remote sensing

MDPI

Article

Robust 3D Semantic Segmentation Method Based on Multi-Modal Collaborative Learning

Peizhou Ni [1], Xu Li [1,*], Wang Xu [1], Xiaojing Zhou [1], Tao Jiang [2] and Weiming Hu [3,4]

[1] School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China; 230198300@seu.edu.cn (P.N.); 220223306@seu.edu.cn (W.X.); xiaojingzhou@seu.edu.cn (X.Z.)
[2] Xuzhou XCMG Automobile Manufacturing Co., Ltd., Xuzhou 221112, China; jiang_tao@xcmg.com
[3] China Automotive Engineering Research Institute Company Ltd., Chongqing 401122, China; huweiming@caeri.com.cn
[4] School of Transportation, Southeast University, Nanjing 211189, China
* Correspondence: lixu.mail@163.com

Abstract: Since camera and LiDAR sensors provide complementary information for the 3D semantic segmentation of intelligent vehicles, extensive efforts have been invested to fuse information from multi-modal data. Despite considerable advantages, fusion-based methods still have inevitable limitations: field-of-view disparity between two modal inputs, demanding precise paired data as inputs in both the training and inferring stages, and consuming more resources. These limitations pose significant obstacles to the practical application of fusion-based methods in real-world scenarios. Therefore, we propose a robust 3D semantic segmentation method based on multi-modal collaborative learning, aiming to enhance feature extraction and segmentation performance for point clouds. In practice, an attention based cross-modal knowledge distillation module is proposed to effectively acquire comprehensive information from multi-modal data and guide the pure point cloud network; then, a confidence-map-driven late fusion strategy is proposed to dynamically fuse the results of two modalities at the pixel-level to complement their advantages and further optimize segmentation results. The proposed method is evaluated on two public datasets (urban dataset SemanticKITTI and off-road dataset RELLIS-3D) and our unstructured test set. The experimental results demonstrate the competitiveness of state-of-the-art methods in diverse scenarios and a robustness to sensor faults.

Keywords: 3D semantic segmentation; multi-modal fusion; collaborative learning; knowledge distillation

Citation: Ni, P.; Li, X.; Xu, W.; Zhou, X.; Jiang, T.; Hu, W. Robust 3D Semantic Segmentation Method Based on Multi-Modal Collaborative Learning. *Remote Sens.* **2024**, *16*, 453. https://doi.org/10.3390/rs16030453

Academic Editor: Dong Chen

Received: 30 November 2023
Revised: 18 January 2024
Accepted: 22 January 2024
Published: 24 January 2024

1. Introduction

With the continual progression of intelligent driving technology, the safety of intelligent vehicles (IVs) has attracted significant attention and interest [1]. In the field of intelligent driving technology, efficient, effective, and robust environmental perception serves as a foundational prerequisite for subsequent tasks, such as the precise positioning, reliable planning, and secure control of these IVs [2]. As the core module of environmental perception methods, 3D semantic segmentation is able to densely allocate specific semantic labels to individual pixel points, including entities like drivable surfaces and backgrounds, and has emerged as a focal point of concern in recent years.

LiDAR-only semantic segmentation approaches utilize a diverse set of techniques to fully harness geometric information and have managed to achieve competitive results in structured scenarios, such as indoor environments and typical urban traffic scenes [3]. Nevertheless, when confronted with complex and dynamically changing surroundings characterized by sparse and visually similar geometric attributes, these methods encounter limitations inherent to LiDAR sensors, leading to below-expectation performance [4].

A promising method to overcome such limitations lies in the incorporation of camera images, which provide a wealth of dense semantic features, including color and texture

information. Consequently, LiDAR–camera fusion is a strategic approach to enhance the accuracy and robustness of 3D semantic segmentation methods in challenging environmental conditions [5,6]. Utilizing sensor calibration matrices, current LiDAR–camera fusion approaches typically adopt one of two primary strategies: either projecting image pixels onto LiDAR coordinates and performing feature fusion approaches within the sparsely populated LiDAR domain [7–11], or projecting point clouds onto image planes using perspective projection to merge corresponding multi-modal features [4,12–16].

Despite the significant advantages offered by multi-modal fusion, these methods still have following inherent limitations that cannot be circumvented:

(a) Field of View Disparity:

The LiDAR and camera sensors typically possess differing field-of-view characteristics, with only a small overlap area (as depicted in Figure 1). Consequently, it becomes infeasible to establish point-to-pixel mapping for point clouds located outside this overlap area, which significantly restricts the broader application of fusion-based methods.

(**a**) (**b**)

Figure 1. A sample field-of-view difference between LiDAR and camera. (**a**) Original point cloud; (**b**) point cloud in camera field-of-view shown in red.

(b) Dependency on Precise Paired Data:

Fusion-based methods critically rely on the availability of accurately paired data, specifically precise point-pixel mapping between LiDAR and camera data. This mapping is crucial for both the training and inference stages. Thus, any data error or sensor malfunction could have detrimental impacts on segmentation performance and might even lead to algorithm failures. Figure 2 illustrates this vulnerability; for example, cameras are susceptible to light interference, which can result in issues like image confusion, blurriness, overexposure, and other anomalies, while LiDAR sensors can be affected by weather conditions like rain, snow, and fog, leading to phenomena such as "ghost" points or a significant reduction in the amount of point cloud data.

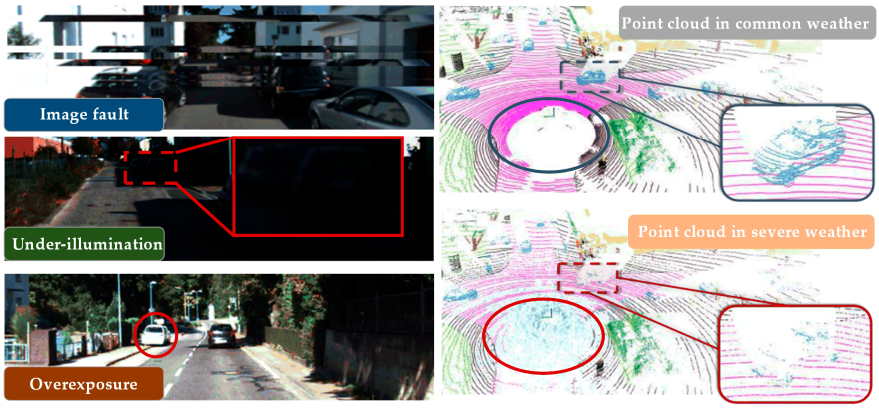

Figure 2. A sample sensor fault.

(c) Resource-Intensive Processing:

Fusion-based approaches require the simultaneous processing of both point cloud and image data, leading to increased demands on computing resources and storage space. Even if efforts have been made to mitigate these challenges through multi-tasking or cascading, such resource demands can pose a substantial burden, especially on devices with limited resources, when deploying real-time applications.

In order to address the aforementioned challenges, this paper presents a robust 3D semantic segmentation method based on multi-modal collaborative learning. It comprehensively considers the complementarity between point cloud and image data at the feature level and output level during training, overcoming the limitation of LiDAR-only methods; benefiting from multi-modal collaborative learning, it can conduct 3D semantic segmentation without image inputs during inference, overcoming the limitations of multi-modal fusion-based methods. Extensive evaluations were conducted across diverse datasets, including the urban dataset SemanticKITTI [17], the off-road dataset RELLIS-3D [18], and our unstructured test set with sparse LiDAR points. The experimental results affirm that, by leveraging the synergies between point cloud and image data, our proposed method can achieve efficient, accurate, and robust 3D semantic segmentation performance in diverse and complex scenarios, especially when the raw data are corrupted. The main contributions of this paper are summarized as follows:

(a) This paper proposes a robust 3D semantic segmentation method based on multi-modal collaborative learning, which effectively deals with the limitations and restrictions of fusion-based 3D semantic segmentation methods.
(b) An attention-based cross-modal knowledge distillation module is proposed to assist 3D feature extraction using 2D image features with higher contributions, which further helps distill multi-modal knowledge to single point-cloud modality for accurate and robust semantic segmentation.
(c) A late fusion strategy guided by a confidence map is proposed to emphasize the strengths of each modality by dynamically assigning per-pixel weights of outputs and further optimizing segmentation results.

The rest of this paper consists of the following sections: Section 2 reviews the related works, Section 3 presents the methodology, Section 4 presents and analyzes the experiments, and Section 5 is the work's conclusion.

2. Related Works

2.1. LiDAR-Based 3D Semantic Segmentation Methods

In general, LiDAR-based 3D semantic segmentation methods can be divided into the following three categories based on distinct data representations.

