections I and IV.

B. Data Preprocessing

Vaihingen 3D and LASDU both cover a relatively large geographical area. Each dataset contains an enormous number of data points. It is easy to cause *CUDA Out Of Memory* problem if we directly send the original point cloud data into the neural network. As a result, we divide the whole dataset into many blocks. The division method for the large point cloud scene is along the *X*–*Y* plane while the data points in *Z* direction in each block are preserved. With regard to the setting for block size, our principle is to try to make the number of points in each block as close as possible. Under this consideration, we set the width and length of the block in Vaihingen 3D to 30 and 30 m, respectively, while those in LASDU are set to 50 m $\times$ 50 m. Aiming to ensure the size consistency of each sample in mini-batch processing, we randomly select 4096 points in each divided block. Here, for the blocks with less than 4096 points, we adopt a sampling with replacement approach. Otherwise, random sampling is conducted without replacement.

Point clouds in LASDU dataset includes RGB color spectrum information; so we can directly utilize them for fusion processing. However, for Vaihingen 3D dataset, point clouds have only position attributes. They do not contain color spectrum attribute. In order to acquire these important information, we leverage bilinear interpolation method, which is similar to [33], to extract the corresponding colors based on the geo-referenced

Fig. 7. Each region in LASDU dataset is shown separately. Sections II and III are training sets. Sections I and IV are used for test purpose. The label legend corresponding to each point can be referenced to the bottom of this figure. Best viewed in color.

true orthophoto of Vaihingen. Note that the band of the color in Vaihingen is near infrared, red (R), and green (G), and not RGB.

C. Evaluation Indicator

In order to quantitatively evaluate the performance of the proposed method, we employ two widely used indicators, i.e., overall accuracy (OA) and mean F_1-score (mF_1).

OA is a very classic evaluation metric. It can be written as

$$\mathrm{OA} = \frac{N_{\text{correct}}}{N_{\text{total}}} \tag{4}$$

where N_{correct} and N_{total}, respectively, stand for the quantity of correctly classified points and the total number of points. Generally, a relatively large OA value suggests a higher performance of the network.

With regard to mF_1, it reflects the comprehensive performance of all semantic categories. Consequently, we first calculate the F_1-score for each category. Through the confusion matrix, we can obtain the value of true positives (TP), false negatives (FN), and false positives (FP), separately. TP means the number of correctly classified positive sample points. For FN, it represents that the number of positive sample points which are recognized as negatives. Regarding FP, it indicates negative sample points which are misclassified as positives. Then, *Precision* and *Recall* can also be calculated by the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

and

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6}$$

After that, the F_1-score of each category can be acquired through

$$F_1 = (1+\beta) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision+Recall}}, \quad \beta = 1. \tag{7}$$

Ultimately, we average all the F_1-scores and have

$$\mathrm{m}F_1 = \frac{1}{C_m} \sum_{j=1}^{C_m} F_{1_j} \tag{8}$$

where C_m is the total of annotated categories. Consistent with OA indicator, a higher mF_1 typically indicates that the model is able to achieve a more superior performance.

D. Training Settings

We leverage PyTorch [54] framework that is an open-source library for deep learning research to implement the proposed approach. During the training procedure of our CSF-Nets, the learning rate is initialized to be 0.001 and it gradually decreases following a cosine annealing strategy [55]. We feed point cloud data to the network in a mini-batch fashion and batch size is set to 8. Totally, we train all models from scratch for 500 epochs on a single GeForce RTX 2080 Ti GPU. Here, the learnable parameter initialization method is subject to kaiming uniform [56] which is suitable for neural networks with ReLU activation function. As regard to the optimization for weights and biases of the model, we select the efficient Adam optimizer [57]. Additionally, we also configure the value of dropout rate to be 0.5, which aims to prevent neural network overfitting problem.

E. Experimental Results

We first conduct experiments to determine the optimal fusion strategy for CSF-Nets. The classification results on Vaihingen 3D dataset and LASDU dataset are reported in Tables III and IV, respectively. Here, the baseline we used in the experiments is the classical PointNet++ architecture. Figs. 8 and 9 give the corresponding confusion matrices of the three CSF-Nets obtained on these two datasets.

It can be seen from Tables III and IV that the introduction of color spectrum information is conducive to the improvement of classification performance. And the proposed CSF-Nets can

TABLE III
FUSION PERFORMANCE ON VAIHINGEN 3D

Network	Power	Low	Imper	Car	Fence	Roof	Facade	Shrub	Tree	OA	mF_1
Baseline	57.90	79.60	90.60	66.10	31.50	91.60	54.30	41.60	77.00	81.20	65.60
Input Fusion	59.04	79.59	89.62	73.47	28.58	92.64	56.93	42.99	81.39	81.73	67.14
CSF-Net (E)	**73.94**	**80.49**	**91.22**	75.03	**46.30**	93.36	60.19	48.18	81.46	83.08	**72.24**
CSF-Net (D)	70.95	80.23	90.82	77.91	41.52	**93.49**	**61.70**	48.30	**81.98**	**83.15**	71.88
CSF-Net (ED)	65.12	80.46	90.77	**79.01**	40.83	92.42	59.39	**48.93**	80.23	82.30	70.79

The first column lists the name of different fusion strategies. The next nine columns present the F_1 indicator (%) for each category in Vaihingen 3D. The last two columns give the value of OA (%) and mF_1 (%).

