remote sensing

MDPI

Review

The Fusion Strategy of 2D and 3D Information Based on Deep Learning: A Review

Jianghong Zhao [1,2,3,4,5], Yinrui Wang [2,4,*], Yuee Cao [6], Ming Guo [2,4], Xianfeng Huang [5], Ruiju Zhang [2,4], Xintong Dou [2,4], Xinyu Niu [2,4], Yuanyuan Cui [2,4] and Jun Wang [7]

1 State Key Laboratory of Geo-Information Engineering, Xi'an 710054, China; zhaojiangh@bucea.edu.cn
2 School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; guoming@bucea.edu.cn (M.G.); zhangruiju@bucea.edu.cn (R.Z.); 2108160219006@stu.bucea.edu.cn (X.D.); 2108570020080@stu.bucea.edu.cn (X.N.); 2108570020088@stu.bucea.edu.cn (Y.C.)
3 Key Laboratory of Modern Urban Surveying and Mapping, National Administration of Surveying, Mapping and Geoinformation, Beijing 102616, China
4 Beijing Key Laboratory for Architectural Heritage Fine Reconstruction & Health Monitoring, Beijing 102616, China
5 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430072, China; huangxf@whu.edu.cn
6 School of Environment and Geographical Sciences, Shanghai Normal University, Shanghai 200234, China; caoyuee@shnu.edu.cn
7 Culture Development Research Institute, School of Humanities, Beijing University of Civil Engineering and Architecture, Beijing 102616, China; wangjun@bucea.edu.cn
* Correspondence: 2108521519005@stu.bucea.edu.cn

Citation: 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. https://doi.org/10.3390/rs13204029

Academic Editors: Domenico Visintini and Filiberto Chiabrando

Received: 25 August 2021
Accepted: 1 October 2021
Published: 9 October 2021

Abstract: Recently, researchers have realized a number of achievements involving deep-learning-based neural networks for the tasks of segmentation and detection based on 2D images, 3D point clouds, etc. Using 2D and 3D information fusion for the advantages of compensation and accuracy improvement has become a hot research topic. However, there are no critical reviews focusing on the fusion strategies of 2D and 3D information integration based on various data for segmentation and detection, which are the basic tasks of computer vision. To boost the development of this research domain, the existing representative fusion strategies are collected, introduced, categorized, and summarized in this paper. In addition, the general structures of different kinds of fusion strategies were firstly abstracted and categorized, which may inspire researchers. Moreover, according to the methods included in this paper, the 2D information and 3D information of different methods come from various kinds of data. Furthermore, suitable datasets are introduced and comparatively summarized to support the relative research. Last but not least, we put forward some open challenges and promising directions for future research.

Keywords: fusion strategy; deep learning; segmentation; detection

1. Introduction

Thanks to the rapid development of deep learning [1,2] and various sensors, the techniques of the real-world scene sensing, analysis, and management are improved constantly, which potentially boosts the development of autonomous driving [3], robotic [4], remote sensing [5], medical science [6,7], the internet of things [8], etc. Therefore, the task of segmentation [9,10] and detection [11,12], the basic tasks of scene understanding, have achieved great improvements recently. However, the disadvantages of the methods based on a single kind of data have emerged gradually. For example, the image includes abundant 2D texture information but fails to represent the geometric information. Even though the 3D information can be acquired by post-processing the data collected by some mature visual-based perceptions, such as the mono- [13] and stereo camera [14], the quality of the

geometric information are not reliable enough and the sensors always struggle with the light and weather conditions. Therefore, various 3D scanners have emerged to acquire more reliable geometric information. For example, LiDAR [15,16] is known as the long range of detecting, accuracy, and the robustness to different light and weather conditions. However, the point cloud struggles with the lack of fine texture due to the irregularity and the sparsity. To make the full use of the complementary feature of 2D and 3D data, information fusion is worth more research. The main challenges of 2D and 3D information fusion based on deep learning techniques are the complex correspondence between different data and their feature maps, and the suitable fusion strategy.

Since the development of the sensor technologies, a range of methodologies have emerged to make full use of the data collected by different sensors. Therefore, several papers have been presented to review the methods that achieve multi-modal data fusion. Zhang et al. [17] published a review in 2016 introducing some methodologies that integrate the optical imagery and LiDAR point cloud for various applications, such as registration, orthophotographs, pan-sharpening, classification, recognition, etc. Wang et al. [18] have reviewed and discussed some strategies of integrating data acquired by the radar, LiDAR, camera, ultrasonic, GPS, IMU, and V2X for automatic driving. Moreover, Debeunne et al. [19] reviewed a series of hybridized solutions of visual-LiDAR SLAM for performance improvement. With the rapid development of deep learning techniques, there are two reviews aiming to introduce the current situation of the method that achieve 2D and 3D information fusion based on deep learning. Fayyad et al. [20] review the deep-learning-based multi-sensor fusion methods for the autonomous vehicle. Not only the environmental perception but also the localization system, such as GNSS, INS, IMU, RGB camera, thermal camera, LiDAR, and radar, are integrated for the object detection and tracking. Additionally, Cui et al. [21] have reviewed the camera-LiDAR fusion methods for depth completion, object detection, semantic segmentation, object tracking, and online cross-sensor calibration based on deep learning techniques in the complex and dynamic driving environment.

Compared with the previous reviews, the contributions of this paper are summarized in the following lines:

(1) This is the first comprehensive review of deep-learning-based fusion strategies that integrate the 2D and 3D information for segmentation and detection.
(2) Providing a novel taxonomy for the fusion strategies categorization.
(3) Including the suitable datasets as comprehensively as possible, which covers the RGBD datasets, fine-grained 3D model dataset, and the dataset including the registered point cloud and images.
(4) Including the most up-to-date (2004–2021) methods and their comparative summaries.
(5) Providing open challenges and promising directions for future research.

This paper aims to introduce up-to-date and representative fusion strategies for 2D and 3D information fusion based on deep learning. In Section 2, we briefly introduce the basic terminologies and background of the deep-learning techniques for classification, detection, and segmentation. In Section 3, the representative datasets and their data details are introduced with comparative analysis, which may help the researchers to select the suitable ones. In Section 4, we introduce and comparatively analyze the fusion strategies for 2D and 3D information integration based on deep learning techniques. Additionally, we gather and analyze the shared codes of those fusion strategies on GitHub. In Section 5, we summarize the popular trends and put forward some open challenges and promising future directions to researchers for reference. In Section 6, a brief conclusion of this paper is delivered.

2. Terminology and Background

Some classical networks may become the backbone or play an important role on 2D and 3D feature learning, respectively, in the existing fusion strategies. Therefore, it is necessary to figure out how those classical methods work and review their terminologies and background. Therefore, we will briefly introduce the classical deep learning techniques in this section.

2.1. Deep Learning Based on Image

2.1.1. CNN-Based Image Classification

With the constant development of deep learning techniques, the neural network has become one of the most outstanding achievements. Since most of the neural networks consist of the input layer, hidden layer, and output layer, the convolutional neural network (CNN) utilizes the convolution as the feature learning operator in a hidden layer. Even though the CNN-based networks presented in these years have become more and more complex and manage to obtain more and more accurate results, some classical networks are still the backbone or footstone of the recently proposed networks. Therefore, those classical networks are worth being briefly introduced in the following lines. The LeNet [22] is known as the beginning of the development of CNN networks. It consists of the convolutional layers, pooling layers, and fully connected layers, which are all basic modules of following networks. Thanks to the successful application of GPU, ReLU, dropout, max pooling, and LRN, the AlexNet [23] achieves the top-five best performance in ILSVRC-2012. The GoogLeNet [24] had the best and the VGG [25] the second-best performance in ILSVRC-2014; both of them play important roles for the following proposed methods. The GoogLeNet is known for the multi-scale feature learning by multi-scale convolution operations. The VGG modifies the AlexNet and utilizes the combination of 3×3 convolution and 2×2 maxpooling to broaden the receptive fields and increase the number of channels layer by layer. Although the VGG did not win the first, its basic structure becomes one of the most important backbones of the feature extractor. For example, the fully convolution network (FCN) [26] takes the backbone of VGG as the feature extractor to become one of the most important networks for the task of segmentation. In ILSVRC-2015, ResNet [27] was first because of its residual blocks, which successfully overcame the degradation problem and enlarged the depth of the network. The residual blocks ensure that each layer will learn new features that are different from the input feature map.

2.1.2. Detection

Compared with classification, which only needs to extract the global feature of the holistic input, the task of detection needs not only the class information but also the spatial location and the size of the bounding box. Moreover, the object detection methods can be grouped into the region-proposal-based methods and the regression/classification-based methods.

1. Region-proposal-based Method

The pipeline of the region-proposal-based method is generating the region proposals of the objects first, then refining the proposals as the bounding boxes, finally predicting the category of each bounding box.

Fast RCNN [28] first takes the selective search to generate series of regions of interest (RoI) for each images. It then learns the ROI-wise features to achieve the categorizing and bounding box regression at the same time. However, the computation of generating the regional proposals is complex. To break the bottleneck of efficiency, the region proposal network (RPN) is introduced in Faster RCNN [29]. The RPN effectively improves the efficiency by predefining k different sizes of anchor boxes. Based on Faster RCNN, the FCN is selected as the class score generator of R-FCN [30] for a better performance. In addition, the feature pyramid (FP) is proposed for the multi-scale feature extraction in FPN [31] by sacrificing the efficiency and memory consumption. Additionally, the Mask R-CNN [32]

achieves instant segmentation by adding an additional branch that manages to predict the pixel-wise semantic masks.

2. Regression/Classification-based Method

Although the region-proposal-based method has achieved an outstanding accomplishment, the precision still highly depends on the quality of the region proposals since the task of the object detection can be regarded as a regression or classification problem. The one-step networks that map the feature to the classified bounding box directly are proposed. The AttentionNet [33] includes an iterative approximation method which scans the image from the top-left to the bottom-right for bounding box regression. Moreover, YOLO [34] becomes one of the most famous detection network. YOLO predicts the bounding boxes, confidence scores, and class probabilities for each grid cell of image by an end-to-end network. However, it fails to detect all of the objects with multiple aspect ratios and scales. The SSD [35] manages to detect the objects in different aspect ratios and scales using a set of default boxes and a multi-scale feature learning structure. With the aid of BN, multi-scale feature learning, jointly training, and Darknet-19 feature extractor, YOLO v2 [36] is more accurate than YOLO and faster than SSD. YOLO v3 [37] is as accurate as SSD but three times faster than it by adopting some small changes and updating the Darknet-19 to Darknet-53. Moreover, the YOLO v4 [38] obtains higher accuracy and speed because the authors designed the CSPDarknet53 as the backbone, utilized the SPP and PAN as the neck, adopted the YOLO v3 as the header, and used a series of new feature operators. Shortly after the release of the YOLO v4, the YOLO v5 with two modified CSP modules in the backbone was put forward. It is light, fast to detect large-scale objects, and flexible for the application.

2.1.3. Segmentation

Compared with classification and detection, segmentation needs the most elaborate feature representation. The fine-grained semantic-feature-extraction process is key to pixel-level labeling.

