ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023) 385–404

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and LiDAR point clouds

Yameng Wang, Yi Wan *, Yongjun Zhang *, Bin Zhang, Zhi Gao

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Keywords:
Land-cover semantic segmentation
Multi-modal data
LiDAR
Aerial imagery

ABSTRACT

Despite the good results that have been achieved in unimodal segmentation, the inherent limitations of individual data increase the difficulty of achieving breakthroughs in performance. For that reason, multi-modal learning is increasingly being explored within the field of remote sensing. The present multi-modal methods usually map high-dimensional features to low-dimensional spaces as a preprocess before feature extraction to address the nonnegligible domain gap, which inevitably leads to information loss. To address this issue, in this paper we present our novel **I**mbalance **K**nowledge-**D**riven Multi-modal **Net**work (**IKD-Net**) to extract features from multi-modal heterogeneous data of aerial images and LiDAR directly. IKD-Net is capable of mining imbalance information across modalities while utilizing a strong modal to drive the feature map refinement of the weaker ones in the global and categorical perspectives by way of two sophisticated plug-and-play modules: the **G**lobal **K**nowledge-**G**uided (**GKG**) and **C**lass **K**nowledge-**G**uided (**CKG**) gated modules. The whole network then is optimized using a joint loss function. While we were developing IKD-Net, we also established a new dataset called the **N**ational Agriculture Imagery Program and **3**D Elevation Program **C**ombined dataset in **California (N3C-California)**, which provides a particular benchmark for multi-modal joint segmentation tasks. In our experiments, IKD-Net outperformed the benchmarks and state-of-the-art methods both in the N3C-California and the small-scale ISPRS Vaihingen dataset. IKD-Net has been ranked first on the real-time leaderboard for the GRSS DFC 2018 challenge evaluation until this paper's submission. Our code and N3C-California dataset are available at https://github.com/wymqqq/IKDNet-pytorch.

1. Introduction

With the rapid development of sensors like optical cameras, radar, and 3D scanners, the era of big data has arrived; and multi-modal data for earth observation has emerged as a research frontier in remote sensing (RS), especially for land-cover semantic segmentation tasks (Ghamisi et al., 2019; Li et al., 2022a; Yang et al., 2021). Multi-modal data analysis is demonstrating that it can break through the performance bottleneck of unimodal semantic segmentation by synthesizing the advantages of each data source in order to obtain more diverse feature information.

Many past studies have focused on the joint use of three-dimensional (3D) airborne LiDAR point clouds and two-dimensional (2D) aerial images. To eliminate the structure difference between the two modalities, recent researchers mostly have mapped 3D point cloud data to 2D image spaces to get products like digital surface models (DSMs) or intensity images and then extracting the 2D features for analysis and classification. CMGFNet (Hosseinpour et al., 2022) proposed a gated fusion network to achieve multi-level feature fusion between very high resolution (VHR) images and DSM. GRRNet (Huang et al., 2019) stacked the NIR-Red-Green images and the normalized DSM (nDSM) as four-channel input and utilized five gated feature labeling units to fuse the features from the encoder and decoder. MultiModNet (Liu et al., 2022) extracted features from a NIR-Red-Green image and the nDSM with the pyramid attention and the gated fusion unit, which then were joined before placing them into the decoder. Although superior to unimodal methods, these cross-modal learning methods inevitably fail to fully explore the content of each modality because the prior operation of mapping the point cloud data from 3D to 2D does some irreparable harm to important characteristics, especially the geometric structure information. To the best of our knowledge, few researchers have focused on the issue of maintaining the complete information of the raw heterogeneous data in

* Corresponding authors.
E-mail addresses: ymw@whu.edu.cn (Y. Wang), yi.wan@whu.edu.cn (Y. Wan), zhangyj@whu.edu.cn (Y. Zhang), bin.zhang@whu.edu.cn (B. Zhang), gaozhinus@whu.edu.cn (Z. Gao).

https://doi.org/10.1016/j.isprsjprs.2023.06.014
Received 7 December 2022; Received in revised form 24 June 2023; Accepted 27 June 2023
Available online 6 July 2023

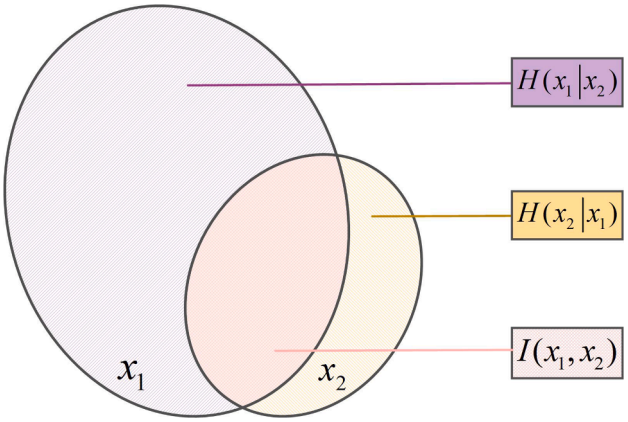

Fig. 1. Information diagram. The lilac circle x_1 and light yellow circle x_2 stand for LiDAR and the aerial imagery, respectively. $H(x_1|x_2)$ and $H(x_2|x_1)$ are the conditional entropy between the two modalities. $I(x_1, x_2)$ is their mutual information. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the multi-modal land-cover semantic segmentation task.

The information imbalance phenomenon also cannot be neglected in the joint analysis and semantic segmentation of multi-modal data. We use the 3DEP-QL1 LiDAR data (3D Elevation Program LiDAR of Quality-Level-1) and NAIP (National Agriculture Imagery Program) image data to demonstrate this issue in this paper. Each tile of 3DEP point cloud data (covering an area of about 0.6 km^2) has a data size of about 120 M and satisfies the QL1 accuracy standards (Heidemann, 2012). Specifically, the aggregate nominal pulse density (ANPD) is more than 8 pls/m^2 (8 points per square meter) and the aggregate nominal pulse spacing (ANPS) is less than 0.35 m. The absolute vertical accuracy (over the nonvegetated ground) is less than 0.1 m and the attributes of return number and intensity values always be included. The NAIP imagery covering the same area with the above point cloud tile has a data size of only about 5 M. It has a 0.6-meter ground sample distance (GSD) and affords RGB three-channel attributes. Also, the spatial resolution satisfies the standard digital orthoimage standards (Rufe, 2014). Fig. 1 illustrates the relationship of the amount of information that the two modalities provide in the joint analysis, where x_1 and x_2 stand for the modality of LiDAR and the imagery, respectively. Except for the mutual information $I(x_1, x_2)$, there may be several orders of magnitude more unique information of LiDAR than imagery, namely, $H(x_1|\ x_2) \gg H(x_2|\ x_1)$.

This phenomenon has attracted attention in multi-modal fusion, self-supervised learning, contrastive learning, and many other fields in the computer vision community. P4Contrast (Liu et al., 2020) leverages "pairs of point-pixel pairs" to provide extra flexibility in creating hard negatives and avoid the networks learning features only from the more discriminating one of different modalities. TupleInfoNCE (Liu et al., 2021c) proposed a tuple disturbing strategy to prevent networks from largely ignoring weak modalities while only focusing on strong modalities when learning multi-modal representation. However, most of these methods concentrate on reducing the influence of imbalanced information between modalities during sample selection rather than exploiting them within network optimization.

Despite the increasing interest in multi-modal learning, multi-modal datasets in the remote sensing community are scant (Yang et al., 2021), with just two widely used datasets containing point clouds and images (i.e., the ISPRS Vaihingen semantic dataset (Rottensteiner et al., 2014) and the GRSS DFC 2018 dataset (Xu et al., 2019)). Both datasets merely cover urban land in small areas and are insufficient to quantitatively evaluate multi-modal algorithms. Thus, large-scale multi-sensor datasets are urgently needed. To fill this gap, we presented our **N**ational Agriculture Imagery Program and **3**D Elevation Program **C**ombined dataset in **California** (**N3C-California**). The LiDAR in our dataset is from 3DEP public data and the aerial imagery is from NAIP public data. We performed geometric registration and cropping on the above data and published the corresponding four categories' pixel-level labels, thus providing a quantitative evaluation of the algorithms in the field of multi-modal earth observation. The pixel-level labels were mapped from the classification attributes of the dense point clouds, which were manually annotated with very high precision.

We synthesized the information discarded by previous methods in the preprocessing stage and fully exploited the inherent differences between modalities. To tackle the issues we discovered, we propose **I**mbalance **K**nowledge-**D**riven Multi-modal **Net**work (**IKD-Net**) to enables weak modality to get affluent information from stronger ones and promotes the modality synergy substantially.

The main contributions of this paper can be summarized as follows:

1. A specialized benchmark dataset called **N3C-California** for quantitative evaluation in multi-modal joint segmentation tasks. N3C-California is the largest coverage area annotated LiDAR-imagery dataset to date.
2. A novel efficient architecture called **IKD-Net**, which extracts features from raw multi-modal data directly rather than from their abridged derivatives. Its end-to-end disentangled dual-stream backbone helps to keep the information of heterogeneous modalities intact. The detailed ablation analysis and extensive comparative experiments in this paper on N3C-California and two other multi-modal datasets validated the design logic and superiority of IKD-Net.
3. Two plug-and-play gated modules **G**lobal **K**nowledge-**G**uided (**GKG**) and **C**lass **K**nowledge-**G**uided (**CKG**) that take advantage of the inherent imbalance information between two RS modalities. These modules provide new insight into multi-modal data interaction.
4. A well-designed **joint loss function** that consists of two single-task loss functions and a pixel-wise similarity loss to maintain the balance of the parameter flow in each branch during network optimization.

2. Related work

2.1. Semantic segmentation for 3D LiDAR point clouds

Point clouds are essentially low-resolution resamplings of the 3D physical world. The design of learning-based semantic segmentation methods for point clouds is closely related to the data structure of 3D representations (Guo et al., 2020).

Some methods first convert 3D point clouds into intermediate regular structures and extract the features with mature 2D or 3D convolution thereafter. The segmentation results then are finally projected back to the original point clouds. These semantic segmentation methods are classified as projection-based and discrete-based. Projection-based approaches can be divided into multi-view representation (Audebert et al., 2016; Boulch et al., 2017; Lawin et al., 2017; Tatarchenko et al., 2018) and spherical representation (Iandola et al., 2016; Milioto et al., 2019; Wu et al., 2018; Wu et al., 2019) according to the projection process. Discretization-based approaches convert point clouds into discrete representations, which include dense discretization representation (Huang and You, 2016; Long et al., 2015; Meng et al., 2019; Tchapmi et al., 2017) and sparse discretization representation (Choy et al., 2019; Graham et al., 2018; Rosu et al., 2019; Su et al., 2018). These methods unfortunately fail to take full advantage of the underlying geometric and structural information as the projection step inevitably leads to missing information.

Qi et al. (2017a) introduced the pioneering work PointNet, which directly processes irregular point clouds by extracting features point-by-point using a shared multilayer perceptron (MLP). Building upon this, PointNet++ (Qi et al., 2017b), PointSIFT (Jiang et al., 2018) and PointWeb (Zhao et al., 2019a) further enhance the ability to encode

neighboring information in point clouds. VD-LAB (Li et al., 2022b) integrates three novel modules into the U-Net network structure, significantly enhancing the model's generalization ability.

In addition to MLP, some studies have been devoted to specific point convolution operators. Hua et al. (2018) established a standardized approach for conducting convolution operations on point clouds. Wang et al. (2018) utilized a parametric continuous function to represent the convolution kernel, making it well-suited for unstructured point clouds. Thomas et al. (2019) introduced a deformable convolution kernel, called KPConv, which enables adaptive learning while minimizing memory consumption. Engelmann et al. (2020) extended the receptive field of point convolution operators by combining them with dilated convolutions.

Some approaches have explored the use of recurrent neural networks (RNN) in point clouds. For instance, Huang et al. (2018) embedded a novel slice pooling layer and a slice unpooling layer into an RNN framework. 3P-RNN (Ye et al., 2018) effectively fuses context information by sequentially employing pointwise pyramid pooling and two-direction hierarchical RNNs. DAR-Net (Zhao et al., 2019b) employs a convolutional-recurrent network to dynamically aggregate local and global features for improved performance.

Additionally, some graph-based methods have been employed for point cloud analysis. Landrieu and Boussaha (2019) utilized the local cloud embedder and graph-structured contrastive loss to compute a point cloud oversegmentation. Wang et al. (2019) introduced the graph attention convolution, which dynamically learned attention weights for different nodes. PointGCR (Ma et al., 2020) employed an undirected graph representation to learn global contextual information across the channel dimension. AF-GCN (Zhang et al., 2023) captures local features and long-range contexts by graph convolutions and the Graph Attentive Filter (GAF), respectively.

Most point cloud algorithms can handle only small ranges of point clouds, where the process of chunking may destroy the overall geometry of the point clouds. There are a few algorithms for large-scale point clouds, but they have computationally costly pre- or post-processing steps (Chen et al., 2019; Landrieu and Simonovsky, 2018; Rethage et al., 2018). RandLA-Net (Hu et al., 2020), on the other hand, adopted a random sampling strategy to continuously downsample large-scale point clouds, which greatly reduces the computational effort and preserves the complex geometric structure in large-scale point clouds using a local feature aggregation module.

2.2. Multi-modality learning

According to the time point of feature fusion, multi-modal learning can be roughly divided into three categories (early, middle, and late) which correspond to data-level, feature-level, and decision-level fusion, respectively.

The early fusion strategy performs data fusion at the front end of the network and further inputs the merging layer into a single branch network for segmentation. Nahhas et al. (2018) concatenated three LiDAR-derived features (DSM, DEM, and nDSM), seven shape-derived features, three image-spectral-derived features, and eight image-texture-derived features together and then reduced the dimension by an autoencoder before using CNN to abstract the deep features. Huang et al. (2019) put the near-infrared (NIR), red, and green bands from images and nDSM from LiDAR into a modified residual learning network and a gated feature labeling (GFL) process to extract buildings. Gadzicki et al. (2020) conducted experiments showing that early fusion methods outperform late fusion methods in human activity recognition tasks. The above early fusion strategies only treat LiDAR data as supplementary information to images, however, and ignore the discrepancies between the two modalities.

The middle fusion strategy focuses on the inter-feature combination and interaction. Fusion-FCN (Xu et al., 2019) employs a 1×1 conv to fuse three intermediate features extracted from the merging band of VHR image and LiDAR intensity raster data, nDSM, and high spectral data, respectively. HAFNet (Zhang et al., 2020) uses parallel structures to extract unimodal features from RGB and DSM. An attention-aware fusion block combines corresponding layer outputs for multi-modal feature learning. Zhang et al. (2017) proposed an improved FCN model with parallel encoders for images and 2D elevation features from LiDAR. The feature maps from both streams are fused after each convolution module and concatenated before feeding into the decoder of the FCN. MFNet (Sun et al., 2021) utilizes intra-modal, inter-modal, and multilevel feature fusion modules to integrate context information across modalities. S2ENet (Fang et al., 2021) enhances the interaction between hyperspectral data and LiDAR using spatial and spectral enhancement modules within the network. CMGFNet (Hosseinpour et al., 2022) employs a gated fusion module to combine features from VHR images and DSM, as well as a top-down strategy to fuse high-level and low-level features. MDL_RS (Hong et al., 2021) utilizes parallel Ex-Net to extract features from two modalities, followed by feeding the acquired results into the unified Fu-Net. Similarly, EndNet (Hong et al., 2022) also utilizes parallel FE-Nets to extract features from hyperspectral and LiDAR data individually. Subsequently, the acquired results are fed into the unified F-Net, enabling efficient fusion of cross-modal information. He et al. (2023) used gating and self-attention modules to fuse features extracted from multispectral and SAR images at multiple stages for flood detection. However, middle-level fusion is a challenging direction. Besides abstracting discriminative unimodal features, the feature-level fusion strategies should be able to distinguish the inter-modal differences and balance each modality's contribution to synthesize the high-level cross-modal features.