TABLE IV
FUSION RESULTS ON LASDU

Network	Ground	Building	Tree	Low vegetation	Artifact	OA	mF_1
Baseline	89.45	95.09	84.92	65.04	27.71	85.96	72.44
Input Fusion	89.87	94.89	83.56	66.08	37.36	86.26	74.35
CSF-Net (E)	**91.22**	95.22	**87.02**	**72.53**	44.20	**87.56**	**78.04**
CSF-Net (D)	90.83	95.30	85.36	66.79	**46.39**	87.10	76.93
CSF-Net (ED)	90.79	**95.36**	83.84	65.82	45.58	86.71	76.28

The first column lists the name of different fusion strategies. The next five columns present the F_1 indicator (%) for each category in LASDU. The last two columns give the value of OA (%) and mF_1 (%).

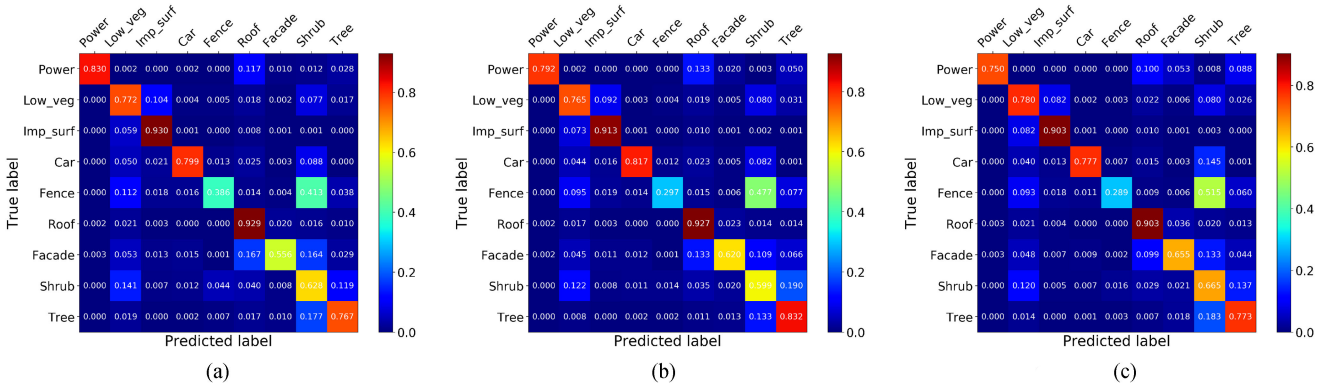

Fig. 8. Confusion matrices generated by CSF-Net (E), CSF-Net (D), and CSF-Net (ED), respectively, on Vaihingen 3D dataset. Best viewed in color. (a) CSF-Net (E). (b) CSF-Net (D). (c) CSF-Net (ED).

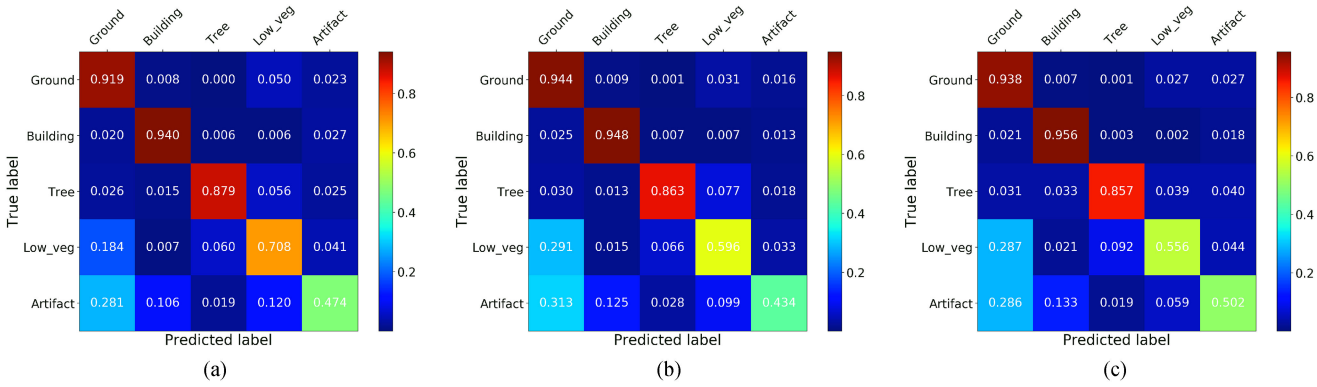

Fig. 9. Confusion matrices generated by CSF-Net (E), CSF-Net (D), and CSF-Net (ED), respectively, on LASDU dataset. Best viewed in color. (a) CSF-Net (E). (b) CSF-Net (D). (c) CSF-Net (ED).