1. Post Processing

To achieve the semantic segmentation, DeepLab (2014) [39] utilizes the fully connected pairwise CRF [40] proposed by Krhenbühl and Koltun as post-processing to refine the segmentation result. Thanks to this method, both the short- and long-range interaction between the pixels are taken into account, no matter how far they are apart with each other.

2. Hierarchical Feature Fusion Strategy

To take the advantages of CNNs to learn the hierarchical feature, some networks fuse and take full use of the features from each level of the sequential layer for global and local feature fusion. The existing fusion strategies can be grouped as the early fusion strategy and the late fusion strategy. The early fusion strategy transforms the local feature maps from early stages to the same size with the global feature maps from the later stages, and vice versa for the late fusion strategy.

(1) Early fusion strategy

ParseNet [41] first extracts the global feature by the previous layer with the global pool operation. Secondly, it further unpools the global feature vector into the same size with the feature map of the next layer. Finally, it fuses the feature maps from the different layers by the concatenation operation. Furthermore, SharpMask [42] puts forward a progressive refinement module that transforms and integrates the feature maps from the previous and next layer. These two networks provide a brand-new idea of fusing hierarchical features.

(2) Late-fusion strategy

The late-fusion strategy transforms the feature maps from the later stage to the same spatial size with the feature maps from the earlier stages. The fully convolutional network (FCN) [26] progressively recovers the resolution of feature maps using the deconvolution [43,44], which consists of the convolution and upsampling operation. At the same time, it gradually fuses the recovered feature maps and the feature maps from the former layers by the skip structure. This hierarchical feature fusion strategy not only integrates the global and local features but also takes the advantages of hierarchy. Inspired by the FCN, SegNet [45] introduces the encoder and decoder. The spatial size of feature maps from each layer of the encoder and decoder are symmetrical. Therefore, lots of researchers benefit from the encoder and decoder. For example, U-Net [46] is formed as an elegant shape of "U", and the feature maps in the same spatial size from the encoder and decoder are fused by the skip concatenation.

3. Dilated Convolutions

Since the convolution has achieved great success, some researchers aim to improve it. The main challenge of the hierarchical feature fusion is how to overcome the resolution loss due to the pooling operation. Dilated convolution is one of the most popular solutions, which expands the receptive fields of convolution. The 2018 version of DeepLab [47] and the real-time ENet [48] all make good use of the dilated convolution with different dilation rates.

2.2. Deep Learning on Point Cloud

Since the point cloud is one of the most popular real 3D data, the methods of processing point clouds based on the deep learning techniques spring up. However, the irregularity and sparsity become the main challenges. Recently, deep-learning-based technologies have experienced a rapid development. Therefore, we will introduce some classification, detection, and segmentation methods according to a taxonomy, which has been summarized by [49–51], in the following paragraphs.

2.2.1. Classification

1. Multi-View-based Method

It is reasonable for the researchers to project the point cloud into the multi-view images and then apply the mature 2D networks for global feature extraction. MVCNN [52] is a pioneer, which encodes each voxel with the global features maxpooled from the multi-view images. Then, the mutual relationship between the view groups is further considered by Yang et al. [53] for more discriminative 3D features. With the advantage that the graph represents the relationship between the nodes better, the View-GCN [54] utilizes the directed graph to model the relationship between multi-view images for local and non-local message passing. However, this kind of method is limited by the 2D representation, which may cause the geometric information loss.

2. Voxel-based Method

To better use the geometric information, researchers transform the unordered point cloud into the ordered intermediate data. Then, it is possible to generalize the mature 2D CNN works in 3D. At first, the VoxNet [55] takes the volumetric occupancy grid generated by point cloud as the input. However, the computation and memory grow cubically when the resolution of voxel grows. Therefore, OctNet [56] and 3dcontextnet [57] adopt the octree as a 3D grid index to reduce the computational and memory costs.

3. Point-based Method

Although the voxel retains the geometric information to a certain extent, the resolution of the voxel may cause information loss. To address this problem, point-based methods are proposed to process the points directly. The existing methods utilize different feature extractors, such as point-wise MLP, point convolution, and graph-based convolution. PointNet [58] is one of the most important networks, which utilizes the point-wise MLP

and the maxpool to extract the global feature. Additionally, the T-Net is introduced to achieve the permutation invariance. Moreover, the 3D continuous point convolution is proposed. The PointConv [59] introduces a point convolution based on the Monte Carlo estimation, which consists of the weighted function and density function. Furthermore, the SpiderCNN [60] defines the SpiderConv with the step function for coarse geometric information extraction and the Taylor expansion for fine-grained intrinsic local feature learning. With the development of the graph convolution, the graph-based method becomes emergent because it manages to represent the inter-relationship between points. Simonovsky et al. [61] first proposed the edge-conditioned convolution (ECC), which processes the graph whose vertexes represent each point and directed edge connects each vertex with its neighbors. The DGCNN [62] employs a novel EdgeConv to learn the peripheral feature. So, it learns not only the point-wise feature but also the local feature for each point and achieves permutation invariance at the same time.

2.2.2. Detection

Similar to the 2D detection methods, the 3D detection techniques can be categorized into the proposal-based and proposal-free method. The proposal-based methods generate a series of proposals first and then prune them into the proper size. The proposal-free methods predict the class possibility and regress the 3D bounding boxes at the same time by an end-to-end network.

1. Proposal-based Method

The Point RCNN [63] firstly segments the foreground for the proposal generation. The semantic and local spatial features are then fused for the 3D bounding box regression. Based on the Point RCNN, the Point RGCN [64] introduces a graph convolution for the detection process to obtain better bounding box. Moreover, STD [65] pre-defines the spherical anchors to improve the proposal generation. The semantic score and proposal-wise features are then extracted for the redundant proposals removal and the bounding box regression. Additionally, to take advantage of the mature 2D detector, the patch refinement method [66] back-projects the BEV detection results to the point cloud as 3D frustums, which will be refined to the bounding boxes by the local refinement network. Moreover, Qi et al. introduced the VoteNet [67], which generates the Hough votes for 3D detection. The VoteNet votes for the visual center of object and then aggregates the feature for bounding box regression.

2. Proposal-free Method

Thanks to the 2D FCN, Li et al. proposed a VeloFCN [68], which transforms the point cloud feature map into 2D and then utilizes the 2D FCN for further feature extraction. The VoxNet [55] takes voxels as its input and utilizes the regional proposal network for object detection. Instead of transforming the point cloud into other intermediate data, the 3DSSD [69] is the first single short 3D detection method that processing the unordered point cloud directly. This network consists of the fusion sampling strategy, the Feature-FPS and the candidate generation layer, which helps to exploit the representative points and achieve the anchor-free regression.

2.2.3. Segmentation

The aforementioned classification methods have successfully extracted the global features of the point clouds; some of them can be modified for segmentation by recovering the resolution of feature maps or fusing the global and local features, such as PointNet, PointConv, PointSpider, etc. Similar to the 2D segmentation methods, the 3D segmentation networks benefit from the hierarchical feature learning.

1. Semantic Segmentation

(1) Multi-view-based Method

The DeePr3SS [70] first projects the point cloud to the multi-view images for semantic segmentation. Then, the pixel-wise scores of each view are fused for 3D segmentation. Moreover, an end-to-end network named SqueezeNet [71] achieves the fast segmentation and projects the point cloud as spherical representation.

(2) Voxel-based Method

The volumetric representation also plays a crucial role in 3D segmentation. Huang et al. first utilized a fully 3D-CNN to segment the occupancy voxels. To achieve better segmentation, the SEGCloud [72] applies the deterministic trilinear interpolation for resolution recovery and the CRF as the post-processing to enforce the spatial consistency of the segmentation result. Thanks to the good performance of 3D-CNN, the fully convolutional point network (FCPN) [73] first achieves hierarchical feature extraction using the 3D convolution and weighted average pooling, which manages to learn both the global and local features.

(3) Point-based Method

PointNet not only achieves an outstanding performance on object classification but also concatenates the global feature and point-wise features for semantic segmentation. Although PointNet [58] manages to extract the point-wise feature, it fails to learn the local feature of each point with its neighbors. To address this problem, PointNet++ [74] hierarchically researches the regions of neighbors for each point for local feature learning, which helps to improve the accuracy of segmentation. Moreover, CNN also plays an important role in the point-based segmentation methods. For example, PointCNN [75] introduces an X-convolution transformation which maintains the permutation invariance of points by calculating the relative distance between each point and each of its neighbors. Moreover, the graph is utilized to model the point cloud. The DGCNN [76] generates a graph whose nodes represent each point first and then takes the graph as input for both classification and segmentation. SPG [77] defines a novel supergraph. Its node is named superpoint and represents a group of points, and its directed edge is called superedge; these are encoded with attributes for the representation of geometric and context information.

2. Instance Segmentation

Even though the semantic segmentation has already achieved elaborate point-wise prediction, it fails to distinguish each instance. In particular, the instances belonging to the same semantic class may be confounded as one object when they are placed close to each other. To meet the need of accurate 3D understanding of the real environment, the instance segmentation methods are worthy of further research.

(1) Proposal-based Method

The generative shape proposal network (GSPN) [78] first generates the 3D proposals by the region-based PointNet (R-PointNet) and then removes the redundant proposals by reinforcing the geometric understanding. Based on the instance-level proposals, the point-wise semantic mask are predicted for the final instance segmentation. Moreover, the one-stage and end-to-end 3D-BoNet [79] without predefined anchors directly generates the 3D bounding box and binary semantic mask for each instance with the aid of a multi-criteria loss function.

(2) Proposal-free Method

To avoid the dependency of proposals generation, lots of proposal-free methods are gradually being proposed. The similarity group proposal network (SGPN) and associatively segmenting instances and semantic (ASIS) module are two important networks that represent different ideas of proposal-free instance segmentation. The SGPN [80] directly groups the points into instance based on the semantic feature map, the pair-wise feature similarity matrix, and the heuristic and non-maximal suppression method. The ASIS [81]

achieves semantic and instance segmentation at the same time and makes the semantic and instance features support each other.

3. Dataset

Since the benchmark dataset is one of the most important parts of the methods based on deep learning techniques, we aim to comprehensively introduce the existing dataset, which may be suitable for segmentation and detection based on fusion of the 2D and 3D information. Although there are some popular datasets that have been widely used, we still want to introduce more suitable datasets as comprehensively as possible because both the quality and quantity of data will influence the performance of networks. Even though there are some widely used dataset, such as SUN RGBD, S3DIS, and KITTI, some newly proposed datasets in better quality and quantity are worthy of more attention. Moreover, different datasets focus on different types of data and scenes. For example, the indoor scene includes the living space, workplace, study place, shopping mall, etc., and the outdoor scenes may be collected on highway roads, urban scenes, rural scenes, etc. To the best of our knowledge, there is no dataset possessing all types of scenes without any bias on the categories. Therefore, to ensure the researchers can choose the dataset easily based on their requirements, we introduce the existing datasets, which consist of both 2D and 3D data, as much as possible in this section. The datasets can be grouped as the RGBD dataset and the 3D dataset registered with 2D data. The details of each dataset are shown in the following paragraphs.