Point-based methods [19–21] directly process unordered point clouds using MLP-based (multi-layer-perceptron-based) techniques. These methods have demonstrated excellent segmentation results on small-scale and dense point clouds. However, their application to sparse inputs in large-scale scenarios is often limited by factors such as poor locality, high computational costs, and substantial memory requirements, resulting in lower accuracy and slower reasoning speeds.

Voxel-based methods [22–26] transform point clouds into dense voxels and employ 3D convolution to extract and reconstruct the features in each voxel, which achieves superior segmentation results. However, the redundancy in dense voxel representation and the computational inefficiency of 3D convolution contribute to an exponential increase in the complexity of these methods, leading to poor real-time performance.

Projection-based methods can keep a good balance between segmentation performance and real-time performance, benefiting from the compactness of inputs and the lightness of 2D CNNs. They typically employ top-down or spherical projection for point cloud preprocessing, resulting in the formation of Bird's-Eye Views (BEVs) [27–29] and Range

Views (RVs) [30–32]. Nevertheless, the BEV approach remains sparse while preserving the size of objects, and the RV approach disrupts the original topological relationships.

Considering the demand of accurate feature extraction and real-time application, we utilized our former work [33] as the 3D branch backbone, which is a multi-projection fusion method and leverages rich complementary information between different views.

2.2. Knowledge Distillation Methods

Knowledge distillation (KD) was originally proposed for network model compression [34], that is, transferring rich hidden information from complex and large teacher networks to lightweight and compact student networks, aiming to reduce the performance gap between the two models. It was initially designed for image classification tasks, taking various forms of knowledge as distillation targets, including intermediate outputs [35,36], visual attention maps [37,38], interlayer similarity maps [39], and sample-level similarity maps [40,41].

Recent advancements have extended knowledge distillation to semantic segmentation tasks for intermediate feature extraction. For instance, ref. [42] simultaneously extracts pixel-level knowledge, paired-similarity knowledge, and global knowledge, achieving high-order consistency between fine-grained and comprehensive network outputs. Ref. [43] facilitates student model learning by reinterpreting the teacher network's output as a new potential domain and proposes an affinity distillation module to capture the long-term dependencies of the teacher network. Ref. [44] introduces a point-to-voxel knowledge distillation method and a difficulty-sensing sampling strategy to enhance distillation efficiency.

With the rapid progress of multi-modal computer vision technology, more and more research has applied knowledge distillation to the prior information on transmission between different modalities. For example, refs. [45–47] use additional two-dimensional image information during training to enhance algorithm performance in the inferring stage; ref. [48] introduces 2D-assisted pre-training; ref. [49] expands 2D convolution into 3D convolution; and ref. [50] proposes a dense foreground-guided feature imitation method and sparse instance distillation method to transfer spatial knowledge from LiDAR to multiple camera images for 3D target detection.

Nevertheless, in contrast to dense and regular camera images, LiDAR point clouds have inherent characteristics like sparsity, randomness, and variable density. This substantial disparity between the two modalities poses a formidable challenge for knowledge distillation across modalities. The direct application of knowledge distillation between the two modalities will pollute the specific modal information. Therefore, we propose an attention-based cross-modal knowledge distillation module, enhancing the feature extraction of the 3D branch without losing its specific modality information.

3. Methods

In this section, we introduce a robust 3D semantic segmentation method based on multi-modal collaborative learning, as shown in Figure 3. First, the efficient semantic segmentation backbone (including 2D and 3D branches) is utilized to leverage rich complementary information and offer reliable intermediate features for later multi-modal collaborative learning. Then, a cross-modal knowledge distillation module is proposed to enhance the feature representation of the 3D branch in multiple scales using prior feature information from the 2D branch. Finally, a late fusion strategy driven by confidence mapping is proposed to weight the prediction results of the two modal branches in a direct and explicit manner, which highlights the advantages of each modal branch while weakening the interference of incorrect data inputs, so as to generate the final accurate and robust prediction results.

In the following subsections, the architecture of the proposed method will be described in detail.

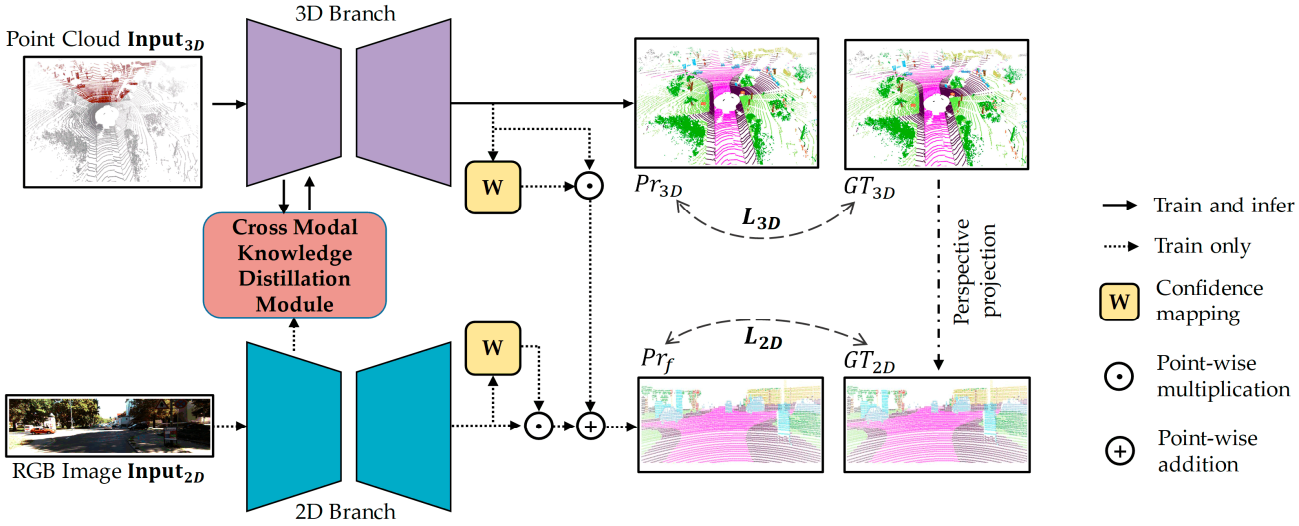

Figure 3. Overall architecture of the proposed method, comprising two key components: the cross-modal knowledge distillation module and the late fusion strategy driven by confidence mapping.

3.1. Semantic Segmentation Backbone

There are two primary objectives of 2D and 3D semantic segmentation backbones: first, to offer reliable semantic information and geometrical features to the latter proposed cross-modal knowledge distillation module; second, to utilize their outputs to further constrain and enhance the final results of 3D semantic segmentation.

To achieve these goals, we simply employ HRNet [51] as the 2D branch and our former work [33] as the 3D branch for efficient and effective feature extraction and semantic segmentation. To be specific, the 2D branch, HRNet, maintains high-resolution representations through the whole process, providing semantically richer and spatially more precise 2D semantic features; the 3D branch combines RV and BEV at both the feature-level and output-level, which significantly mitigates information loss during the projection.

3.2. Attention-Based Cross-Modal Knowledge Distillation Module

The proposed cross-modal knowledge distillation module (see Figure 4) first fuses the paired 2D and 3D features $\{\mathbf{F}_C, \mathbf{F}_R, \mathbf{F}_B\}$ based on the attentional mapping (AM); then, it distills the enhanced fusion features $\mathbf{F}_C^{fe}$ and the enhanced 3D features $\mathbf{F}_R^e$, $\mathbf{F}_B^e$ in a unidirectional alignment. In this manner, we can transfer the comprehensive information from multi-modal data into the LiDAR model for its feature enhancement, while retaining its specific characteristics. Below, we take Image-RV as an example to analyze the process.

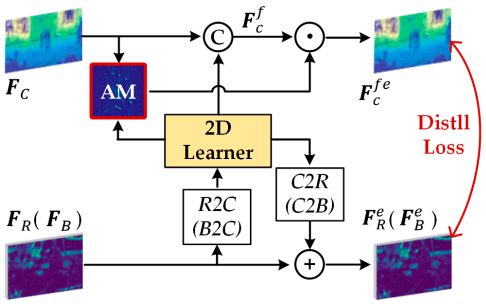

Figure 4. The architecture of the cross-modal knowledge distillation module.

3.2.1. Feature Alignment

The feature alignment between camera images and LiDAR RV images is introduced to generate pairwise matching features of the two modalities, so as to facilitate the subsequent knowledge distillation. It is implemented by calculating the geometric transformation

matrices $\mathbf{M}_{C2R}$ (mapping from camera image to range image) and $\mathbf{M}_{R2C}$ (mapping from range image to camera image). The two metrics are inverses of each other: $\mathbf{M}_{C2R}$ and $\mathbf{M}_{R2C}$. Here, we take $\mathbf{M}_{C2R}$ as an example to analyze the process of fusion.

