PointBoost: LiDAR-Enhanced Semantic Segmentation of Remote Sensing Imagery

Yongjun Zhang, *Member, IEEE*, Yameng Wang, Yi Wan, Wenming Zhou, and Bin Zhang

***Abstract*—Semantic segmentation of imagery is typically reliant on texture information from raster images, which limits its accuracy due to the inherently 2-D nature of the plane. To address the nonnegligible domain gap between different metric spaces, multimodal methods have been introduced that incorporate Light Detection and Ranging (LiDAR) derived feature maps. This converts multimodal joint semantic segmentation between 3-D point clouds and 2-D optical imagery into a feature extraction process for the 2.5-D product, which is achieved by concatenating LiDAR-derived feather maps, such as digital surface models, with the optical images. However, the information sources for these methods are still limited to 2-D, and certain properties of point clouds are lost as a result. In this study, we propose PointBoost, an effective sequential segmentation framework that can work directly with cross-modal data of LiDAR point clouds and imagery, which is able to extract richer semantic features from cross-dimensional and cross-modal information. Ablation experiments demonstrate that PointBoost can take full advantage of the 3-D topological structure between points and attribute information of point clouds, which is often discarded by other methods. Experiments on three multimodal datasets, namely N3C-California, ISPRS Vaihingen, and GRSS DFC 2018, show that our method achieves superior performance with good generalization.**

***Index Terms*—Light Detection and Ranging (LiDAR), remote sensing imagery, semantic segmentation.**

I. Introduction

REMOTE sensing data exhibit diversity in sources and representations [1], [2], [3], [4]. Different types of remote sensing data are suitable for individual application, with optical imagery and Light Detection and Ranging (LiDAR), which are prominent examples [5], [6], [7]. Optical imagery provides grayscale information in multiple spectral bands, such as visible light and certain infrared bands. Its well-developed acquisition technology allows for high resolution and distinct visual characteristics, making it a favored choice for remote sensing interpretation [8]. However, optical imagery also has some inherent limitations, such as strict imaging requirements and sensitivity to adverse weather conditions, such as rain, clouds, and low light, which can significantly affect image quality [4], [9], [10]. Additionally, the fact that different objects can have the same spectral characteristics, or the same spectrum can correspond to different objects, presents a significant challenge for remote sensing interpretation [11].

In contrast, the LiDAR point cloud represents an active remote sensing technology [12], [13], [14], [15]. Unlike imagery confined to a 2-D space, LiDAR data capture and provide 3-D structural information of the targets, unaffected by lighting conditions. LiDAR offers distinct advantages in vegetation identification. Due to the fact that a cluster of light beams emitted by LiDAR generates multiple different returns when interacting with stacked leaves, it allows for the delineation of intricate tree canopy structures. Nevertheless, LiDAR data lack texture information and experience intensity tradeoff, which limits its application in remote sensing interpretation to some extent.

Complementing the strengths and weaknesses of each other, the combination of optical imagery and LiDAR data has emerged as a popular solution to overcome the performance limitations of single-modal data in the field of remote sensing [16], [17], [18], [19], [20]. However, the integration of 2-D optical images and 3-D point clouds poses a challenge due to the notable domain gap that exists between their distinct metric spaces. The two modalities differ significantly in terms of spatial dimensions, data structures, and types of characteristics. Specifically, point clouds provide detailed 3-D structural information about objects, while images offer 2-D plane features. Point clouds exhibit characteristics, such as disorder, sparsity, and uneven distribution, whereas images are typically presented in the form of regular grids. Furthermore, point clouds primarily reflect the material characteristics of objects, while images portray the texture and other visual attributes. These inherent disparities make it difficult to process these two cross-modal heterogeneous data using a unified approach.

The integration of 2-D optical images and 3-D point clouds is challenged by the nonnegligible domain gap between their different metric spaces.

Previous studies primarily utilized elevation and its derivations as auxiliary data for optical images, discarding other LiDAR fields, which turns the multimodal joint semantic segmentation between 3-D point clouds and 2-D optical imagery into feature extraction of the 2.5-D product. Sherrah [21] exploited a fully convolutional network (FCN) [22] to handle digital surface model (DSM) derived from LiDAR and a VGG16 [23] to handle

Manuscript received 30 March 2023; revised 23 May 2023; accepted 12 June 2023. Date of publication 16 June 2023; date of current version 28 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 42030102, Grant 42192583, Grant 42001406, and Grant 62102268; in part by the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under Grant 2020CFA003; and in part by the Major Special Projects of Guizhou [2022]001. *(Corresponding author: Yi Wan.)*

Yongjun Zhang, Yameng Wang, Yi Wan, and Bin Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: zhangyj@whu.edu.cn; ymw@whu.edu.cn; yi.wan@whu.edu.cn; bin.zhang@whu.edu.cn).

Wenming Zhou is with the China Railway Design Corporation, Tianjin 300308, China (e-mail: zhouwenming@crdc.com).

Digital Object Identifier 10.1109/JSTARS.2023.3286912

satellite images. The resulting outputs are concatenated and put into the fully connected layers for dense semantic labeling. FuseNet [24] integrates DSM from LiDAR with optical data using a multiscale FCN. Liu et al. [25] preprocessed the LiDAR data to generate hand-crafted features, such as DSM, height variation, and surface norm. An FCN and a logistic regression are then used to obtain two initial probabilistic results from optical imagery and the hand-crafted features, respectively. Finally, these two probabilistic predictions are combined within a higher order conditional random field framework. Sun et al. [26] concatenated the derivations from LiDAR (Difference of Gaussians (DoG) and DSM) and images before putting them into a multifilter CNN classifier.

In conclusion, to address the dimensional discrepancy between 2-D optical imagery and 3-D LiDAR data, these methods project the 3-D LiDAR data onto several 2-D feature maps and concatenate them with the optical imagery along the channel dimension. However, this process inevitably ruins the topological relationship between points in 3-D space.

In this work, we propose a *PointBoost* as an innovative solution to address the domain gap between cross-dimensional multimodal data. Unlike previous approaches that compromise one dimension of the point cloud, PointBoost fully utilizes both LiDAR and image data. PointBoost is composed of three key stages: LiDAR feature extraction, 3D-to-2D feature projection and concatenation, and joint-feature extraction.

The main contributions are as follows.

1) An effective and flexible *2D–3D-joint segmentation framework* called PointBoost is proposed, which can fully preserve the attributes of multimodal data and maintain the topological relationship between points in 3-D space.
2) PointBoost implements *feature-level cross-modal mapping* for LiDAR point clouds and imagery, allowing for comprehensive and accurate feature extraction.