The late fusion strategy generally uses individual branches to extract the features of each modality, and the results are fused directly in the decision phase. Marmanis et al. (2018) proposed the Holistically-Nested Edge Detection (HED) network to fuse the boundary prediction of the separate streams. Gialampoukidis et al. (2021) proposed a method to merge feature maps from K modalities into a K-order tensor for image retrieval. Despite having a high degree of flexibility, relatively little research has been done for late fusion because it discards cross-modal interactions and modalities cannot be adequately interrelated.

2.3. Attention and gating mechanism

It is commonly believed that the human eye can quickly locate the key things that are meaningful from a cluttered picture. Researchers apply this thinking in deep learning and in response have proposed the concept of attention mechanisms. Remarkable results have been achieved in natural language processing (NLP) (Galassi et al., 2020), speech recognition (Chorowski et al., 2015), and image perception (Fu et al., 2019). In 2014, a Google Mind team (Mnih et al., 2014) used an attention mechanism based on reinforcement learning (RL) in a recurrent neural network (RNN), which not only looks at the image as a whole but also extracts the necessary information from the local area. Their approach achieved excellent performance on image classification tasks and was the beginning of a trend toward widespread application of attention mechanisms. Bahdanau et al. (2014) introduced the attention mechanism concept to the NLP field for the first time. Yin et al. (2016) suggested three alternatives for employing attention mechanisms in CNNs and conducted an early exploration of their application in CNNs. Hu et al. (2018) designed the squeeze-and-excitation (SE) block to generate a channel weight distribution vector to realign the correlation between feature channels. On this basis, using the selective kernel network (SKNet), Li et al. (2019) introduced lightweight multi-channel multi-scale channel attention to obtain channel-boosted features. With the convolutional block attention module (CBAM), Woo et al. (2018) sequentially applied the attention mechanism to the input feature map in both the channel and space dimensions to produce a refined feature. The dual attention network (DANet) (Fu et al., 2019) utilized two parallel branches to produce position and channel-enhanced feature maps,

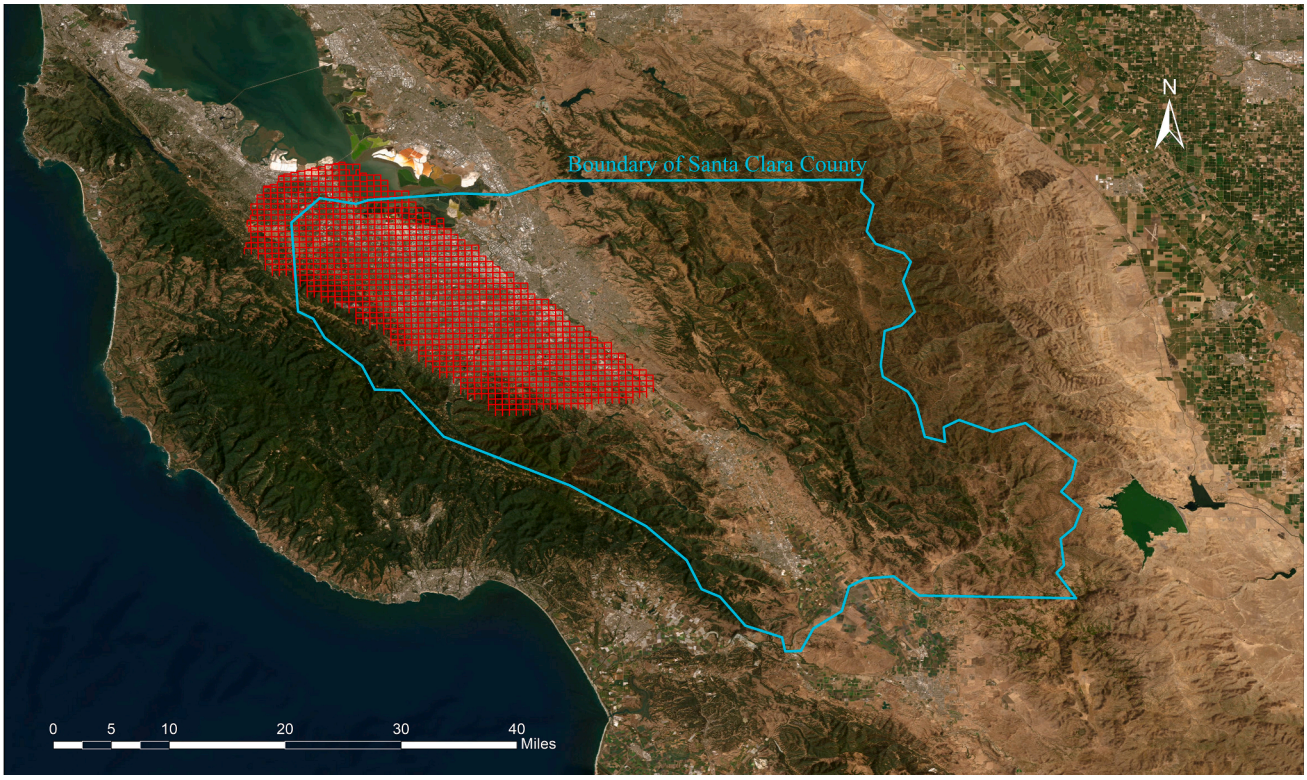

Fig. 2. The coverage area (red grids) of the N3C-California dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

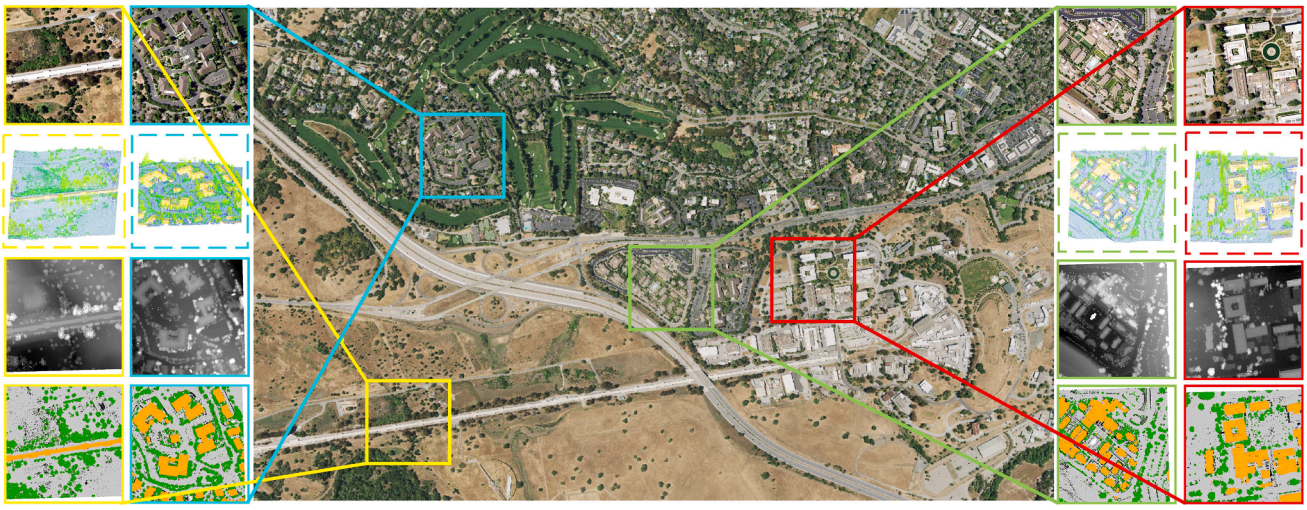

Fig. 3. Several samples in N3C-California, from top to bottom: aerial images, point clouds, DSM, and ground truth.

which were summed and then convolved to derive the image segmentation results. In conclusion, most of the existing attention mechanism approaches can obtain the weight distribution map from a single input and then act on the input itself. These methods are limited by the inherent ceiling on the amount of information in the input itself and can only fine-tune a feature map based on the contextual relationships between the pixels within the image. Our novel approach introduces multi-modal data into the process and uses strong modal point clouds to generate a weight distribution map to "teach" the feature redistribution of weak modality aerial images and thereby break the bottleneck of unimodal information.

3. N3C-California dataset

Despite the rapid development of earth observation methods for multi-modal data, there are unfortunately only a few LiDAR-imagery multi-modal datasets that are dedicated to remote sensing tasks, of which ISPRS Vaihingen (Rottensteiner et al., 2014) and GRSS DFC 2018 (Xu et al., 2019) are the most commonly used. The ISPRS Vaihingen dataset provides aerial imagery in the 2D semantic labeling contest and LiDAR in the 3D semantic labeling contest. As it was not designed as a unified multi-modal benchmark, the number of categories of aerial imagery does not correspond to the number of LiDAR point clouds. The GRSS DFC 2018 dataset provides only 14 pairs of aerial imagery and

Table 1
Attribute comparison of N3C-California, ISPRS Vaihingen, and GRSS DFC 2018.

	N3C-California	ISPRS Vaihingen	GRSS DFC 2018
Number of tiles	1212	33	14
LiDAR ANPD (pls/m^2)	≥8	4	10
Image dimension (px)	1304 × 1304 (avg)	2493 × 2063 (avg)	11,920 × 12,020
GSD (cm/pixel)	100	9	5
Coverage (km^2)	725.72 (urban & rural)	1.36 (urban)	5.01 (urban)
Classes	4	6/9 (image/LiDAR)	20

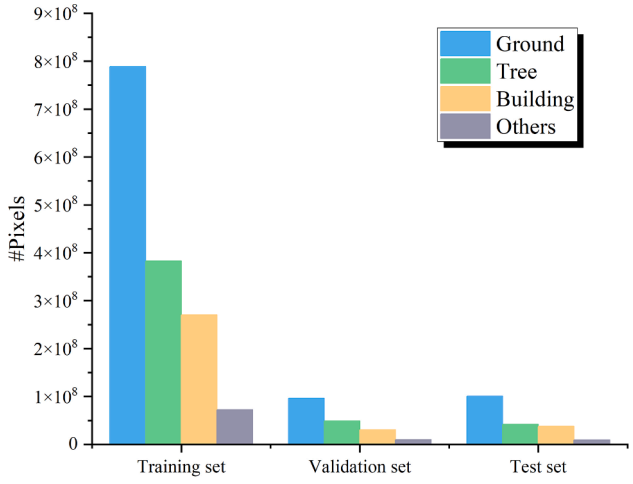

Fig. 4. Number of annotated pixels in N3C-California.

LiDAR data. Both of the above datasets cover a limited range of urban areas, and the landforms are relatively simple.

To fill the gap, we introduce here our **N**ational Agriculture Imagery Program and **3**D Elevation Program **C**ombined dataset in **California** (**N3C-California**) to address the need for a benchmark specifically for multi-modal joint land-cover segmentation tasks. As shown in Fig. 2, N3C-California covers most of residential areas Santa Clara County, California and contains 1,212 pairs of LiDAR, DSM, and aerial image tiles. The DSM is obtained by projecting the elevation of the point cloud. To facilitate the downstream tasks of remote sensing, N3C-California provides four semantic categories (ground, tree, building, and others). Fig. 3 shows several samples from N3C-California.

Table 1 presents an attribute comparison between N3C-California, ISPRS Vaihingen, and GRSS DFC 2018, highlighting the significant advantages of N3C-California in terms of quantity and coverage. Specifically, our dataset contains over 36 times and 86 times more tiles than ISPRS Vaihingen and GRSS DFC 2018, respectively. Regarding LiDAR ANPD, our dataset is over twice larger than ISPRS Vaihingen and comparable to GRSS DFC 2018. However, the point cloud data of GRSS DFC 2018 lacks most of the 20 classes shown in the corresponding images. Although the GSDs of the three datasets are 100, 9, and 5 cm/pixel, respectively, N3C-California is much more extensive than the other two datasets in terms of total area, as it covers not only urban areas but also rural regions. Our N3C-California dataset offers four semantic categories for remote sensing downstream tasks. However, there is a mismatch between the number of categories in the aerial imagery (6 classes) and the LiDAR point clouds (9 classes) in the ISPRS Vaihingen dataset. The point cloud data of GRSS DFC 2018 only contains 5 of the 20 object categories that appear in the imagery. As a result, the category mismatch in both ISPRS Vaihingen and GRSS DFC 2018 datasets makes them less suitable for multi-modal learning tasks.

For the convenience of model training, we cropped the data into 10,800 image patches with 512 × 512 pixels of 20 % overlaps. The training set, validation set, and test set were randomly divided according to the ratio of 8:1:1, as illustrated in Fig. 4. By contrast, the division of ISPRS Vaihingen (11 samples for training, five samples for validation, and 17 samples for testing) and GRSS DFC 2018 (four samples for training, none for validation, and 10 samples for testing) do not exactly correspond to the general setting of deep network training, as they are not specifically multi-modal deep benchmarks.

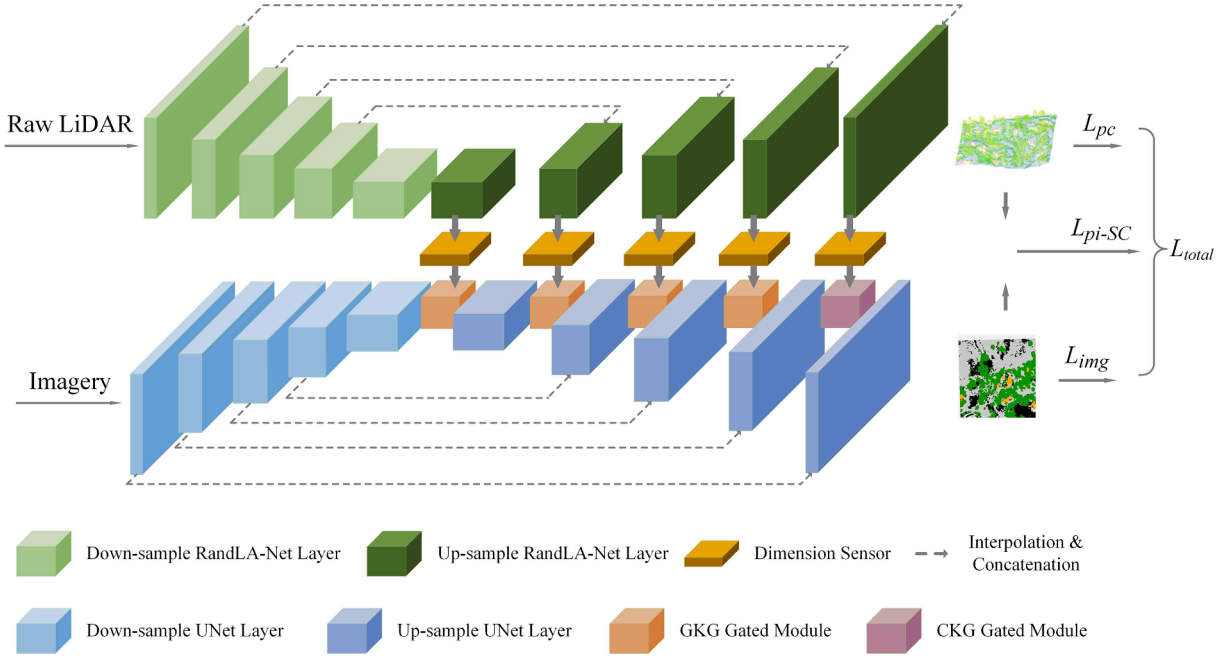

Fig. 5. The workflow of IKD-Net.