3.1. RGBD Dataset

Thanks to the development of structure light sensors, more RGBD datasets are proposed for deep-learning research. Since structure light sensors perform better in indoor scenes, most of the RGBD datasets have been collected in indoor scenarios. The RGBD image is a well-performing form of representing both the 2D appearance and 3D geometric information, in which the RGB and depth values are pixel-wise aligned. A single RGBD image cannot reflect the holistic scenes due to the limit of visual perception field; however, there are some exceptions. For example, the SUN3D represents the holistic scenes by point cloud generated by the registered RGBD sequences. Moreover, the DIODE is collected in both indoor and outdoor scenes, and the depth values of an RGBD image are acquired by the LiDAR scanner. The existing meaningful RGBD datasets will be briefly introduced in the following lines.

3.1.1. Indoor Dataset

1. Dataset Proposed by Lai et al. [82] for 2D Object Detection

This is a large-scale, hierarchical multi-view RGB-D object dataset, which is acquired by the RGB-D camera. Additionally, a Point Grey Research Grasshopper camera provided the RGB images with higher resolution, which were calibrated with the RGB-D images by the Camera Calibration Toolbox for Matlab. This dataset included about 250,000 multi-view RGB-D images of 300 common everyday objects organized into 51 categories by the hierarchical category structure WorldNet. The data are labeled with the ground truth bounding box by the annotation method via the 3D reconstruction method proposed by this article, so it reduces the workload effectively.

2. Dataset Proposed by Koppula et al. [83] for 3D Semantic Segmentation

This 3D indoor point clouds dataset was collected by the Kinect, and the point cloud was generated by multi-view RGB-D images. This dataset included 52 3D scenes of homes and offices, which were composed of about 550 views and 2495 segments labeled into 27 categories.

3. Berkeley Dataset Proposed by Janoch et al. [84] for Object Detection

Considering the success of the Kinect depth sensor, this dataset aimed to "put the Kinect to work" for computer vision. This dataset was continually updated by crowdsourcing. The initial version named B3DO provided 849 RGBD images of 75 different scenes in the workspace. Moreover, the images were taken from variable viewpoints and distances

from the objects, which were frequently partially occluded and possessed a great diversity of appearance. The objects were labeled with a bounding box into 50 categories by the Amazon Mechanical Turk workers.

4. The MPII Multi-Kinect Dataset Proposed by Susanto et al. [85] for 3D Object Detection

The MPII consisted of 2240 pairs of RGB and depth images and 3D point cloud from the registered multi-view RGBD images acquired by the Kinect in 33 different scenes of kitchens. Each object of this dataset was annotated with the bounding box and grouped into nine classes of common issues.

5. NYU [86] and NYU v2 [87] Proposed by Silberman et al. for Semantic Segmentation and Scene Classification

NYU was the first indoor dataset which consists of RGB-D-I (RGB, depth, intensity) images acquired by the Microsoft Kinect and the dense manual annotation, which is suitable for semantic segmentation and scene classification. A total of 108,671 frames were collected in 64 different scenes, 2347 of which were manually labeled. The category of scenes included bathroom, bedroom, bookstore, cafe, kitchen, living room, and office. Additionally, there were 12 kinds of common objects organized by Worldnet, which consisted of bed, blind, bookshelf, cabinet, ceiling, floor, picture, sofa, table, television, wall, window, and background.

NYU Depth Dataset v2 extended the quantity and refined the semantic labels of original NYU. It contained 1449 registered RGB-D images captured from 464 diverse real-world indoor scenes across 26 scene classes taken from three cities with detailed per-pixel labeling of categories and physical relationships. There were 35,046 objects in the scenes and the images were manually selected from 435,103 video frames captured by the Kinect.

6. Dataset Proposed by Zhang et al. [88] for 2D Object Detection

This RGB-D dataset contained 900 objects in complex environments, including a notebook PC, drink box, basket, bucket, and bicycle in the scenes, which were captured from 33, 36, 36, 67, and 92 scenes, respectively. The images were all casually captured by the Kinect with different scales, textures, and rotations with hand-cropping or aligning.

7. SUN3D Proposed by Xiao et al. [89] for both 2D and 3D Object Detection and Segmentation

SUN3D is a large-scale RGB-D video database with both 2D and 3D semantic annotation and bounding box labeling. This dataset contains RGB-D images and point clouds generated by the RGB-D images. Traditionally, the previous datasets, such as Berkeley 3-D Object and NYU Depth, only provided the view-based scene. However, SUN3D first provides the 3D holistic large-scale place-centric point clouds reconstructed by the SfM and the constrains of object correspondences. Additionally, the annotation is propagated from the frames to the holistic scene. The original 415 sequences in 254 different scenes are captured by the ASUS Xtion PRO LIVE sensor in 41 buildings. The scenes are grouped into 10 classes and the objects were organized into 12 categories.

8. Synthetic RGB-D Scenes Dataset [90] Generated by CAD Models with Fine-Grained Texture for Segmentation, Which Proposed by Lai et al.

Since obtaining a real-world dataset with the pixel-wise ground truth label is expensive and time consuming, this paper explores how to generate a synthetic dataset using Trimble 3D Warehouse, which is one of the largest online-sourced CAD model repository provided by the hobbyists and professionals.

The original dataset named the RGB-D Scenes Dataset includes eight scenes of kitchen and office environments with common tabletop objects. There are three to twelve objects in each scene and the objects placing on the same plane and close or occlude each other. The latest RGB-D Scenes Dataset v2 extents the original one to 14 scenes indoor scenes with the tabletop objects and large furniture pieces recorded from a lounge, coffee room, meeting area, and office. There are nine kinds of synthetic objects placed in the automatic generated virtual scenes. To ensure practicability, this dataset simulates the sensor noise by adding Gaussian noise. Additionally, the sampled objects are scaled between 0.85 to 1 unit of the

largest length and rotated randomly to enlarge the diversity. The dataset also provides the voxel representation encoded with the grayscale intensity, RGB value, binary occupancy, and surface normal vector.

9. SUN RGB-D Proposed by Song et al. [91] for Scene Classification, 2D and 3D Object Detection, Semantic Segmentation

SUN RGB-D is an RGB-D large-scale indoor scenes understanding benchmark suit captured by the Intel RealSense, Asus Xtion, Kinect v1, and Kinect v2. This dataset provides 10335 RGB-D images and 800 kinds of objects captured from 47 kinds of scenes with dense 2D and 3D annotation. There are 3D point clouds generated by the RGB-D images, which are registered by the SIFT- and RANSAC-initialized ICP and aligned with the gravity direction. The main contribution of this dataset is the high-quality annotation. The authors hired workers to label the objects and room layers, and also to design a series of mechanisms to ensure the quality. There are four evaluation metrics included, which are suitable for scene categorization, semantic segmentation, object detection, object orientation, room layout estimation, total scenes understanding, and cross sensor task. The second important contribution is that the scale of this dataset is the largest so far because it calibrates the data from NYU depth v2, Berkley B3DO, and SUN3D.

10. ViDRILO Proposed by Martinez-Gomez et al. [92] for Scene Classification, Object Detection, Semantic Segmentation, Localization, 3D Reconstruction, and Data Compression.

The dataset ViDRILO is captured by the Microsoft Kinect device loaded on the Powerbot robot in two buildings with similar structure but diverse objects and room layouts. All rooms need artificial lighting, so both light and dark rooms are included. There are 5 RGBD sequences captured in 10 kinds of scenes, which were generated to 307,200 3D colored point cloud based on the mutual relationship between sequences. There are 15 types of objects, which include bench, extinguisher, computer, table, chair, board, printer, bookshelf, urinal, sink, hand drier, screen, trash, phone, and fridge.

11. SceneNN Proposed by Hua et al. [93] for Segmentation (Intrinsic Decomposition) and Shape Complement

Due to the lack of comprehensive and fine-grained annotation of the RGB-D dataset at that time, SceneNN provides the triangle meshes reconstructed by the RGB-D images captured in 100 indoor scenes. Moreover, both the 2D and 3D annotations include bounding box, per-pixel and per-vertex labeling, fine-grained information, axis-aligned bounding box, oriented bounding box, and object poses. To enrich the texture information, the author assigned each vertex of mesh with image.

12. SceneNet Proposed by Handa et al. [94] for Semantic Segmentation; SceneNet RGB-D Proposed by McCormac et al. [95] for Semantic Segmentation, Instance Segmentation, Object Detection, Optical Flow, Depth Estimation, Camera Pose Estimation, and 3D Reconstruction

SceneNet consists of an open-source repository of synthetic indoor scenes and online 3D CAD model repositories. The scene repository is formed by the author, but the object models are acquired from online repositories hosted by robotvault.bitbucket.org. The complexity of each scene is controlled by the algorithm, which helps to increase the diversity. To make sure the effectiveness of applying the algorithm to real-world scenes trained by synthetic data, the simulated Kinect noise is added for better image rendering.

Theoretically, SceneNet RGB-D can provide large virtual scene configurations with detailed annotation to overcome the scale limitation of the previous real-world datasets. ScenesNet RGB-D provides 5 M RGB-D images captured in the synthetic layouts. Therefore, this dataset is suitable to pre-train the data-driven computer vision techniques for performance improvement.

13. Multiview RGB-D Dataset proposed by Georgakis et al. [96] for 2D and 3D object detection

This dataset focuses on small-scale objects and provides densely sampled multi-view RGB-D images captured from nine kitchen scenes focusing on hand-held household objects with both 2D and 3D bounding boxes. The data are collected by the hand-held Kinect sensor and the subset of objects in the BigBird dataset.

14. Matterport3D proposed by Chang et al. [97] for 2D and 3D semantic segmentation

The Matterport Dataset is a large and diverse RGBD dataset that includes 194,400 RGBD images and 10,800 panoramic images acquired by the Matterport camera in 90 building-scale scenes. To the best of our knowledge, this is the only dataset including the building-level scene, which is more suitable for indoor navigation because it includes not only good quality, high diversity, and a large quantity of data, but also the precise global alignment between every room and floor in each building.

3.1.2. Hybrid Data

There are two datasets including both the indoor and outdoor RGBD images.

1. Stanford dataset proposed by Tombari et al. [98] for segmentation

This dataset is built up for the machine learning techniques which consists of indoor object data from the 3D Scanning Repository, Aim@Shape Watertight, a self-collected indoor dataset, and an outdoor dataset named New York City (NYC) acquired by the Kinect. The categories of the indoor data include packets, biscuits, juice bottles, coffee cans, boxes of salt of different brands and color, an armadillo, Asian dragon, Thai statue, bunny, happy Buddha, and dragon. The NYC dataset includes the outdoor building facades, vegetation, and vehicles.

2. DIODE proposed by Vasiljevic et al. [99] for semantic and object detection

DIODE is the first large real-world dataset captured in both indoor and outdoor scenes using the framework of RGB camera and FaroFocusS350 scanner. There are high-quality RGBD panoramas, point clouds, and the accurate surface normals in the dataset. The diversity of the scans reflects on the differences of each scene and their composition. The categories of the indoor scenes include: homes, offices, lecture halls, communal spaces; the outdoor scenes include: city streets, parking lots, parks, forest, and riverbanks.