The rest of this article is organized as follows. In Section II, we briefly describe the multimodal feature learning methods of point clouds and imagery. Then, we introduce the details of PointBoost in Section III. The experiments are presented and analyzed in Section IV. Finally, Section V concludes this article.

II. Related Work

In recent years, a significant number of research articles has been dedicated to the fusion of optical imagery and point clouds due to their complementary characteristics [27], [28], [29], [30], [31], [32], [33]. The emergence of deep learning has greatly improved the performance and efficiency of multimodal feature learning. As a result, it has become a standard practice to directly project 3-D point clouds onto several 2-D feature maps and extract hybrid features using a 2-D deep learning network.

Jiang et al. [34] implemented two distinct encoder parts for RGB and DSM, respectively. Each layer's depth feature map is fused with the corresponding image feature map. The resulting fusion map is subsequently passed through five residual layers for further processing. Nahhas et al. [35] initially extracted 2-D feature maps from orthophoto and LiDAR data, specifically RGB bands, DSM, DEM, nDSM, and the number of returns. Subsequently, they employed three sequential CNN modules to generate low-level features, compressed features, and high-level features for the final segmentation. Zhang et al. [36] extracted features separately from images and DSM derived from point clouds using two improved FCN models. The outputs of the middle three convolution modules in the encoder part are upsampled to the same size, summed up, and passed through a softmax function to produce the segmentation result. Pan et al. [37] proposed an encoder–decoder structure for fine segmentation of aerial imagery and LiDAR. In the encoder stage, color-infrared images and DSMs from LiDAR are processed separately using VGG16-based branches. The decoder stage uses the subpixel convolution layers to upscale the coarse outputs from the encoder in an adaptive manner. Huang et al. [38] developed a modified residual learning network accompanied by a gated feature labeling unit to generate multilevel features from the red band, green band, near-infrared bands, and nDSM data. Arief et al. [39] proposed an advanced fusion approach by integrating a deep layer into the stochastic atrous network. This technique effectively merges both image-based and LiDAR-derived features. Eitel et al. [40] devised two separate streams followed by a late fusion network that utilizes RGB and depth image pairs as inputs. Hazirbas et al. [24] also presented a dual-branch network that leverages RGB and depth images. In this method, the features are combined on the deeper network. Xu et al. [41] designed a three-branch segmentation network, which propelled them to the first place in the 2018 IEEE GRSS Data Fusion Contest. The first branch of their model utilizes very high-resolution (VHR) images and LiDAR intensity data as inputs, while the second branch leverages DSM. The third branch, located in the middle of the network, utilizes hyperspectral images. Zhang et al. [20] presented a hybrid attention-aware fusion network comprising three streams: RGB-specific, DSM-specific, and cross-modal streams. During the encoder stage, the outputs from these three streams are fed through attention-aware multimodal fusion blocks and then passed into the subsequent convolution modules of the cross-modal stream. In the decoder stage, the segmentation results from the three streams are merged using an attention-aware multimodal fusion block to generate the final result. Chen et al. [42] proposed a novel approach for building extraction that employs a multimodal adaptive iterative strategy. First, two contour maps are detected from DSM and high spatial resolution imagery (HSRI), respectively, and two sets of hierarchical results are obtained through adaptive iterative segmentation. Next, the vegetation detected from the HSRI is eliminated from the nonground region identified from DSM. The outcome of this process is considered as the initial building segmentation result. Finally, the outputs of the previous steps are fused, and hierarchical overlay analysis and morphological operations are performed to achieve the final building extraction result. Marmanis et al. [43] introduced a holistically nested edge detection network, which enables semantically informed edge detection. The resulting object boundaries are subsequently merged with imagery and DEM and processed using three parallel CNN modules. The final prediction is generated by averaging the outputs from these three modules.

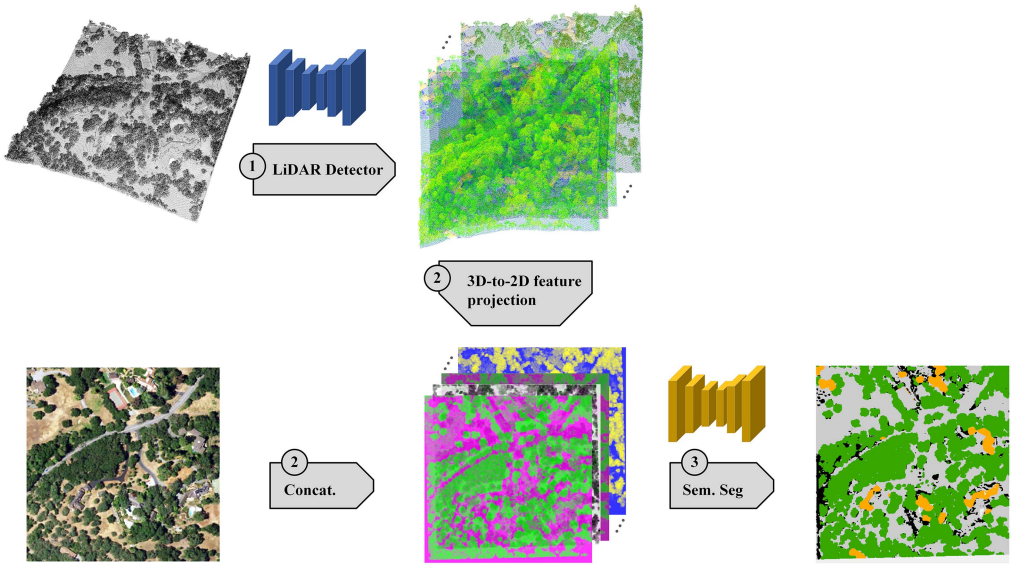

Fig. 1. PointBoost overview. The PointBoost framework comprises of three main stages. (1) LiDAR-based feature extraction network. (2) 3D-to-2D feature projection and concatenation. (3) Semantics network for 2-D joint-feature extraction.

Although these methods have demonstrated their effectiveness in addressing the task, they still encounter challenges in bridging the domain gap between 3-D point clouds and 2-D optical imagery. The main lies in the projection operations applied in preprocessing, which can negatively compromise the interpoint spatial topological relationships and point attributes, such as intensity and number of returns.

III. Methodology and Conceptual Framework

In this section, we provide a comprehensive explanation of the PointBoost framework. As depicted in Fig. 1, PointBoost is a straightforward and flexible framework consisting of three primary steps. First, a classic 3-D extractor is utilized to learn feature from unordered LiDAR points, which leverages the topological relationship in 3-D space to its fullest potential. Unlike other methods that rely on condensed derivatives, such as DSM, our approach focuses on the raw point data. In the second step, a spatial coordinate transformation matrix is established based on the projection relationship from 3-D to 2-D space. Subsequently, the point-level feature vectors obtained from the previous step are projected onto the imagery space through this transformation. Finally, the LiDAR-derived features are combined with the corresponding images, and the results are fed into an optical CNN structure to generate the 2-D labeling results.