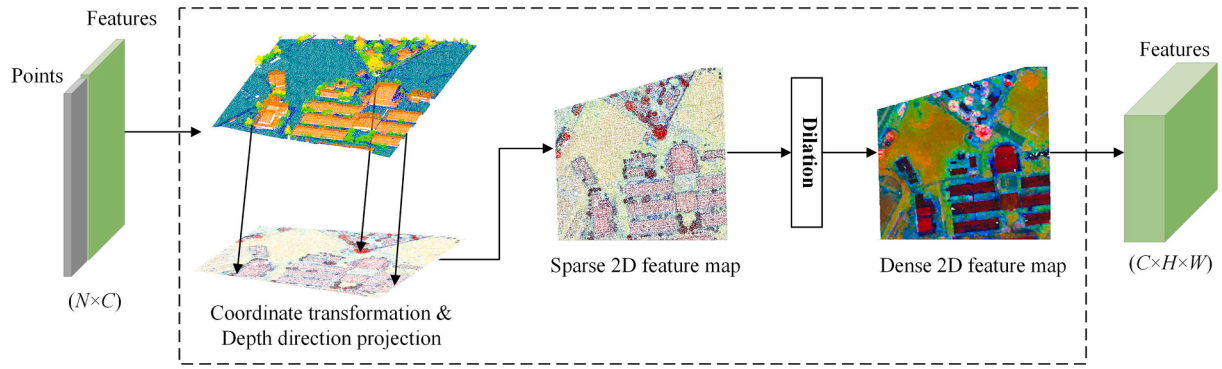

Fig. 6. The structure of the dimension sensor module. The dashed box illustrates an example where the number of channels is 1. In this case, the features of each point are projected onto the corresponding pixel by employing coordinate transformation and depth direction projection. This process produces a sparse 2D feature map. Afterwards, a dense 2D feature map is generated by applying dilation operations.

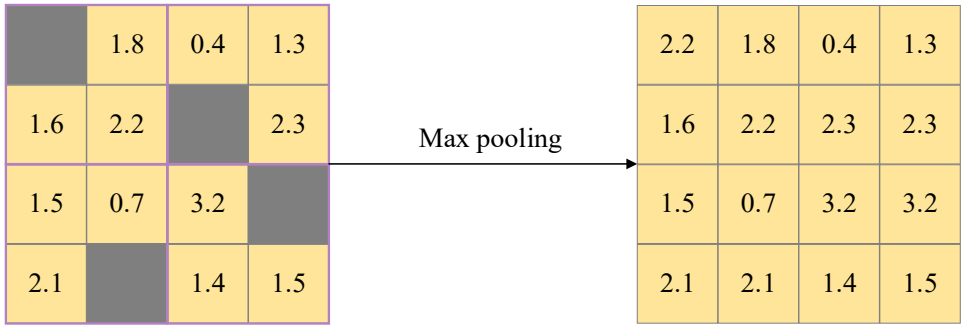

Fig. 7. Dilation operation.

4. Methodology

4.1. Heterogeneous network

As point clouds and aerial images belong to different dimensional spaces, the common strategy of combining the abridged 2D feature maps from LiDAR with images before insertion in the network will sacrifice the rich information of the former and fail to exploit the multi-modal features fully. To this end, we propose our novel **I**mbalance **K**nowledge-**D**riven Multi-modal **Net**work (**IKD-Net**), which can extract heterogeneous features from LiDAR point clouds and aerial image data in parallel. Fig. 5 presents the workflow of the disentangled dual-stream heterogeneous network. The two branches were crafted to be similar encoder-decoder structures, thereby making it possible to obtain similar size feature maps at the same branch depth. The LiDAR stream utilizes RandLA-Net (Hu et al., 2020) as the backbone network to extract 3D features, while the image stream utilizes UNet (Ronneberger et al., 2015) for 2D feature extraction. Both branches can access individual knowledge, such as geometry in 3D space for LiDAR and texture and color information in 2D space for images. Unlike previous approaches that treat features from different modalities as homogeneous and design symmetric feature interaction modules, we exploit the affluent knowledge of LiDAR (strong modality) to drive the refinement of feature maps from aerial images (weaker modality) with our **G**lobal **K**nowledge-**G**uided (**GKG**) gated module and **C**lass **K**nowledge-**G**uided (**CKG**) gated module in the decoder parts. Four GKG gated modules obtain the global feature distribution from the LiDAR features at different resolutions to guide the image features at the same network depth to focus on the region of interest (ROI). The CKG gated module, which is applied at the end of the dual-stream architecture, provides the performance evaluation of each category from a global perspective and achieves the coarse-to-fine segmentation. Before the feature interactions, a front-end projection transformation module called the **D**imension **S**ensor (**DS**) is performed.

4.2. Dual-stream feature extraction

While the dual-stream network design aims to fully preserve the information of unimodal data and effectively leverage multi-modal features to facilitate feature interaction, the dual branches at the same time must obtain feature maps of the same size at the same depth of the network. If the input image size is 512×512, the number of input LiDAR points should be on the order of 2×10^5. Therefore, the 3D branch must have an exceptional ability to handle large-scale point clouds. To meet this need, we selected RandLA-Net as the 3D backbone and designed the corresponding encoder-decoder 2D network. In the encoder part, RandLA-Net first uses a linear transformation layer to expand the feature dimension to 8. The specially-designed local feature aggregation (LFA) and random sampling modules are repeated in the subsequent four downsampling layers. The LFA module consists of two crucial blocks (local spatial encoding and attention pooling).

Local spatial encoding embeds the local geometric pattern for each individual point. Specifically, it first finds the K nearest points around each point with the *k*-nearest neighbors (*k*-NN) algorithm. Then, within the above K points, the information of the coordinates is aggregated. In this paper, K is set to 16. The aggregated calculation formula is as follows:

$$r_i^k = MLP\left(p_i \oplus p_i^k \oplus \left(p_i - p_i^k\right) \oplus \left\|p_i - p_i^k\right\|\right) \quad (1)$$

where r_i^k represents the coordinate encoding value of the *i*-th point and its *k*-th neighbor point. *MLP* denotes the multilayer perceptron. The four terms in the outermost bracket are the coordinates of the center point, the coordinates of its *k*-th neighbor point, the relative coordinate difference, and the relative distance between the two points. $\oplus$ stands for concatenation.

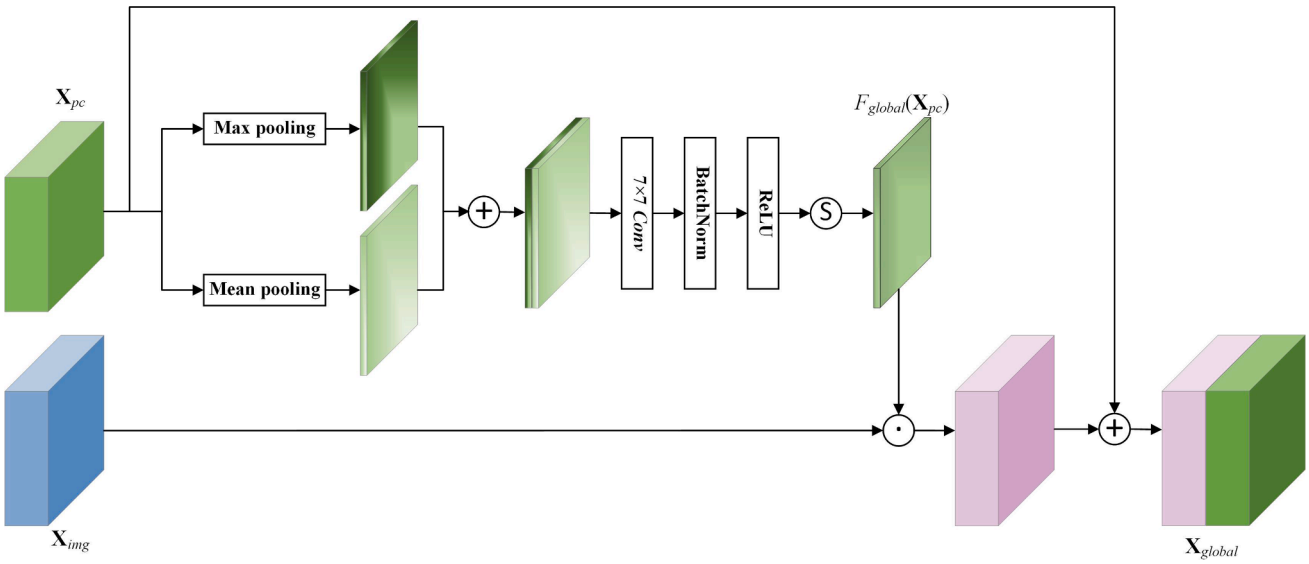

Fig. 8. The structure of the GKG gated module. +, · and S represent concatenation, multiplication along the channel, and sigmoid function, respectively.

Secondarily, the transformed encoding r_i^k is concatenated with the corresponding feature f_i^k:

$$\widetilde{f}_i^k = r_i^k \oplus f_i^k \tag{2}$$

The attention pooling module further refines the feature encodings obtained in the previous step $\widetilde{\mathrm{F}}_i = \{\widetilde{f}_i^1 \cdots \widetilde{f}_i^k \cdots \widetilde{f}_i^K\}$(i.e., the feature maps are multiplied at the pixel level with the weight distribution maps generated on them). Ultimately, the results are accumulated to obtain the aggregated features $\widehat{f}_i$ of the individual points.

We use features from the strong modal point clouds in the decoder to guide the aerial images for feature map refinement. The refined images then use interpolation and convolution blocks to restore the image size, and the LiDAR stream does likewise.

4.3. Dimension sensor

We designed our plug-in DS module for the purpose of aligning the high-dimensional features to the low-dimensional ones, as shown in Fig. 6. The dashed box contains an example of a case where the number of channels is 1. First, we performed coordinate transformation and depth direction projection on each 3D point to produce a feature map in the same metric space as the image. However, there are bound to be pixels not covered by points in feature maps (i.e., hole pixels), which introduce significant inaccuracies when the images are superimposed. Thus, we executed a dilation operation with the trick of max pooling, which fills the hole pixels without increasing the complexity of the network, as depicted in Fig. 7. The principle of using max pooling to fill holes is fundamentally similar to the morphological dilation of binary images. The max pooling operation selects only the maximum value in each rectangular subregion, which represents the most responsive part of the feature map. Using the max pooling operation effectively eliminates noise such as hole pixels, rather than being affected by it, and preserves as much useful information as possible. After passing through the DS, the point cloud features can be converted to the image space with a high degree of fit.

4.4. Global knowledge-guided gated module

Owing to the nature of spotlighting the neighborhoods of convolutional kernels, the information flow in convolutional neural networks is restricted to local areas (Zhao et al., 2018). A spatial attention mechanism can generate an overall probability map to focus on the ROI, thereby extending the global contextual understanding of complex scenes. The probability distribution of the existing methods is derived from the input itself, but due to the limitation of unimodal data information capacity, there is a ceiling to this refinement. To tackle this problem, we proposed a GKG gated module to provide a global probability distribution map utilizing feature maps from strong modality point clouds to guide the further refinement of the image feature map. The structure of our GKG gated module is shown in Fig. 8.

The global attention map from point cloud $F_{global}(\mathbf{X}_{pc})$ is defined as:

$$F_{global}(\mathbf{X}_{pc}) = \sigma\left(g^{7\times7}\left(\left[AvgPool(\mathbf{X}_{pc}), MaxPool(\mathbf{X}_{pc})\right]\right)\right) \tag{3}$$

The avg- and max-pooling operations generate compact feature representations in the spatial dimension. $g^{7\times7}$ is a sequence operation of 7×7 *conv*, batch normalization, and ReLU. σ denotes the sigmoid function. By this sequential operation, the feature map of the strong modal point cloud is compressed into a spatial-wise weight distribution map. We drew inspiration from the spatial attention module of CBAM (Woo et al., 2018) and improved upon its framework by incorporating our knowledge-driven ideas. Specifically, the input and output of the original spatial attention module are based on the same feature map, while our GKG module is designed for multi-modal data. It obtains a spatial weight map from the strong modality and uses it to guide the optimization of the response values in the feature map of weak modality.

Then, $F_{global}(\mathbf{X}_{pc})$ is applied to drive the attention boosting of the weak modalities. The spatial attention-boosted feature map is obtained from the aerial image. Finally, the spatially enhanced image feature maps are concatenated with the point cloud feature maps. In summary, the output of the GKG gated module is:

$$\mathbf{X}_{global} = \left[F_{global}(\mathbf{X}_{pc}) \odot \mathbf{X}_{img}, \mathbf{X}_{pc}\right] \tag{4}$$

where $\odot$ represents multiplication along the channel.

The GKG gated module leverages the higher-level semantic information of the strong modality to provide guidance on the global distribution for the weak modality, refining the latter's understanding of the global context.

4.5. Class knowledge-guided gated module

Besides global information, inter-class variability also plays an influential role in segmentation tasks. Assuming that there is a fixed-length encoding for each category (i.e., the theoretical class center),

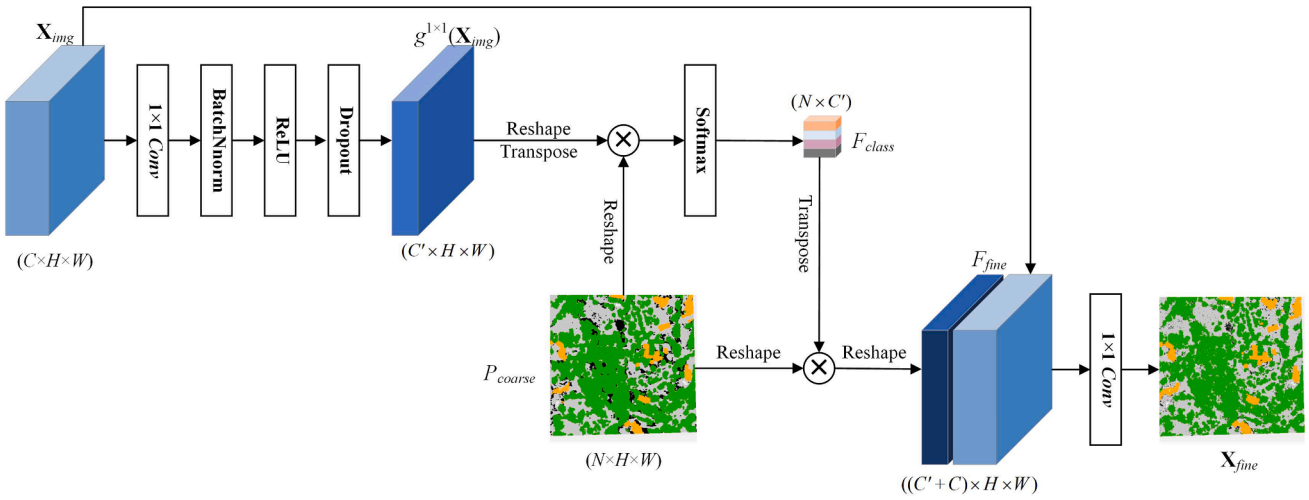

Fig. 9. The structure of the CKG gated module. x stands for the matrix multiplication.