During the transformation, we utilize the original point cloud as an intermediary agent. We first calculate the matrix $\mathbf{M}_{C2P} \in Z^{H_c \times W_c}$ in Equation (1), which aligns the features of camera images to original point clouds, as shown in Figure 5.

$$\mathbf{M}_{C2P} = \begin{bmatrix} n_{(0,0)} & \cdots & n_{(0,W_c\text{-}1)} \\ \vdots & \ddots & \vdots \\ n_{(H_c\text{-}1,0)} & \cdots & n_{(H_c\text{-}1,W_c\text{-}1)} \end{bmatrix} \tag{1}$$

where (H_c, W_c) are the width and height of the 2D camera images, and $\left\{ n_{(i,j)} | 0 \le i \le H_c\text{-}1, 0 \le j \le W_c\text{-}1 \right\}$ is the $n_{(i,j)th}$ point which projects on (i, j) coordinates. Then, the transformation matrix $\mathbf{M}_{P2R} \in Z^{N \times 2}$ from original points to RV images is formed as follows:

$$\mathbf{M}_{P2R} = \begin{bmatrix} \mathbf{r}_0 & \cdots & \mathbf{r}_{N\text{-}1} \end{bmatrix} = \begin{bmatrix} u_0 & \cdots & u_{N\text{-}1} \\ v_0 & \cdots & v_{N\text{-}1} \end{bmatrix} \tag{2}$$

where N is the number of points, and $\{\mathbf{r}_k = (u_k, v_k) | 0 \le k \le N\text{-}1\}$ represents the projected pixel coordinates of the 2D RV image, corresponding to the k_{th} point. By calculating the $\mathbf{M}_{C2P}$ and $\mathbf{M}_{P2R}$, we obtain the geometric transformation matrix $\mathbf{M}_{C2R} \in Z^{H_r \times W_r \times 2}$:

$$\mathbf{M}_{C2R} = \begin{bmatrix} \mathbf{r}_{n_{(0,0)}} & \cdots & \mathbf{r}_{n_{(0,W_c\text{-}1)}} \\ \vdots & \ddots & \vdots \\ \mathbf{r}_{n_{(H_c\text{-}1,0)}} & \cdots & \mathbf{r}_{n_{(H_c\text{-}1,W_c\text{-}1)}} \end{bmatrix} \tag{3}$$

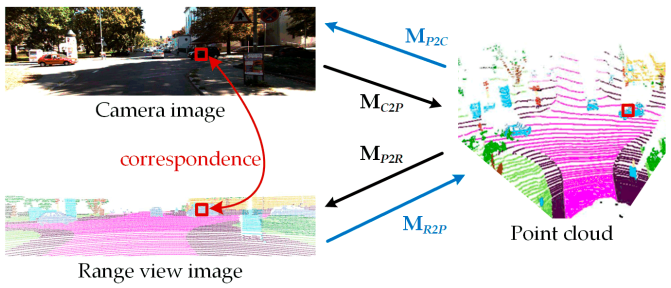

Figure 5. The architecture of feature alignment.

3.2.2. Fusion and Distillation

After feature alignment, we can utilize paired features from the 2D and 3D branches for the fusion and distillation block.

Considering the huge feature gap introduced by different modal networks, it is inappropriate to fuse 3D features and their corresponding 2D features directly. Therefore, we design a 2D-Learner based on MLP to narrow the gap between different modal features. It can be formulated as follows:

$$\mathbf{F}_{convert} = \text{2D\_Learner}(\mathbf{M}_{R2C} \cdot \mathbf{F}_R) \tag{4}$$

Then, we design a fusion method based on spatial attention to achieve the enhanced fusion features, which could improve the feature representation by focusing on important features and suppressing unimportant features. It can be formulated as follows:

$$\begin{aligned} \mathbf{F}_C^{fe} &= A \odot \mathbf{F}_C^f \\ \mathbf{F}_C^f &= \mathbf{F}_{convert} \& \mathbf{F}_C \end{aligned} \tag{5}$$

where $\odot$ represents point-wise multiplication; & represents channel concatenation; and A represents the attentional map which takes 2D and 3D features into comprehensive consideration. This can be formulated as follows:

$$\begin{aligned} A &= \frac{N(\mathbf{F}_C)+N(\mathbf{F}_{convert})}{2} \\ N(F) &= H_F W_F \text{softmax}(\frac{P(F)}{\tau}) \\ P(F)_{i,j} &= \frac{1}{Ch}\sum_{ch=1}^{Ch}\left|F_{ch,i,j}\right| \end{aligned} \tag{6}$$

where $F \in R^{Ch\times H_F\times W_F}$ represents the feature map; $P(F) \in R^{H_F\times W_F}$ represents the result of average-pooling the absolute values along the channel dimensions of F; $N(F) \in R^{H_F\times W_F}$ represents the attention derived from softmax standardization of values at all spatial locations; and τ represents the hyperparameter that regulates the distribution entropy.

After that, we operate feature distillation between enhanced fusion feature $\mathbf{F}_C^{fe}$ and enhance 3D feature $\mathbf{F}_R^e = \mathbf{M}_{C2R}\cdot\mathbf{F}_{convert} + \mathbf{F}_R$. Specifically, we design a feature-level distillation loss $Loss_{dis}$ as the supplement of the segmentation task loss, which comprises multi-scale feature imitation loss $Loss_{fea}$ and attention imitation loss $Loss_{att}$. Feature imitation loss aims at narrowing distribution differences between the two modal features. Attention imitation loss aims at enabling $\mathbf{F}_R^e$ to learn and generate attention patterns similar to $\mathbf{F}_C^{fe}$, thus focusing more attention on spatial positions that $\mathbf{F}_C^{fe}$ considers more important. The overall distillation loss is expressed as follows:

$$\begin{aligned} Loss_{dis} &= Loss_{fea}\left(\mathbf{F}_C^{fe}, \mathbf{F}_R^e\right) + \lambda Loss_{att}\left(\mathbf{F}_C^{fe}, \mathbf{F}_R^e\right) \\ Loss_{fea}(X1, X2) &= \sum_{ch=1}^{Ch}\sum_{i=1}^{H_F}\sum_{j=1}^{W_F}\left(X1_{ch,i,j}\text{-}X2_{ch,i,j}\right)^2 \\ Loss_{att}(X1, X2) &= \sum_{i=1}^{H_F}\sum_{j=1}^{W_F}\left|P(X1)_{i,j}\text{-}P(X2)_{i,j}\right| \end{aligned} \tag{7}$$

where λ represents the hyperparameter that controls the relative importance between the two loss functions and balances them at the same scale.

Through the above analysis, we can see that $\mathbf{F}_{convert}$ is generated from 3D point cloud features, while also be influenced by the 2D image branch with enhanced fusion features $\mathbf{F}_C^{fe}$ as input. Therefore, as the intermediary between enhanced fusion features and 3D point cloud features, the 2D-Learner could effectively prevent the image modality from contaminating specific information on point cloud modality in the distillation process, while simultaneously providing rich color, texture, and semantic information for the point cloud modality.

In addition, the fusion branch is adopted only in the training stage, and the 2D image branch can be discarded in the inference stage. Compared with multi-modal-fusion-based methods, our method could process raw point clouds, avoiding the large blind area of image field-of-view, and effectively avoiding additional computational burdens in practical applications.

3.3. Confidence-Map-Driven Late Fusion Strategy

Through the above multi-modal knowledge distillation module, the 3D branch can learn additional semantic features from the 2D branch. However, the advantages of these features may not be fully reflected in the segmentation results, that is, the predicted results of the fusion methods are often not as good in some aspects as predictions based solely on images. For example, image-only segmentation methods have an absolute advantage in small target segmentation and object contour extraction in complex backgrounds, but the performance may decrease when fused with sparse point clouds.

Furthermore, influenced by the diversity of scene elements and the accuracy of sensor devices, the data-quality level of multi-modal fusion inputs is uneven. For example, cameras are susceptible to lighting interference, leading to phenomena such as image blur and overexposure; LiDAR is prone to the impact of weather conditions like rain or snow, resulting in a sharp decrease in the number of point clouds. This unevenness is also reflected in the output results of their respective modal branches. Therefore, when the image quality is low, it is advisable to rely more on the geometric and depth information from point clouds to mitigate the interference of erroneous color and texture information. Conversely, low-quality point clouds often struggle to accurately represent the spatial geometric information of the scenario.