The PointBoost process can be illustrated by the following formula:

$$r = G\left(\Gamma\left(F\left(x\right)\right) \oplus y\right) \tag{1}$$

where r represents the final 2-D labeling results. x and y stand for LiDAR patches and the corresponding images, respectively. F presents the LiDAR feature extraction stage, G symbolizes the joint-feature extraction stage, and Γ denotes the 3-D to 2-D transformation stage from LiDAR space to imagery space. The symbol $\oplus$ represents the concatenation.

A. LiDAR Feature Extraction

Compared with imagery, LiDAR is a more complex type of data that describes the real world in a more detailed and precise manner as more features are captured within its representation. In addition, the topological relationships between LiDAR points offer a structured description of remote sensing objects, serving as a robust enhancement to the 2-D texture information. For example, although tall trees and low shrubs may exhibit similar texture characteristics and 2-D structures in imagery, their topologies in 3-D space are vastly different. However, there is an inherent dimensional disparity between LiDAR and imagery data, resulting in differences in their data organization. To address this challenge, we first extracted features from the LiDAR point cloud and then aligned the feature map with the image, effectively minimizing the loss of information from the point cloud data and balancing the data organization.

The LiDAR feature extraction method is not mandatory and can be chosen based on the specific requirements of the application. In this article, we utilized PointNet++ [44], an optimized version of PointNet [45], to handle point clouds. PointNet consists of multiple multilayer perceptrons' layers and a max pooling layer that is stacked to generate discriminative features. To maintain the point-to-point relationships and boost the network's capacity for integrating local information, PointNet++ employs sampling and grouping modules with varying parameters to extract features at various scales. It also improves upon the concatenation strategy of PointNet by incorporating an encoder–decoder structure equipped with skip connections. The network is optimized by cross-entropy loss function. When extracting point cloud features, we preserved 128-D feature vectors by discarding the last convolution layer.

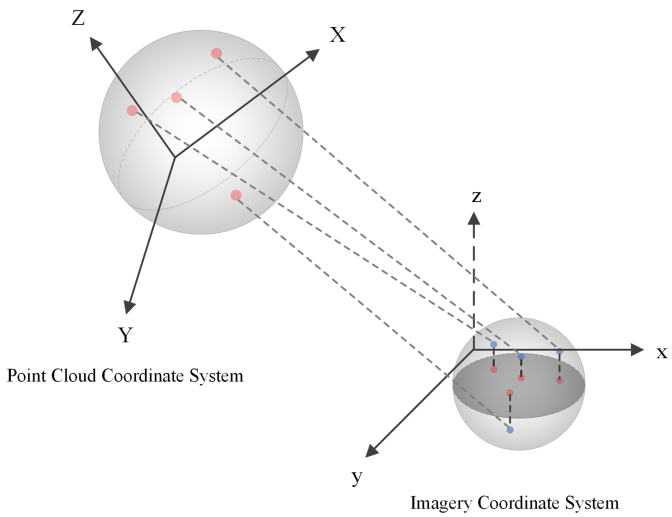

Fig. 2. Process of projecting LiDAR-derived features from 3-D space onto the 2-D imagery space.

B. 3D-to-2D Feature Projection and Concatenation

Point clouds and images not only occupy different dimensional spaces but also use distinct coordinate systems. LiDAR-derived features are based on a ground object system, while images utilize a 2-D imaging system. The LiDAR-derived features can be projected from 3-D space onto the 2-D imagery space through a combination of an affine transformation and an orthographic transformation, as illustrated in Fig. 2.

The overall process of projection can be expressed by the formula as follows:

$$x_{\text{img}} = S_o \cdot (S_a \cdot T \cdot R_X R_Y R_Z \cdot x_{\text{pc}}) \tag{2}$$

where $x_{\text{img}} = [x, y, 1, 1]^{\text{T}}$ and $x_{\text{pc}} = [X, Y, Z, 1]^{\text{T}}$ are the coordinates in imagery and point cloud space, respectively. S_o is the scaling ratio of orthographic transformation and S_a is that of affine transformation. T is the translation matrix. R_X, R_Y, and R_Z are the rotation matrices around the X-axis, Y-axis, and Z-axis, respectively, whose determinants' forms are as follows:

$$R_X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_X & -\sin\theta_X & 0 \\ 0 & \sin\theta_X & \cos\theta_X & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3}$$

$$R_Y = \begin{pmatrix} \cos\theta_Y & 0 & \sin\theta_Y & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta_Y & 0 & \cos\theta_Y & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4}$$

$$R_Z = \begin{pmatrix} \cos\theta_Z & -\sin\theta_Z & 0 & 0 \\ -\sin\theta_Z & \cos\theta_Z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5}$$

where θ_X, θ_Y, and θ_Z are the rotation angles around the X-axis, Y-axis, and Z-axis, respectively.

For occluded points, we averaged the feature vectors extracted from points projected to the same pixel and assigned this averaged value to the corresponding pixel. This strategy allows us to integrate the information carried by all occluded points while preserving the 3-D structure within small areas before LiDAR feature extraction.

C. Joint-Feature Extraction

In the third stage, we merged the features extracted from LiDAR after cross-modal projection with the corresponding images of the same area. Similar to the first step, the joint-feature extraction method in our proposed framework is also flexible and can be easily adjusted to suit various multimodal segmentation tasks. In this article, we utilized a U-shaped network [46] to extract joint features from these modalities, which consists of an encoder that maps the input data to a compact feature representation, followed by a decoder that reconstructs the original input from the feature representation. The encoder comprises multiple layers of convolutional and pooling operations, which downsample the input data to capture high-level semantic features. The decoder, on the other hand, consists of multiple layers of deconvolutional and upsampling operations, which restore the spatial resolution of the feature maps and reconstruct the original input. The network is optimized using a cross-entropy loss function.

This stage can facilitate the integration of spatial and geometric information from point clouds with the rich visual information from images, resulting in a more comprehensive and discriminative feature representation.

IV. Experiments and Analysis

In this section, we provide a detailed account of our experimental methodology. First, we introduce the three datasets used in our experiments. Next, we describe the experimental setup that we employ. Finally, we present the results of our experiments, demonstrating the superior quantitative and qualitative performance of our proposed method over the benchmarks and other state-of-the-art (SOTA) methods on the three datasets.

