A CASCADED CROSS-MODAL NETWORK FOR SEMANTIC SEGMENTATION FROM HIGH-RESOLUTION AERIAL IMAGERY AND RAW LIDAR DATA

Yameng Wang [1], *Bin Zhang* [1], *Yi Wan* [1], *Yongjun Zhang* [1*]

[1] School of Remote Sensing and Information Engineering, Wuhan University, 430079, China
*Corresponding author: zhangyj@whu.edu.cn

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883824

ABSTRACT

As various sensors appear, extracting information from multimodal data becomes a prominent topic. Current multimodal approaches for image and LiDAR normally discard the point-to-point topology relationship of the latter to keep the dimension matched. To tackle this task, we propose a cascaded cross-modal network (CCMN) to extract the joint-features from high-resolution aerial imagery and LiDAR point directly, instead of their abridged derivatives. Firstly, point-wise features are extract from raw LiDAR data by a forepart 3D extractor. Subsequently, the LiDAR-derived features are executed spatial reference conversion to project and align to the imagery coordinate space. Finally, the cross-modal compounds containing the obtained feature maps and the corresponding images are placed into a U-shape structure to generate segmentation results. The experiment results indicate that our strategy surpasses the popular multimodal method by 6% on mIoU.

Index Terms— Semantic segmentation, aerial images, LiDAR, multimodal data, convolutional neural network

1. INTRODUCTION

Semantic segmentation, i.e., pixel-level visual interpretable classification, has long been one of the basic tasks for the computer vision community, and has been widely used in fields such as autonomous driving [1], medical image recognition [2], and so on. As deep learning methods shine in the field of computer vision, convolutional neural networks have also been introduced into semantic segmentation tasks. Canonical methods such as FCN [3] and UNet [4] provide robust baselines for subsequent research. Although a variety of researches have been emerging, there is still a long way toward human-level high-precision scene descriptions.

Semantic segmentation of remote sensing data is an important prerequisite for remote sensing information extraction, and of great significance for various automated monitoring tasks. Compared with natural images in computer vision, a prominent character of remote sensing data is the diversity of sensors and data types, such as optical image, infrared image, synthetic aperture radar (SAR), Light Detection and Ranging (LiDAR), etc. Each type of data exhibits individual characteristic and is suitable for different application scenarios. High-resolution aerial images have rich texture information and continuous features, but there are the phenomena of different objects with the same spectra characteristics and the same spectrum with different objects, which bring difficulties to image interpretation. Infrared remote sensing plays a critical role in geological structure detection and pollution monitoring, but the quality of the obtained data is greatly affected by the weather. As an active detection method, LiDAR is not restricted by natural conditions such as light and weather, and performs outstandingly in harsh and complex environments. To overcome the limitations of the single data source, multimodal data is used for remote sensing semantic segmentation. Among them, the combination of optical image and LiDAR can access more informative features due to the complementary perception of their stereo structure and texture mode.

Since the 3D point clouds and the 2D optical images belong to different metric spaces, it is not convenient to integrate both seamlessly for the nonnegligible domain gap. Some studies have discarded some LiDAR fields and only exploit the elevation and its derivations as the auxiliary data for images. Sherrah [5] concatenated the optical images and its corresponding digital surface model (DSM) as an enhanced input for FCN to get the pixel-wise semantic labels. Audebert et al. [6] designed a dual-stream SegNet architecture with IRRG images (near-infrared, red, and green channel) and the abridged LiDAR features (DSM and the normalized DSM (NDSM)) as double inputs. Furtherly, they proposed two schemes, i.e. early fusion and late fusion, to obtain the multimodal joint-features [7]. Liu et al. [8] proposed an unsymmetrical structure where one branch handled the optical imagery and the other dealt with the hand-crafted LiDAR feature such as height, height variations, and surface norm. Sun et al. [9] split the original LiDAR data into three channels of NDSM, intensity, and number of returns, which all had the same size as the image counterpart. These channels concatenated with the difference of Gaussians derived from the source image were then put into the multi-filter CNN together to utilize information from the two modals. These studies compressed the dimension of LiDAR crudely, leading to the