continue to improve the performance with just about an increment of 0.1 M parameters. This also fully suggests that the three extended fusion strategies are all effective, especially for CSF-Net (E) which achieves the optimal performance indicated by mF_1-score. And there are about an increment of 7.6% and 5.0% in terms of mF_1-score, compared with simple attribute attachment fusion method in the input layer. Furthermore, as shown in Figs. 8 and 9, we can find that almost all *energy* of the confusion matrices are concentrated on the principal diagonal. Most categories reach an acceptance performance. Additionally,

TABLE V
COMPARISON RESULTS ON VAIHINGEN 3D

Fusion model	Power	Low	Imper	Car	Fence	Roof	Facade	Shrub	Tree	OA	mF_1
PointNet++* [22]	59.04	79.59	89.62	73.47	28.58	92.64	56.93	42.99	81.39	81.73	67.14
PointSIFT [58]	55.70	80.70	90.90	77.80	30.50	92.50	56.90	44.40	79.60	82.20	67.70
PointCNN* [59]	62.51	79.59	89.48	71.93	30.75	90.34	45.92	45.81	**82.44**	80.95	66.53
DensePoint* [60]	56.37	79.46	90.07	73.92	29.74	92.07	53.44	43.15	80.94	81.66	66.57
RS-CNN* [61]	65.30	79.38	90.56	**79.80**	31.15	93.39	55.51	38.21	78.70	81.71	68.00
DGCNN* [34]	44.60	71.20	81.80	42.00	11.80	93.80	**64.30**	46.40	81.70	78.30	59.70
PointConv* [23]	67.28	79.75	90.65	72.12	37.49	92.98	55.90	40.54	81.30	82.78	68.67
KPConv [58]	63.10	**82.30**	91.40	72.50	25.20	**94.40**	60.30	44.90	81.20	83.70	68.40
PosPool* [62]	57.31	79.11	91.14	74.55	33.02	93.07	56.46	45.72	80.06	82.15	67.83
ConvPoint [63]	58.84	80.88	90.70	65.86	34.35	90.28	52.38	39.11	76.97	81.47	65.49
D-FCN [58]	70.40	80.20	91.40	78.10	37.00	93.00	60.50	46.00	79.40	82.20	70.70
DANCE-Net [64]	68.40	81.60	**92.80**	77.20	38.60	93.90	60.20	47.20	81.40	**83.90**	71.20
Ours	**73.94**	80.49	91.22	75.03	**46.30**	93.36	60.19	**48.18**	81.46	83.08	**72.24**

*means that the color spectrum information is used in the input layer. The first column lists the name of different models. The next nine columns present the F_1 indicator (%) for each category in Vaihingen 3D. The last two columns show OA (%) and mF_1 (%).

TABLE VI
COMPARISON RESULTS ON LASDU

Fusion model	Ground	Building	Tree	Low vegetation	Artifact	OA	mF_1
PointNet++* [22]	89.87	94.89	83.56	66.08	37.36	86.26	74.35
PointCNN* [59]	88.94	91.16	83.12	62.24	37.33	83.58	72.56
DensePoint* [60]	89.77	94.43	84.50	67.72	36.61	86.10	74.61
RS-CNN* [61]	90.36	94.88	84.15	69.50	41.96	86.64	76.17
DGCNN* [34]	90.95	94.24	84.76	68.41	41.54	86.21	75.98
PointConv* [23]	89.73	94.58	84.11	70.06	39.55	86.35	75.60
KPConv [65]	89.12	93.43	83.22	59.70	31.85	83.71	71.47
GACNet* [66]	90.51	94.46	**87.27**	69.10	43.85	87.07	77.04
PosPool* [62]	86.76	93.89	83.47	57.00	36.95	82.09	71.61
HAD-PointNet++ [67]	88.74	93.16	82.24	65.24	36.89	84.37	73.25
ConvPoint [63]	**91.37**	94.15	83.36	71.63	42.36	86.49	76.57
Ours	91.22	**95.22**	87.02	**72.53**	**44.20**	**87.56**	**78.04**

* means that the color spectrum information is used in the input layer. The first column lists the name of different models. The next five columns present the F_1 indicator (%) for each category in LASDU. The last two columns give the value of OA (%) and mF_1 (%).

the recognition accuracy of *Impervious surfaces* and *Roof* in Vaihingen 3D dataset and *Ground* and *Building* in LASDU dataset is even more than 90%. This reveals the superiority of our CSF-Nets as well.

Apart from that, we observe that CSF-Net (D) and CSF-Net (ED) are slightly inferior than CSF-Net (E), about a decline of one to two points on the value of mF_1 on both Vaihingen 3D dataset and LASDU dataset, in spite of the performance improvement. We think that earlier fusion can make features propagate to deeper layers of the network. It is relatively beneficial to extract more generic and advanced information for classification model. However, excessive fusion probably leads to the redundancy and thereby interferes with the perception of the model. For instance, the OA metric of CSF-Net (E) and CSF-Net (D) are on a par with each other, whereas there exists an obvious degradation for CSF-Net (ED).