A. Dataset Description

To substantiate the superiority and generalizability of our method, we conducted experiments on three distinct datasets comprising both imagery and point clouds covering the same area. The datasets are N3C-California [47], ISPRS Vaihingen [48], and GRSS DFC 2018 [41] datasets.

N3C-California is a unified LiDAR-imagery benchmark specifically for multimodal joint land-cover segmentation tasks. The dataset comprises 1212 pairs of LiDAR and images' tiles, covering over 725 km^2 of both urban and rural areas in Santa Clara County, California. Each image tile has a size of 1304 $\times$ 1304 pixels with a resolution of 100 cm/pixel. The N3C-California dataset is categorized into four typical remote sensing semantic categories, namely, ground, tree, building, and others. The dataset is divided into training, validation, and test sets at a ratio of 8:1:1.

ISPRS Vaihingen is an open dataset released by the ISPRS Test Project on Semantic Labeling. It covers a village area with many detached and small multistory buildings. The dataset comprises 33 tiles of VHR image and corresponding DSMs, with an average size of 2493 × 2063 pixels and a resolution of 9 cm/pixel. The images are classified into six land-cover classes, including building, tree, low vegetation, background, cars, and impervious surfaces. The training set comprises tiles of image numbered 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, and 37. The validation set consists of tiles numbered 11, 15, 28, 30, and 34. The test set includes tiles numbered 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38.

GRSS DFC 2018 Dataset is provided by the 2018 IEEE GRSS Data Fusion Contest for urban land use and land-cover classification. It comprises three types of optical remote sensing data, including multispectral-LiDAR point cloud data, hyperspectral data, and VHR RGB imagery. The multispectral-LiDAR point cloud data have a point density of 10 pls/m^2 and covers approximately 5.01 km^2 of urban area. The VHR RGB data includes 14 image tiles, with a size of 11 920 × 12 020 pixels and a resolution of 5 cm/pixel. The dataset is classified into 20 categories of natural land-cover and man-made objects, including healthy grass, stressed grass, artificial turf, evergreen trees, deciduous trees, bare soil, water, residential buildings, nonresidential buildings, roads, sidewalks, crosswalks, major thoroughfares, highways, railways, paved parking lots, unpaved parking lots, cars, trains, and stadium seats. The dataset only includes training and test set with a ratio of 2:5.

The N3C-California dataset exhibits exceptional data precision, while ISPRS Vaihingen and GRSS DFC 2018 datasets have undergone geometrically corrected with high precision. Therefore, all three multimodal datasets demonstrate outstanding geometric consistency and fulfill the requirements for joint processing of multisource heterogeneous data.

B. Experimental Setting and Evaluation Metrics

All experiments were conducted on a NVIDIA GeForce RTX 3090 24G GPU using the PyTorch deep learning framework. For the LiDAR feature extraction stage, we randomly selected 8192 points as one sample from each patch. The network is trained for eight epochs using the Adam optimizer with an initial learning rate of 0.001 and a decay rate of 10^{-4}. The batch size is set to 16 for this stage. In the joint-feature extraction stage, we used input patches of size 512 × 512. The network is trained for 80 000 iterations using SGD with an initial learning rate of 0.01 and cosine annealing strategy, with a batch size of 8.

We evaluated the methods by overall accuracy (OA), mean accuracy (Mean Acc), Cohen's Kappa (Kappa), intersection over union (IoU), and $F1$-score.

OA measures the ratio of the number of correct predictions p_{correct} to the total number of pixels p_{all}

$$\text{OA} = \frac{p_{\text{correct}}}{p_{\text{all}}}. \tag{6}$$

The other metrics are based on the confusion matrix, where the true positive number, the true negative number, the false positive number, and the false negative number for the kth class are denoted by TP^k, TN^k, FP^k, and FN^k. The symbol K represents the number of categories.

Mean Acc and Kappa are as follows:

$$\text{Mean Acc} = \sum_{k=1}^{K} \frac{\text{TP}^k}{\text{TP}^k + \text{FP}^k} \tag{7}$$

$$\text{Kappa} = \frac{\text{OA} - p_e}{1 - p_e} \tag{8}$$

where

$$p_e = \frac{\sum_{k=1}^{K} (\text{TP}^k + \text{FP}^k)(\text{TP}^k + \text{FN}^k)}{p_{\text{all}}^2}. \tag{9}$$

IoU and $F1$-score for the kth class are defined as follows:

$$\text{IoU}^k = \frac{\text{TP}^k}{\text{TP}^k + \text{FP}^k + \text{FN}^k} \tag{10}$$

$$F1 - \text{score}^k = \frac{2 \times \text{precision} \times \text{recall}}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} \tag{11}$$

where

$$\text{precision} = \frac{\text{TP}^k}{\text{TP}^k + \text{FP}^k} \tag{12}$$

$$\text{recall} = \frac{\text{TP}^k}{\text{TP}^k + \text{FN}^k}. \tag{13}$$

C. N3C-California Dataset Results

1) Ablation Study: Our method's biggest highlight lies in its direct utilization of cross-modal data. To demonstrate the effectiveness of our approach, we conducted a series of ablation experiments on the N3C-California dataset. Specifically, we converted several attributes of the point cloud data, such as DSM, intensity, and number returns, into feature maps of the image domain. We then concatenated these maps with RGB and inputted the resulting combination into the same joint-feature extraction network of PointBoost, which uses cross-dimensional multimodal data directly. Their performances were compared. Additionally, we evaluated the 3-D segmentation of point clouds using PointBoost's LiDAR detector. The results of these ablation experiments are presented in Table I, where the best results for each metric are shown in bold.

The results, as presented in Table I, demonstrate that our variant PointBoost, which utilizes LiDAR-derived feature maps (rows 3–6), significantly outperforms the 2-D single-modal method using only imagery (row 1) across all aggregate metrics. Moreover, the variant PointBoost outperforms 3-D single-modal methods that rely solely on point clouds (row 2) in terms of OA and Kappa. The standard PointBoost model (row 7) outperforms the variant PointBoost (rows 3–6). These results suggest that the standard PointBoost model, which preserves all the 3-D properties of LiDAR data, is able to extract richer semantic features from cross-dimensional and cross-modal information.