In summary, the impact of different modalities on the prediction results should not be equal. Therefore, inspired by the idea of decision-level fusion, we propose a late fusion strategy based on confidence mapping. This strategy directly and explicitly weights the prediction results of the two modal branches, highlighting the respective advantages of each modality branch while mitigating the interference of erroneous data inputs, so as to output the final accurate and robust predictions. Specifically, a pixel-by-pixel confidence weight map is calculated using the probability of the predicted segmentation results, which is used to measure the reliability of the output segmentation results of each modality branch. For a segmentation network, its output consists of *Class* channels, each representing the probability that a pixel belongs to a particular category in the *Class* categories. The category with the highest probability is chosen as the final segmentation result. Generally, when the prediction for a pixel has an extremely high probability for a particular category and low probabilities for the others, this prediction can be considered with high confidence; conversely, when the probability distribution between categories is close to uniform, it indicates low confidence in the prediction for that pixel. Inspired by this, we designed the calculation of output confidence as follows:

$$\begin{gathered} \mathrm{Pr_f} = \mathrm{sigmoid}(\mathrm{W_{2D}} \odot \mathrm{Pr_{2D}} + \mathrm{W_{3D}} \odot \mathrm{Pr_{3D}}) \\ \mathrm{W_{2D}} = \sqrt{\frac{1}{\text{Class-1}} \sum_{\mathrm{i}=1}^{\text{Class}} \left(\mathrm{Pr_{2D}(i)\text{-}max\{Pr_{2D}(j)\ |j \in [1, Class]\}}\right)^2} \\ \mathrm{W_{3D}} = \sqrt{\frac{1}{\text{Class-1}} \sum_{\mathrm{i}=1}^{\text{Class}} \left(\mathrm{Pr_{3D}(i)\text{-}max\{Pr_{3D}(j)\ |j \in [1, Class]\}}\right)^2} \end{gathered} \tag{8}$$

where $\mathrm{W_{2D}}$ and $\mathrm{W_{3D}}$ represent the confidence maps of the 2D branch and 3D branch, respectively; $\mathrm{Pr_{2D}}$ and $\mathrm{Pr_{3D}}$ represent the segmentation results of the 2D branch and 3D branch, respectively; and $\mathrm{Pr_f}$ represents the final fusion results.

3.4. Joint Learning

In the optimization of the 2D branch, traditional supervised learning methods are not unsuitable due to the lack of dense image annotations. Consequently, we adopt the concept of transfer learning and introduce a 2D semantic discriminator $\mathrm{D_s}$ to differentiate between predicted semantic labels and ground truth (GT) semantic labels. Specifically, $\mathrm{D_s}$ incorporates both global and Markov discriminators, which enables the consideration of local texture information as well as ensuring global consistency. The adversarial loss $\mathrm{Loss_{2D}}$ can be formulated as follows:

$$\begin{gathered} \mathrm{Loss_{2D}} = \mathrm{L_{global}} + \mathrm{L_{Markov}} \\ \mathrm{L(D_S, Pr_f, GT_{2D})} = \mathrm{E_{pr}[\log(D_s(Pr_f))]} + \mathrm{E_{GT}[\log(1\text{ -}D_s(GT_{2D}))]} \end{gathered} \tag{9}$$

where $\mathrm{GT_{2D}}$ represents the sparse 2D GT semantic labels generated by projecting the corresponding 3D GT semantic labels using point-to-pixel mapping, $\mathrm{Pr_f}$ is the corresponding predicted probability, and E represents the expectation operation.

For 3D branch optimization, we combine the weighted cross-entropy loss and Lovasz-softmax loss to optimize the point cloud branch:

$$\mathrm{Loss_{3D}} = -\sum_{i} \frac{1}{\sqrt{v_i}} \mathrm{GT_{3D}(i)} \log \mathrm{Pr_{3D}(i)} + \frac{1}{\mathrm{Class}} \sum_{k} \mathrm{J(e(class_k))} \tag{10}$$

where v_i is the frequency of each category (the number of points in each category), GT_{3D} and Pr_{3D} are the GT and corresponding predicted probability, J is the Lovasz extension of IoU (Intersection-over-Union), and $e(class_k)$ is the vector of errors for category $class_k$.

We amalgamate the loss functions from the two segmentation branches and the distillation loss to optimize the entire network through end-to-end training, aiming to maximize the IoU index for each category. The final loss function can be formulated as follows:

$$\mathrm{Loss_{total}} = \mathrm{Loss_{2D}} + \mathrm{Loss_{3D}} + \mathrm{Loss_{dis}} \tag{11}$$

4. Experiments and Results

4.1. Dataset

To assess the effectiveness of our method at improving accuracy and robustness, we utilized the urban dataset SemanticKITTI and the off-road dataset RELLIS-3D, which provide diverse scenarios allowing us to comprehensively evaluate the performance of our method. The details are as follows:

SemanticKITTI is a widely used benchmark dataset for semantic segmentation tasks in autonomous driving. The dataset provides images and point clouds with semantic-level 3D annotations. It contains 19,130 frames for training, 4071 frames for validity, and 24,892 frames for testing. We treated train–valid–test sequences and 19 categories which are consistent with the benchmark algorithms.

RELLIS-3D was collected from three unpaved roads on the Texas A&M University RELLIS campus, containing images and point clouds with semantic-level 3D annotations. It contains 7800 frames for training, 2413 frames for validity, and 3343 frames for testing. We treated train–valid–test sequences and 14 categories which are consistent with the benchmark algorithms.

Additionally, we collected 100 frames of the unstructured scene, where outdoor parking lots and roads without clear road boundaries or lacking marking lines are considered as unstructured scenes. The test set was gathered using a vehicle equipped with a Velodyne 32-line LiDAR and a forward-view monocular camera for visualization. Notably, the two sensors were synchronized in time, but external parameter calibration was not conducted. The lack of sensor calibration matrices disabled the implement of fusion-based methods, and the sparsity in beam numbers simulated the fault of LiDAR inputs. These distinctive characteristics make our unstructured test set particularly suitable for testing the segmentation performance and robustness of these methods.

4.2. Implement Details

Cross-modal knowledge distillation was applied to the middle and the last layer of the encoders. We set the spatial-attention-related hyper-parameters as $\tau = 0.5$, referring to [34,52,53], and the loss related hyper-parameters as $\lambda = 2.5 \times 10^{-3}$, referring to [34,54].

Our network was trained for 50 epochs with a batch size of 16. We utilized stochastic gradient descent (SGD) as the optimizer with a weight decay of 0.001, a momentum of 0.9, and an initial learning rate of 0.02. All experiments were on NVIDIA RTX 3090 GPUs.

We verified the performance of the proposed methodology using the common evaluation index (IoU and mIoU) in semantic segmentation tasks.

4.3. Comparative Results and Discussion of SemanticKITTI

We compared the results of our proposed method with typical and representative LiDAR segmentation methods on the SemanticKITTI benchmark. To be specific, RandLA-

Net and KPConv were on behalf of SOTA point-based methods, while SPVNAS and Cylinder3D were on behalf of voxel-based methods; PolarNet and SalsaNext were on behalf of single-projection-based methods; and MPF, GFNet, AMVNet, and our 3D Branch were on behalf of multi-projection-based methods. These methods are top algorithms in their respective fields on the SemanticKITTI benchmark. Moreover, three representative and open-access LiDAR–camera fusion segmentation methods (RGBAL, xMUDA, and PMF) were used as the comparison.

The quantitative comparison results are shown in Tables 1 and 2, where the **bold** numbers indicate the best results, and the **green bold** numbers indicate the second-best results.

Table 1. Results and time of typical LiDAR-only methods and our method on SemanticKITTI test set.