V. DISCUSSION

In this section, we further compare our optimal CSF-Net with other SOTA models on Vaihingen 3D dataset and LASDU dataset. The quantitative and qualitative results achieved by the proposed approach are both shown in Sections V-A and V-B, respectively.

A. Performance Comparison on Vaihingen 3D

Table V lists the comparative experiment results on Vaihingen 3D dataset. From Table V, we can find that our CSF-Net surpasses all comparison models on mF_1, which is one percentage point higher than the best-performed method, DANCE-Net [64]. Moreover, for OA indicator, CSF-Net is also competitive, which is second only to DANCE-Net [64] and KPConv [58]. It is noteworthy that CSF-Net is able to reach convergence with a faster speed, compared to the complex models in Table V. For example, CSF-Net just takes 500 epochs to converge while the number of epochs is 1000 for DANCE-Net [64]. In addition, the proposed CSF-Net obtains outstanding performance on the categories that are very difficult to distinguish, such as *Powerline*, *Fence*, and *Shrub*.

Fig. 10 shows the visualization performance and the error map of our CSF-Net. It achieves a very satisfactory effect. We can observe in Fig. 10 that the proposed model is able to correctly classify most of the points. And in Fig. 11, our CSF-Net and PointConv [23] are qualitatively evaluated. We use red ellipses to mark the parts that our model correctly predicts while PointConv [23] classifies incorrectly. PointConv mispredicts some *Fence* points for *Tree* and *Shrub*. Due to the similar geometric distribution and structure, it is relatively easy to confuse these categories for classification model. However,

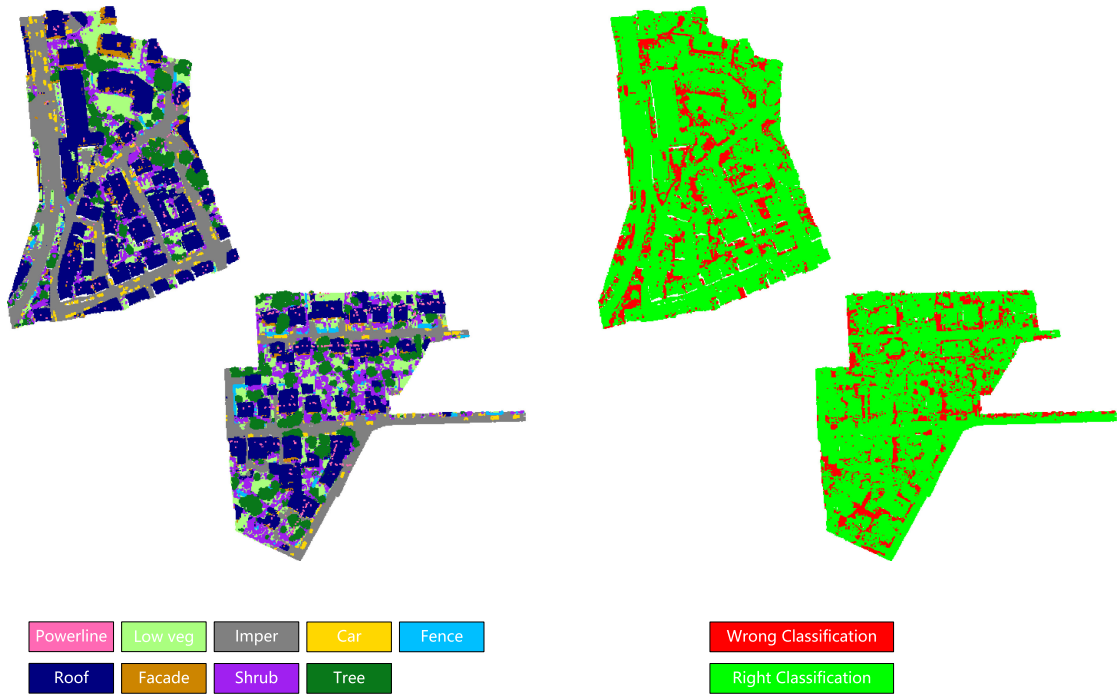

Fig. 10. Left image presents the visualization effect of our CSF-Net on Vaihingen 3D test set. The right image shows the error map. Best viewed in color.