Table 2
Ablation study for different number of GKG gated modules. CAT represents the simple feature concatenation operations.

Backbone	#GKG	#CAT	OA	Mean Acc	Kappa	mIoU
UNet	–	–	86.35	67.17	77.22	59.43
IKD-Net-	–	1	87.07	84.87	79.17	63.45
IKD-Net-	–	2	90.59	87.02	84.46	68.38
IKD-Net-	–	3	91.67	89.20	86.38	70.81
IKD-Net-	–	4	92.18	90.77	87.30	71.97
IKD-Net-	1	–	88.25	85.63	80.92	65.33
IKD-Net-	2	–	91.00	88.06	85.25	69.49
IKD-Net-	3	–	91.70	89.49	86.42	70.87
IKD-Net-	4	–	**92.75**	**91.42**	**88.22**	**72.89**

Table 3
Ablation study for CKG gated module.

Backbone	CKG	Loss	OA	Mean Acc	Kappa	mIoU
IKD-Net- (GKG-4)	–	L_{img}	92.75	**91.42**	88.22	72.89
IKD-Net- (GKG-4)	✓	L_{img}	92.35	90.75	87.54	72.07
IKD-Net- (GKG-4)	✓	$L_{img} + L_{pc}$	**93.19**	91.36	**88.86**	**73.47**

Table 4
Ablation study for different loss functions.

Backbone	Loss	OA	Mean Acc	Kappa	mIoU
IKD-Net	L_{img}	92.35	90.75	0.88	72.07
IKD-Net	$L_{img} + L_{pc}$	93.19	**91.36**	0.89	73.47
IKD-Net	$L_{img} + L_{pc+} L_{bi\text{-}kl}$	90.64	87.99	0.85	69.80
IKD-Net	$L_{img} + L_{pc+} L_{kl}$ (ours)	**93.81**	90.61	**0.90**	**75.50**

Table 5
Ablation study for different dilation functions.

Backbone	Dilation operation	OA	Mean Acc	Kappa	mIoU
IKD-Net	–	89.73	84.73	0.83	66.93
IKD-Net	Median interpolation	90.69	85.17	0.85	68.45
IKD-Net	Average interpolation	90.56	86.25	0.84	68.66
IKD-Net	Max interpolation	90.34	85.33	0.84	67.90
IKD-Net	Median pooling	88.49	84.20	0.81	65.49
IKD-Net	Average pooling	92.61	91.86	0.88	72.82
IKD-Net	2D power-average pooling	93.13	**91.98**	0.89	73.43
IKD-Net	Max pooling (ours)	**93.81**	90.61	**0.90**	**75.50**

Table 6
Ablation study for different attention functions. K-G represents knowledge-driven mechanism.

Backbone	Attention module	OA	Mean Acc	Kappa	mIoU
IKD-Net	–	92.31	90.60	0.87	71.95
IKD-Net	SE layer	92.95	89.66	0.88	72.66
IKD-Net	SE K-G layer	93.12	89.79	0.89	73.04
IKD-Net	SK layer	92.69	88.86	0.89	73.38
IKD-Net	SK K-G layer	93.18	90.53	0.89	73.23
IKD-Net	Self-attention structure	92.80	86.90	0.88	72.14
IKD-Net	Self-attention K-G structure	92.94	88.28	0.89	72.62
IKD-Net	Spatial attention module	93.47	89.58	0.89	73.63
IKD-Net	Global K-G gated module (GKG)	**93.81**	**90.61**	**0.90**	**75.50**

all the pixels in the global scene belonging to that category should make a contribution (Zhang et al., 2019). Further, these encodings can in turn optimize the category attribution of each pixel in the scene. Accordingly, we added a coarse-to-fine structure called the CKG gated module at the output part of IKD-Net, as shown in Fig. 9.

First, the CKG gated module distills the contextual information and generates the category probability map F_{class} with the coarse segmentation result P_{coarse} from the point clouds and the feature map $\mathbf{X}_{img}$ of the corresponding image. Each row of F_{class} provides the performance evaluation of a category from a global perspective by converging the feature vectors of all the pixels belonging to the category.

In detail, we put $\mathbf{X}_{img}$ through a sequence operation $g^{1\times1}$ of 1 × 1 *conv*, batch normalization, and ReLU to reduce the channel dimension from C to C'. Then, a reshape operation is applied to $g^{1\times1}(\mathbf{X}_{img})$ and P_{coarse} respectively:

$$g^{1\times1}\left(\mathbf{X}_{img}\right) \in \mathbb{R}^{C'\times H\times W} \rightarrow g^{1\times1}\left(\mathbf{X}_{img}\right) \in \mathbb{R}^{C'\times HW} \tag{5}$$

$$P_{coarse} \in \mathbb{R}^{N\times H\times W} \rightarrow P_{coarse} \in \mathbb{R}^{N\times HW} \tag{6}$$

where N is the category number.

The category probability map $F_{class} \in \mathbb{R}^{N\times C'}$ is calculated as:

$$F_{class} = \mathrm{softmax}\left(P_{coarse}\left(g^{1\times1}\left(\mathbf{X}_{img}\right)\right)^{\mathrm{T}}\right) \tag{7}$$

Second, taking the coarse class distribution of each pixel as the mediator, the attentional class feature vector of each pixel is obtained by multiplying the coarse segmentation result P_{coarse} and the category probability map F_{class}. The attentional class feature map is:

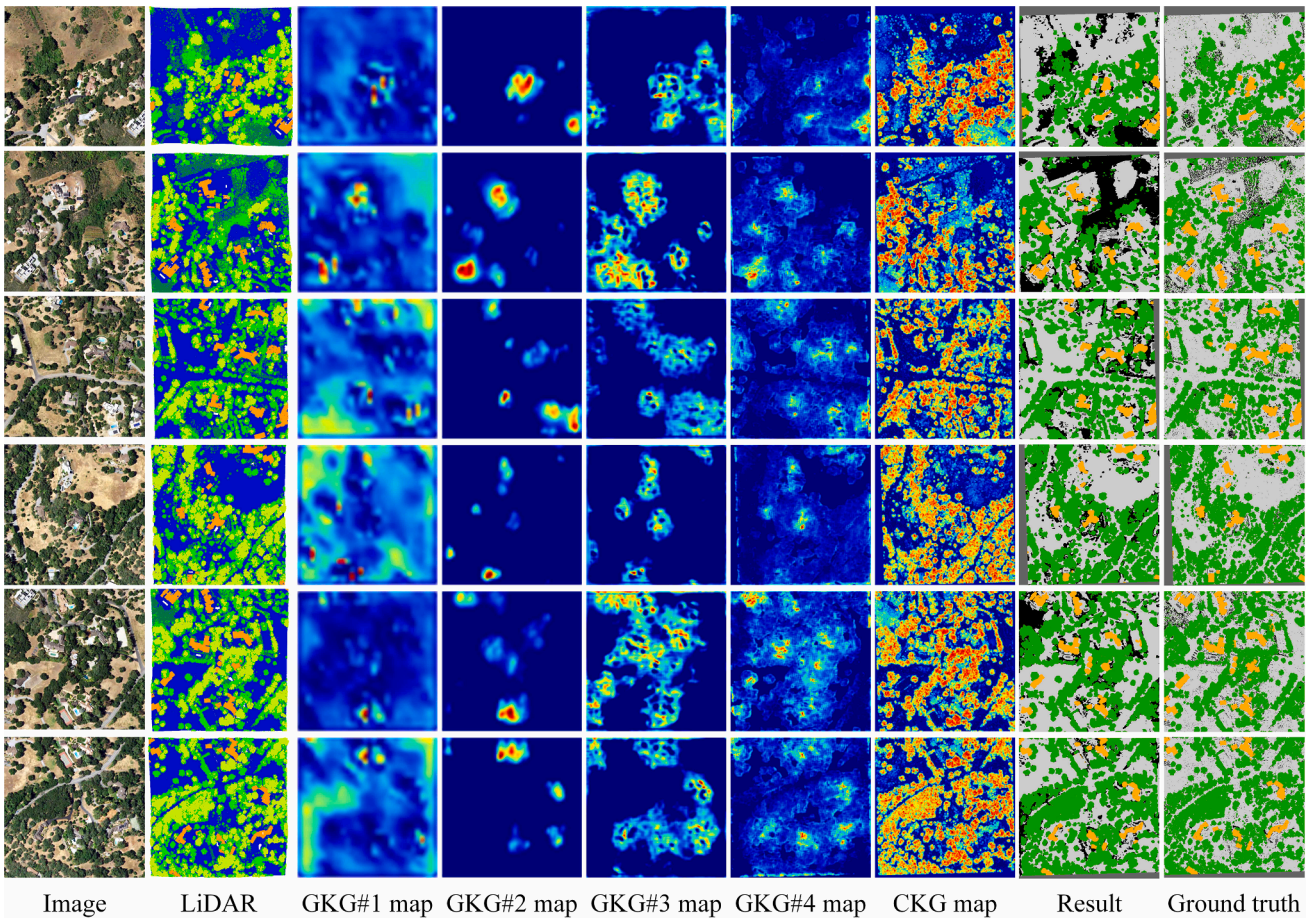

Fig. 10. Visualization of the feature maps after applying each module.

Table 7
Quantitative comparison of IKD-Net, the baseline methods, the multi-modal benchmark method, the visual classical semantic segmentation methods, and the SOTA multi-modal segmentation networks on N3C-California dataset.

Method	Input	OA	Mean Acc	Kappa	IoU				
					Others	Ground	Tree	Building	Mean
UNet (baseline)	RGB	86.35	67.17	0.77	2.46	80.92	73.33	81.01	59.43
RandLA-Net (baseline)	LiDAR	87.49	85.78	0.82	40.06	88.32	**84.19**	69.98	70.64
Hybri-UNet	RGB + DSM	89.00	71.76	0.82	12.98	85.86	77.03	87.13	65.75
UperNet	RGB + DSM	85.74	78.90	0.77	10.58	90.34	51.20	86.80	59.73
HRNet	RGB + DSM	91.90	80.39	0.86	20.94	91.01	80.92	83.89	69.19
vFuseNet	RGB + DSM	86.11	75.47	0.75	**49.73**	81.99	57.58	77.43	66.68
MultifilterCNN	RGB + DSM + intensity +number returns + DoG	88.97	76.44	0.82	25.77	84.33	77.63	82.63	67.59
MFNet	RGB + DSM +Slope angle + DoG	91.00	74.85	0.86	14.29	87.36	82.09	89.87	68.40
S^2ENet	RGB + DSM + intensity +number returns	92.63	77.89	0.87	27.07	89.68	76.20	91.19	71.03
MDL_RS	RGB + DSM + intensity +number returns	90.99	72.55	0.85	16.46	87.08	74.82	86.57	66.23
JSH-Net	RGB + DSM + intensity +number returns	91.59	75.46	0.86	23.56	88.00	75.18	88.23	68.74
EndNet	RGB + DSM + intensity +number returns	88.29	66.58	0.80	3.11	8.34	72.23	78.87	59.40
CMGFNet	RGB + DSM + intensity +number returns	92.90	75.95	0.88	22.69	**93.65**	80.53	92.90	72.44
IKD-Net (ours)	RGB + LiDAR	**93.81**	**90.61**	**0.90**	30.68	93.11	82.35	**95.87**	**75.50**

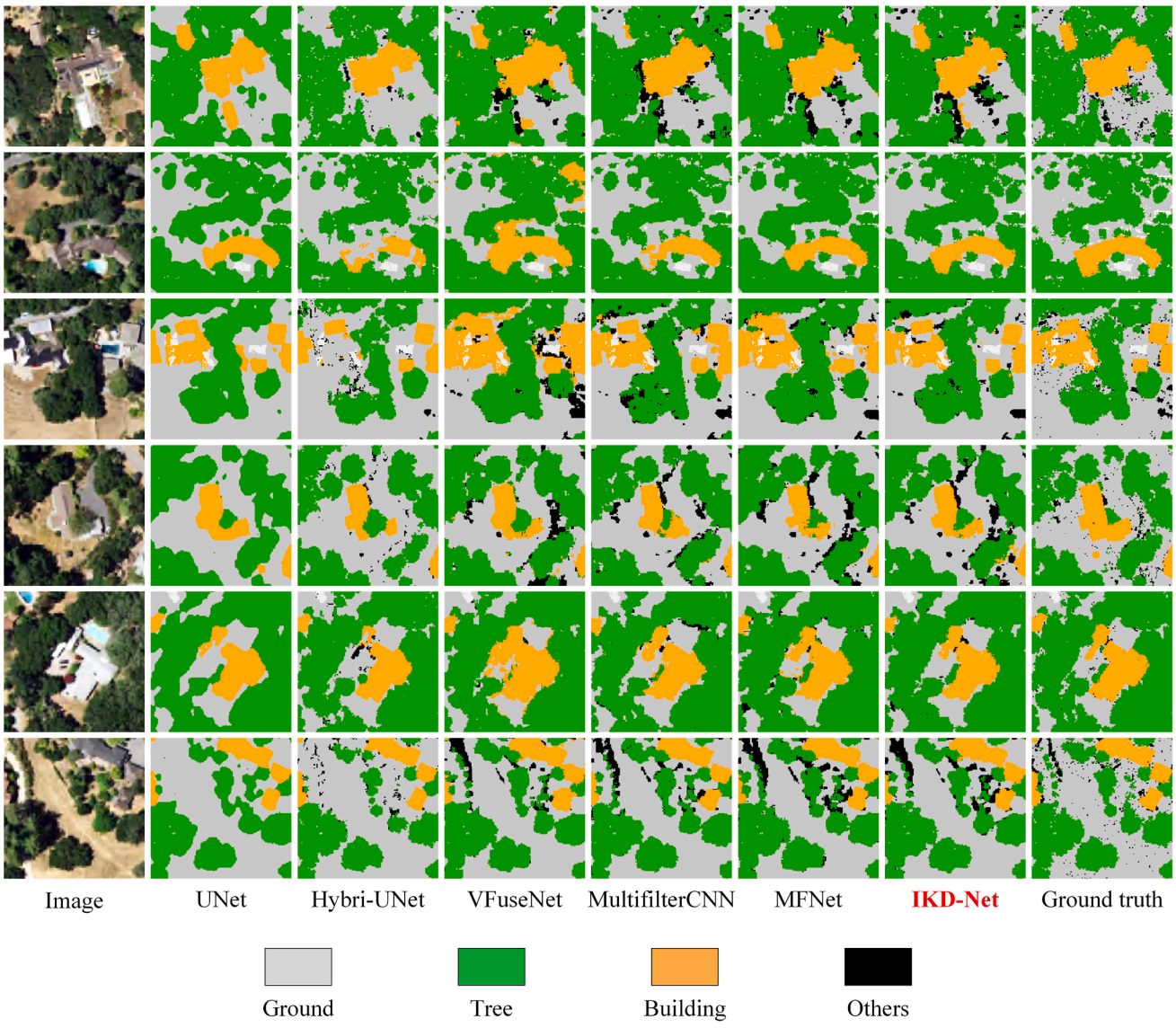

Fig. 11. Qualitative comparison of the baseline, the multi-modal benchmarks, the SOTA multi-modal segmentation networks, and IKD-Net (ours) on N3C-California.

$$F_{fine} = P_{coarse}^{\mathrm{T}} F_{class} \tag{8}$$

For the subsequent operation, F_{fine} is reshaped:

$$F_{fine} \in \mathbb{R}^{C \times HW} \rightarrow F_{fine} \in \mathbb{R}^{C \times H \times W} \tag{9}$$

Finally, the concatenation of F_{fine} and $\mathbf{X}_{img}$ is put into a 1×1 $conv\, f^{1\times1}$ to obtain the class boosting feature map:

$$\mathbf{X}_{fine} = f^{1\times1}\left[\mathbf{X}_{img}, F_{fine}\left(P_{coarse}, \mathbf{X}_{img}\right)\right] \tag{10}$$

The CKG gated module guides each pixel in the high-level feature maps of an image (weak modality) to adaptively approach the theoretical class centers according to the segmentation results of the point clouds (strong modality).