Compared with the prior dataset, there are three main contributions. The depth information and the point cloud are accurate because they are captured by the LiDAR sensor. This is the first dataset covering a similar quantity of both indoor and outdoor data captured by same sensor framework. The RGB camera is placed very near to the LiDAR sensor, so the mutual relationship between the point cloud and the panorama images is known.

3.2. 3D Dataset with 2D Information

Since the development of the techniques for spatial sensing, more and more 3D datasets are proposed based on the light detection and ranging (LiDAR). The metadata of LiDAR is a point cloud, which is represented as the unordered points encoded with their X, Y, and Z coordinates. Therefore, the 3D point-cloud datasets are popular for the 3D segmentation and detection. Moreover, there are also some other forms of 3D data, such as CAD and mesh models. Although the LiDAR performs well in both indoor and outdoor scenes, the total scenes area of the dataset is still limited due to the range of LiDAR. Hence, the urban-level dataset needs to be acquired by the photogrammetry based on the aerial images. Moreover, the 2D information collected by the cameras should be registered with the 3D data.

3.2.1. Indoor

1. S3DIS proposed by Armeni et al. [100] for semantic segmentation and 3D object recognition

This is one of the most popular indoor datasets for understanding scenes. S3DIS contains over 215 million colored point clouds acquired by the LiDAR scanner in five large-scale indoor scenes of three different buildings covering 6000 square meters in total. The categories of scenes include office areas, educational spaces, exhibition spaces, conference rooms, personal offices, restrooms, open spaces, lobbies, stairways, and hallways. Additionally, the data are organized into 12 semantic categories: ceiling, floor, wall, beam, column, window, door, table, chair, sofa, bookcase, board.

2. Joint 2D–3D-Semantic proposed by Armeni et al. [101] for scene classification, semantic segmentation, and 3D object detection

Joint 2D–3D-Semantic expends the original S3DIS with a series of registered 2D, 2.5D, 3D data, and the instance-level annotation across all modalities. Hence, this dataset is suitable for developing the learning models that seamlessly transcend across domains. There are more than 70,000 images with RGB, depth, normal, semantic annotation, global XYZ, and camera metadata. The 3D data include 3D mesh, voxel, and 695,878,620 point clouds. The RGBD data are captured by the Matterport Camera with a 360-degree rotation at each scan location.

3. ScanNet proposed by Dai et al. [102] for scene classification, object detection, and semantic segmentation

ScanNet is a richly annotated real-world RGB-D dataset including 2.5 M images of 1513 scans in 707 diverse spaces. The annotations include the semantic category, estimated calibration parameters between RGB and depth information, and camera pose. Additionally, 3D surface reconstruction, textured meshes, and aligned CAD models are added to the dataset.

4. Dataset proposed by Sun et al. [103] for visual place recognition and localization

This dataset provides the camera poses in the real-world coordinate system, which reflects the correspondences of the images and point clouds. The existing IBL (image-based localization) research is limited by the SfM (structure from motion) or the SLAM (simultaneous localization and mapping) and lack of evaluation methodologies and the dataset with accurate ground truth. Therefore, this dataset provides the point clouds acquired by the Riegl VZ-400 3D laser scanner. Moreover, the images are captured by two DSLR (digital single lens reflex) cameras and the query images are collected by seven cell phones at random positions at different times to simulate real user scenarios.

5. ShapeNet proposed by Chang et al. [104] for geometric analysis

ShapeNet is a comprehensive and richly annotated large-scale dataset which consists of more than 3,000,000 3D CAD models observed in the real world, and 220,000 models out of that are organized by WorldNet to 3135 categories. The ShapeNet possesses information-rich annotation, which includes language-related annotations, geometrics annotations (rigid alignment, parts and key points, symmetry, object size), functional annotations (functional parts, affordances), physical annotations (surface material, weight). Moreover, the whole ShapeNet is divided into two subsets: ShapeNetCore and ShapeNetSem. ShapeNetCore contains 51,300 unique 3D model covering 55 common categories with manual labelling. ShapeNetSem possesses only 12,000 models but denser annotation which includes the real-world dimension, material composition, and total volume and weight. Additionally, the number of categories is increased to 270. However, the main disadvantage is the strong bias of categories because the dataset contains more rigid man-made artifacts than natural objects. This is because the creators of CAD models are more interested in artificial objects and modeling natural objects is more difficult.

6. ScanObjectNN proposed by Uy et al. [105] for object classification

ScanObjectNN is a special 3D object dataset that consists of both real-world and synthetic data. The main contribution of this dataset is that it helps to analyze the robustness of the classification algorithm and the gap between the synthetic and real-scene dataset. This dataset inspires the researchers to consider whether classification method trained in the synthetic dataset preforms as well as the real-world data. The synthetic data is acquired from the ModelNet40, which contains complete, well-segmented, and noise-free models. Moreover, there are 15 kinds of real-world objects selected from 700 scenes in SceneNN and ScanNet. To explore the robustness of classification methods, a series of perturbations are added in the well-segmented objects with various degrees of background and partiality. Training a method by the synthetic data and then testing it by real-world data is effective for the robustness analysis and the evaluation of the gap between the synthetic and the real-world dataset.

3.2.2. Outdoor

1. KITTI [106,107] for 3D semantic segmentation, object detection, stereo and optical flow estimation, and 3D visual odometry/SLAM

KITTI is one of the most popular outdoor datasets acquired by the MLS system equipped with synchronized cameras, a Velodyne HDL-64E laser scanner, and a localization system that consists of GPS, GLONASS, IMU, and RTK. Therefore, both the images and point cloud are registered and the positions of each sensor are known. There are 389 stereo and optical flow image pairs, 39.2 km of the stereo visual odometry/SLAM sequence, and more than 200,000 manually labeled 3D object bounding boxes in the dataset. With the growing needs for semantic segmentation, the labor-intensive point-wise semantic labels can be gradually achieved by tons of researchers. Ros et al. [108] manually labelled 216 images with eight categories: vegetation, sidewalk, building, fence, road, car, sky, and pole. Then, Zhang et al. [109] enlarged the semantic annotation to 252 images from eight sequences and added two more categories (pedestrian and cyclist). Finally, Behley et al. provided the largest amount and diversity of the semantic annotation in SemanticKITTI [110], which labeled 23,201 images, 4549 points, and 28 classes.

2. nuScenes proposed by Caesar et al. [111] for 3D object detection and tracking

nuScenes is the first outdoor dataset that consists of the data acquired by cameras, radar, and LiDAR with a 360-degree field of view. The data collecting platform is a vehicle equipped with six cameras, five radars, and one LiDAR, and all the sensors are placed closely to each other. The annotation of 3D bounding boxes are encoded with the semantic category and eight attributes (visibility, activity, and pose). Compared with the popular KITTI dataset, the amount of annotation is 7 times, and the images are 100 times as many as the KITTI. Therefore, both the quantity and diversity outperform.

3. Swiss3DCities proposed by Can et al. [112] for 3D semantic segmentation

Swiss3DCities is a new outdoor urban 3D point clouds dataset generated by photogrammetry based on images acquired by the UAV equipped with high-resolution cameras in three Swiss cities. The point clouds in this dataset are more uniform and denser than the point clouds acquired by the ground LiDAR. The total area of the scene is 2.7 square kilometers. After the acquisition of aerial images, the georeferencing is acquired by the GCPs. The sparse point clouds are then generated based on the georeference and Structure-from-Motion (SfM). Once the data is aligned and georeferneced, the dense mesh is constructed for denser point-cloud generation. The final density of the point clouds is about 500 K to 15 M per tile, and each point is encoded with x, y, z, and RGB values.

4. A2D2 proposed by Geyer et al. [113] for 3D object detection, semantic segmentation, and instance segmentation

The A2D2 (Audi Autonomous Driving Dataset) is an outdoor dataset that comprises simultaneously acquired images and 3D point clouds with the manual annotation of bounding box and semantic and instant labeling. There are six cameras and five LiDAR scanners equipped on the data-collecting platform, which provide data with a 360-degree field of view. The parameters of the corresponding relationship between LiDAR to LiDAR and LiDAR to cameras are provided in the dataset. Moreover, A2D2 provides 41,277 non-sequential frames with semantic labeling and 12,497 frames with both 2D and 3D bounding box annotation of objects. However, there are still 392,556 sequential frames without annotation.

5. Toronto-3D proposed by Tan et al. [114] for 3D semantic segmentation

Toronto-3D is a large-scale urban dataset with a colored point clouds dataset acquired by the mobile laser scanning (MLS) systems and point-wise semantic labels of eight object classes (road, road marking, natural, building, utility line, pole, car, fence, unclassified). There are approximately 78.3 million points in the dataset which cover 250 m of road. The main contribution is that each point is attached with 10 attributes (x, y, z, RGB, intensity, GPS time, scan angle rank, label).

6. Semantic3D.Net proposed by Hackel et al. [115] for 3D semantic segmentation

The outdoor scenes dataset at that time are mostly acquired by the LiDAR scanners equipped on the mobile mapping car (such as the Sydney Urban Objects dataset, Paris-rue-Madame) which provides a lower point density than the typical static scanner. Semantic3D.net (accessed on 29 September 2021) is the first outdoor dataset consisting of dense colored point cloud captured by the static terrestrial laser scanner, which is known for its high measurement resolution and long measurement range. There are over 4 billion labeled points organized into eight semantic categories. Both various natural and man-made scenes are included in the dataset to prevent overfitting of the classifier. The detailed colorization is operated by the post-processing of deploying the high-resolution cube map. Additionally, the annotations are manually created, which is labor-intensive but effectively avoids inheriting errors. The point clouds are firstly labeled and then projected to 2D images. The categories include man-made terrain, natural terrain, high vegetation, low vegetation (lower than 2 m), buildings (churches, city halls, stations, tenements, etc.), remaining hardscape, scanning artifacts, cars, and trucks.

7. CSPC-Dataset proposed by Tong et al. [116] for semantic segmentation

CSPC is an outdoor dataset with colored point clouds scanned by the wearable mobile mapping robot with six kinds of manual point-wise labels (ground, buildings, vehicles, bridges, vegetation, poles) for semantic segmentation. The author claims that the scenes are more complete, the density of point is relatively uniform, the diversity and complexity of objects outperform, and there is a high discrepancy between different scenes. There are approximately 68 million points in the dataset.