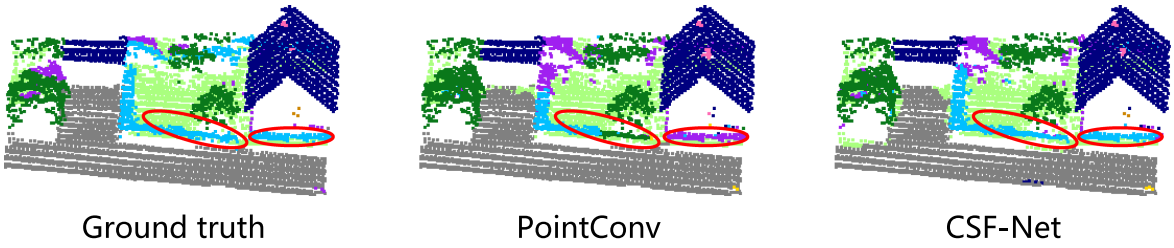

Fig. 11. Qualitative results on Vaihingen 3D dataset. Best viewed in color.

the proposed CSF-Net which fully utilizes the color information can correctly classify almost all of the points with these challenging categories in this scene.

B. Performance Comparison on LASDU

We further perform comparison experiments on LASDU dataset. Table VI provides the quantitative results on test scenes, i.e., Sections I and IV of this dataset. The CSF-Net reaches the highest F_1 on three of the total five categories, that is, *Building*, *Low vegetation*, and *Artifact*. More specially, the quantity of points in *Artifact* category is relatively small. It is not easy to learn and perceive this category for classification network. Yet, CSF-Net is still able to get an acceptable effect on *Artifact*, which fully shows that our method has advantages in the face of class-imbalanced problem. In terms of the overall performance, CSF-Net achieves OA and mF_1-score of 87.56% and 78.04%, respectively, which is higher than all comparison models. This suggests that our method can adapt different point cloud scenes and it possesses a relatively better generality.

As for the qualitative performance, we can see from Fig. 12 that the visual effects of the test scenes in LASDU are relatively better and the boundaries of different objects are relatively clear as well. Only a few points are scattered in the red area, namely, the misclassified parts. Furthermore, compared with the best performed GACNet [66] in Fig. 13, CSF-Net also has good performance in the area of building boundary and low vegetation. GACNet [66] incorrectly predicts some points belonging to *Building* and *Low vegetation* as *Artifact*, whereas the results obtained by CSF-Net are more close to the ground truth.

VI. Conclusion

In this article, we explore more fusion strategies for ALS point clouds and color spectrum information on semantic labeling task and present a CSF-Net. More concretely, we carefully develop a CSF module as an auxiliary branch to extract deep semantic information of color spectrum. And, we attach the proposed CSF module to a classical encoder–decoder architecture. Then, we expand three possible fusion strategies and accordingly construct three network architectures, i.e., CSF-Net (E), CSF-Net

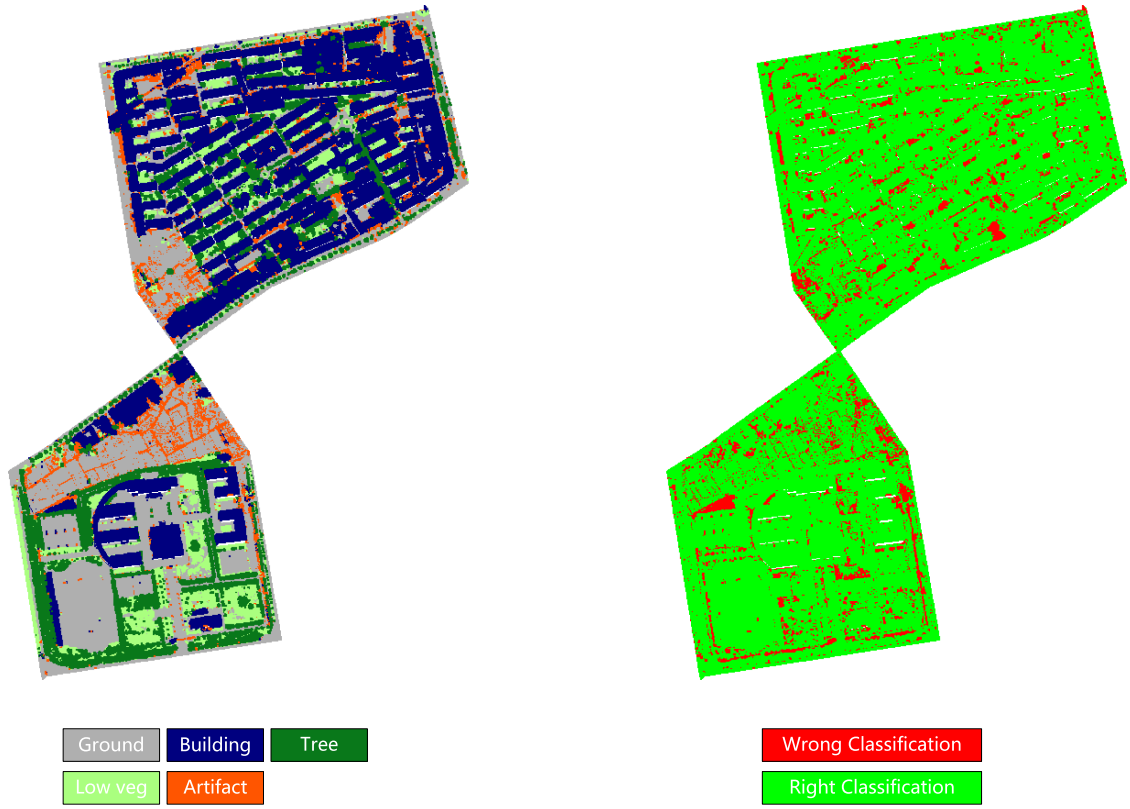

Fig. 12. Left image presents the visualization results of our CSF-Net on Sections I and IV of LASDU dataset. The right image shows the error map. Best viewed in color.