Methods	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign	mIoU (%)	Time (ms)
RandLA-Net	94.2	26.0	25.8	40.1	38.9	49.2	48.2	7.2	90.7	60.3	73.7	20.4	86.9	56.3	81.4	61.3	66.8	49.2	47.7	53.9	769
KPConv	96.0	30.2	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	31.6	**95.0**	64.2	84.8	69.2	69.1	56.4	47.4	58.8	263
SPVNAS	**97.3**	51.5	50.8	**59.8**	**58.8**	65.7	65.2	43.7	90.2	67.6	75.2	16.9	91.3	65.9	**86.1**	**73.4**	**71**	64.2	66.9	66.4	256
Cylinder3D	**97.1**	**67.6**	**64.0**	**59.0**	58.6	**73.9**	**67.9**	36.0	91.4	65.1	75.5	32.3	91.0	66.5	85.4	71.8	68.5	62.6	65.6	**67.8**	179
PolarNet	93.8	40.3	30.1	22.9	28.5	43.2	40.2	5.6	90.8	61.7	74.4	21.7	90.0	61.3	84.0	65.5	67.8	51.8	57.5	54.3	63
SalsaNext	91.9	48.3	38.6	38.9	31.9	60.2	59.0	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1	59.5	**42**
MPF	93.4	30.2	38.3	26.1	28.5	48.1	46.1	18.1	90.6	62.3	74.5	30.6	88.5	59.7	83.5	59.7	69.2	49.7	58.1	55.5	**35**
GFNet	96.0	53.2	48.3	31.7	47.3	62.8	57.3	**44.7**	93.6	72.5	**80.8**	31.2	94.0	**73.9**	85.2	71.1	69.3	61.8	68.0	65.4	100
AMVNet	96.2	59.9	54.2	48.8	45.7	**71.0**	65.7	11.0	90.1	71.0	75.8	32.4	92.4	69.1	85.6	71.7	**69.6**	62.7	67.2	65.3	-
Our 3D Branch	96.4	54.3	51.2	48.1	49.1	65.3	64.8	36.2	**93.8**	**74.5**	78.2	**33.7**	93.2	70.6	**86.2**	72.1	69.0	**65.2**	**69.3**	66.9	83
Ours	**97.1**	**66.8**	**58.8**	54.3	**59.0**	66.4	**66.3**	**80.0**	**93.7**	**79.8**	**79.1**	**40.2**	**95.1**	**74.8**	85.9	**73.7**	68.7	**66.5**	**69.5**	**72.4**	83

Table 2. Results and time of typical LiDAR–camera fusion methods and our method on SemanticKITTI validity set, where we only used point clouds in camera field-of-view.

Methods	Car	Bicycle	Motorcycle	Truck	Other Vehicle	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other Ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic Sign	mIoU (%)	Time (ms)
RGBAL	87.9	36.8	26.4	63.8	54.6	58.0	**72.0**	0.0	94.6	43.6	77.5	**0.7**	78.9	52.7	84.3	60.8	**72.9**	56.1	39.5	55.9	**12**
xMUDA	93.2	11.9	33.3	**72.6**	51.8	68.0	61.3	0.0	**95.7**	36.4	78.6	0.1	86.2	**57.1**	78.7	**69.0**	**74.5**	59.5	37.2	56.1	33
PMF	**94.6**	**49.0**	**62.0**	68.2	**72.5**	**68.7**	66.1	0.0	**96.3**	**44.3**	**80.4**	0.0	**87.8**	56.8	**87.6**	**71.2**	72.7	**64.8**	**43.5**	**62.4**	27
Ours	**97.5**	**40.4**	**77.4**	**94.2**	**73.4**	**73.3**	**93.4**	0.0	94.1	**51.1**	**82.1**	**0.2**	**91.6**	**64.9**	**87.2**	67.3	71.7	**65.9**	**49.0**	**67.1**	**21**

It is obvious that our method outperformed all the methods in terms of mIoU; to be specific, there was a 4.6% improvement from the best LiDAR-only method, Cylinder3D, and 4.7% from the best fusion-based method, PMF. Moreover, our method could still guarantee the real-time performance, since an 83 ms processing time is less than 100 ms (calculated from 10 Hz LiDAR collection frequency).

When compared with LiDAR-only methods (see Table 1), our method achieved the best performance in 9 of all 19 categories and the second-best performance in 6 of all categories. This outperformance shows that our method can effectively merge comprehensive multi-modal features (including the dense semantic feature information of images) into point clouds to make up for the deficiency of the performance, especially in those categories with sparse features (e.g., motorcyclist, fence, trunk, pole, traffic sign) or similar geometric features (e.g., building, parking, sidewalk), which cannot be effectively distinguished solely using LiDAR geometric features. The same conclusions can be drawn from Figure 6 with fewer error points, where correct/incorrect predictions are painted in gray/red, respectively, to highlight the differences.

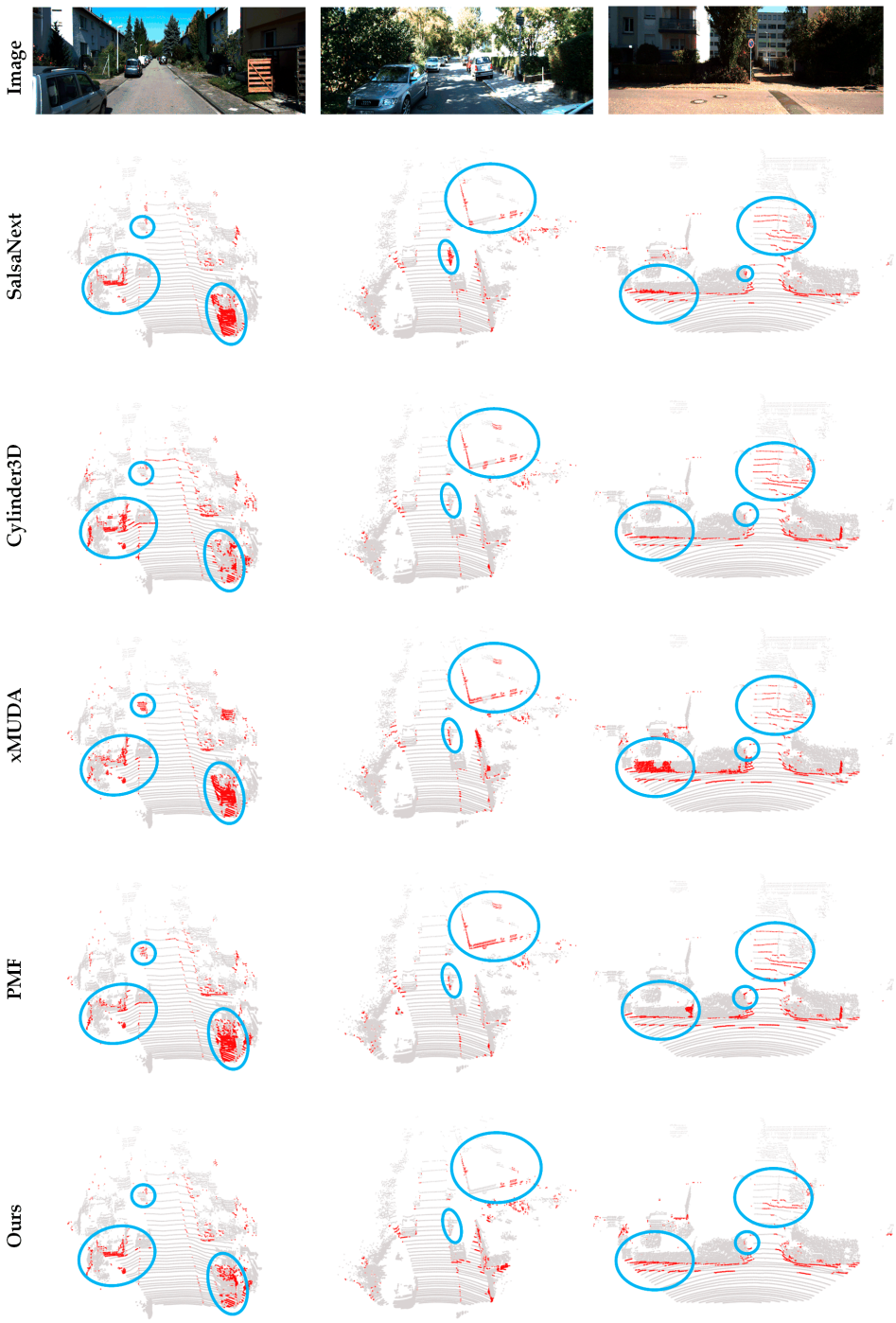

Figure 6. Qualitative comparisons with SOTA methods on SemanticKITTI.

When compared with fusion methods (see Table 2), our multi-modal collaborative learning method can effectively integrate multi-modal features and eliminate the extreme sensitivity of image modality to complex and variable environments (e.g., diverse illumination intensity, similar color textures), resulting in the best performance in most categories (12 of all 19 categories). Additionally, the absence of dense GT labels for 2D semantic segmentation renders fusion-based methods less adept at distinguishing small, irregular objects. Optimization based on transfer learning compensates for this limitation, achieving a 1.1~21.4% improvement from the second-best performance in those small-scale categories like person, bicyclist, fence, pole, and traffic sign. The same conclusions can be drawn from Figure 6 with fewer error points. For example, the white car/wall in the sunlight

and objects in the shadow were misclassified by fusion-based methods, while our method could accurately distinguish them.