8. All-In-One Drive proposed by Weng et al. [117] for 3D object detection, tracking, trajectory prediction, semantic and instant segmentation, depth estimation, and long-range perception

AIODrive is a unique large-scale dataset providing various data, annotations and environment variations acquired by comprehensive sensors. The sensors include RGB, stereo and depth cameras, LiDAR, SPAD-LiDAR, radar, IMU, and GPS. The high-density and long-range point clouds encoded with (x, y, z, I) are obtained by the combination of LiADRs, SPAD-LiDAR, and a depth camera with a 360-degree horizontal field of view (FoV). In particular, the SPAD-LiDAR was first used in a public perception dataset. AIODrive provides the most diverse annotations for multiple mainstream perception tasks, which includes 2D–3D bounding boxes for object detection, 2D–3D semantic and instant labels

for segmentation and some annotation above the mainstream (motion data for all agents, fine-grained object class labels, vehicle control signals, city map, and road structure). Additionally, there are some environment variations generated by the simulator to improve the robustness of perception systems in rare driving scenarios, such as highly crowded scenes, high-speed driving, diverse weather and lighting, car accidents, vehicles running through a red light, speeding and changing lanes aggressively, and children and adults jogging and running.

9. Argoverse proposed by Chang et al. [118] for D objects detection, tracking, and trajectory prediction

Argoverse is a large-scale dataset that supports various autonomous driving perception tasks and focuses on 3D tracking and motion forecasting. There are synchronized multiple data, including 360-degree images acquired by seven high-resolution ring cameras and two front-facing stereo cameras, point clouds collected by two roof-mounted VLP-32 LiDAR and 6-DOF localization of each timestamp calculated by the combination of the GPS-based and sensor-based localization. Argoverse also contains the vector map of lane centerlines, a rasterized map of the ground height, drivable area, and region of interest (ROI). The annotation includes the manually annotated 3D vehicles bounding boxes and their fleet logs and 3D trajectories. There are three main advantages of this dataset. Firstly, this is the first publicly available dataset that provides the semantic vector map of both the road infrastructure and traffic rules. Secondly, the amount of the 3D trajectory annotation is ten times more than the KITTI. Lastly, the dataset includes diverse scenarios, for example: managing an intersection, slowing for a merging vehicle, accelerating after a turn, stopping for a pedestrian on the road, etc.

10. ApolloScape proposed by Huang et al. [119] for 3D object detection, semantic and instance segmentation, and 3D reconstruction

According to the tasks of autonomous driving, we understand the environments based on 3D semantic HD map, D perceptual system, the on-the-fly self-localization system and the path of each target. Compared with the existing dataset, such as SOTA, KITTI, Cityscape, and so on, ApolloScape dataset is an outdoor large-scale point-clouds dataset with larger, denser, and richer labeling, which includes semantic and instant labeling for point clouds, stereo driving videos and point clouds, accurate 6 DoF camera pose, and per-pixel lane mark labeling. Moreover, there are about 70 K 2D and 3D instance-level labeled cars.

11. Virtual KITTI proposed by Gaidon et al. [120] for 3D object detection, tracking, and semantic and instance segmentation

The existing scene benchmarks are limited by the high cost of data acquisition and accurate labels. This article proves that the deep learning methods pre-trained on the mixture of real and virtual data behave similarly with the method only trained by the real-world data. Therefore, the author proposes an efficient real-to-virtual world cloning method, which enlarges the KITTI dataset with the synthetic data called Virtual KITTI.

Firstly, the original scenes in KITTI are decomposed into different visual components. Unity is then used to create the virtual scenes that are closed to the KITTI. To enlarge the diversity of scenes, secondary roads and some background objects such as trees and buildings are manually placed into the virtual scenes. Additionally, the direction and brightness of light sources are manually controlled to create more light conditions. In addition, the texture of objects is generated by the unlit shaders in Unity. Consequently, the per-pixel and instance-level ground-truth semantic labels and the 2D and 3D bounding box are also automatically generated.

12. Dataset proposed by Fang et al. [121] for 3D object classification, detection, semantic segmentation

This dataset is a simulation dataset generated by the data from the real environment and traffic flow for autonomous driving vehicles, which provides dense point clouds, registered color images, and semantic annotation. Thanks to the novel LIDAR simulation framework of this article, this simulation dataset is efficient. This dataset outperforms compared with the state-of-the-art simulation dataset and can gradually approach the effectiveness of the real-scenes dataset by increasing the quantity. Additionally, using the simulation data to pre-train the model and the real-world data for fine tuning has already been proved more efficient than using the same amount of real-world data only. The real-world data of this dataset are acquired by the Riegl VMX-1HA and pre-segmented by PointNet++ for efficiency, then the wrong parts of the segmentation results are manually corrected. Based on the annotation, the clean static background is obtained by removing the movable obstacles. To generate the synthetic scenes that are similar to the real traffic scenario, a data-driven method consisting of a probability map for the obstacle placement and model selection strategy are proposed.

3.3. Comparative Analysis

In the following lines, we are going to make a comparative analysis of the aforementioned datasets and other popular datasets. According to the tabular forms that include a series of comparison results, it may be easy for researchers to figure out which dataset is suitable.

3.3.1. Indoor

We categorized the indoor datasets as object-level and scene-level datasets. The object-level datasets only include the individual objects, which is suitable for the task of object classification and part segmentation. The existing datasets are all composed of multiple resources, such as online open-sourced 3D repositories, self-collected data, existing datasets, etc. With the increasing demand for object-level datasets for the tasks of classification and part segmentation based on deep learning techniques, the quality and quantity of the dataset has improved rapidly. For example, there are only hundreds of objects included in the early Stanford repository in 2011, but the latest PartNet possesses almost one million objects. Moreover, the diversity of categories differs a lot. The common datasets only include 6 to 55 categories. However, the ShapeNet includes a remarkable 3135 classes and its subset ShapeNetSem has 270 classes. The comparative summary of indoor object-level datasets is shown in Table 1.

Table 1. Comparative summary of existing indoor object-level dataset (Seg: segmentation; Rec: recognition; Cla: classification; Comp: completion).

	Name	Object	Class	Data	Data Type	Source	Task
[98]	Stanford	- 400 31	6 20 6	3D Model RGBD	Real-world	Stanford 3D Scanning Repository Shape Watertight Database Kinect	Semantic Seg
[104]	ShapeNet ShapeNetCore ShapeNetSem	3 M 51.3 K 12 K	3135 55 270	CAD Model	Real-world	Online Open-source	Semantic Seg
[122]	ShapeNet Core55	16.9 K	16 55 -	Point Cloud Parts Voxel Model	Real-world	Online Open-source	Semantic Seg Part Seg 3D Rec
[123]	PartNet	573.6 K	24	Point Cloud	Real-world	Online Open-source	Herarchical Seg
[124]	ModelNet	48 K	40	RGBD, CAD Model, Voxel	Synth-etic	3D Warehouse Yobi3D SUN Database Princeton Shape	Cla Shape Comp

The scene-level datasets represent the holistic scene by the multi-view RGBD images, point cloud, or 3D model. However, the scale of holistic scenes differs a lot between different datasets. For example, the popular SUN 3D and SUN RGB-D datasets display the room-scale scenes. However, the S3DIS and its augmented version 2D-3D-S and Matterport dataset include the holistic floor-scale scenes. Moreover, the Matterport dataset is the only building-scale dataset, and every floor of same building is aligned. Table 2 shows the comparative summary of indoor scene-level datasets.

Table 2. Comparative summary of indoor scene-level datasets (Seg: segmentation; Rec: recognition; Cla: classification; Comp: completion).

	Name	Size (m^2)	Amount	Object	Class	Data	Data Type	Sensor	Task
[82]	-	-	250 K	300	51	RGBD	Real-world	Kinect, Camera	Semantic Seg
[83]	-	-	-	2.5 K	27	RGBD	Real-world	SLAM, Kinect	Semantic Seg
[86]	-	-	-	-	127	RGBDI	Real-world	Kinect	Semantic Seg, Scene Cla
[84]	Berkley B3DO	-	849	-	50	RGBD	Real-world	Kinect	Object Det
[125]	-	-	111	-	6	RGBD	Real-world	Kinect XtionPRO	Unknow Seg
[85]	-	-	2.2 K -	-	9	RGBD Point Cloud	Real-world	Kinect	Object Det
[87]	NYU v2	-	1.4 K	35.0 K	-	RGBD	Real-world	Kinect	Instance Seg
[88]	-	-	-	900	5	RGBD	Real-world	Kinect	Object Det
[89]	SUN3D	-	415 -	-	12	RGBD, Camera Pose Camera Pose	Real-world	Xtion PRO	Semantic Seg
[90]	RGBD Scenes v.2	-	-	-	10	RGBD Point Cloud Voxel Model CAD Model	Synthetic	Trimble 3D Warehouse	Semantic Seg
[91]	SUN RGB-D	-	10.3 K	-	800	RGBD Point Cloud	Real-world	Intel Real-Sense, Asus Xtion, Kinect	Object Det
[92]	ViDRILO	-	22.5 K 0.3 M	-	10	RGBD Point Cloud	Real-world	Kinect	Scene Cla
[93]	SceneNN	2124	-	1.5 K	-	RGBD 3D Mesh Camera Pose	Real-world	Kinect, XtionPRO	Semantic Seg, Shape Comp, 3D Rec
[94]	SceneNet	-	-	3.7 K	-	RGBD 3D Secene	Synthetic	3D CAD Model Repositories	Semantic Seg
[95]	SceneNet RGBD	-	5 M -	-	255	RGBD 3D Secene Camera Pose	Synthetic	3D CAD Model Repositories	Semantic Seg, Object Det
[96]	-	-	7.5 K 96.6 M	118	11	RGBD Point Cloud	Real-world	Kinect, Big Bird Da-taset	Object Det
[99]	DIODE	-	11.5 K 170 M -	-	-	RGBD Point Cloud Normal	Real-world	FARO Focus S350	Semantic Seg
[100]	S3DIS	6020	- 965 M -	-	12	RGBD Image Colored Point Cloud Mutual Relationship	Real-world	Matterport Camera	Semantic Seg
[101]	2D-3D-S	6020	70.5 K 695 M -	-	13	RGB Image Colored Point Cloud 3D Mesh Normal Camera Pose Mutual Relationship	Real-world	Matterport Camera, Structured-light Sensors	Semantic Seg
[102]	ScanNet	34453	2.5 M -	36.2 K	19	RGBD 3D Mesh Camera Pose CAD Model	Real-world	Kinect	Semantic Seg, 3D Rec
[103]	-	5000	682 67 M -	-	-	RGB Image Point Cloud Mutual Relationship	Real-world	Riegl, DSLR Cameras	Image Localization

Table 2. *Cont.*

	Name	Size (m^2)	Amount	Object	Class	Data	Data Type	Sensor	Task
[126]	Robust-PointSet	-	-	12.3 K	40	CAD Model	Synthetic	ModelNet40	Robustness Evaluation
[105]	Scan-ObjectNN	-	-	2.9 K	15	CAD Model	Synthetic	ModelNet40	Semantic Seg, Object Det
[97]	Matterport	219 K	0.2 M -	50.8 K	40	RGBD Point Cloud 3D Mesh	Real-world	Panorama Camera, SceneNN, ScanNet	Semantic Seg, Scene Cla

3.3.2. Outdoor

The outdoor datasets can be grouped based on their scale as object-level, road-level and urban-level. Since the outdoor scenes are extremely large compared with the indoor scenes, the combination of camera and LiDAR is more suitable for the outdoor data collection. Moreover, almost all of the object-level and road-level datasets are acquired by the mobile laser scanning (MLS) system, which is usually a vehicle equipped with a series of sensors, such as camera, LiDAR, radar, GPS/IMU, etc. There are two popular object-level datasets—New York City and Sydney Urban Objects dataset—which can be utilized for the part segmentation and classification. However, the New York City dataset only includes point-cloud data without labeling. Therefore, the Sydney Urban Objects dataset is more suitable for the deep-learning-based task of classification. Table 3 shows a comparative summary of outdoor object-level datasets.