4.6. Loss function

In order to maintain the balance of the parameter flow in each branch during network optimization, we proposed a joint loss function, which consists of three types of supervision: two single-task loss functions and a pixel-wise similarity loss.

Joint Loss. Serving as the whole objective function, the joint loss enables the network to be trained in an end-to-end manner, which is summarized as:

$$L_{total} = L_{CE}\left(P\left(x_{img}\right)\right) + L_{CE}\left(P\left(x_{pc}\right)\right) + L_{pi-SC}\left(P\left(x_{img}\right), P\left(x_{pc}\right)\right) \tag{11}$$

where x_{img} and x_{pc} denotes the input image and LiDAR, respectively; $P(*)$ represents the final probability distribution map; L_{CE} is the segmentation cross-entropy loss; and $L_{pi\text{-}SC}$ is the pixel-wise similarity loss.

Single-task Loss. For each branch of semantic segmentation, given predict $P(x)$ and ground truth y, we use cross-entropy loss to optimize, which is as follows:

$$L_{CE}(P(x)) = -\sum_{i=1}^{N}\sum_{c=1}^{M} y_{ic}\log(P(x_{ic})) \tag{12}$$

where N represents the number of pixels or points and M denotes the number of categories.

Pixel-wise Similarity Constraint. To make the convergence spaces of the images and point clouds of the same scene as close as possible, we straightforwardly add similarity constraints. Further, inspired by knowledge distillation (Hinton et al., 2015; Liu et al., 2019), we consider the class distribution of the point clouds (strong modality) $P(x_{pc})$ as soft targets to guide the images (weak modality) $P(x_{img})$, thus improving the accuracy of 2D semantic segmentation.

Table 8
Quantitative comparison of IKD-Net, the baseline methods, the benchmark competitors, and the recent SOTA multi-modal segmentation networks on ISPRS Vaihingen dataset.

Method	OA	F1 Score					
		Imp surf	Building	Low veg	Tree	Car	Mean
UNet (baseline)	84.5	86.7	90.1	76.9	83.9	62.9	80.1
RandLA-Net (baseline)	85.5	87.6	92.4	78.5	83.7	00.2	68.5
SVL_3	84.8	86.6	91.0	77.0	85.0	55.6	79.0
HUST	85.9	86.9	92.0	78.3	86.9	29.0	74.6
RIT	86.3	88.1	93.0	80.5	87.2	41.9	78.1
UOA	87.6	89.8	92.1	80.4	88.2	82.0	86.5
ADL_3	88.0	89.5	93.2	82.3	88.2	63.3	83.3
DST_1	88.7	90.3	93.5	82.5	88.8	73.9	85.8
DLR_8	89.2	90.4	93.6	83.9	89.7	76.9	86.9
UFMG_4	89.4	91.1	94.5	82.9	88.8	81.3	87.7
ONE_7	89.8	91.0	94.5	84.4	89.9	77.8	87.5
CASIA2	91.1	93.2	96.0	84.7	89.9	86.7	90.1
CCANet	91.1	93.3	94.3	82.0	88.6	86.6	89.0
BANet	90.5	92.2	95.2	83.8	89.9	86.8	89.6
HCANet	90.3	92.5	95.0	84.2	89.4	84.0	89.0
HECR-Net	91.5	93.6	95.5	85.8	90.4	89.1	90.9
MAResU-Net	90.2	92.2	94.8	79.1	90.0	85.9	88.5
ESANet	90.6	91.4	95.7	77.2	90.5	85.5	88.2
BoTNet	90.2	92.2	94.5	84.0	89.6	82.9	88.6
MANet	91.0	93.0	95.5	84.6	90.0	89.0	90.4
UNetFormer	91.0	92.7	95.3	84.9	90.6	88.5	90.4
JSH-Net	91.4	93.3	96.3	85.0	90.0	90.4	91.0
CMFNet	91.4	92.4	**97.2**	80.4	90.8	85.5	89.5
HMANet	91.4	93.5	95.9	85.4	90.4	89.6	91.0
MFNet	91.7	92.2	96.3	84.7	89.1	89.7	90.4
SPANet	91.8	93.5	96.2	86.8	90.9	90.6	91.6
LoG-CAN	91.9	93.7	96.6	85.9	90.9	90.2	91.4
IKD-Net (ours)	**92.1**	**96.1**	90.5	**87.2**	**92.0**	**92.5**	**91.6**

We use Kullback-Leibler divergence to implement this pixel-wise similarity constraint, which is formulated as follows:

$$L_{pi-SC}\left(P\left(x_{img}\right),P\left(x_{pc}\right)\right)=\frac{1}{W\times H}\sum_{i=1}^{W\times H}\mathrm{KL}\left(P_i\left(x_{img}\right)\|P_i\left(x_{pc}\right)\right) \tag{13}$$

where

$$\mathrm{KL}\left(P_i\left(x_{img}\right)\|P_i\left(x_{pc}\right)\right)=P_i\left(x_{img}\right)\log\left(\frac{P_i\left(x_{img}\right)}{P_i\left(x_{pc}\right)}\right) \tag{14}$$

5. Experiments

In this section, we explain the experimental setups and evaluation metrics. We conducted sufficient ablation studies to verify the rationality of our sophisticated modules and the overall structure of IKD-Net. We visualize the outcomes of each module here. Then, we compare our IKD-Net results with those of the state-of-the-art (SOTA) methods on the N3C-California, ISPRS Vaihingen, and GRSS DFC 2018 datasets. Finally, we demonstrate how our method achieved outstanding performance on all three datasets. The best values for each specific metric are highlighted in bold in the following tables.

5.1. Experimental setting

5.1.1. Implementation details

All the experiments were conduct on a Linux PC equipped with an NVIDIA GeForce RTX 3090 24G GPU. The code of our own architecture and the code we reproduced are based on the PyTorch deep learning framework. During training, the batch size was set to 2 for the experiments on all the datasets. Each epoch had 1,000 iterations, and the maximum number of epochs was always 50. On the N3C-California dataset, the Adam algorithm with a 0.001 learning rate was employed for optimization. For the other datasets, the SGD method with a 0.01 learning rate, 0.0001 wt decay, and 0.9 momentum was chosen. The input images were 512 × 512 pixel in size. To ensure that point clouds cover as many pixels as possible and balance memory consumption, we randomly selected a total of 131,072 points from the LiDAR patch covering the same area and fed them into the networks simultaneously. This number of points represents half of the pixel-number of a 512 × 512 image.

5.1.2. Evaluation metrics

The results were evaluated by overall accuracy (OA), mean accuracy (Mean Acc), Cohen's Kappa (Kappa), mean intersection over union (mIoU) and F1 Score.

OA is defined as the ratio of the number of correctly classified pixels p_{correct} to the total number of pixels p_{all}.

$$\mathrm{OA}=\frac{p_{\mathrm{correct}}}{p_{\mathrm{all}}} \tag{15}$$

OA is simple to calculate but is easily dominated by a large number of samples in the case of unbalanced samples, which can be addressed by three other metrics.

We assume TP^k, FP^k, TN^k, FN^k represent the true positive number, the false positive number, the true negative number, and the false negative number for k-th class, respectively, in the confusion matrix. Accuracy and IoU for k-th class (Acc^k and IoU^k) are defined as:

$$\mathrm{Acc}_k=\frac{\mathrm{TP}^k}{\mathrm{TP}^k+\mathrm{FP}^k} \tag{16}$$

$$\mathrm{IoU}^k=\frac{\mathrm{TP}^k}{\mathrm{TP}^k+\mathrm{FP}^k+\mathrm{FN}^k} \tag{17}$$

For total K categories, Mean Acc and mIoU are defined as:

$$\text{Mean Acc}=\frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{TP}^k}{\mathrm{TP}^k+\mathrm{FP}^k} \tag{18}$$

$$\mathrm{mIoU}=\frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{TP}^k}{\mathrm{TP}^k+\mathrm{FP}^k+\mathrm{FN}^k} \tag{19}$$

The formula for Kappa is:

$$\mathrm{Kappa}=\frac{\mathrm{OA}-\mathrm{p_e}}{1-\mathrm{p_e}} \tag{20}$$

where

$$\mathrm{p_e}=\frac{\sum_{k=1}^{K}(\mathrm{TP}^k+\mathrm{FP}^k)(\mathrm{TP}^k+\mathrm{FN}^k)}{p_{\mathrm{all}}^2} \tag{21}$$

The F1 score is the harmonic mean of Acc^k and the recall rates. The F1 score for for k-th class (F1 score ^{k}) is calculated as follows:

$$\text{F1 - Score}^k=\frac{2\times\mathrm{Acc}^k\times\mathrm{recall}}{\frac{1}{\mathrm{Acc}^k}+\frac{1}{\mathrm{recall}}} \tag{22}$$

where

$$\mathrm{recall}=\frac{\mathrm{TP}^k}{\mathrm{TP}^k+\mathrm{FN}^k} \tag{23}$$

5.2. Results on N3C-California dataset

In the ablation studies section, we marked the dual-stream backbone of our IKD-Net as IKD-Net-, which indicates that we discarded the GKG and CKG gated modules and it now was equipped only with one single-task loss function (the image segmentation loss function).

5.2.1. Ablation study for GKG gated module

The four structured GKG gated modules in the two-branch structure

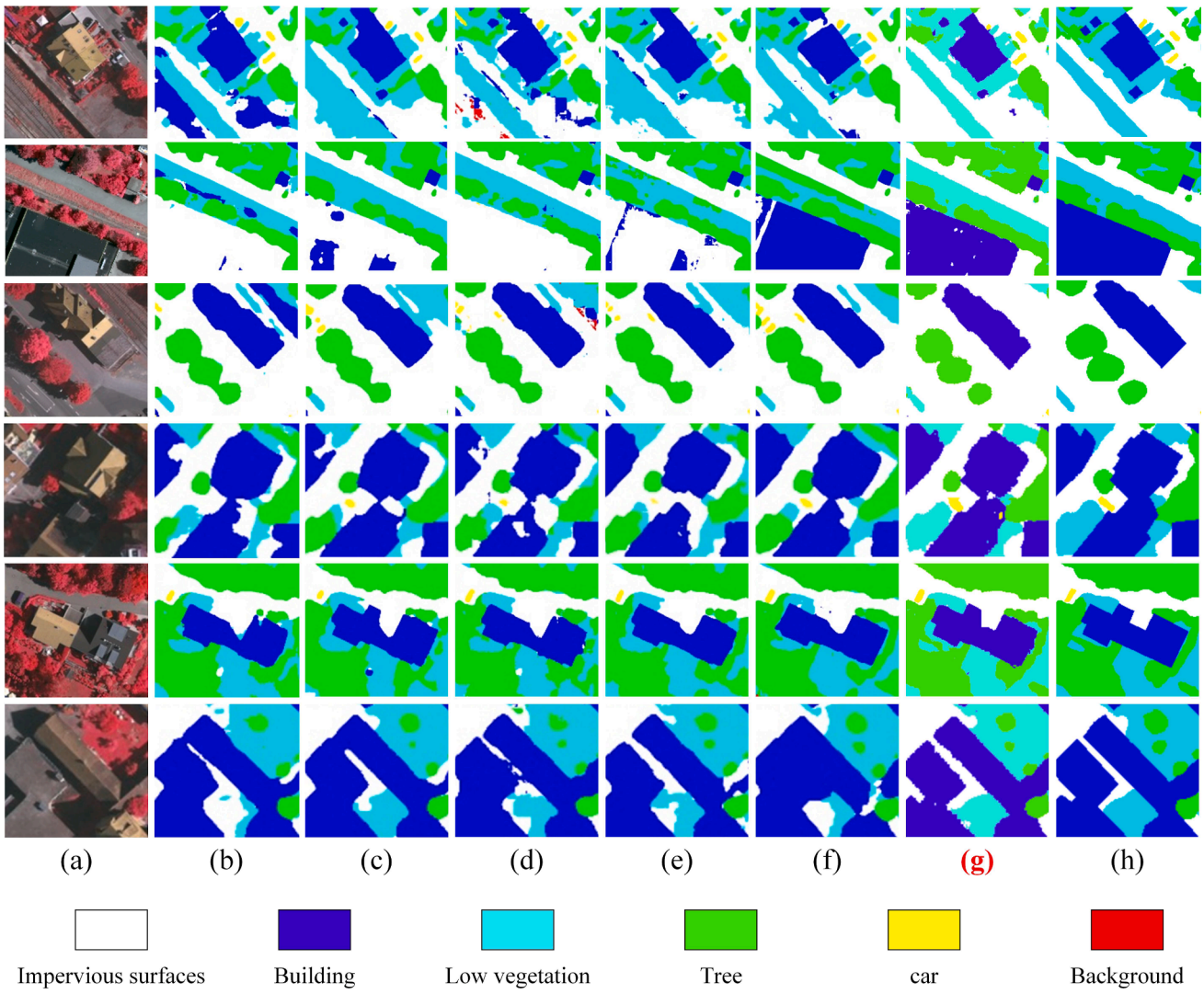

Fig. 12. Qualitative comparison of IKD-Net with the benchmark competitors on ISPRS Vaihingen dataset: (a) Imagery, (b) DST_1, (c) DLR_8, (d) UFMG_4, (e) ONE_7, (f) CASIA2, (g) IKD-Net (ours), and (h) Ground truth.

provided multi-resolution weight redistribution maps for top-to-down weak modality feature map refinement under strong modality guidance. To demonstrate the effect of the structured GKG modules, we gradually increased the number of GKG modules on the backbone IKD-Net-. The experimental results are shown in Table 2. In rows 2–5, we replaced the GKG gated modules with the simple feature concatenation operations at the same positions.

At least four primary conclusions can be drawn from Table 2. First, as implied in rows 1–5, the supplemental feature maps from the LiDAR stream greatly improved the image segmentation, and the more information that was provided from the former the greater the accuracy improvement. Simply overlaying four multi-resolution feature maps from the LiDAR stream (row 5), our IKD-Net- backbone outperformed UNet by nearly 0.06 in OA, over 0.23 in Mean Acc, over 0.1 in Kappa, and over 0.12 in mIoU. Second, upgrading the simple concatenation operations with GKG gated modules (row 2–5 vs. row 6–9) further enhanced the ability to jointly exploit the multi-modal features by driving the refinement of the feature distribution of the weak modality with the affluent knowledge from the strong modality. Third, as we gradually increased the number of GKGs, the accuracy steadily improved, indicating that the effects of our GKGs were cumulative. Eventually, IKD-Net- equipped with four structured stacked GKG gated modules (row 9) surpassed the baseline by over 0.06 in OA, over 0.24 in Mean Acc, 0.11 in Kappa, and over 0.13 in mIoU.