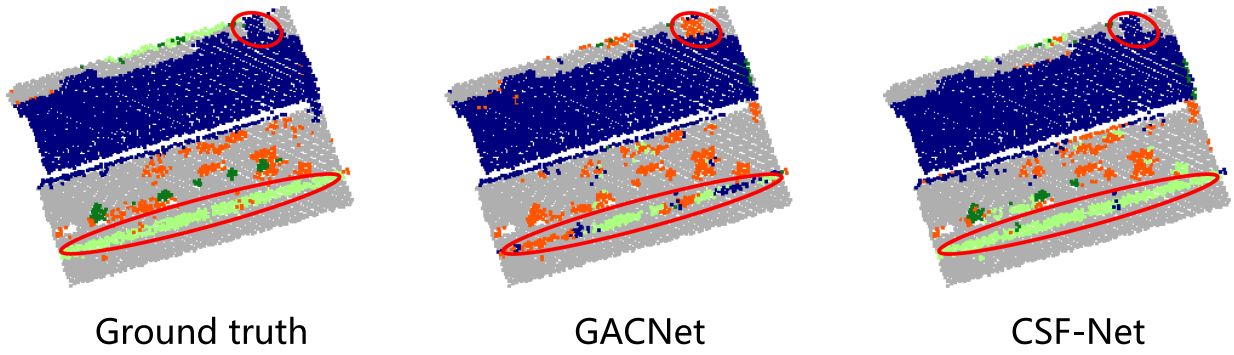

Fig. 13. Qualitative results on LASDU dataset. Best viewed in color.

(D), and CSF-Net (ED). The introduction of CSF module only brings negligible extra network parameters (about 0.1 M) and all the three CSF-Nets are still able to be trained in an end-to-end manner. Finally, we utilize a WCE loss to optimize the learnable parameters until convergence.

Experiments on two challenging large-scene datasets, Vaihingen 3D and LASDU, demonstrate that the presented three fusion strategies all can boost the classification performance, especially for CSF-Net (E) which achieves the highest mF_1-score, i.e., 72.24% and 78.04%, respectively. We think that earlier fusion for different modal data is relatively conducive to extract more superior and generic features. Furthermore, compared with other excellent methods, our CSF-Net is also very competitive. It is able to perform well on both quantitative and qualitative experiments and surpasses most of the models, which confirms the superiority of our CSF-Net. With regard to future work, we will continue to focus on the fusion of different modal data and attempt to construct the fusion network by leveraging some automatic architecture search approaches to further promote the classification performance.

ACKNOWLEDGMENT

The authors would like to thank ISPRS Working Group (WG) II/4, and the College of Surveying and Geo-informatics, Tongji University and The Department of Photogrammetry and Remote

Sensing, Technical University of Munich (TUM) to provide Vaihingen 3D dataset and LASDU dataset, respectively. The authors would also like to thank all colleagues in the laboratory for the fruitful discussions about multimodal fusion. These communications are of great benefit to this research. The authors would also like to sincerely show their appreciation for all anonymous reviewers for their quite helpful comments and suggestions.

REFERENCES

[1] M. Zhang, W. Li, and Q. Du, "Diverse region-based CNN for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.

[2] R. Tao, X. Zhao, W. Li, H.-C. Li, and Q. Du, "Hyperspectral anomaly detection by fractional Fourier entropy," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 4920–4929, Dec. 2019.

[3] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.

[4] H. Jing, X. Sun, Z. Wang, K. Chen, W. Diao, and K. Fu, "Fine building segmentation in high-resolution SAR images via selective pyramid dilated network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6608–6623, 2021, doi: 10.1109/JSTARS.2021.3076085.

[5] H. Kim and G. Sohn, "Random forests based multiple classifier system for power-line scene classification," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 38, pp. 253–258, 2011.

[6] C. Ippolito, K. Krishnakumar, and S. Hening, "Preliminary results of powerline reconstruction from airborne LiDAR for safe autonomous low-altitude urban operations of small UAS," in *Proc. IEEE Sensors*, 2016, pp. 1–3.

[7] S. Pu *et al.*, "Real-time powerline corridor inspection by edge computing of UAV LiDAR data," *Int. Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. 4213, pp. 547–551, 2019.

[8] S. Malihi, M. V. Zoej, M. Hahn, M. Mokhtarzade, and H. Arefi, "3D building reconstruction using dense photogrammetric point cloud," in *Proc. Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 3, 2016, pp. 71–74.