2) Comparing With SOTA Methods: We compared our PointBoost with several popular methods in the field of remote sensing on the N3C-California dataset. These included the baseline

TABLE I
ABLATION STUDY USING DIFFERENT INPUT DATA WITH THE SAME JOINT-FEATURE EXTRACTION NETWORK ON THE N3C-CALIFORNIA DATASET

Method	Input	OA	Mean Acc	Kappa	IoU				
					Others	Ground	Tree	Building	Mean
UNet	RGB	86.35	67.17	77.22	2.46	80.92	73.33	81.01	59.43
PointNet++	LiDAR	86.82	**88.44**	78.48	22.45	80.73	84.58	85.64	68.35
Hybrid-UNet	RGB+DSM	89.00	71.76	81.66	12.98	85.86	77.03	87.13	65.75
Hybrid-UNet	RGB+DSM+intensity	89.08	81.36	80.78	15.52	83.08	85.90	76.04	65.14
Hybrid-UNet	RGB+DSM+number returns	89.91	80.72	81.86	18.15	80.96	85.55	67.90	63.14
Hybrid-UNet	RGB+DSM+intensity +number returns	89.22	81.52	81.00	15.82	83.20	85.98	76.82	65.46
PointBoost (ours)	RGB+LiDAR	**92.52**	82.20	**88.02**	**34.67**	**89.44**	**87.65**	**91.87**	**75.91**

TABLE II
QUANTITATIVE COMPARISON OF POINTBOOST AND RELATED SOTA MODELS ON N3C-CALIFORNIA DATASET

Method	Input	OA	Mean Acc	Kappa	IoU				
					Others	Ground	Tree	Building	Mean
UNet (baseline)	RGB	86.35	67.17	77.22	2.46	80.92	73.33	81.01	59.43
Hybrid-UNet	RGB+DSM	89.00	71.76	81.66	12.98	85.86	77.03	87.13	65.75
vFuseNet	RGB+DSM	86.11	75.47	74.68	**49.73**	81.99	57.58	77.43	66.68
MultifilterCNN	RGB+DSM+intensity +number returns+DoG	88.97	76.44	81.68	25.77	84.33	77.63	82.63	67.59
MFNet	RGB+DSM +Slope angle+DoG	91.00	74.85	85.72	14.29	87.36	82.09	89.87	68.40
PointBoost (ours)	RGB+LiDAR	**92.52**	**82.20**	**88.02**	34.67	**89.44**	**87.65**	**91.87**	**75.91**

method UNet [46], the multimodal benchmark method Hybrid-UNet [21], and the SOTA multimodal segmentation networks vFuseNet [49], MultifilterCNN [26], and MFNet [19]. The results are presented in Table II.

In Table II, we observe that the IoU for the "Others" category is significantly lower than the other three categories across all methods. This can be attributed to the fact that the number of pixels belonging to the "Others" category is only about 5% of the other classes, which makes it challenging for networks to learn the discriminative features.

Regarding the IoU for each category, our method shows the most significant improvement in the "Tree" category. We hypothesize that this is because the "Tree" category possesses more complex and significant 3-D structural features, which are often destroyed by other multimodal algorithms during the preprocessing stage.

Overall, our method outperforms all other methods in both aggregate and subdivision metrics for meaningful categories. With respect to OA, our method shows an improvement of over 6% compared with the baseline and over 1.5% compared with the most powerful SOTA multimodal segmentation network. PointBoost also significantly improves Mean Acc, with improvements of 15%, 10%, and 7% or more over the baseline, multimodal benchmark, and best-performing SOTA, respectively. Furthermore, our PointBoost achieves notable improvements on the kappa indicator, with accuracy increases ranging from 2.3% to 10.2% compared with other methods and a substantial quantitative improvement in the IoU indicator.

We present the visualization results of our PointBoost and other methods for six scenes in Fig. 3. In the first row, it is evident that the single-modal method is prone to making some confusing judgments due to the phenomenon of the same spectrum that corresponds to different objects (as seen in the lower right corner), while the multimodal method can make correct identifications. In the fourth scene, completing buildings that are shaded by tall trees is difficult with only the texture features of the image (as shown in row 4, column 2). Even the SOTA multimodal method has extremely limited efficacy (row 4, columns 3–6). Our method ensures the accuracy of tree boundaries while also preserving the integrity of the buildings (row 4, column 7). This is because our method works directly on point clouds and images, preserving the 3-D structure of point cloud clusters, almost all attributes of points and the texture of images simultaneously. In all six scenarios, our method predicts the finest edges for both trees and buildings.

D. ISPRS Vaihingen Dataset Results

In this section, we compare PointBoost with the top-ranked algorithms published on the ISPRS Vaihingen Dataset Challenge webpage, including SVL_3 [50], HUST [51], RIT [52], UOA [53], ADL_3 [54], DST_1 [21], DLR_8 [43], UFMG_4 [55],

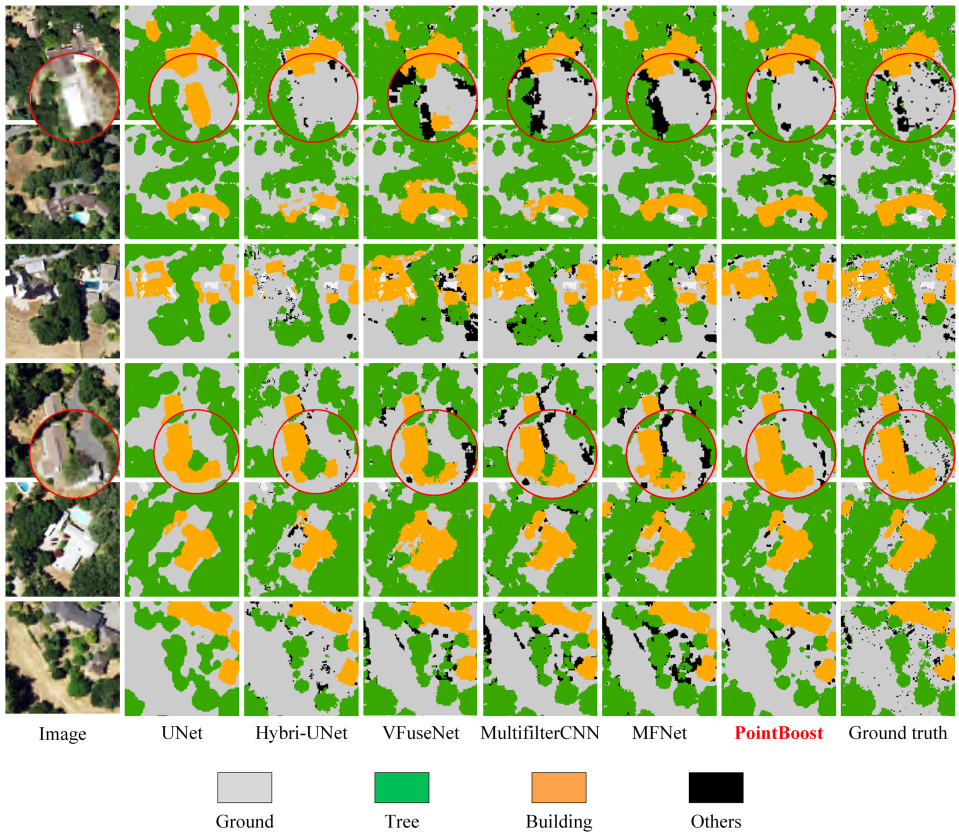

Fig. 3. Visualization comparison of the baseline, the multimodal benchmarks, the SOTA multimodal segmentation networks, and PointBoost (ours) on N3C-California.