5.2.2. Ablation study for CKG gated module

The CKG gated module simultaneously distilled the contextual information of the strong and weak modalities to obtain the category-wise feature map, the so-called class centers. The class centers were then exploited to guide the refinement of the feature maps of the weak modal images from coarse to fine. We observed the effect of adding the CKG gated module based on the optimal structure in the last section (marked as IKD-Net- (GKG-4)), as indicated in Table 3.

Row 2 in Table 3 contains the results of IKD-Net- (GKG-4) with the addition of the CKG gated module. The accuracy decreased slightly with respect to simple IKD-Net- (GKG-4) (row 1) on all the metrics. This decrease may have been due to CKG depending largely on the coarse segmentation process, which is not fully optimized by the single image segmentation loss function because the backward-propagation route is too circuitous for the LiDAR stream. Therefore, we incorporated an additional cross-entropy loss function to the LiDAR stream, as shown in row 3. By adding the CKG gated module to IKD-Net- (GKG-4), it eventually exceeded its counterpart without CKG on three metrics.

5.2.3. Ablation study for loss function

The joint loss function L_{total} takes into account both the independent

Table 9
Quantitative comparison of IKD-Net, the baseline methods, the top ranked teams, and the recent SOTA multi-modal segmentation networks on GRSS DFC 2018 dataset.

class	UNet (baseline)	RandLA-Net (baseline)	XudongKang	Gaussian	IPIU	challenger	AGTDA	dlrpba	CEGCN	NLCaps -Net	EB-CNN	CAG	CAGU	IKD-Net (ours)
1	88.36	72.27	–	–	–	–	–	–	61.30	28.24	51.50	–	–	**80.37**
2	74.82	86.30	–	–	–	–	–	–	61.53	72.15	74.62	–	–	**95.61**
3	00.74	–	–	–	–	–	–	–	63.71	6.78	21.87	–	–	**99.61**
4	93.36	–	–	–	–	–	–	–	61.38	5.92	81.22	–	–	**97.21**
5	60.89	–	–	–	–	–	–	–	61.38	12.04	24.05	–	–	**94.02**
6	11.06	6.17	–	–	–	–	–	–	62.48	2.26	14.75	–	–	**81.77**
7	10.48	80.23	–	–	–	–	–	–	59.02	0.45	72.01	–	–	**99.88**
8	36.51	–	–	–	–	–	–	–	22.56	42.60	50.69	–	–	**88.73**
9	84.19	–	–	–	–	–	–	–	33.08	86.22	**90.64**	–	–	82.07
10	85.24	–	–	–	–	–	–	–	14.39	32.67	41.99	–	–	**91.84**
11	71.78	–	–	–	–	–	–	–	9.32	20.68	45.93	–	–	**85.05**
12	34.66	–	–	–	–	–	–	–	19.99	0.86	3.79	–	–	**72.74**
13	1.56	–	–	–	–	–	–	–	11.73	30.91	**55.71**	–	–	28.54
14	0.02	–	–	–	–	–	–	–	59.60	39.92	**90.64**	–	–	75.22
15	**95.52**	–	–	–	–	–	–	–	60.30	18.07	31.53	–	–	95.12
16	00.04	–	–	–	–	–	–	–	**58.90**	27.01	51.44	–	–	17.50
17	0.0	**77.27**	–	–	–	–	–	–	58.56	1.29	1.34	–	–	48.81
18	15.74	–	–	–	–	–	–	–	58.60	11.51	33.92	–	–	**97.94**
19	18.95	–	–	–	–	–	–	–	60.38	9.86	87.16	–	–	**92.67**
20	**63.81**	–	–	–	–	–	–	–	60.88	14.04	40.07	–	–	39.76
AA	43.10	64.39	71.26	71.66	74.40	75.99	76.15	76.32	59.64	25.81	47.75	67.39	77.04	**78.22**
OA	45.93	75.22	76.45	**80.78**	79.23	77.90	79.79	80.74	60.80	32.75	63.57	70.28	80.72	78.28
Kappa	0.43	0.60	0.75	0.80	0.78	0.77	0.79	0.80	0.59	0.26	0.55	0.68	**0.81**	0.77

Table 10
Category numbers and the corresponding category names of GRSS DFC 2018 dataset.

#	Class	#	Class
1	Healthy grass	11	Sidewalks
2	Stressed grass	12	Crosswalks
3	Artificial turf	13	Major thoroughfares
4	Evergreen trees	14	Highways
5	Deciduous trees	15	Railways
6	Bare earth	16	Paved parking lots
7	Water	17	Unpaved parking lots
8	Residential buildings	18	Cars
9	Non-residential buildings	19	Trains
10	Roads	20	Stadium seats

optimizations of each branch and the synergy between them. Table 4 indicates the superposition effect of the three terms in L_{total}, which are the single image segmentation loss function, single point cloud segmentation loss function, and pixel-wise similarity constraint. Additionally, a comparison between the bi-directional K-L loss and the K-L loss using the point clouds coarse segmentation results as soft labels is presented.

As explained in the last section, the joint use of two single-task loss functions facilitated the performance of the CKC module. Furthermore, the addition of a pixel-wise similarity constraint further promoted the improvement of image segmentation accuracy, thereby surpassing its counterpart with a single image segmentation loss of nearly 1.5 % in OA, over 2 % in Kappa, and over 3.4 % in mIoU. The use of bi-directional K-L loss significantly reduced the accuracy of the segmentation results. This could be attributed to the fact that using the weak modality image's coarse segmentation result as a soft label to guide the redistribution of the strong modality point clouds' feature map does not refine the latter. This confirmed the validity of our knowledge-driven mechanism.

5.2.4. Ablation study for dilation functions

In the dilation step of the DS module, we chose to use the max pooling operation to address the potential occurrence of hole pixels when projecting 3D points onto a 2D space with almost no increase in the complexity or computational load of the network. To further investigate the effectiveness of different interpolation and pooling methods, we conducted a comparison in Table 5, including three interpolation methods: median interpolation, average interpolation, and max interpolation, as well as four pooling methods: median pooling, average pooling, 2D power-average pooling, and max pooling.

Table 5 illustrates that utilizing dilation methods other than median pooling yields higher final performance than without dilation (row 1), indicating that the dilation operation plays a crucial role in optimizing the feature map. Compared to interpolation methods (row 2–4), which employ the same filling value within the same channel, pooling methods (row 5–8) that employ filling values for each local area can better account for the differences between different sub-regions. This leads to performance improvements of approximately 2 %, 5 %, and more than 4 % on 0A, Mean Acc, and mIoU, respectively. While the median pooling method (row 5) substantially weakens the impact of extreme response values, the average-based pooling method (row 6–7) is inevitably influenced by missing pixels in the local areas. As a result, the max pooling method (row 8) outperforms the others by eliminating noise such as missing pixels.

5.2.5. Ablation study for attention modules

The GKG module is built upon the spatial attention module of CBAM (Woo et al., 2018) and incorporates our proposed knowledge-driven mechanism. We also applied the knowledge-driven mechanism to other common attention modules, including SE (Hu et al., 2018), SK (Li et al., 2019), and Self-attention (Vaswani et al., 2017). Ablation experiments are conducted, and the results are presented in Table 6.

Table 6 demonstrates that all methods utilizing attention modules (row 2–9) exhibit improvements in OA, Kappa, and mIoU. Moreover, for each attention module (row 2–3, row 4–5, row 6–7, and row 8–9), the introduction of our knowledge-driven mechanism led to further improvements in accuracy across all indicators. This indicates that under the guidance of the knowledge-driven mechanism, the strong modality optimizes the redistribution of the feature map of weak modality. Notably, the GKG module, which incorporates the knowledge-driven mechanism into the spatial attention module, achieves the highest accuracy across all four indicators.

5.2.6. Visualization results of GKG and CKG gated modules

In order to qualitatively analyze the effect of each module on the features, we visualized the feature maps after applying each module, as shown in Fig. 10. We took the mean value in the channel direction for the high-dimensional feature maps to normalize and stretch the

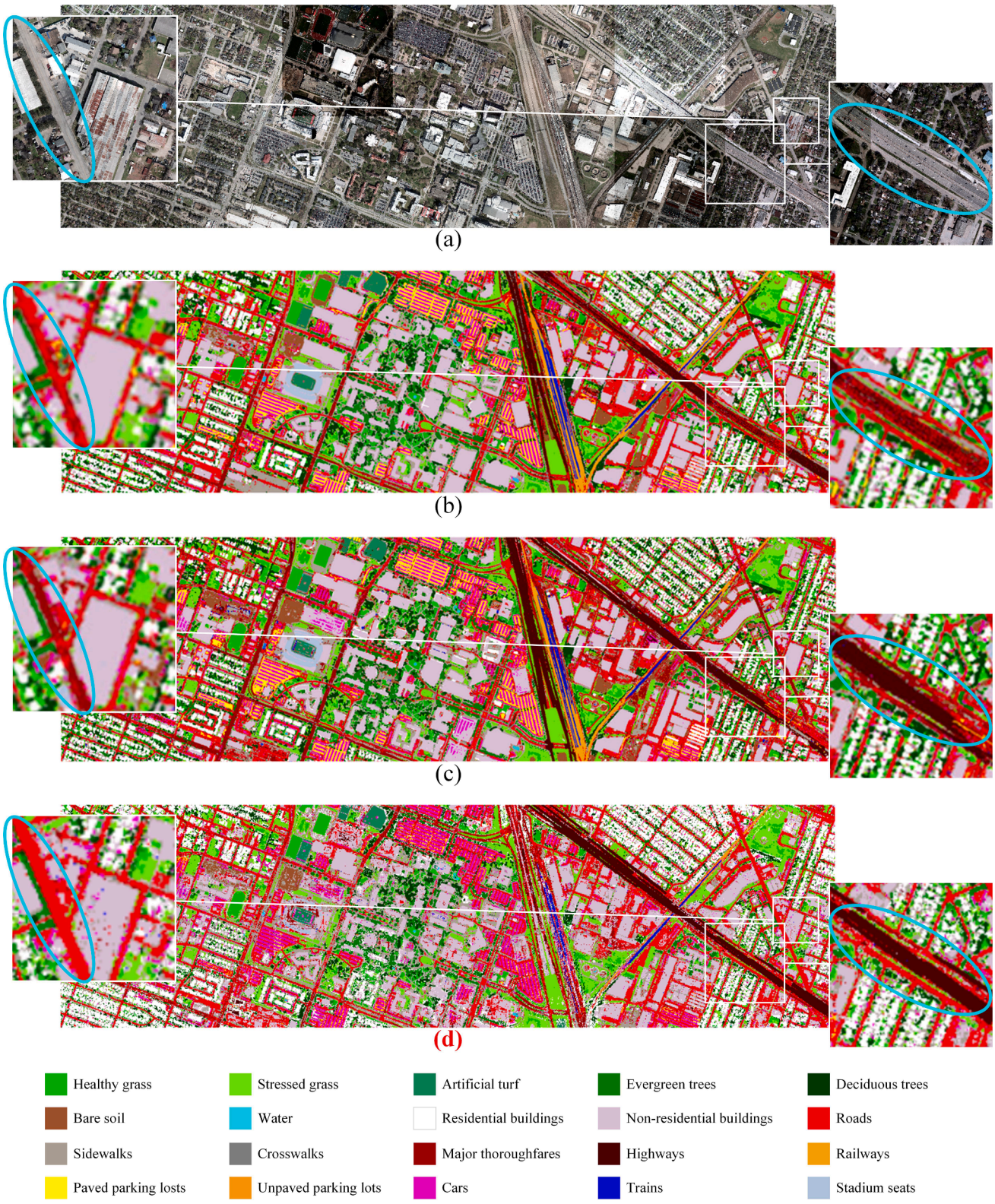

Fig. 13. Classification map over the entire scene of GRSS DFC 2018 dataset: (a) Imagery, (b) AGTDA, (c) dlrpba, and (d) IKD-Net (ours).

obtained single-channel 2D feature map to 0–255. Finally, we performed a pseudo-color transformation to obtain the feature maps that facilitated visual interpretation. The ground truth was generated by projecting the "classification" attributes of the LiDAR data onto a 2D space. However, due to the inherent distortion that occurs when projecting LiDAR patches from 3D to 2D, some areas along the edges of the final output may lack coverage. To address this, we filled the pixels without labels with dark gray in the "Result" and "Ground truth" images.

As is evident from each row of Fig. 10, the segmentation results were progressively detailed after each module was applied. We observed that as the number of GKG modules increased, their effect of fusing global information became more pronounced, mainly in the Building category.

Table 11
Multi-class semantic segmentation on N3C-California dataset. IoU is calculated for each category.

Backbone	Strong modality	OA	Mean Acc	Kappa	IoU				
					Others	Ground	Tree	Building	Mean
IKD-Net	–	93.22	**91.06**	0.89	29.02	91.17	79.46	95.55	73.80
IKD-Net	Image	90.90	85.99	0.85	25.58	88.39	67.91	94.54	69.10
IKD-Net (ours)	LiDAR	**93.81**	90.61	**0.90**	**30.68**	**93.11**	**82.35**	**95.87**	**75.50**

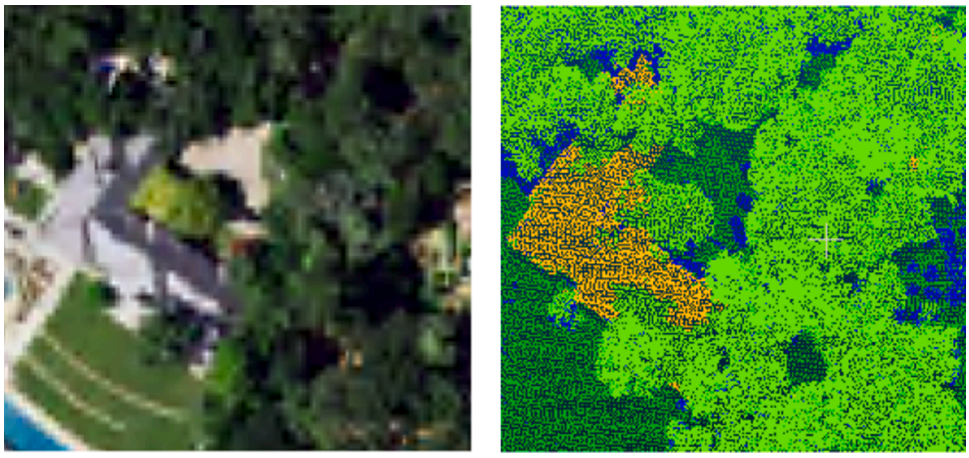

Fig. 14. The area where the building is obstructed by trees in imagery and LiDAR.

Table 12
Binary semantic segmentation of individual class on N3C-California dataset.

Class	Backbone	Strong modality	OA	Mean Acc	Kappa	IoU		
						Background	Foreground	Mean
Ground	IKD-Net	–	96.12	96.05	0.92	91.58	93.28	92.43
	IKD-Net	Image	94.65	94.62	0.89	88.64	90.80	89.72
	IKD-Net (ours)	LiDAR	**96.34**	**96.30**	**0.93**	**92.05**	**93.63**	**92.84**
Tree	IKD-Net	–	96.44	95.32	0.88	95.70	82.77	89.24
	IKD-Net	Image	94.25	89.23	0.80	93.25	72.14	82.69
	IKD-Net (ours)	LiDAR	**96.83**	**95.56**	**0.90**	**96.17**	**84.37**	**90.27**
Building	IKD-Net	–	96.93	95.79	0.90	96.29	84.85	90.57
	IKD-Net	Image	98.67	97.68	0.96	98.26	94.60	96.43
	IKD-Net (ours)	LiDAR	**98.90**	**98.59**	**0.97**	**98.56**	**95.61**	**97.08**

Table 13
Multi-class semantic segmentation on ISPRS Vaihingen dataset. F1 Score is calculated for each category.