Table 3. Comparative summary of outdoor object-level dataset.

	Name	Point	Object	Class	Data	Data Type	Sensor	Task
[98]	New York City	-	15	3	Point Cloud	Real-world	Lidar sensor (MLS)	Semantic Seg
[127]	Sydney Urban Objects Dataset	2.3 M	588	26	Point Cloud	Real-world	Velodyne (MLS)	Cla

Since the outdoor datasets are almost collected by the MLS due to the extremely large scale of the outdoor scenes, the outdoor datasets are usually composed by the static and dynamic objects on and near the road. So, the scale of the scenes can be measured by the length of the road where the data is collected. However, Semantic3D.Net is an exception that is collected by static LiDAR, which manages to obtain the point cloud in a higher density. KITTI is one of the most popular outdoor dataset, which has experienced a long range of development. The raw data is collected in 2012, but the full annotation is completed in 2020. During that time, more and more outdoor datasets are gradually proposed, which achieve a constant improvement of quality, quantity, and diversity. For example, the AIODrive contains the highest diversity of data, annotation, and driving scenario. Moreover, some virtual datasets are proposed to enlarge the labor-exhausted real-world dataset, such as VirtualKITTI and two dataset proposed by Wang et al. and Fang et al. The comparative summary of outdoor road-level datasets is shown in Table 4.

Due to the extremely large scale of the whole urban area, collecting an urban-level dataset by traditional MLS or static LiDAR is impossible. Therefore, there are some urban-level datasets collected by aerial LiDAR scanning (ALS) or unmanned aerial vehicle (UAV) photogrammetry. From 2013 to 2021, the latest SensatUrban dataset obtained the largest quantity of data with detailed manual annotation for the task of segmentation, which will effectively help to boost the development of smart cities. Table 5 includes the comparative summary of outdoor road-level datasets.

Table 4. Comparative summary of outdoor road-level datasets (Seg: segmentation; Rec: recognition; Cla: classification).

	Name	Size (km)	Amount	Object	Class	Data	Data Type	Sensor/ Data Source	Task
[106,107]	KITTI	39.2	12 K 1799 M	200 K	2	RGB Image, Point Cloud	Real-world	Velodyne Camera	3D & 2D Det SLAM
[108]	KITTI (Ros)	-	216	-	11	Labeled Image	Real-world	KITTI	2D Seg
[109]	KITTI (Zhang)	-	252	-	10	Labeled Image			
[110]	Semantic-KITTI	-	4549 M	-	28	Point Cloud			3D Seg
[120]	Virtual KITTI	-	17 K	-	13	RGB Image	Synthetic		2D Seg, 2D Det, Track, Optical Flow
[128]	Paris-Rue-Madame	0.16	20 M	642	17	Point Cloud	Real-world	Velodyne Riegl	3D Seg
[129]	IQmulus	10	300 M	-	50	Colored Point Cloud	Real-world	Stereopolis II	3D Seg
[115]	Semantic3D.net		4 B	-	8	Point Cloud	Real-world	TLS	3D Seg
[130]	Paris-Lille-3D	1.94	143 M	-	50	Point Cloud	Real-world	Velodyne	3D Seg
[121]	-	-	100 K	-	5	Point Cloud	Synth-etic	Riegl 3D Model	3D Seg,3D Det,3D Cla
[131]	-	-	20 K	-	300	CAD models	Synth-etic	Stanford-Cars CompCars	3D Object Pose Estimation
[118]	Argoverse	290	107 K -	10.6 K	17	RGB Image Point Cloud	Real-world	VLP-32,GPS, Stereo Camera GPS/IMU	3D Det, Object Track
[119]	Apollo-Scape	-	144 K 120 K -	-	35	RGBD Image, Point Cloud Sensor Pose	Real-world	Riegl VMX-CS6 Camera GPS/IMU	3D & 2D Seg, 3D & 2D Det, 3D Localize, 3D Rec
[111]	nuScenes	242	400 K 1.4 M -	-	23	Point Cloud, Map Radar Signal, Sensor Pose	Real-world	LiDAR Scanner Radar Scanner Camera GPS/IMU	3D & 2D Det, Track
[113]	A2D2	-	41,277 -	-	38	RGB Image Point Cloud, Mutual Relationship	Real-world	LiDAR Scanner Camera	3D & 2D Seg, 3D & 2D Det, Depth Estimation, Optical Flow
[114]	Toronto-3D	1	78.3 M -	-	8	Colored Point Cloud HD Map	Real-world	Teledyne Optech Maverick Camera GNSS	3D Seg
[116]	CSPC	-	68 M	-	6	Colored Point Cloud		Velodyne Lady Bug 5 GPS/IMU	3D Seg
[117]	AIODrive	-	250 K -	26 M		RGBD Image Point Cloud, Sensor Pose	Real-world	LiDAR Scanner Spad-LiDAR Scanner Radar Scanner RGBD Camera GPS/IMU	3D Seg, 3D Det, Track, Trajectory Prediction, Depth Estimation
[132]	PC-Urban	-	4.3 B	-	25	Point Cloud	Real-world	Ouster LiDAR Scanner	3D Seg

Table 5. Comparative summary of outdoor road-level datasets (Seg: segmentation; Rec: recognition; Cla: classification).

	Name	Size (km^2)	Amount	Class	Data	Data Type	Sensor	Task
[133]	ISPRS	0.15	- 1.2 M	2	RGB Image, Point Cloud	Real-world	ALS	3D Det, 3D Rec
[134]	DublinCity	2	4471 1.4 B	13	RGB Image, Point Cloud	Real-world	NIKON D800E, Leica RCD30 (ALS)	3D Seg, 3D Rec
[112]	Swiss3D-Cities	2.7	- 226 M	5	RGB Image, Point Cloud	Real-world	UAV	3D Seg
[135]	SensatUrban	7.64	- 2847 M	13	RGB Image, Point Cloud	Real-world	SODA Camera (UAV), RTK GNSS	2D & 3D Seg
[136]	DALES	330	505 M	8	Point Cloud	Real-world	Riegl Q1560 (ALS)	3D Seg, 3D Rec
[137]	LASDU	1.02	3 M	5	Point Cloud	Real-world	ALS	3D Seg, 3D Rec
[138]	Campus3D	1.58	937 M	24	Colored Point Cloud	Real-world	UAV	3D Seg, Instance Seg
[139]	Waycom	76	250 K 12 M	4	RGB Image, Point Cloud	Real-world	MLS	2D & 3D Det Track

4. Fusion Strategy

This section aims to review the fusion strategy of 2D and 3D information for two common tasks of scene understanding: segmentation and object detection. Since the 2D image contains more appearance information and 3D point cloud, model, or depth contains more accurate geometric information, it is necessary to study the 2D and 3D information fusion strategy for the feature complementation. According to our conclusion, we put forward a novel taxonomy for the existing 2D and 3D fusion strategy categorization. The categories include the non-feature-based and the feature-based fusion strategy. Moreover, the non-feature-based fusion strategy can be further divided into the data-based, result-based, and data–result-based strategy. In addition, the feature-based strategy can be further grouped into the one-stage, multi-stage, and cross-level fusion strategy. In the following paragraphs, we will introduce the details of the existing fusion strategies and their representative methods.

4.1. Segmentation

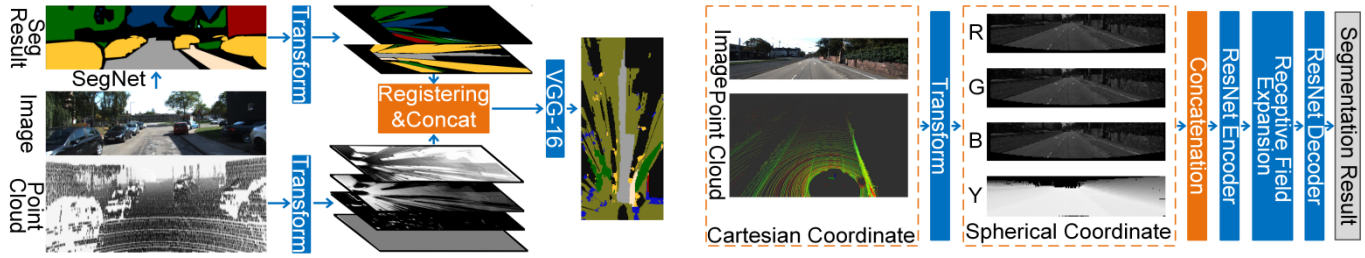
4.1.1. Data-based Fusion Strategy

The data-based fusion strategy aims to transform the 2D and 3D data into the same kind of data that contains both 2D and 3D information for data integration. However, the main challenge is that 2D and 3D data are organized by different data structures. For example, the images are an ordered 2D grid encoded with the optical digital values, but the point cloud is a set of unordered points encoded with their coordinates. The common method is colorizing the point cloud by the registered images. However, the sparsity of point cloud may cause the appearance information loss. Therefore, integrating the 2D and 3D information by intermediate data is a useful strategy. The comparative summary of some segmentation methods using the data-based fusion strategy is shown in Table 6; they will be introduced in detail in the following paragraphs.

Table 6. Comparative summary of segmentation methods using data-based fusion strategy (P: point cloud; Data 2: intermediate data).

	Name	Data	Data 2	Data Set	Acc	Class Acc	AP	IoU	F1-Score	Scene	Code
[140]	UGrid-Fused	RGB, P	UGrid UView	KITTI	_	_	89.5 90.0	_	93.8 93.1	Outdoor	No
[141]	_	RGB, P	Grid BEV	KITTI	81	49.4	_	69.8	_	Outdoor	No
[142]	StdnDSN	RGB, P	StdnDSN	Self-collected	91.5	_	_	_	_	Outdoor	No
[143]	_	RGB, P	Spherical Image	KITTI	_	_	89.63	_	94.3	Outdoor	No

The Bird's Eye View (BEV) is a compact 2D occupancy representation that is suitable for projecting the point cloud onto a 2D plane. Hence, the BEV grid is a suitable intermediate data for data fusion. Wulff et al. [140] propose a multi-dimensional occupation grid representation based on the BEV named UGrid-Fused, which can be imported into the FCN for semantic segmentation. Each cell in UGrid-Fused contains 15 statistics and constitutes a 40 m $\times$ 20 m area. The occupancy of the grid map includes a binary map, count map, obstacle map, six height measurement maps, and six reflectivity intensity maps. This modified BEV helps to calibrate the advantages of image and point clouds. Moreover, Erkent et al. [141] propose a hybrid approach (shown in Figure 1a) combining the advantages of Bayesian filtering and DNN for semantic segmentation based on the BEV. The occupancy of BEV is encoded with the 2D segmentation result generated by SegNet, dynamic state, statical state, current state probabilities and update probabilities generated by the point cloud using the Bayesian filter. Then, the BEV will be imported to the VGG-16 for densely semantic segmentation.