TABLE III
QUANTITATIVE COMPARISON OF POINTBOOST AND THE TOP-RANKED ALGORITHMS PUBLISHED ON THE ISPRS VAIHINGEN DATASET CHALLENGE WEBPAGE

Method	OA	*F*1-score					
		Imp surf	Building	Low veg	Tree	Car	Mean
SVL_3 [50]	84.8	86.6	91.0	77.0	85.0	55.6	79.0
HUST [51]	85.9	86.9	92.0	78.3	86.9	29.0	74.6
RIT [52]	86.3	88.1	93.0	80.5	87.2	41.9	78.1
UOA [53]	87.6	89.8	92.1	80.4	88.2	82.0	86.5
ADL_3 [54]	88.0	89.5	93.2	82.3	88.2	63.3	83.3
DST_1 [21]	88.7	90.3	93.5	82.5	88.8	73.9	85.8
DLR_8 [43]	89.2	90.4	93.6	83.9	89.7	76.9	86.9
UFMG_4 [55]	89.4	91.1	94.5	82.9	88.8	81.3	87.7
ONE_7 [56]	89.8	91.0	94.5	84.4	**89.9**	77.8	87.5
CASIA2 [57]	91.1	93.2	**96.0**	84.7	**89.9**	86.7	90.1
PointBoost (ours)	**91.9**	**96.7**	91.1	**89.5**	83.0	**93.9**	**90.8**

ONE_7 [56], and CASIA2 [57]. The evaluation criteria are consistent with those published on the website, focusing on the OA and *F*1-score, accurate to one decimal place, as shown in Table III. The best results are indicated in bold.

Our method achieves the best results in both OA and mean *F*1-score, with improvements of 0.8% and 0.7%, respectively, over the suboptimal method. However, we notice that PointBoost does not demonstrate an advantage in the "Building" and "Tree" categories, which performs well on the N3C-California dataset. This could be attributed to the relatively sparse point cloud of the ISPRS Vaihingen dataset, with a point density of only 4 pts/m^2, which is less than half of the N3C-California dataset. This limited point density results in relatively limited 3-D structural information that the point cloud can provide.

We display the results of the top-performing five algorithms and PointBoost in Fig. 4. It can be observed that the "Low vegetation" category in this dataset is often mistaken for other categories. Our method accurately distinguishes it from the "Building" and "Tree" categories, whereas other methods tend to make errors, as seen in the upper left corner of the fourth scene (row 4) and the bottom left corner of the fifth scene (row 5).

E. GRSS DFC 2018 Dataset Results

We compare the accuracy of PointBoost and the winning algorithms [41] on the GRSS DFC 2018 Dataset using mean Acc, OA, and Kappa indicators as per the competition rules. The results are presented in Table IV, where the best results are highlighted in bold.

It should be noted that the listed are the best-performing methods on the GRSS data and algorithm standard evaluation.

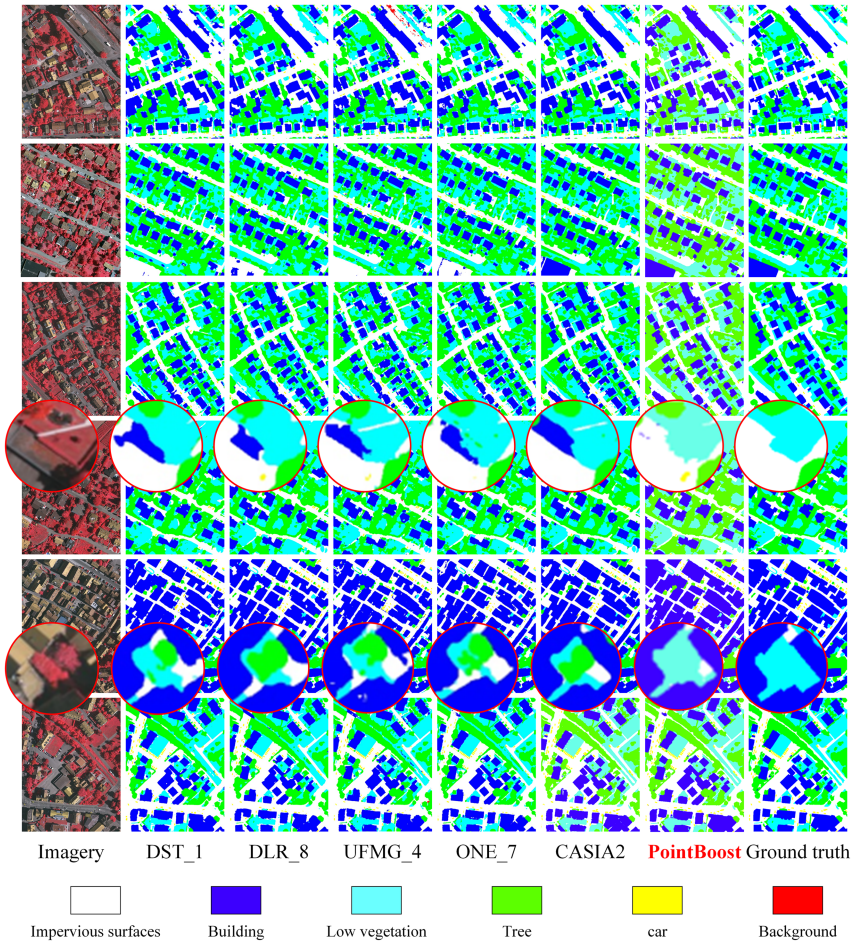

Fig. 4. Visualization comparison of the top-ranked algorithms and PointBoost (ours) on the ISPRS Vaihingen dataset.

TABLE IV
QUANTITATIVE COMPARISON OF POINTBOOST AND THE WINNING ALGORITHMS [41] ON THE GRSS DFC 2018 DATASET

Team	Mean Acc	OA	Kappa
XudongKang	71.26	76.45	0.75
Gaussian	71.66	80.78	**0.80**
IPIU	74.40	79.23	0.78
challenger	75.99	77.90	0.77
AGTDA	76.15	79.79	0.79
dlrpba	**76.32**	80.74	**0.80**
PointBoost (ours)	75.79	**80.88**	0.79

However, PointBoost still achieves comparable performance with them and performs 0.1% better than the second-best method in terms of OA metric improvement.