4.4. Comparative Results and Discussion of RELLIS-3D

The models were also evaluated on RELLIS-3D and the comparison results on the test set are shown in Table 3.

Table 3. Results and time of typical methods and our method on RELLIS-3D. * indicates results using point clouds in camera field-of-view.

Methods	Grass	Tree	Pole	Water	Vehicle	Log	Person	Fence	Bush	Concrete	Barrier	Puddle	Mud	Rubble	mIoU (%)	Time (ms)
RandLA-Net	62.1	76.7	42.8	0	34.8	6.3	82.9	10.2	69.8	72.8	54.3	17.6	8.5	1.7	38.6	769
Cylinder3D	64.9	76.5	**63.8**	0	**50.3**	5.5	**87.0**	11.8	71.8	**80.4**	**80.5**	**33.0**	10.6	2.3	**45.6**	149
SPVNAS	64.1	76.1	54.4	0	**48.6**	15.1	85.8	11.0	71.5	70.0	64.8	22.4	8.5	5.2	42.7	167
SalsaNext	65.3	79.6	44.3	0	26.6	22.2	84.4	13.5	**73.6**	77.9	63.7	26.0	11.6	6.6	42.5	25
GFNet	64.3	76.4	49.8	0	34.4	21.5	83.1	12.6	72.4	73.5	79.3	24.2	10.1	7.2	43.6	83
Our 3D Branch	**66.1**	**80.3**	50.4	0	32.4	**24.6**	85.7	**15.2**	72.4	73.8	78.3	24.6	10.4	**9.2**	44.5	71
RGBAL *	63.7	68.7	49.9	0	28.9	12.4	82.3	6.3	72.9	70.2	64.2	23.7	10.9	6.8	40.1	11
xMUDA *	64.4	71.6	54.3	0	23.3	15.1	82.7	6.0	73.3	70.8	64.2	24.4	11.2	5.4	40.5	29
PMF *	65.9	80.1	60.4	0	26.4	12.8	85.4	10.4	**73.5**	**82.0**	67.8	23.6	**11.9**	7.6	43.4	28
Ours	**66.8**	**80.7**	**66.3**	0	31.2	**27.3**	**86.5**	**19.6**	**73.5**	**82.0**	**81.6**	**28.6**	**15.7**	**9.6**	**47.8**	71

It is obvious that there were sharp decreases in the IoU and mIoU for the existing methods due to the dataset's complexity and the similarity of available feature information. The reasons are as follows: When feature information in the dataset is highly similar, it means that distinguishing characteristics among different objects or regions might be subtle. If the features that define different categories are not well discriminated, the methods may struggle to precisely delineate object boundaries, thus making it challenging for them to accurately differentiate between categories, leading to lower IoU and mIoU scores.

However, our method still excelled in overall performance, benefitting from the effectiveness of the proposed multi-modal collaborative learning approach in combining comprehensive multi-modal features. Specifically, it outperformed the best LiDAR-only method, Cylinder3D, by 3.2%, and the best fusion-based method, PMF, by 4.4%.

Furthermore, our method exhibited superior performance in 9 out of 14 classes, particularly excelling in small objects (e.g., pole, log, and fence) and classes with similar geometric features (e.g., grass, concrete, mud, and rubble), where LiDAR point features are typically insufficient. This conclusion is further supported by the qualitative comparison results shown in Figure 7.

4.5. Comparative Results and Discussion of Our Test Set

To demonstrate the robustness and adaptivity of our method in various and complex scenarios, especially with LiDAR faults, we performed extra experiments on our unstructured test set with sparse LiDAR data inputs. We tested Cylinder3D (voxel-based), SalsaNext (single-projection-based), and our 3D Branch (multi-projection-based) as representative methods.

The results show that the models are most severely affected by sparse inputs; specifically, SalsaNext was the most affected because its predictions are mainly determined by the size of dense range image inputs, while the impact to Cylinder3D was relatively low since it is specially optimized for sparse point clouds. However, none of these methods could extract accurate objects and continuous flats. In contrast, our method could accurately classify the cars and performed better on drivable areas, e.g., the predicted drivable area in these methods was either discontinuous or misclassified as vegetation (see the red circles for each scenario in Figure 8). This validates that our method can effectively integrate the advantages of both modalities and achieve the best performance in robustness evaluation.

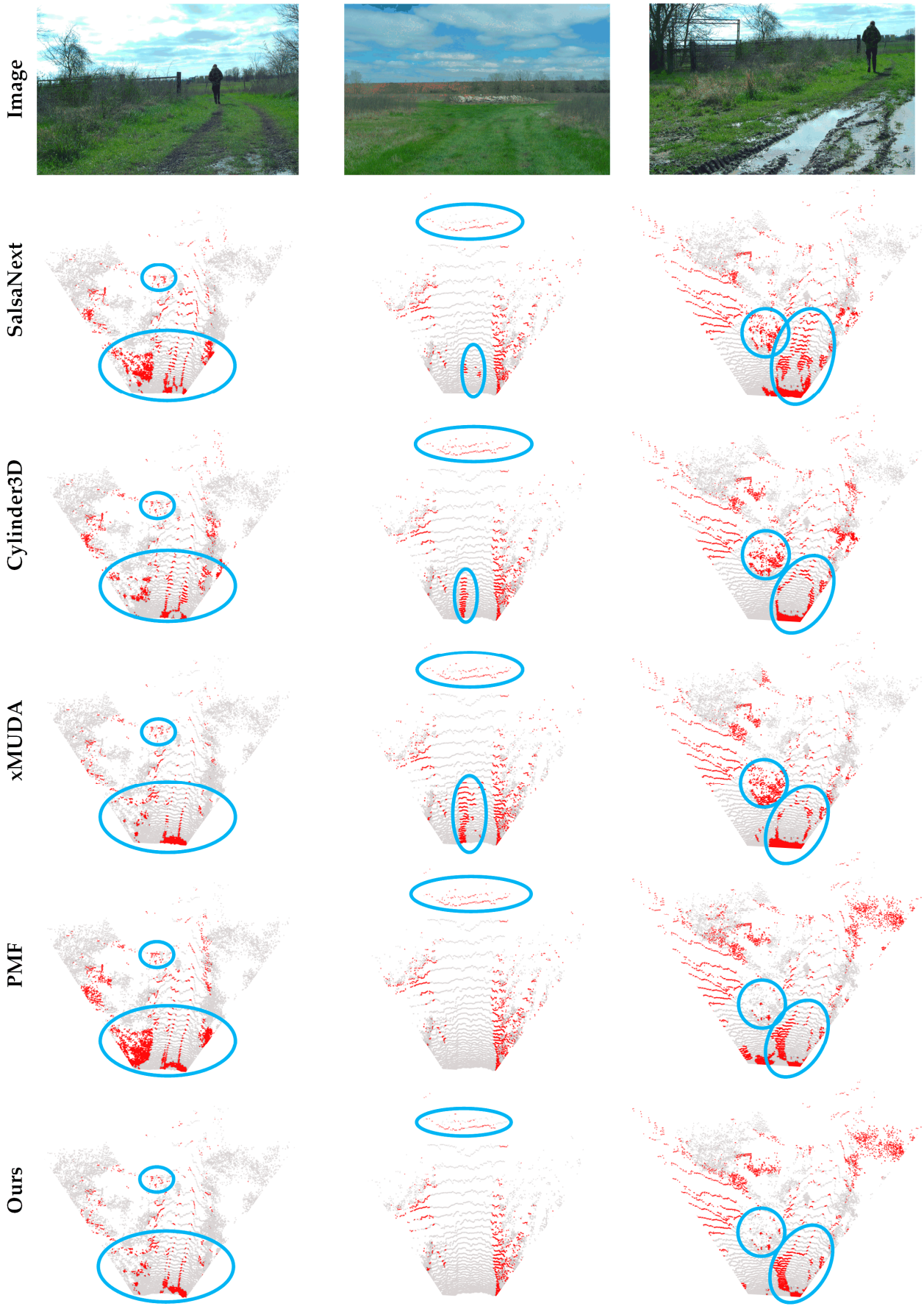

Figure 7. Qualitative comparisons with SOTA methods on RELLIS-3D.