[9] L. Zhang, Z. Li, A. Li, and F. Liu, "Large-scale urban point cloud labeling and reconstruction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 138, pp. 86–100, 2018.

[10] C.-J. Liu, V. A. Krylov, P. Kane, G. Kavanagh, and R. Dahyot, "Im2elevation: Building height estimation from single-view aerial imagery," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2719.

[11] B. Yunfei *et al.*, "Classification of LiDAR point cloud and generation of DTM from LiDAR height and intensity data in forested area," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 37, no. 7, pp. 313–318, 2008.

[12] M. Rutzinger, B. Höfle, M. Hollaus, and N. Pfeifer, "Object-based point cloud analysis of full-waveform airborne laser scanning data for urban vegetation classification," *Sensors*, vol. 8, no. 8, pp. 4505–4528, 2008.

[13] B. Höfle, M. Hollaus, and J. Hagenauer, "Urban vegetation detection using radiometrically calibrated small-footprint full-waveform airborne lidar data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 67, pp. 134–147, 2012.

[14] S. Solberg *et al.*, "Mapping Lai in a Norway spruce forest using airborne laser scanning," *Remote Sens. Environ.*, vol. 113, no. 11, pp. 2317–2327, 2009.

[15] L. T. Ene, E. Næsset, T. Gobakken, O. M. Bollandsås, E. W. Mauya, and E. Zahabu, "Large-scale estimation of change in aboveground biomass in miombo woodlands using airborne laser scanning and national forest inventory data," *Remote Sens. Environ.*, vol. 188, pp. 106–117, 2017.

[16] J. Liu *et al.*, "Extraction of sample plot parameters from 3D point cloud reconstruction based on combined RTK and CCD continuous photography," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1299.

[17] Z. Huang, W. Li, X.-G. Xia, H. Wang, F. Jie, and R. Tao, "LO-Det: Lightweight oriented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vo. 60, pp. 1–15, doi: 10.1109/TGRS.2021.3067470.

[18] Z. Huang, W. Li, X.-G. Xia, X. Wu, Z. Cai, and R. Tao, "A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, 2021, Art. no. 5601920.

[19] J. Secord and A. Zakhor, "Tree detection in urban regions using aerial LiDAR and image data," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 2, pp. 196–200, Apr. 2007.

[20] K. Khoshelham and S. Oude Elberink, "Role of dimensionality reduction in segment-based classification of damaged building roofs in airborne laser scanning data," in *Proc. Int. Conf. Geographic Object Based Image Anal.*, 2012, pp. 7–9.

[21] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[22] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108, doi: 10.1109/CVPR.2019.00985.

[23] W. Wu, Z. Qi, and L. Fuxin, "PointConv: Deep convolutional networks on 3D point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9613–9622.

[24] M. Khodadadzadeh, J. Li, S. Prasad, and A. Plaza, "Fusion of hyperspectral and LiDAR remote sensing data using multiple feature learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2971–2983, Jun. 2015.

[25] J. Zhang and X. Lin, "Advances in fusion of optical imagery and LiDAR point cloud applied to photogrammetry and remote sensing," *Int. J. Image Data Fusion*, vol. 8, no. 1, pp. 1–31, 2017.

[26] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham, "Towards semantic segmentation of urban-scale 3D point clouds: A dataset, benchmarks and challenges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4977–4987.

[27] Y. Zhu, W. Li, M. Zhang, Y. Pang, R. Tao, and Q. Du, "Joint feature extraction for multi-source data using similar double-concentrated network," *Neurocomputing*, vol. 450, pp. 70–79, 2021.

[28] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information fusion for classification of hyperspectral and LiDAR data using IP-CNN," *IEEE Trans. Geosci. Remote Sens.*, to be published, doi: 10.1109/TGRS.2021.3093334.

[29] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "LiDAR-camera fusion for road detection using fully convolutional neural networks," *Robot. Auton. Syst.*, vol. 111, pp. 125–131, 2019.

[30] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 1887–1893.

[31] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer, "SqueezeSegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2019, pp. 4376–4382.

[32] G. Krispel, M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "Fuseseg: LiDAR point cloud segmentation fusing multi-modal data," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2020, pp. 1874–1883.

[33] M. Yousefhussien, D. J. Kelbe, E. J. Ientilucci, and C. Salvaggio, "A multi-scale fully convolutional network for semantic labeling of 3D point clouds," *ISPRS J. Photogrammetry Remote Sens.*, vol. 143, pp. 191–204, 2018.

[34] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, 2019.

[35] H. Su *et al.*, "SplatNet: Sparse lattice networks for point cloud processing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2530–2539.

[36] M. Jaritz, J. Gu, and H. Su, "Multi-view pointnet for 3D scene understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 3995–4003.

[37] F. Jahan, J. Zhou, M. Awrangjeb, and Y. Gao, "Inverse coefficient of variation feature and multilevel fusion technique for hyperspectral and LiDAR data classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 367–381, 2020, doi: 10.1109/JSTARS.2019.2962659.