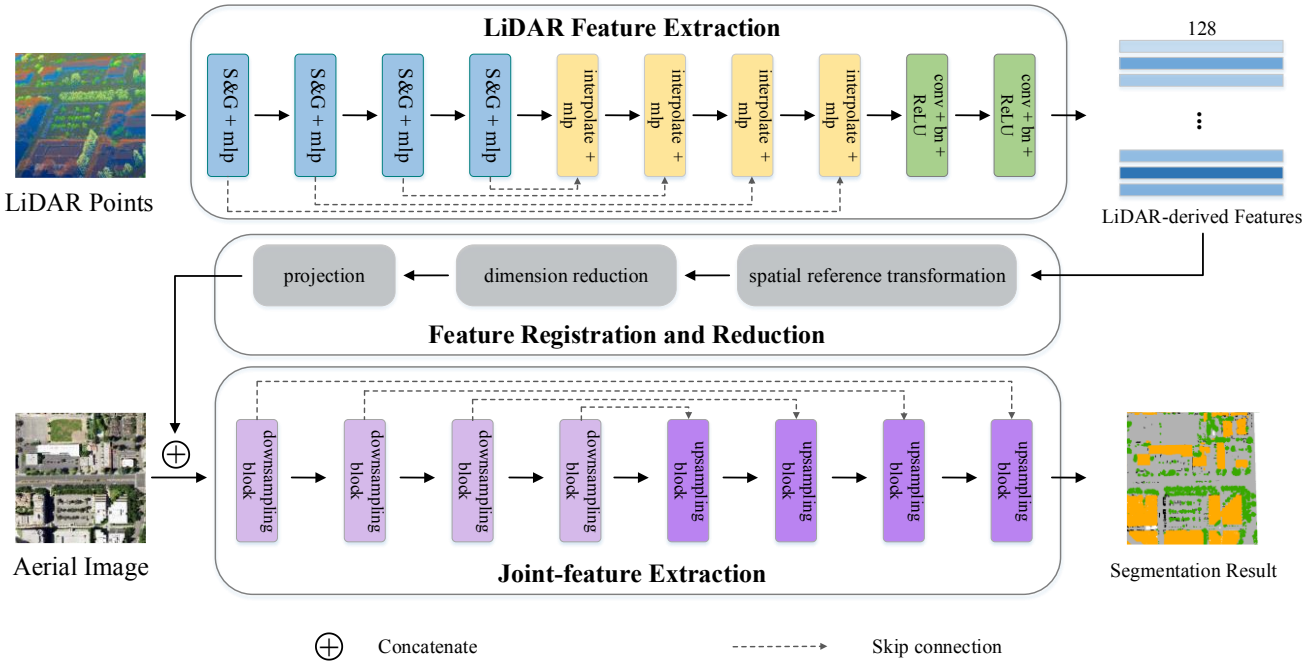


Figure 1. Illustration of CCMN. S&G represents sampling and grouping operation.

degeneration of the topological relationship between points in 3D space.

To take full advantage of the respective information of multimodal data, we propose a cascaded cross-modal network (CCMN) which can be roughly divided into three stages to utilize imagery and raw LiDAR points directly. In the first stage, the raw LiDAR points with x, y and z coordinates are put into a 3D network. Subsequently, the point-wise feature vectors are registered to the corresponding pixels and executed dimensionality reduction by principal component analysis (PCA). In the last stage, the feature maps derived from LiDAR are concatenated with their image counterparts and put into the U-shape network to extract joint-features for the final segmentation. Our method preserves all the attributes of LiDAR data and realizes feature-level cross-modal mapping. The experimental results show that our proposed CCMN obviously surpasses the single-modal and the popular multimodal methods.

2. METHODOLOGY

Our CCMN consists of three parts as Figure 1 shown, which can also be formulated as follow.