Method	Strong modality	OA	F1 Score					
			Imp surf	Building	Low veg	Tree	Car	Mean
IKD-Net	–	85.4	79.1	85.2	61.7	71.7	55.2	70.6
IKD-Net	Image	83.7	75.1	81.5	60.1	71.3	37.5	65.1
IKD-Net (ours)	LiDAR	**92.1**	**96.1**	**90.5**	**87.2**	**92.0**	**92.5**	**91.6**

Intuitively, in terms of the global scene, the characteristics of the buildings became more distinct compared to the ground and trees. Columns 3–6 show that the scattered buildings gradually became more distinguishable from the other categories as the boundaries became increasingly more precise. After the fourth GKG module (column 6) was applied, a large area of trees now has clear demarcation lines from the ground. The CKG module performed class-weighted refinement on the feature maps, which improved the inter-class difference and intra-class similarity, as illustrated in column 7. It further subdivided the large areas that were misclassified into the same type by the previous modules, and the boundaries between the trees (orange) and buildings (turquoise) also became sharper.

5.2.7. *Comparing with SOTA methods*

Finally, we compared the complete IKD-Net with the baseline method (UNet (Ronneberger et al., 2015), RandLA-Net (Hu et al., 2020)), the multi-modal benchmark method (Hybri-UNet (Sherrah, 2016)), the visual classical semantic segmentation methods (HRNet (Wang et al., 2020) and UperNet (Xiao et al., 2018)), and the SOTA multi-modal segmentation network in RS field (vFuseNet (Audebert et al., 2018), MultifilterCNN (Sun et al., 2018), MFNet (Sun et al., 2021), S2Enet (Fang et al., 2021), MDL_RS (Hong et al., 2021), JSH-Net (Zhang et al., 2022), EndNet (Hong et al., 2022), and CMGFNet (Hosseinpour et al., 2022)). The same experimental hyperparameters were used for all the methods. The results are shown in Table 7. If the original methods included instructions on the type of input data, we followed those

Table 14
Multi-class semantic segmentation on GRSS DFC 2018 dataset. Accuracy is calculated for each category.

	Method Strong modality	IKD-Net -	IKD-Net Image	IKD-Net (ours) LiDAR
1	Healthy grass	68.17	77.09	**80.37**
2	Stressed grass	**98.66**	95.71	95.61
3	Artificial turf	90.38	80.11	**99.61**
4	Evergreen trees	78.71	82.94	**97.21**
5	Deciduous trees	70.42	73.34	**94.02**
6	Bare earth	71.77	**85.68**	81.77
7	Water	98.85	98.98	**99.88**
8	Residential buildings	73.08	87.56	**88.73**
9	Non-residential buildings	52.18	79.6	**82.07**
10	Roads	57.64	63.64	**91.84**
11	Sidewalks	79.56	66.81	**85.05**
12	Crosswalks	**80.50**	51.43	72.74
13	Major thoroughfares	**79.97**	57.38	28.54
14	Highways	17.39	64.60	**75.22**
15	Railways	72.25	91.77	**95.12**
16	Paved parking lots	6.73	11.86	**17.50**
17	Unpaved parking lots	18.21	0.0	**48.81**
18	Cars	85.69	71.89	**97.94**
19	Trains	49.39	44.76	**92.67**
20	Stadium seats	30.15	**56.83**	39.76
	AA	63.99	67.10	**78.22**
	OA	62.14	73.75	**78.28**
	Kappa	0.57	0.68	**0.77**

guidelines. If no specific limitations were mentioned, we opted for input data consisting of RGB, DSM, intensity, and number of returns.

For each method in Table 7, the IoU of the Others category was much lower than the other three categories, which was caused by two factors. First, the Others category contained three subcategories (low vegetation, water, and road surface), which made it more difficult to obtain a unified feature description. Second, the number of pixels belonging to the Others category was only about five percent of the other classes, making it very difficult for networks to learn the discriminative features.

Compared to the multi-modal benchmark method Hybri-UNet, the visual classical semantic segmentation method UpperNet didn't exhibit significant advantages, and in fact, the accuracy in OA and Kappa indicators even decreased. This suggests that visual classical semantic segmentation methods may not always be applicable to RS data.

Most of the multi-modal methods exceeded the baseline method UNet in OA, Mean Acc, Kappa, and mIoU, indicating that introducing the other modality indeed improved the effect of 2D semantic segmentation. However, almost none of them achieved a higher mIoU than the baseline method RandLA-Net, which may have been due to their inevitable information loss when mapping the 3D point cloud data to the 2D image space in the preprocessing stage.

Our IKD-Net significantly surpassed the current SOTA multi-modal segmentation methods in all the metrics. In particular, our IoU in the Building category reached over 0.95, laying a better foundation for the downstream RS tasks. The excellent outcome of IKD-Net mainly can be attributed to its heterogeneous networks and well-designed feature interaction module that directly extract features from the raw data source and utilize the imbalance information between them. It is worth noting that the 2D products from 3D point cloud data (DSM, intensity image, etc.) are generated from dense point clouds with approximately 2×10^6 points in every LiDAR patch while our IKD-Net randomly selected only 131,072 points from each LiDAR patch for a compromise with the computer memory. Nevertheless, even the relatively sparse point clouds still provided a powerful knowledge-driven effect for the aerial images, dramatically improving the segmentation accuracy.

Fig. 11 displays the qualitative comparison results on six scenes. The unimodal method, UNet, barely distinguished the Others category, which revealed that multi-modal data offers a distinct advantage in the segmentation of the more ambiguous categories. For the second scene, it is evident that IKD-Net outperformed the other SOTA multi-modal strategies in terms of completeness and edge conformity for building segmentation. Furthermore, our method accurately outlines the edges of the two connected buildings on the right side of the third scene, effectively restoring their connected form. Although the edges of the Tree category in all six scenes were very irregular and had many scattered small areas, IKD-Net outstandingly reconstructed its rough boundary lines.

5.3. Results on ISPRS Vaihingen dataset

As shown in Table 8, we compared our IKD-Net with the baseline methods (UNet, RandLA-Net), the benchmark competitors (SVL_3 (Gerke, 2014), HUST (Quang et al.), RIT (Piramanayagam et al., 2016), UOA (Lin et al., 2016), ADL_3 (Paisitkriangkrai et al., 2015), DST_1 (Sherrah, 2016), DLR_8 (Marmanis et al., 2018), UFMG_4 (Nogueira et al., 2019), ONE_7 (Audebert et al., 2016), and CASIA2 (Liu et al., 2018)), and the recent SOTA multi-modal segmentation networks in RS field (CCANet (Deng et al., 2021), BANet (Wang et al., 2021), HCANet (Zhang et al., 2022), HECR-Net (Liu et al., 2021b), MAResU-Net (Li et al., 2021a), ESANet (Seichter et al., 2021), BoTNet (Srinivas et al., 2021), MANet (Li et al., 2021b), UnetFormer (Wang et al., 2022), JSH-Net (Zhang et al., 2022), CMFNet (Ma et al., 2022), HMANet (Niu et al., 2022), MFNet (Sun et al., 2021), SPANet (Hou et al., 2023), and LoG-CAN (Ma et al., 2023)) using the ISPRS Vaihingen dataset. The benchmark competitors on the challenge evaluation website were measured only on their OA and F1 score rounded to three decimal places so we indicate the same criteria in Table 8. The best value under a certain metric is bolded.

Our IKD-Net ranked first among all the excellent methods on the OA and the mean F1 and achieved the best F1 score in four of the five categories. We believe the superiority of IKD-Net is due to its ability to treat the two modalities distinctly and then leverage the strong modality to drive the feature learning of the weak modality.

When using RandLA-Net for point cloud single-modal classification, the accuracy of the car category is close to 0. However, multi-modal classification using our IKD-Net can increase single-image classification accuracy by about 30 %. The LiDAR data of ISPRS Vaihingen dataset is acquired through row scanning, resulting in a sparse distribution with very few points belonging to the Car category, making it challenging to determine this category. Nevertheless, our approach of point clouds guide image feature redistribution can still improve remote RS segmentation accuracy, even when the point cloud quality is relatively poor. This is due to the relatively large amount of information contained in each single point in LiDAR data.

Fig. 12 displays the qualitative results of IKD-Net and five excellent benchmark methods on the ISPRS Vaihingen dataset. All the methods delivered exceptional performance, while our IKD-Net excelled in integrity and accurately identified the boundaries of buildings. Moreover, only IKD-Net was able to separate the tree objects of the third scene while the other methods joined them together.

5.4. Results on GRSS DFC 2018 dataset

In this section, we review our experiments on the GRSS DFC 2018 dataset to further demonstrate the superiority of our method. The GRSS DFC 2018 dataset was provided by the Image Analysis and Data Fusion Technical Committee for the 2018 IEEE GRSS Data Fusion Contest (DFC). Table 9 lists the baseline methods (UNet, RandLA-Net), the best ranked teams in the data fusion classification challenge track (Xu et al., 2019), and the recent SOTA multi-modal segmentation networks in RS field (CEGCN (Liu et al., 2021a), NLCaps-Net (Lei et al., 2021), EB-CNN (Lu et al., 2022), CAG (Cai and Wei, 2022), and CAGU (Lin et al., 2022)). The best value under a certain metric is bolded. Table 10 shows the category numbers and corresponding category names.

Most of the categories in the images of the GRSS DFC 2018 dataset do not have corresponding points in the point cloud data, which poses a

great challenge to multi-modal learning. As indicated in Table 9, our IKD-Net ranked highest on Mean Acc and achieved comparable results on OA and Kappa to that of the best performing approaches. It is worth noting that all the best ranked teams adopt post-processing and some of them further employ object detection techniques, which boosted their accuracy by around 15 %. However, we still achieved higher scores than the previous winners of the competition and has been ranked first in the real-time leaderboard for the challenge evaluation until this paper's submission. We therefore conclude that our IKD-Net has shown that it is extremely efficient in information utilization and is able to extract deep features from raw multi-modal data and jointly use them according to their inherent characteristics.

Fig. 13 shows the imagery and the classification results of the two winning teams and our IKD-Net. Our method excelled in the connectivity of the longest highways (dark-brown), as depicted in the enlargement on the right. There was less confusion between the roads (red) and major thoroughfares (reddish-brown) on our results while team AGTDA and team dlrpba were unable to discriminate these two categories very well, as depicted in the enlargement on the left. There were also two obvious minor differences in our approach compared to the top two methods. First, some of the pixels located in the non-residential buildings (lavender) were misclassified as roads (red); and second, some pixels of paved parking lots (yellow) were confused with those of cars (pink).

6. Discussion

We conducted class-based experimental analysis and discussion on the proposed knowledge-guided mechanism using the N3C-California, ISPRS Vaihingen, and GRSS DFC 2018 dataset.

Firstly, we performed ablation assessments on an individual class basis in the N3C-California dataset from two perspectives: multi-class and binary semantic segmentation.

The results of multi-class semantic segmentation are presented in Table 11, where IoU is calculated for each category. The first row displays the results of replacing all GKG and CKG modules in IKD-Net with direct concatenation. In the second row, we replaced the guidance modality in the GKG and CKG modules with aerial imagery. Finally, the third row shows the segmentation result of the IKD-Net with LiDAR as the guidance modality.

Table 11 demonstrates that, across all four categories, the results of LiDAR guidance strategy are better than direct concatenation and imagery guidance strategies, and the magnitude of improvement is **class dependent**. For the Ground category, the IoU of LiDAR guidance strategy is still nearly 2 % higher than the direct concatenation strategy even when their accuracy both exceeds 90 % and nearly 5 % higher than the imagery guidance strategy. For the Tree category, the IoU of imagery guidance strategy is much lower than the direct concatenation strategy. This indicates that for Tree category the image not only fails to refine the feature map distribution of the point cloud modality but even has a detrimental effect. The underlying reason may be that the Ground category exhibits spectral features that are very similar to vegetation. However, when observed through LiDAR modality, the Ground and Tree categories reveal distinct structural characteristics. In the case of Tree category, point clouds undergo multiple reflections, enabling the capture of the blade's shape outline. Conversely, the Ground category typically involves only a single reflection. Regarding Building category, the advantage of the LiDAR guidance strategy is not so significant compared to the imagery guidance strategy. One possible reason is that the multi-modal segmentation accuracy of the Building category is already high, leaving limited room for improvement. Another reason could be that point clouds and images offer different recognition benefits for the Building category. For example, in regions where buildings are obstructed, point clouds can penetrate occluding objects like trees, whereas images can provide more detailed information about the planar shape of buildings, as shown in Fig. 14. Consequently, even if the material properties of Building category are highly distinctive in the point clouds, the advantages of using LiDAR guidance may be less apparent.

Table 12 presents the results of binary semantic segmentation, which demonstrate the difference in the role played by the knowledge-guided mechanism in different categories without interference from other categories. For each class, the other three classes are merged into Background category and the same three experiments as those in Table 11 are conducted.

The results of binary semantic segmentation on individual classes and multi-class semantic segmentation are largely consistent, indicating that the role of the knowledge-guided mechanism is on class dependent. Overall, LiDAR data plays a positive role as the strong modality. Compared to the imagery guidance strategy, the LiDAR guidance strategy provides a more significant improvement in the Ground and Tree categories. However, in the Building category, the advantage of the LiDAR guidance strategy is still relatively small.

We conducted class-based experiments on two datasets with a relatively large number of categories: ISPRS Vaihingen (5 categories) and GRSS DFC 2018 (20 categories). Table 13 and Table 14 show the results of multi-class semantic segmentation using IKD-Net with direct concatenation strategy, IKD-Net with imagery guidance strategy, and IKD-Net with LiDAR guidance strategy on ISPRS Vaihingen and GRSS DFC 2018 dataset, respectively.

IKD-Net with LiDAR guidance strategy still demonstrates a clear advantage on the ISPRS Vaihingen dataset. For Car category, the accuracy of the direct concatenation strategy and the imagery guidance strategy are relatively low. However, our IKD-Net still greatly improves the accuracy.

There are many missing categories in the LiDAR data of GRSS DFC 2018 dataset. Nevertheless, IKD-Net with LiDAR guidance strategy shows a large and stable improvement in most of the 20 categories. This may be due to the fact that even if some types of points are missing in the LiDAR data, each point contains more information than a single pixel, providing rich additional information for category judgment. However, for two categories, major thoroughfares and stadium seats, the accuracy of the imagery guidance strategy far exceeds that of the LiDAR guidance strategy.

In summary, we believe that the LiDAR guidance strategy is superior to the direct concatenation and imagery guidance strategies in terms of effectiveness, with the extent of improvement varying across different classes. Specifically, the LiDAR guidance strategy demonstrates substantial enhancements on most of categories. In the case of rare categories that may exhibit lower accuracy, targeted post-processing can be employed to further improve results. Compared to the performance optimization achieved by IKD-Net, the cost is very small.