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

Figure 1. The representative segmentation methods using data-based fusion strategy. (**a**) The structure of the hybrid approach combining the advantages of Bayesian filtering and DNN; (**b**) The spherical coordinate transformation method.

At the same time, there are other representations especially designed for data fusion. Zhou et al. [142] first proposed a standard normalized digital surface model (StdnDSN) to achieve the fusion of point clouds and very-high-spatial-resolution images (VHSRIs). The StdnDSN is then segmented by the combination of a grey-level co-occurrence matrix (GLCM) and multi-resolution segmentation (MRS) and classified into the land-cover objects by a CNN. The segmentation result is used to generate the votes by the regional majority voting strategy to accelerate the procedure of classification. Lee et al. [143] introduce a spherical coordinate transformation method (shown in Figure 1b) for data fusion which projects the point clouds and images onto the same spherical coordinate based on their mutual relationship. The intermediate data is a four-channel image in spherical coordinate encoded with the transformed RGB information and the height information transformed along the Y axis. Then, the intermediate data will be imported into the modified SegNet with a novel receptive field expansion structure placed between the encoder and the decoder to learn a broader range of feature.

4.1.2. Result-Based Fusion Strategy

Although the existing data-based fusion strategies manage to fuse both 2D and 3D information, it depends on the quality of data projection and the intermediate data which may cause the information loss. Therefore, result-based fusion strategies are proposed to alleviate this problem. Result-based fusion strategies conduct the processes of 2D and 3D segmentation, respectively, and then integrate the 2D and 3D segmentation results based on their mutual relationship. The comparative summary of segmentation methods using the data-based fusion strategy is shown in Table 7, and their details will be introduced in the following lines.

The conditional random field is a commonly used method for the segmentation result refinement. Therefore, the mutual relationship between 2D and 3D segmentation results are encoded by the CRF to achieve the 2D and 3D result fusion and refinement. Gu et al. [144] put forward an inverse-depth-aware fully convolutional (IDA-FCNN) network for image feature learning and a line-scanning strategy for geometric feature learning based on the inverse-depth histogram of point clouds. Then, the result-based fusion for road segmentation is achieved by a CRF. The pairwise potential in the energy function of CRF penalizes the different image-based and point-cloud-based segmentation results. Gu et al. [145] also used the CRF to integrate the 2D and 3D result of road segmentation (shown in Figure 2b). The point cloud is segmented based on the probability map describing the flatness of each point. At the same time, an FCN is selected for the camera-based road segmentation. Then, the 3D segmentation result is projected on the images. Finally, the result of camera-based and LiDAR-based segmentation are integrated by the CRF fusion strategy whose energy function contains 2D unary potential, 3D unary potential, and 2D–3D pair-wise potential. The PanopticFusion [146] archives the holistic scenes understanding based on the RGBD images. This system first acquires the panoptic labeled images by fusing both the semantic segmentation result from PSPNet and instant segmentation result from Mask R-CNN. Then, the panoptic labeled images are integrated with the volumetric map generated by the depth information. The mutual relationship between panoptic labeled images and volumetric maps is reconstructed by the SLAM. In addition, the author proposed a fully connected CRF model consisting of a novel unary potential approximation and a map division for the map regularization respecting to the panoptic labels.

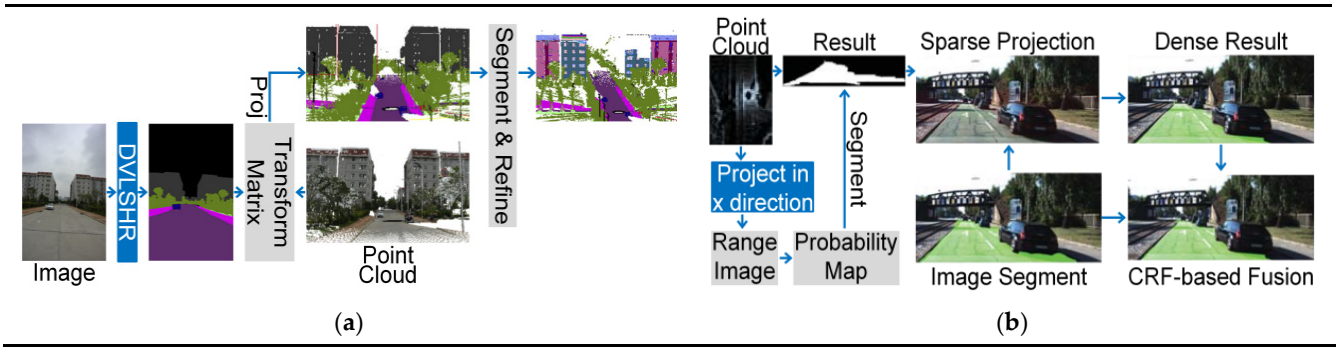


Figure 2. The representative segmentation methods using result-based fusion strategy. (**a**) The CRF-free 3D semantic segmentation result refinement method; (**b**) The CRF-based road segmentation method integrating the 2D and 3D result of road segmentation.

However, Zhang et al. [147] propose a CRF-free 3D semantic segmentation result refinement method (shown in Figure 2a) based on the 2D segmentation result. Firstly, the images are segmented by the ImageNet-pretrained Deeplabv2-VGG16 (DVLSHR model). Secondly, the 2D segmentation result is mapped to the 3D point clouds based on the camera's internal parameters and the external azimuth elements calculated according to the collinear conditions. Based on the mapped 2D result, the outline of each class is segmented. Thirdly, the further refinement of segmenting the physical planes of buildings based on the 3D features is achieved by optimizing the coarse segmentation result. The

proposed FC-GHT algorithm takes the advantages of random Hough transformation and the GHT coarsely segments the physical planes based on the normal vector angle and the Euclidean distance. Then, the coarsely segmented patches are optimized by the patch merging and re-judgment of coplanar points.

Table 7. Comparative summary of segmentation methods using result-based fusion strategy (P: point cloud; Data 2: intermediate data).

	Name	Raw Data	Data 2	Data Set	Acc	Class Acc	AP	IoU	F1-Score	Scene	Code
[148]	_	RGB,P	Mesh	CamVid	_	61.1	_	_	_	Outdoor	No
[144]	IDA-FCNN	RGB,P	BEV	KITTI	_	_	92.7	_	96.4	Outdoor	No
[147]	DVLSHR	RGB,P	DVLSHR	City-Scapes	74.9	_	_	64.2	_	Outdoor	No
[145]	LC-CRF	RGB,P	BEV	KITTI	_	_	92.1	_	97.1	Outdoor	No
[146]	PanopticFusion	RGBD	Voxel	ScanNet	_	52.9	_	_	_	Indoor	No

Moreover, the task of 3D model segmentation also needs the result-based fusion strategy. The existing methods are time consuming because they often extract the image and geometric features from every image, so the reduction of the magnitude of images will help to accelerate the labeling. Riemenschneider et al. [148] aimed to predict the best view to reduce the view redundancy for SfM/MVS reconstruction. The best view selection and labeling are achieved based on the proposed semantic cues, view redundancy, and scene coverage. The semantic cues including a heavy feature rely on the image context, which takes a longer time to be computed; a lightweight feature relies on geometric information, which takes a shorter time. The heavy feature (16-dimensional feature vector) consists of the CIELAB Lab color component, eight responses of the MR8 filter bank, height from ground plane, depth of dominant plane, and surface normal (optional: dense SIFT). The lightweight feature includes the area of the mesh surface, 2D projection in a specific camera, ratio between the former two elements, mesh surface and its projection, and the angle between the mesh surface and its projection. The lightweight feature is used to view the redundancy reduction, and then the heavy feature is used for semantic classification. Finally, the observation importance is introduced to reflect the importance of images which helps to accelerate the surface-wise classification and mesh labeling.

4.1.3. Feature-Based Fusion Strategy

According to our knowledge, the existing feature-based fusion strategies can be grouped into three categories: one-stage, multi-stage, and cross-level. The comparative summary of segmentation methods using the feature-based fusion strategy is shown in Table 8, and their details will be introduced in the following paragraphs.

1. One-stage Fusion Strategy

The one-stage feature-based fusion strategy only fuses the 2D and 3D information once in the network. In the following lines, we will introduce the methods processing the RGBD image or the point cloud registered with the RGB images, respectively.

(1) RGBD images

Some of the methods first extract the 2D appearance feature based on the RGB information and then project the 2D feature to 3D as initialization for further 3D feature learning. Qi et al. [149] propose an end-to-end 2D semantic segmentation method for the fusion of appearance and geometric information. It combines the CNN and the GNN (graph neural network). The nodes of the graph represent the points and the edges link each node with its nearest neighbors in 3D. The image features extracted by the CNN (modified VGG-16 and global pooling) are used to initialize each node of the graph, and then each node will be iteratively updated by a recurrent function and neighbor information through the edges from their neighbors. The neighbor information is first computed by feeding the hidden state to MLP and average operator, then updated by the vanilla RNN or LSTM (similar performance). With four propagation steps, the accuracy reaches its best.

Instead of initializing the 3D feature extraction process with 2D features, which may lead to feature bias, some networks utilize the 2D and 3D features equally. According to the method put forward by Gupta et al. [150], the region proposals are generated by the multi-scale combinatorial grouping (MCG) based on the normal gradients, geocentric pose, and soft edge map. The fine-tuning R-CNN is then utilized for the proposal-wise RGB feature and depth feature learning. The main contribution of this work is the HHA representation generated by the depth values. The HHA representation is encoded with the horizontal disparity, the height above the ground, and the angle between the gravity direction. Since the range of the values of HHA representation are scaled from 0 to 255, the R-CNN is effectively generalized to extract the feature of depth. Then, the depth and RGB features are imported to a SVM for the bounding box generation. Finally, based on the result of object detection, the instance mask predicting and semantic segmentation are achieved at the same time. This method is shared with a good quality code at https://github.com/s-gupta/rcnn-depth (accessed on 29 September 2021). Jaritz et al. [151] propose a novel network (shown in Figure 3a) achieving depth completion and semantic segmentation by the late fusion manner. The features are extracted from the depth and RGB images, respectively, by two different NASNet encoders. Then, a channel-wise concatenation operator followed by series of convolutions is used for the features fusion. However, the original depth images are sparse. Therefore, the depth images need to be adjusted to a similar resolution with the feature map of RGB images for element-wise feature fusion. 3DMV [152] (shown in Figure 3b) is a joint two-stream network for the 3D semantic segmentation. The 2D stream uses the modified ENet without proxy loss score layer to extract the 2D feature based on multi-view images. Additionally, a voxel max-pooling operation is utilized to integrate the 2D features from multiple views. The 3D stream takes the voxel encoded with the two-channel binary as input for 3D feature extraction using the 3D convolutions. To manage the joint 2D–3D feature fusion, the author comes up with a differentiable back-projection layer to project the 2D features into the 3D volumetric representation based on the known 6-DoF pose alignments. The 2D and 3D feature vector fusion is achieved by the concatenation operation. According to the ablation study and the evaluation result, the fusion strategy of joint 2D–3D features effectively improves the performance compared with the geometry-only, image-only, and voxel-color-only segmentation. This method is shared with a good quality code at https://github.com/angeladai/3DMV (accessed on 29 September 2021).