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

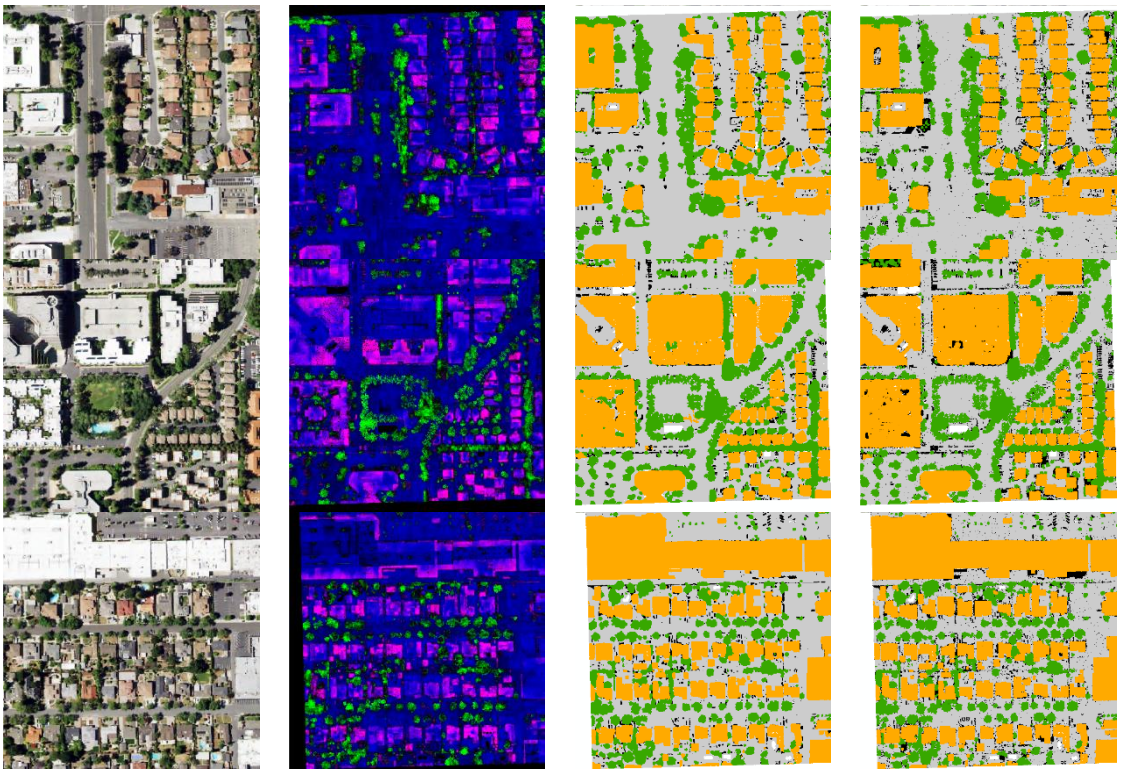
where r, x, and y are the final segmentation results, LiDAR points and image patches respectively. F stands for LiDAR feature extraction stage and G represents joint-feature extraction stage. Γ is the coordinate transformation stage from LiDAR space to imagery space. $\oplus$ is the symbol of concatenation.

2.1. LiDAR feature extraction

Compared with imagery, LiDAR is more complex and describes the real world in a detailed way as more features are encoded within. The difference of dimensions between these two types of data results in conceptual unbalance. A feasible way to solve this is to extract the 2D feature maps from LiDAR first. In our method, we select PointNet++ [10], an optimized version of PointNet [11], to handle point clouds. PointNet stacks several multi-layer perceptron (MLP) layers and a max pooling layer to generate discriminative features. In order to preserve the point-to-point relevance and increase the network's ability to integrate local information, PointNet++ uses sampling and grouping modules with different parameters to obtain the features of different scales and changes the concatenation strategy of PointNet to encoder-decoder structure equipped with skip connection. When extracting point cloud features, we discard the last convolution layer and keep the 128-dimensional feature vectors as the output.

2.2. Feature registration and reduction

The LiDAR-derived features obtained in section 2.1 are based on the geodetic coordinate system, while the aerial image of the same region is based on the projection coordinate system. We perform spatial reference conversion on LiDAR-derived features to align them with the corresponding images. To balance the information quantity of two types of data, we reduce the dimension of LiDAR-derived features from 128 to 3 through PCA.



Aerial image LiDAR-derived feature Segmentation result Ground truth

Figure 2. Visualization of three qualitative results.

2.3. Joint-feature extraction

At the third stage, we connect features from LiDAR after the cross-modal projection with the images of the same area and use an U-shape network to extract joint-features. The downsampling phase of the network consists of four stacked modules, each of which contains two convolution layers of 3×3 kernel size, one batch normalization layer, one activation function ReLU layer, and one max pooling layer. The output feature dimensions of each module are 64, 128, 256, 512 in sequence. After each max pooling layer, the size of the feature map becomes half of the original. The upsampling phase is composed of four inverse processes corresponding to those of the downsampling stage. The downsampling and upsampling layers of the same level are connected by skip connection.

3. EXPERIMENTAL RESULTS AND DISCUSSIONS

3.1. Datasets and metrics

To the best of our knowledge, there has been no image-LiDAR combined dataset up till now. To fill this gap, we make an experimental united dataset where the aerial imagery is downloaded from National Agriculture Imagery Program (NAIP) and the corresponding LiDAR point clouds are from the United States Geological Survey (USGS). The study area covers most part of Santa Clara County, California, U.S. The proportion of samples in training, validation, and testing dataset is 8:1:1. The LiDAR point clouds are voxelized and subsequently split into 1234 blocks each containing less than 65536 points. The aerial images are cut into 512×512 pieces with 20% overlap. The annotations of point clouds and images all come from the "classification" attribute of LiDAR.