The visualization results of our method and the winning teams AGTDA and dlrpba for 20 types of ground objects are presented in Fig. 5. Due to the large number of categories and intricate details, the segmentation of the GRSS DFC 2018 dataset presents a significant challenge. However, the three methods demonstrate the effective distinction between residential buildings (white) and vegetation (green). PointBoost misclassifies some pixels in the roads (red) category as nonresidential buildings (purple).

V. DISCUSSION

Experiments conducted on three datasets demonstrate the superior segmentation accuracy of our method compared with other multimodal approaches. Notably, our PointBoost exhibits finer accuracy in the "Tree" and "Low vegetation" categories of the N3C-California and ISPRS Vaihingen datasets, respectively, which differ in point cloud densities. This highlights the unique advantage of our method in vegetation classification. As depicted in Fig. 3, row 5, our method accurately identifies the edges of individual trees in the middle of the image, while other methods exhibit various forms of misjudgment. In the upper left corner of the fourth row in Fig. 4, only our method correctly distinguishes the "Low vegetation" category from the "Building" category.

These positive results can be attributed to two key factors. First, the use of LiDAR data directly in our PointBoost framework allows us to leverage the multiple returns generated by the interaction of laser beams with vegetation, enabling the preservation of the hierarchical structure of vegetations. In contrast,

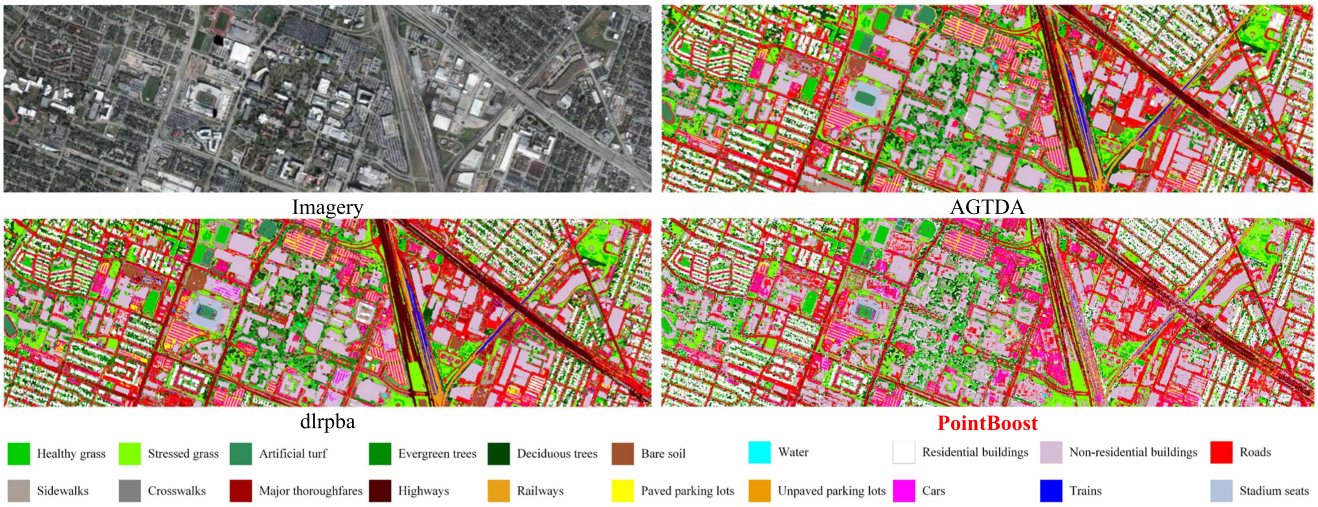

Fig. 5. Visualization comparison of the winning algorithms and PointBoost (ours) on the GRSS DFC 2018 dataset.

common multimodal methods, which convert LiDAR data into 2-D feature maps, such as DSM in the preprocessing stage, fail to exploit this inherent characteristic of LiDAR and, consequently, lose crucial geometric information contained in the 3-D point clouds. Second, in the 3D-to-2D feature projection step, our method employs an averaging technique to integrate the feature vectors of points projected onto the same pixel, which effectively preserves attribute information discarded by other methods. This approach ensures the retention of 3-D structural features between points in small areas. Conversely, other methods only retain the point with the highest elevation among those projected onto the same pixel when converting point clouds into DSM. Hence, our method not only integrates attribute information from multiple points but also maintains the geometric structure of the 3-D point clouds.

VI. Conclusion

In this article, we propose PointBoost, an effective sequential segmentation method that takes advantage of the full potential of cross-modal data of LiDAR point clouds and imagery. Unlike other methods that usually discard the topology between points and some LiDAR attributes, PointBoost is designed to work directly with point clouds and images, providing the more accurate and detailed segmentation results and a richer representation of the real world. Additionally, the PointBoost framework is highly flexible, allowing for arbitrary modifications to its 3-D or 2-D segmentation network. The experiments demonstrate that PointBoost outperforms related SOTA methods. Going forward, our research will prioritize developing an end-to-end solution for cross-modal information fusion as well as plug-in intermodal projection.

References

[1] L. Zhang, Lefei Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[2] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, 2021.

[3] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, 2020, Art. no. 111716.

[4] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, 2021, Art. no. 114417.

[5] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102926.

[6] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, 2020.

[7] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, Nov. 2021, Art. no. 5517010.

[8] E. Alvarez-Vanhard, T. Corpetti, and T. Houet, "UAV & satellite synergies for optical remote sensing applications: A literature review," *Sci. Remote Sens.*, vol. 3, 2021, Art. no. 100019.

[9] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, 2020.

[10] M. S. Moran, Y. Inoue, and E. M. Barnes, "Opportunities and limitations for image-based remote sensing in precision crop management," *Remote Sens. Environ.*, vol. 61, no. 3, pp. 319–346, 1997.

[11] M. Leordeanu and M. Hebert, "A spectral technique for correspondence problems using pairwise constraints," in *Proc. IEEE 10th Int. Conf. Comput. Vis.*, 2005, vol. 2, pp. 1482–1489.

[12] F. Poux, R. Neuville, P. Hallot, and R. Billen, "Smart point cloud: Definition and remaining challenges," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 119–127, 2016.

[13] R. L. Schwiesow, "A comparative overview of active remote-sensing techniques," in *Probing Atmospheric Boundary Layer*. Boston, MA, USA: American Meteorological Society, 1986, pp. 129–138.

[14] D. K. Killinger and A. Mooradian, *Optical and Laser Remote Sensing*. Berlin, Germany: Springer, 2013.