[38] R. Hang, Z. Li, P. Ghamisi, D. Hong, G. Xia, and Q. Liu, "Classification of hyperspectral and LiDAR data using coupled CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4939–4950, Jul. 2020.

[39] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[40] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[41] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[42] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," 2016, *arXiv:1602.07360*.
[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[44] V. Jampani, M. Kiefel, and P. V. Gehler, "Learning sparse high dimensional filters: Image filtering, dense CRFs and bilateral neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4452–4461.
[45] M. Kiefel, V. Jampani, and P. V. Gehler, "Permutohedral lattice CNNs," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[46] Z. Cao *et al.*, "End-to-end DSM fusion networks for semantic segmentation in high-resolution aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1766–1770, Nov. 2019.
[47] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Berlin, Germany: Springer, 2015, pp. 234–241.
[48] P. Kumar, P. Nagar, C. Arora, and A. Gupta, "U-SegNet: Fully convolutional neural network based automated brain tissue segmentation tool," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 3503–3507.
[49] W. Li, F.-D. Wang, and G.-S. Xia, "A geometry-attentional network for ALS point cloud classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 164, pp. 26–40, 2020.
[50] F. Rottensteiner *et al.*, "The ISPRS benchmark on urban object classification and 3D building reconstruction" *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 1, no. 1, pp. 293–298, 2012.
[51] J. Niemeyer, F. Rottensteiner, and U. Soergel, "Contextual classification of LiDAR data and building object detection in urban areas," *ISPRS J. Photogrammetry Remote Sens.*, vol. 87, pp. 152–165, 2014.
[52] Z. Ye *et al.*, "LASDU: A large-scale aerial LiDAR dataset for semantic labeling in dense urban areas," *ISPRS Int. J. Geo- Inf.*, vol. 9, no. 7, p. 450, 2020.
[53] X. Li *et al.*, "Heihe watershed allied telemetry experimental research (HiWATER): Scientific objectives and experimental design," *Bull. Amer. Meteorological Soc.*, vol. 94, no. 8, pp. 1145–1160, 2013.
[54] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.
[55] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[57] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
[58] C. Wen, L. Yang, X. Li, L. Peng, and T. Chi, "Directionally constrained fully convolutional neural network for airborne LiDAR point cloud classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 50–62, 2020.
[59] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on X-transformed points," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 820–830, 2018.
[60] Y. Liu *et al.*, "Densepoint: Learning densely contextual representation for efficient point cloud processing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5239–5248.
[61] Y. Liu, B. Fan, S. Xiang, and C. Pan, "Relation-shape convolutional neural network for point cloud analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8895–8904.
[62] Z. Liu, H. Hu, Y. Cao, Z. Zhang, and X. Tong, "A closer look at local aggregation operators in point cloud analysis," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany: Springer, 2020, pp. 326–342.
[63] A. Boulch, "ConvPoint: Continuous convolutions for point cloud processing," *Comput. Graph.*, vol. 88, pp. 24–34, 2020.
[64] X. Li, L. Wang, M. Wang, C. Wen, and Y. Fang, "Dance-Net: Density-aware convolution networks with context encoding for airborne LiDAR point cloud classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 128–139, 2020.
[65] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "KPCONV: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6411–6420.
[66] L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, "Graph attention convolution for point cloud semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10296–10305.
[67] R. Huang, Y. Xu, D. Hong, W. Yao, P. Ghamisi, and U. Stilla, "Deep point embedding for urban classification using ALS point clouds: A new perspective from local to global," *ISPRS J. Photogrammetry Remote Sens.*, vol. 163, pp. 62–81, 2020.

Jihao Li (Graduate Student Member, IEEE) received the B.Sc. degree in electronic information engineering from Xidian University, Xi'an China, in 2017. He is currently working toward the Ph.D. degree in signal and information processing with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and deep learning, especially on point cloud processing, remote sensing processing, and neural architecture search.

Wenkai Zhang (Member, IEEE) received the B.Sc. degree from China University of Petroleum, Qingdao, China, in 2013, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2018, all in electronic information engineering.

He is an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Wenhui Diao (Member, IEEE) received the B.Sc. degree from Xidian University, Xi'an, China, in 2011, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2016, all in electronic information engineering.

He is currently an Assistant Professor with the Institute of Electronics, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Yingchao Feng (Graduate Student Member, IEEE) received the B.Sc. degree in electronic information engineering from Xidian University, Xi'an, China, in 2017. He is currently working toward the Ph.D. degree in communication and information system with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include computer vision and deep learning, especially on object detection, semantic segmentation, and remote sensing.

Xian Sun (Senior Member, IEEE) received the B.Sc. degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009, all in electronic information engineering.

He is currently a Professor with Institute of Electronics, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Kun Fu (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively, all in electronic information engineering.

He is currently a Professor with Institute of Electronics, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, geospatial data mining, and visualization.