We select accuracy (Acc), mean intersection-over-union (mIoU), and mean F1-score (mF1) as the evaluation measure. Acc indicates the proportion of the pixels correctly predicted. Intersection-over-union (IoU) is the ratio of the intersection and union of pixels predicted as a certain class and those actually belonging to this class. F1-score represents the harmonic mean of precision and recall. The mean values of them on different categories are mIoU and mF1 respectively.

3.2. Implement details

All experiments are completed on a Windows 10 PC equipped with an NVIDIA GeForce RTX 3090 24G GPU and PyTorch deep learning framework. The optimizer is SGD during the whole procedure. In the LiDAR feature extraction stage, 8192 points are randomly sampled from each patch. For the network in this stage, the batch size is scheduled to 16, and a total of 8 epochs are trained. The learning rate decays linearly with initial value and decay rate set to 0.001 and 0.7 respectively. PCA compresses the

channel of feature map from 128 to 3. When extracting joint-features, the patches are cropped into 512×512. The batch size is 8, and the total number of iterations is 80,000. The learning rate is originally set to 0.01 and adjusted by the cosine annealing strategy.

3.3. Results and analysis

We conduct comparison experiments on baseline method UNet, hybrid UNet [5], our CCMN with PCA and CCMN without PCA. We select the network models that get the highest scores for mIoU on the validation dataset to implement inference. The results on the test dataset are summarized in Table 1.

Table 1. Comparison of different methods

Network	Input	Acc	mIoU	mF1
UNet(baseline)	Image	0.8748	0.6267	0.7119
hybrid UNet [5]	Image + DSM	0.9169	0.7080	0.7819
CCMN-PCA (ours)	Image + LiDAR	0.9189	0.7224	0.8023
CCMN (ours)	Image + LiDAR	**0.9314**	**0.7601**	**0.8370**

Table 1 illustrates that our strategy to combine image and LiDAR surpasses the standard practice to cascade image and DSM by large margins, namely 1.5% on Acc, 6% on mIoU and more than 5% on mF1. While using PCA lowers the superiority to a certain extent, it can greatly reduce the calculation.

We visualize the intermediate and final output of our CCMN in Figure 2. The input images are shown in the first column, with LiDAR-derived features, final segmentation results, and ground truths following sequentially. In the last two columns, the ground, tree, buildings, and other classes are marked in gray, green, yellow, and black respectively. Figure 2 displays that many pixels belonging to buildings are misclassified as grounds in the LiDAR-derived feature maps. However, concatenated input of images and LiDAR-derived features significantly improves the segmentation accuracy.

4. CONCLUSION

In this paper, we propose a cascaded cross-modal network to utilize imagery and LiDAR data directly, instead of the derivatives such as DSM. By using raw LiDAR data without additional dimension compression, our method takes full advantage of the properties of point clouds and remains the topological relationship between points in 3D space. The experiment results present that our method surpasses single-modal and popular multimodal methods. Our future research will focus on the end-to-end solution to cross-modal information fusion, as well as the plug-in inter-modal projection.

5. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under project number 42030102 and the Fund for Innovative Research Groups of the Hubei Natural Science Foundation under project number 2020CFA003.

6. REFERENCES

[1] L. Tran and M. Le, "Robust U-Net-based Road Lane Markings Detection for Autonomous Driving", *2019 International Conference on System Science and Engineering (ICSSE 2019)*, pp. 62-66, Jul. 19-21, 2019.

[2] L. Geert et al., "A survey on deep learning in medical image analysis", *Medical image analysis*, vol. 42, pp. 60-88, 2017.

[3] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation", *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440, 2015.

[4] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234-241, 2015.

[5] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery", *CoRR*, 2016.

[6] N. Audebert, B. Le Saux and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks", *ACCV*, pp. 180-196, 2016.

[7] N. Audebert, B. Le Saux and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks", *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20-32, Jun. 2018.

[8] Y. Liu, S. Piramanayagam, S. T. Monteiro and E. Saber, "Dense semantic labeling of very-high-resolution aerial imagery and lidar with fully-convolutional neural networks and higher-order crfs", *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1561-1570, 2017.

[9] Y. Sun, X. Zhang, Q. Xin and J. Huang, "Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data", *ISPRS J. Photogramm. Remote Sens.*, vol. 143, pp. 3-14, 2018.

[10] C. R. Qi, L. Yi, H. Su and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space", *NIPS*, pp. 5105-5114, 2017.

[11] Charles R Qi, Hao Su, Kaichun Mo and Leonidas J Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation", *Proc. Computer Vision and Pattern Recognition (CVPR) IEEE*, vol. 1, no. 2, pp. 4, 2017.