[15] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogramm. Remote Sens.*, vol. 115, pp. 22–36, 2016.

[16] 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.

[17] K. Wang, T. Wang, and X. Liu, "A review: Individual tree species classification using integrated airborne LiDAR and optical imagery with a focus on the urban environment," *Forests*, vol. 10, no. 1, 2018, Art. no. 1.

[18] L.-C. Chen, T.-A. Teo, Y.-C. Shao, Y.-C. Lai, and J.-Y. Rau, "Fusion of LIDAR data and optical imagery for building modeling," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 35, no. B4, pp. 732–737, 2004.

[19] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, Sep. 2021, Art. no. 5404418.
[20] P. Zhang et al., "A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data," *Remote Sens.*, vol. 12, no. 22, 2020, Art. no. 3764.
[21] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*.
[22] 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.
[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[24] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. 13th Asian Conf. Comput. Vis.*, 2017, pp. 213–228.
[25] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 1561–1570.
[26] Y. Sun, X. Zhang, Q. Xin, and J. Huang, "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 143, pp. 3–14, 2018.
[27] Y. Cui et al., "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 722–739, Feb. 2022.
[28] T. H. Nguyen, S. Daniel, D. Gueriot, C. Sintes, and J.-M. L. Caillec, "Coarse-to-fine registration of airborne LiDAR data and optical imagery on urban scenes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 1, pp. 3125–3144, May 2020.
[29] S. Cao, D. Hu, W. Zhao, M. Du, Y. Mo, and S. Chen, "Integrating multiview optical point clouds and multispectral images from ZiYuan-3 satellite remote sensing data to generate an urban digital surface model," *J. Appl. Remote Sens.*, vol. 14, no. 1, 2020, Art. no. 014505.
[30] K. Lahssini, F. Teste, K. R. Dayal, S. Durrieu, D. Ienco, and J.-M. Monnet, "Combining LiDAR metrics and Sentinel-2 imagery to estimate basal area and wood volume in complex forest environment via neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, no. 1, pp. 4337–4348, May 2022.
[31] Z. Kang, J. Yang, and R. Zhong, "A Bayesian-network-based classification method integrating airborne lidar data with optical images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 4, pp. 1651–1661, Apr. 2017.
[32] D. Chai, "A probabilistic framework for building extraction from airborne color image and DSM," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 948–959, Mar. 2017.
[33] E. Maltezos, N. Doulamis, A. Doulamis, and C. Ioannidis, "Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 042620.
[34] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "Rednet: Residual encoder-decoder network for indoor RGB-D semantic segmentation," 2018, *arXiv:1806.01054*.
[35] F. H. Nahhas, H. Z. M. Shafri, M. I. Sameen, B. Pradhan, and S. Mansor, "Deep learning approach for building detection using lidar-orthophoto fusion," *J. Sensors*, vol. 2018, 2018, Art. no. 7212307.
[36] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, and H. Mayer, "Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 52.
[37] X. Pan, L. Gao, A. Marinoni, B. Zhang, F. Yang, and P. Gamba, "Semantic labeling of high resolution aerial imagery and LiDAR data with fine segmentation network," *Remote Sens.*, vol. 10, no. 5, 2018, Art. no. 743.
[38] J. Huang, X. Zhang, Q. Xin, Y. Sun, and P. Zhang, "Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network," *ISPRS J. Photogramm. Remote Sens.*, vol. 151, pp. 91–105, 2019.
[39] H. A. Arief, G.-H. Strand, H. Tveite, and U. G. Indahl, "Land cover segmentation of airborne LiDAR data using stochastic atrous network," *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 973.
[40] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2015, pp. 681–687.
[41] Y. Xu et al., "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.
[42] S. Chen, W. Shi, M. Zhou, M. Zhang, and P. Chen, "Automatic building extraction via adaptive iterative segmentation with LiDAR data and high spatial resolution imagery fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 1, pp. 2081–2095, May 2020.
[43] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, 2018.
[44] 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. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5105–5114.
[45] R. Q. Charles, H. Su, M. Kaichun, 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. 77–85.
[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
[47] Y. Wang, Y. Wan, Y. Zhang, B. Zhang, and Z. Gao, "Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using images and LiDAR point clouds," *arXiv:2303.15777*.
[48] F. Rottensteiner, G. Sohn, M. Gerke, and J. D. Wegner, "ISPRS semantic labeling contest," *ISPRS: Leopoldshöhe, Germany*, vol. 1, no. 4, p. 4, 2014.
[49] N. Audebert, B. L. Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, 2018.
[50] M. Gerke, "Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen)," *Dept. Earth Observ. Sci.*, Univ. Twente, Enschede, The Netherlands, Tech. Rep., 2015, doi: 10.13140/2.1.5015.9683.
[51] N. T. Quang, N. T. Thuy, D. V. Sang, and H. T. T. Binh, "Semantic segmentation for aerial images using RF and a full-CRF," 2015.
[52] S. Piramanayagam, W. Schwartzkopf, F. W. Koehler, and E. Saber, "Classification of remote sensed images using random forests and deep learning framework," in *Proc. Image Signal Process. Remote Sens. XXII*, 2016, vol. 10004, pp. 205–212.
[53] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3194–3203.
[54] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. van-den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 36–43.
[55] K. Nogueira, M. D. Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7503–7520, Oct. 2019.
[56] N. Audebert, B. L. Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 180–196.
[57] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 78–95, 2018.

Yongjun Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University (WHU), Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently a Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, integration of LiDAR point clouds and images, and three-dimensional city reconstruction. Her research interests include remote sensing image processing and machine learning.

Yameng Wang received the B.S. degree in remote sensing science and technology in 2018 from Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

Her research interests include remote sensing image processing and machine learning.

Yi Wan was born in 1991. He received the B.S. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013 and 2018, respectively.

He is currently an Associate Research Fellow with Wuhan University. His research interests include digital photogrammetry, computer vision, three-dimensional reconstruction, and change detection in remote sensing imagery.

Wenming Zhou received the M.S. degree in geodesy and survey engineering from Central South University, Changsha, China, in 2013.

He is currently a Senior Engineer with Surveying, Mapping and Geoinformation Research Institute, China Railway Design Corporation, Tianjin, China. His research interests include aerial photogrammetry, satellite remote sensing technology, and their application in railway engineering.

Bin Zhang received the B.S. degree in remote sensing science and technology from Liaoning Technical University, Fuxin, China, in 2017. He is currently working toward the Ph.D. degree in photogrammetry and remote sensing with Wuhan University, Wuhan, China.

His research interests include high spatial resolution remote sensing image processing, computer vision, and pattern recognition.