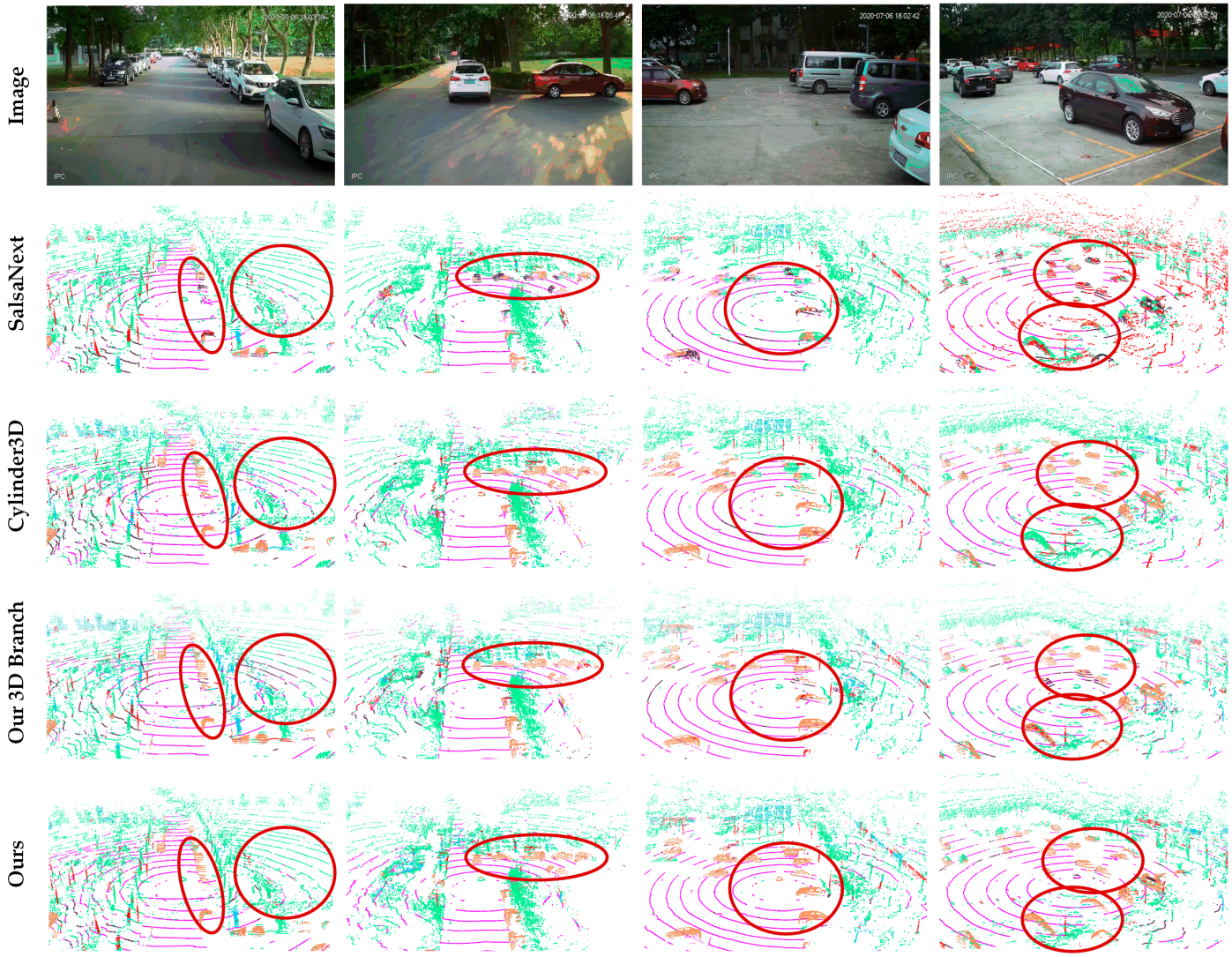

Figure 8. Sample results of representative models on our test set.

4.6. Ablation Study

We performed thorough ablation experiments to investigate the contribution of each component of our method, including the effects of cross-modal knowledge distillation module ("CMKD") and the late fusion strategy ("LF"). We used the same parameters to train these methods. The results are shown in Table 4.

Table 4. Results of ablation studies for network components on SemanticKITTI validity set.

Row	Module					mIoU (%)
	3D Branch	CMKD			LF	
		FA	F and D	AM		
1	√					63.0
2	√	√				62.8
3	√	√	D			63.7
4	√	√	√			65.4
5	√	√	√	√		66.0
6	√	√	√	√	E	66.6
7	√	√	√	√	√	67.1

4.6.1. Effects of Cross-Modal Knowledge Distillation Module

Comparing the first two rows in Table 4, it is evident that the direct introduction of image features after feature alignment ("FA") had a negative impact on the overall performance, resulting in a 0.2% decrease. This indicates that certain image features may interfere with the 3D branch, providing further evidence of the necessity for knowledge distillation.

Comparing the first and fourth rows, a notable improvement of 2.4% was achieved after employing fusion and distillation. This improvement primarily stemmed from the knowledge provided by the more robust fusion prediction. Moreover, comparing the third and fourth rows, the simple distillation ("D") between two modalities without the fusion approach led to a 1.7% mIoU decrease. This is attributed to the contamination between modalities resulting from straightforward distillation, akin to traditional distillation methods.

Comparing the fourth and fifth rows, it is apparent that the attentional map ("AM") improved the mIoU by 0.6% through simple channel concatenation, which validates its effectiveness.

4.6.2. Effects of Late Fusion Strategy

Comparing the fifth and seventh rows, the incorporation of the confidence-map-driven late fusion strategy harnessed the strengths of both the image and point cloud for different categories, resulting in a 1.1% improvement in the mIoU.

Furthermore, comparing the sixth and seventh rows, "LF" achieved a 0.5% higher mIoU compared to the traditional equal weight-based late fusion method ("E"). This improvement highlights the superiority of the "LF" method in autonomously addressing and complementing the advantages and disadvantages of each modality branch.

5. Discussion

Despite the above advantages, there are still areas where our work can be improved. The IoU of our method is not the best when it comes to segmenting some objects with richer 3D geometric information, such as cars, bicycles, motorcycles, and people. This is because the projection-based method loses some of the 3D spatial information. Additionally, the present method relies on high-end GPUs to achieve real-time performance. Therefore, our future work will focus on further optimizing the segmentation accuracy of these categories by introducing point-based or voxel-based branches, as well as optimizing the proposed method from the perspective of model compression and single modality knowledge distillation for applications in resource-constrained intelligent vehicles.

6. Conclusions

This paper introduces a robust 3D semantic segmentation method based on multi-modal collaborative learning and addresses the challenges that impede the performance of fusion-based 3D semantic segmentation methods. The proposed attention-based cross-modal knowledge distillation module leverages attentional fusion to selectively integrate multi-modal features and utilizes feature distillation to enrich 3D point cloud features via 2D image priors. The confidence-map-driven late fusion strategy dynamically assigns weights for both branches to accentuate the strengths of each modality. Through the integration of these modules, our method is capable of acquiring richer semantic and geometric information from multi-modal data, thereby effectively enhancing the performance and robustness of a pure LiDAR semantic segmentation network.

We evaluated our proposed method on three datasets: the urban dataset SemanticKITTI, the off-road dataset RELLIS-3D, and our self-created unstructured test set. Extensive experiments showed that the proposed method is competitive with state-of-the-art methods in diverse scenarios and is more robust to sensor fault conditions. The ablation experiments served to further validate the contributions of our designed modules.

Author Contributions: Conceptualization, P.N. and X.L.; methodology, P.N.; software, P.N. and W.X.; validation, P.N.; formal analysis, P.N. and W.X.; investigation, P.N. and X.Z.; resources, P.N. and T.J.; data curation, P.N. and W.H.; writing—original draft preparation, P.N.; writing—review and editing, P.N.; supervision, X.L.; project administration, X.L.; funding acquisition, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China under Grant 2022YFC3002603; in part by the Primary Research & Development Plan of Jiangsu Province, grant number BE2022053-5; in part by the National Natural Science Foundation of China, grant number 61973079, and in part by the Collective Intelligence & Collaboration Laboratory, grant number QXZ23012201.

Data Availability Statement: The SemanticKITTI Dataset used in this study is available at http://semantic-kitti.org/dataset.html (accessed on 13 May 2021); The RELLIS-3D Dataset employed in this study is available at https://github.com/unmannedlab/RELLIS-3D (accessed on 8 January 2022).