7. Conclusion

In this paper, we proposed a novel end-to-end heterogeneous dual-stream architecture network called IKD-Net for multi-modal land-cover segmentation. Unlike the current mainstream multi-modal approaches in remote sensing, our dual-stream architecture extracts the features from raw multi-modal heterogeneous data directly rather than their abridged derivatives to retain the intact information of both modalities. Our two GKG and CKG plug-and-play gated modules then utilize the strong modal (LiDAR) to drive the feature map refinement of the weak modality (aerial image) in the global and categorical perspective. The whole network is finally optimized by a sophisticated joint loss function. In the course of our work, we also established a new dataset called N3C-California to provide a particular benchmark for multi-modal joint segmentation to address the lack of large-scale annotated LiDAR-imagery datasets dedicated to remote sensing tasks. We conducted not only sufficient ablation studies of the above modules and visualized their effects in this paper but conducted additional experiments as well that demonstrated IKD-Net's ability to exceed the

benchmarks and the SOTA methods on the N3C-California and ISPRS Vaihingen datasets. Furthermore, IKD-Net has been ranked first in the real-time leaderboard on the GRSS DFC 2018 challenge evaluation until this paper's submission.

The biggest limitation of our IKD-Net is that both aerial image and LiDAR require labeling for network optimization, and labeling point clouds can be particularly challenging. In subsequent works, we plan to address this limitation by introducing semi-supervised or contrastive learning strategies to extract features from a limited quantity of labeled data and a larger amount of unlabeled data, thereby alleviating the burden associated with data labeling. Moreover, we aim to enhance the synergy of multiple modalities by designing multi-task networks.

Declaration of Competing Interest

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

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grants 42192583, 42030102, and 42001406), the Fund for Innovative Research Groups of the Hubei Natural Science Foundation (Grant 2020CFA003), and the Major special projects of Guizhou [2022] 001.

References

Audebert, N., Saux, B.L., Lefèvre, S., 2016. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In: Asian Conference on Computer Vision. Springer, pp. 180–196.

Audebert, N., Le Saux, B., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Boulch, A., Le Saux, B., Audebert, N., 2017. Unstructured point cloud semantic labeling using deep segmentation networks. 3dor@ eurographics 3, 1–8.

Cai, W., Wei, Z., 2022. Remote sensing image classification based on a cross-attention mechanism and graph convolution. IEEE Geosci. Remote Sens. Lett. 19, 3026587.

Chen, S., Niu, S., Lan, T., Liu, B., 2019. Large-scale 3d point cloud representations via graph inception networks with applications to autonomous driving. arXiv preprint arXiv:1906.11359.

Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. In: Advances in Neural Information Processing Systems, pp. 28.

Choy, C., Gwak, J., Savarese, S., 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3075–3084.

Deng, G., Wu, Z., Wang, C., Xu, M., Zhong, Y., 2021. CCANet: Class-constraint coarse-to-fine attentional deep network for subdecimeter aerial image semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–20.

Engelmann, F., Kontogianni, T., Leibe, B., 2020. Dilated point convolutions: on the receptive field size of point convolutions on 3d point clouds. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 9463–9469.

Fang, S., Li, K., Li, Z., 2021. S^2ENet: Spatial–spectral cross-modal enhancement network for classification of hyperspectral and LiDAR data. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154.

Gadzicki, K., Khamsehashari, R., Zetzsche, C., 2020. Early vs late fusion in multimodal convolutional neural networks. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION). IEEE, pp. 1–6.

Galassi, A., Lippi, M., Torroni, P., 2020. Attention in natural language processing. IEEE Trans. Neural Networks Learn. Syst. 32, 4291–4308.

Gerke, M., 2014. Use of the stair vision library within the ISPRS 2D semantic labeling benchmark.

Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., 2019. Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art. IEEE Geosci. Remote Sens. Mag. 7, 6–39.

Gialampoukidis, I., Moumtzidou, A., Bakratsas, M., Vrochidis, S., Kompatsiaris, I., 2021. A multimodal tensor-based late fusion approach for satellite image search in sentinel 2 images. In: MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27. Springer, pp. 294–306.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9224–9232.

Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M., 2020. Deep learning for 3d point clouds: a survey. IEEE Trans. Pattern Anal. Mach. Intell. 43, 4338–4364.

He, X., Zhang, S., Xue, B., Zhao, T., Wu, T., 2023. Cross-modal change detection flood extraction based on convolutional neural network. Int. J. Appl. Earth Obs. Geoinf. 117, 103197.

Heidemann, H.K., 2012. Lidar base specification, Techniques and Methods, Version 1.0: Originally posted August 17. 2012; Version 1.1: October 29, 2014; Version 1.2: November 12, 2014; Version 1.3: February 28, 2018 ed, Reston, VA, p. 114.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 2.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2021. More diverse means better: multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Remote Sens. 59, 4340–4354.

Hong, D., Gao, L., Hang, R., Zhang, B., Chanussot, J., 2022. Deep encoder–decoder networks for classification of hyperspectral and LiDAR data. IEEE Geosci. Remote Sens. Lett. 19, 3017414.

Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. CMGFNet: a deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 184, 96–115.

Hou, J., Guo, Z., Feng, Y., Wu, Y., Diao, W., 2023. SPANet: spatial adaptive convolution based content-aware network for aerial image semantic segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 2192–2204.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A., 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11108–11117.

Hua, B.-S., Tran, M.-K., Yeung, S.-K., 2018. Pointwise convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 984–993.

Huang, J., You, S., 2016. Point cloud labeling using 3d convolutional neural network. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 2670–2675.

Huang, Q., Wang, W., Neumann, U., 2018. Recurrent slice networks for 3d segmentation of point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2626–2635.

Huang, J., Zhang, X., Xin, Q., Sun, Y., Zhang, P., 2019. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. ISPRS J. Photogramm. Remote Sens. 151, 91–105.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. arXiv preprint arXiv:1602.07360.

Jiang, M., Wu, Y., Zhao, T., Zhao, Z., Lu, C., 2018. Pointsift: A sift-like network module for 3d point cloud semantic segmentation. arXiv preprint arXiv:1807.00652.

Landrieu, L., Boussaha, M., 2019. Point cloud oversegmentation with graph-structured deep metric learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7440–7449.

Landrieu, L., Simonovsky, M., 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4558–4567.

Lawin, F.J., Danelljan, M., Tosteberg, P., Bhat, G., Khan, F.S., Felsberg, M., 2017. Deep projective 3D semantic segmentation. In: International Conference on Computer Analysis of Images and Patterns. Springer, pp. 95–107.

Lei, R., Zhang, C., Du, S., Wang, C., Zhang, X., Zheng, H., Huang, J., Yu, M., 2021. A non-local capsule neural network for hyperspectral remote sensing image classification. Remote Sens. Lett. 12, 40–49.

Li, X., Wang, W., Hu, X., Yang, J., 2019. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519.

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J., 2022a. Deep learning in multimodal remote sensing data fusion: A comprehensive review. Int. J. Appl. Earth Obs. Geoinf. 112, 102926.

Li, J., Weinmann, M., Sun, X., Diao, W., Feng, Y., Hinz, S., Fu, K., 2022b. VD-LAB: A view-decoupled network with local-global aggregation bridge for airborne laser scanning point cloud classification. ISPRS J. Photogramm. Remote Sens. 186, 19–33.

Li, R., Zheng, S., Duan, C., Su, J., Zhang, C., 2021a. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M., 2021b. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Lin, G., Shen, C., Van Den Hengel, A., Reid, I., 2016. Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3194–3203.

Lin, M., Jing, W., Di, D., Chen, G., Song, H., 2022. Context-aware attentional graph U-Net for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 19, 3069987.

Liu, Y., Chen, K., Liu, C., Qin, Z., Luo, Z., Wang, J., 2019. Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2604–2613.

Liu, Y., Yi, L., Zhang, S., Fan, Q., Funkhouser, T., Dong, H., 2020. P4contrast: contrastive learning with pairs of point-pixel pairs for RGB-D scene understanding. arXiv preprint arXiv:2012.13089.
Liu, Y., Fan, Q., Zhang, S., Dong, H., Funkhouser, T., Yi, L., 2021c. Contrastive multimodal fusion with tupleinfonce. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 754–763.
Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., Pan, C., 2018. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS J. Photogramm. Remote Sens. 145, 78–95.
Liu, Q., Xiao, L., Yang, J., Wei, Z., 2021a. CNN-enhanced graph convolutional network with pixel-and superpixel-level feature fusion for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 59, 8657–8671.
Liu, Q., Kampffmeyer, M., Jenssen, R., Salberg, A.-B., 2022. Multi-modal land cover mapping of remote sensing images using pyramid attention and gated fusion networks. Int. J. Remote Sens. 43, 3509–3535.
Liu, W., Zhang, W., Sun, X., Guo, Z., Fu, K., 2021b. HECR-Net: Height-embedding context reassembly network for semantic segmentation in aerial images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 9117–9131.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.
Lu, Z., Liang, S., Yang, Q., Du, B., 2022. Evolving block-based convolutional neural network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–21.
Ma, Y., Guo, Y., Liu, H., Lei, Y., Wen, G., 2020. Global context reasoning for semantic segmentation of 3D point clouds. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2931–2940.
Ma, X., Ma, M., Hu, C., Song, Z., Zhao, Z., Feng, T., Zhang, W., 2023. LoG-CAN: local-global Class-aware Network for semantic segmentation of remote sensing images. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5.
Ma, X., Zhang, X., Pun, M.-O., 2022. A crossmodal multiscale fusion network for semantic segmentation of remote sensing data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 3463–3474.
Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172.
Meng, H.-Y., Gao, L., Lai, Y.-K., Manocha, D., 2019. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 8500–8508.
Milioto, A., Vizzo, I., Behley, J., Stachniss, C., 2019. Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 4213–4220.
Mnih, V., Heess, N., Graves, A., 2014. Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, pp. 27.
Nahhas, F.H., Shafri, H.Z., Sameen, M.I., Pradhan, B., Mansor, S., 2018. Deep learning approach for building detection using lidar–orthophoto fusion. J. Sens. 2018.
Niu, R., Sun, X., Tian, Y., Diao, W., Chen, K., Fu, K., 2022. Hybrid multiple attention network for semantic segmentation in aerial images. IEEE Trans. Geosci. Remote Sens. 60, 3065112.
Nogueira, K., Dalla Mura, M., Chanussot, J., Schwartz, W.R., Dos Santos, J.A., 2019. Dynamic multicontext segmentation of remote sensing images based on convolutional networks. IEEE Trans. Geosci. Remote Sens. 57, 7503–7520.
Paisitkriangkrai, S., Sherrah, J., Janney, P., Hengel, V.-D., 2015. Effective semantic pixel labelling with convolutional networks and conditional random fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 36–43.
Piramanayagam, S., Schwartzkopf, W., Koehler, F., Saber, E., 2016. Classification of remote sensed images using random forests and deep learning framework. In: Image and Signal Processing for Remote Sensing XXII. SPIE, pp. 205–212.
Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660.
Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems, pp. 30.
Quang, N.T., Thuy, N.T., Sang, D.V., Binh, H.T.T., Semantic Segmentation for Aerial Images using RF and a full-CRF. [Online] Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/results/papers/HUST_details.pdf.
Rethage, D., Wald, J., Sturm, J., Navab, N., Tombari, F., 2018. Fully-convolutional point networks for large-scale point clouds. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 596–611.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, pp. 234–241.
Rosu, R.A., Schütt, P., Quenzel, J., Behnke, S., 2019. Latticenet: Fast point cloud segmentation using permutohedral lattices. arXiv preprint arXiv:1912.05905.
Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., 2014. ISPRS semantic labeling contest. ISPRS: Leopoldshöhe, Germany 1, 4.
Rufe, P.P., 2014. Digital Orthoimagery Base Specification V1. 0.
Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., Gross, H.-M., 2021. Efficient rgb-d semantic segmentation for indoor scene analysis. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 13525–13531.
Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv preprint arXiv:1606.02585.
Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A., 2021. Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16519–16529.
Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.-H., Kautz, J., 2018. Splatnet: Sparse lattice networks for point cloud processing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2530–2539.
Sun, Y., Zhang, X., Xin, Q., Huang, J., 2018. Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data. ISPRS J. Photogramm. Remote Sens. 143, 3–14.
Sun, Y., Fu, Z., Sun, C., Hu, Y., Zhang, S., 2021. Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data. IEEE Trans. Geosci. Remote Sens. 60, 1–18.
Tatarchenko, M., Park, J., Koltun, V., Zhou, Q.-Y., 2018. Tangent convolutions for dense prediction in 3D. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3887–3896.
Tchapmi, L., Choy, C., Armeni, I., Gwak, J., Savarese, S., 2017. Segcloud: Semantic segmentation of 3d point clouds. In: 2017 International Conference on 3D vision (3DV). IEEE, pp. 537–547.
Thomas, H., Qi, C.R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L.J., 2019. Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6411–6420.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30.
Wang, S., Suo, S., Ma, W.-C., Pokrovsky, A., Urtasun, R., 2018. Deep parametric continuous convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2589–2597.
Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J., 2019. Graph attention convolution for point cloud semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10296–10305.
Wang, L., Li, R., Wang, D., Duan, C., Wang, T., Meng, X., 2021. Transformer meets convolution: a bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sens. (Basel) 13, 3065.
Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214.
Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., 2020. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43, 3349–3364.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19.
Wu, B., Wan, A., Yue, X., Keutzer, K., 2018. Squeezeseg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LIDAR point cloud. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 1887–1893.
Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K., 2019. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE, pp. 4376–4382.
Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV), pp. 418–434.
Xu, Y., Du, B., Zhang, L., Cerra, D., Pato, M., Carmona, E., Prasad, S., Yokoya, N., Hänsch, R., Le Saux, B., 2019. 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. Top. Appl. Earth Obs. Remote Sens. 12, 1709–1724.
Yang, M.Y., Landrieu, L., Tuia, D., Toth, C., 2021. Muti-modal learning in photogrammetry and remote sensing. ISPRS J. Photogramm. Remote Sens. 176, 54.
Ye, X., Li, J., Huang, H., Du, L., Zhang, X., 2018. 3d recurrent neural networks with context fusion for point cloud semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 403–417.
Yin, W., Schütze, H., Xiang, B., Zhou, B., 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. Trans. Assoc. Comput. Linguist. 4, 259–272.
Zhang, P., Du, P., Lin, C., Wang, X., Li, E., Xue, Z., Bai, X., 2020. A hybrid attention-aware fusion network (HAFNET) for building extraction from high-resolution imagery and LiDAR data. Remote Sens. (Basel) 12, 3764.
Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F., Han, J., Ding, E., 2019. ACFNET: Attentional class feature network for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6798–6807.
Zhang, N., Pan, Z., Li, T.H., Gao, W., Li, G., 2023. Improving graph representation for point cloud segmentation via attentive filtering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1244–1254.
Zhang, W., Huang, H., Schmitz, M., Sun, X., Wang, H., Mayer, H., 2017. Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling. Remote Sens. (Basel) 10, 52.
Zhang, B., Wan, Y., Zhang, Y., Li, Y., 2022. JSH-Net: joint semantic segmentation and height estimation using deep convolutional networks from single high-resolution remote sensing imagery. Int. J. Remote Sens. 43, 6307–6332.

Zhao, H., Zhang, Y., Liu, S., Shi, J., Loy, C.C., Lin, D., Jia, J., 2018. PSANET: point-wise spatial attention network for scene parsing. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 267–283.

Zhao, H., Jiang, L., Fu, C.-W., Jia, J., 2019a. PointWeb: enhancing local neighborhood features for point cloud processing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5565–5573.

Zhao, Z., Liu, M., Ramani, K., 2019b. DAR-Net: dynamic aggregation network for semantic scene segmentation. arXiv preprint arXiv:1907.12022.