Conflicts of Interest: Author T.J. was employed by the company Xuzhou XCMG Automobile Manufacturing Co., Ltd. Author W.H. was employed by the company China Automotive Engineering Research Institute Company Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Kong, D.; Li, X.; Hu, Y.; Xu, Q.; Wang, A.; Hu, W. Learning a novel LiDAR submap-based observation model for global positioning in long-term changing environments. *IEEE Trans. Ind. Electron.* **2022**, *70*, 3147–3157. [CrossRef]
2. Richa, J.P.; Deschaud, J.-E.; Goulette, F.; Dalmasso, N. AdaSplats: Adaptive Splatting of Point Clouds for Accurate 3D Modeling and Real-Time High-Fidelity LiDAR Simulation. *Remote Sens.* **2022**, *14*, 6262. [CrossRef]
3. Li, Y.; Ma, L.; Zhong, Z.; Liu, F.; Chapman, M.A.; Cao, D.; Li, J. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 3412–3432. [CrossRef]
4. Zhao, L.; Zhou, H.; Zhu, X.; Song, X.; Li, H.; Tao, W. Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. *IEEE Trans. Multimed.* **2023**, *26*, 1158–1168. [CrossRef]
5. Zhao, J.; Wang, Y.; Cao, Y.; Guo, M.; Huang, X.; Zhang, R.; Dou, X.; Niu, X.; Cui, Y.; Wang, J. The fusion strategy of 2D and 3D information based on deep learning: A review. *Remote Sens.* **2021**, *13*, 4029. [CrossRef]
6. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 722–739. [CrossRef]
7. El Madawi, K.; Rashed, H.; El Sallab, A.; Nasr, O.; Kamel, H.; Yogamani, S. Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 7–12.
8. Zhang, R.; Li, G.; Li, M.; Wang, L. Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *143*, 85–96. [CrossRef]
9. Lee, J.-S.; Park, T.-H. Fast road detection by cnn-based camera–lidar fusion and spherical coordinate transformation. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 5802–5810. [CrossRef]
10. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4604–4612.
11. Xu, S.; Zhou, D.; Fang, J.; Yin, J.; Bin, Z.; Zhang, L. Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 3047–3054.
12. Fang, F.; Zhou, T.; Song, Z.; Lu, J. MMCAN: Multi-Modal Cross-Attention Network for Free-Space Detection with Uncalibrated Hyperspectral Sensors. *Remote Sens.* **2023**, *15*, 1142. [CrossRef]
13. Zhuang, Z.; Li, R.; Jia, K.; Wang, Q.; Li, Y.; Tan, M. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 16280–16290.
14. Valada, A.; Mohan, R.; Burgard, W. Self-supervised model adaptation for multimodal semantic segmentation. *Int. J. Comput. Vis.* **2020**, *128*, 1239–1285. [CrossRef]
15. Schieber, H.; Duerr, F.; Schoen, T.; Beyerer, J. Deep Sensor Fusion with Pyramid Fusion Networks for 3D Semantic Segmentation. In Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV), Aachen, Germany, 5–9 June 2022; pp. 375–381.
16. Jaritz, M.; Vu, T.-H.; Charette, R.D.; Wirbel, E.; Pérez, P. Xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12605–12614.

17. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9297–9307.
18. Jiang, P.; Osteen, P.; Wigness, M.; Saripalli, S. Rellis-3d dataset: Data, benchmarks and analysis. In Proceedings of the 2021 IEEE International Conference on robotics and AUTOMATION (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 1110–1116.
19. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
20. Thomas, H.; Qi, C.R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. Kpconv: Flexible and deformable convolution for point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6411–6420.
21. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
22. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
23. Zhao, L.; Xu, S.; Liu, L.; Ming, D.; Tao, W. SVASeg: Sparse voxel-based attention for 3D LiDAR point cloud semantic segmentation. *Remote Sens.* **2022**, *14*, 4471. [CrossRef]
24. Zhu, Z.; Li, X.; Xu, J.; Yuan, J.; Tao, J. Unstructured road segmentation based on road boundary enhancement point-cylinder network using LiDAR sensor. *Remote Sens.* **2021**, *13*, 495. [CrossRef]
25. Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; Han, S. Searching efficient 3d architectures with sparse point-voxel convolution. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 685–702.
26. Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Li, W.; Ma, Y.; Li, H.; Yang, R.; Lin, D. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6807–6822. [CrossRef] [PubMed]
27. Zhang, C.; Luo, W.; Urtasun, R. Efficient convolutions for real-time semantic segmentation of 3d point clouds. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 399–408.
28. Zhang, Y.; Zhou, Z.; David, P.; Yue, X.; Xi, Z.; Gong, B.; Foroosh, H. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9601–9610.
29. Xian, G.; Ji, C.; Zhou, L.; Chen, G.; Zhang, J.; Li, B.; Xue, X.; Pu, J. Location-guided lidar-based panoptic segmentation for autonomous driving. *IEEE Trans. Intell. Veh.* **2022**, *8*, 1473–1483. [CrossRef]
30. Cortinhal, T.; Tzelepis, G.; Erdal Aksoy, E. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In Proceedings of the Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, 5–7 October 2020; Part II 15; pp. 207–222.
31. Yin, X.; Li, X.; Ni, P.; Xu, Q.; Kong, D. A Novel Real-Time Edge-Guided LiDAR Semantic Segmentation Network for Unstructured Environments. *Remote Sens.* **2023**, *15*, 1093. [CrossRef]
32. Chen, T.-H.; Chang, T.S. RangeSeg: Range-aware real time segmentation of 3D LiDAR point clouds. *IEEE Trans. Intell. Veh.* **2021**, *7*, 93–101. [CrossRef]
33. Xu, W.; Li, X.; Ni, P.; Guang, X.; Luo, H.; Zhao, X. Multi-View Fusion Driven 3D Point Cloud Semantic Segmentation Based on Hierarchical Transformer. *IEEE Sens. J.* **2023**, *23*, 31461–31470. [CrossRef]
34. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
35. Hou, Y.; Ma, Z.; Liu, C.; Loy, C.C. Learning to steer by mimicking features from heterogeneous auxiliary networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8433–8440.
36. Hu, J.; Zhao, M.; Li, Y. Hyperspectral image super-resolution by deep spatial-spectral exploitation. *Remote Sens.* **2019**, *11*, 1229. [CrossRef]
37. Hou, Y.; Ma, Z.; Liu, C.; Loy, C.C. Learning lightweight lane detection cnns by self attention distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1013–1021.
38. Chai, Y.; Fu, K.; Sun, X.; Diao, W.; Yan, Z.; Feng, Y.; Wang, L. Compact cloud detection with bidirectional self-attention knowledge distillation. *Remote Sens.* **2020**, *12*, 2770. [CrossRef]
39. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4133–4141.
40. Tung, F.; Mori, G. Similarity-preserving knowledge distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1365–1374.
41. Park, W.; Kim, D.; Lu, Y.; Cho, M. Relational knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3967–3976.
42. Liu, Y.; Chen, K.; Liu, C.; Qin, Z.; Luo, Z.; Wang, J. Structured knowledge distillation for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2604–2613.
43. He, T.; Shen, C.; Tian, Z.; Gong, D.; Sun, C.; Yan, Y. Knowledge adaptation for efficient semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 578–587.

44. Hou, Y.; Zhu, X.; Ma, Y.; Loy, C.C.; Li, Y. Point-to-voxel knowledge distillation for lidar semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8479–8488.
45. Wang, L.; Wu, J.; Huang, S.-L.; Zheng, L.; Xu, X.; Zhang, L.; Huang, J. An efficient approach to informative feature extraction from multimodal data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 5281–5288.
46. Zhao, L.; Peng, X.; Chen, Y.; Kapadia, M.; Metaxas, D.N. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6528–6537.
47. Liu, Z.; Qi, X.; Fu, C.-W. 3d-to-2d distillation for indoor scene parsing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4464–4474.
48. Liu, Y.-C.; Huang, Y.-K.; Chiang, H.-Y.; Su, H.-T.; Liu, Z.-Y.; Chen, C.-T.; Tseng, C.-Y.; Hsu, W.H. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *arXiv* **2021**, arXiv:2104.04687.
49. Xu, C.; Yang, S.; Galanti, T.; Wu, B.; Yue, X.; Zhai, B.; Zhan, W.; Vajda, P.; Keutzer, K.; Tomizuka, M. Image2point: 3d point-cloud understanding with 2d image pretrained models. *arXiv* **2021**, arXiv:2106.04180.
50. Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; Zhao, F. Bevdistill: Cross-modal bev distillation for multi-view 3d object detection. *arXiv* **2022**, arXiv:2211.09386.
51. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
52. Zhang, L.; Ma, K. Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
53. Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; Yuan, C. Focal and Global Knowledge Distillation for Detectors. *arXiv* **2021**, arXiv:2111.11837.
54. Dai, X.; Jiang, Z.; Wu, Z.; Bao, Y.; Zhou, E. General Instance Distillation for Object Detection. *arXiv* **2021**, arXiv:2103.02340.