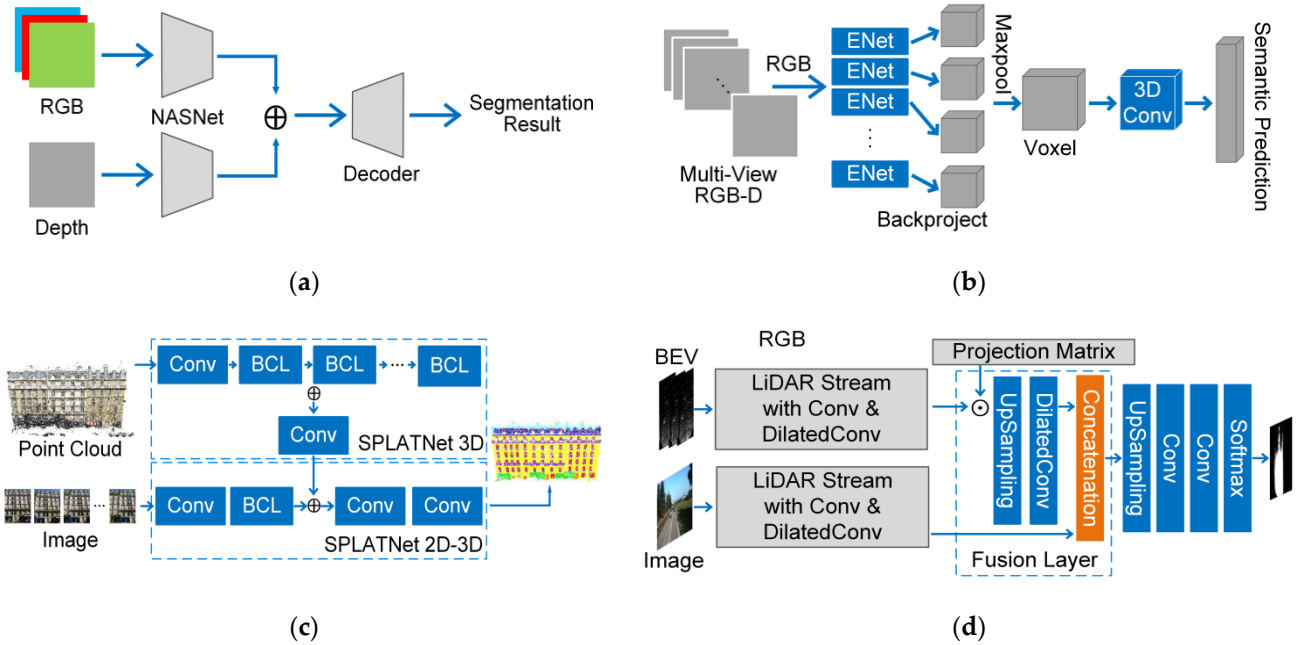


Figure 3. The representative segmentation methods using one-stage feature-based fusion strategy. (**a**) The method achieving depth completion and semantic segmentation by the late fusion manner; (**b**) 3DMV; (**c**) SPLANet; (**d**) TSF-FCN.

(2) Point Cloud Registered with RGB Images

Since the great difference of data structure between point cloud and RGB image, the two-stream network is popular. Moreover, most of them achieve the feature fusion dependent on not only the fusion operation, such as concatenation and summation, but also the intermediate data to overcome the gap between point cloud and image. For example, the BEV represents the point cloud by 2D grid and preserves sorts of geometric information. The TSF-FCN [153] (shown in Figure 3d) is a network consisting of LiDAR stream and RGB stream. The LiDAR stream aggregates the multi-scale contextual information based on the BEV grid whose cells are encoded with the mean height, gap between maximum height and minimum height, and occupancy. The RGB stream extracts the 2D features based on the raw front-view images rather than naively projecting them to the fixed bird-view images. After extracting the features of point clouds and images, respectively, using the dilated convolution, the feature fusion is achieved by the fusion layer. Finally, the fused features are imported into the decoder for further feature extraction, resolution recovery, and pixel-wise semantic labeling. The main contribution of TSF-FCN is the novel fusion layer and the BEV generation method. The feature fusion layer converts the feature map of an image from the encoder of the RGB stream to the same coordinate frame with the LiDAR grid map.

Instead of using the concatenation and summation operation, some researchers utilize the powerful CRF to fuse the segmentation score map from each branch based on the prior knowledge and the mutual relationship between image and point cloud. Yang et al. [154] effectively fused the features of image and point clouds by the BEV representations. Moreover, the authors created a fully connected CRF consisting of a unary potential and a pair-wise potential, which was optimized by the mean-field approximate algorithm. The unary potential consists of both the local and the global feature from the encoders of FCN and PointNet++. To enforce the robustness and maintain the local consistency of color and height, the pair-wise potential consists of the color bilateral kernel, height bilateral kernel and spatial kernel for small, isolated obstacle removal. The height bilaterial kernel is taken from a dense height image generated based on the sparse point clouds by the Markov-based up-sampling method. Since this method aims to segment the road, the height bilateral kernel based on the assumption that the road area is a large flat with high consistency of height helps a lot for the robustness enforcement. Although the BEV performs well, it may cause geometric information loss. Therefore, the especially designed structure for better performance of feature fusion is proposed in the SPLATNet [155] (shown in Figure 3c). It includes a brand-new bilateral convolutional layer (BCL) with the lattice indexing structure for the hierarchical and spatially aware feature learning and joint 2D–3D reasoning. There are finer lattices (larger lattice scales) at an early fusion stage and coarse lattices (smaller lattice scales) when the network goes deeper. Additionally, the end-to-end SPLANet 2D-3D with a "2D-3D Fusion" module maps the 2D pixels into the 3D space and vice versa for both the 2D and 3D semantic segmentation. The 2D features from multi-view images are extracted by the DeepLab, and the 3D features from point clouds are extracted by the SLATNet 3D architecture. The "2D-3D Fusion" firstly concatenates two feature vectors of image and point cloud, then further processes the concatenated features by using a series of 1×1 convolutional layers. This method is shared with a good quality code at https://github.com/NVlabs/splatnet (accessed on 29 September 2021).

However, there is a special method achieving the 2D and 3D feature fusion in a one-stream end-to-end network. MVPNet [156] (Multi-View PointNet) projects the 2D feature map extracted by the video stream to the point cloud. Then, 2D feature vectors are concatenated with the geometry coordinates (X,Y,Z) for each point. Finally, the point cloud with both geometry and appearance information are input into PointNet++ for the semantic segmentation.

1. Multi-Stage Fusion Strategy

Similar to the networks with a one-stage feature fusion strategy, most of the methods with a multi-stage feature fusion strategy are two-stream networks which extract the features of images and point cloud, respectively. However, the multi-stage method fuses the 2D and 3D feature more than once at different positions of the network. The networks are more complicated than the one-stage network, but they may make better use of both 2D and 3D features or achieve multi-tasking.

(1) RGBD Images

Li et al. proposed the LSTM-CF [157] with the long short-term memorized context fusion (LSTM-CF) mode. The long short-term memory layer and a long short-term memorized fusion layer are designed for the global context feature learning. Moreover, instead of simply concatenating the features and overlooking the strong correlation between depth and photometric channels, the fusion layer integrates features from the different channels in a data-driven manner. The feature of HHA and RGB information are extracted, respectively, by a stack of convolutional layers and a long short-term memory layer, and then integrated by the concatenation operation. Then, the proposed memorized fusion layer further integrates the 2D and 3D information by a data-driven adaptive fusion manner, which achieves the bi-directional propagation vertically. Liu et al. [158] propose a two-stream network based on the DCNN. Additionally, they compare the early fusion and late fusion by a series of ablation studies. The early fusion method concatenates the feature of the image and HHA extracted at an early stage. The late-fusion method computes the average or weighted summation of the score maps of RGB and HHA streams by the CRF. The method using the late-fusion strategy with weighted sum is slightly better than the others. 3D-SIS [159] is a detection-based instance segmentation network which jointly learns both the 2D and 3D features (shown in Figure 4b). The pixel-wise 2D features are extracted by series of convolutions and then back-projected to the holistic scenes generated by 3D reconstruction based on the pose alignment. The 3D convolutional feature-extraction backbone consists of the 3D geometry stream and the 3D color stream. Then, the 2D color, 3D color, and 3D geometric features are integrated by the concatenation operation. Finally, the concatenated features map serves the detection box generation and classification, respectively, for the instance segmentation. In addition, the 3D color and geometry featured are secondly fused for the detection box fine-tuning.

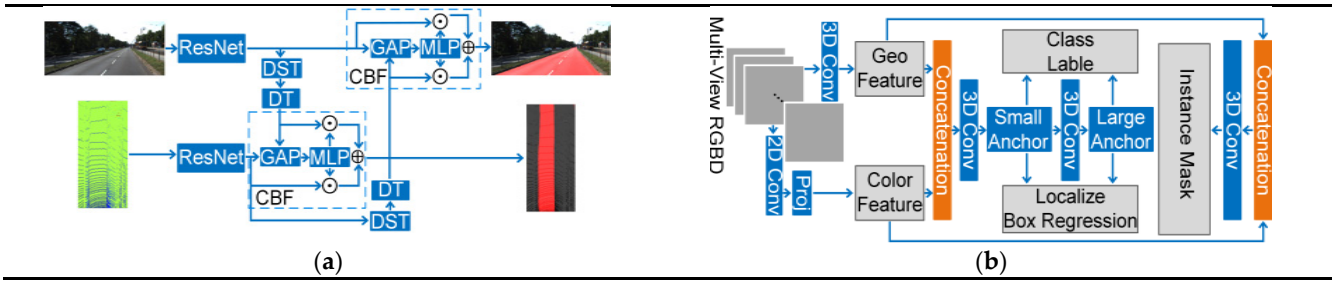


Figure 4. The representative segmentation methods using multi-stage feature-based fusion strategy. (**a**) BiFNet; (**b**) 3D-SIS.

(2) Point Cloud with 2D Images

Yu et al. [160] propose a fully convolutional network with a two-stream encoder for the feature extraction and a multi-stage residual fusion module for the feature fusion. This network takes the BEVs representation generated by the point clouds and images as input. The two-stream encoder extracts the image and point clouds features by five blocks, respectively. There are slight differences between the first blocks of two streams, but the second to fifth layers are all composed of the residual layers and the bottleneck block layers. Rather than limiting the feature fusion at a single early, middle, or late fusion stage, this method takes residual learning for the multi-modal feature fusion. The multi-stage residual fusion module is composed of the residual fusion (ResFuse) module fusing the features

of different sensors and the multi-stage fusion (MSFuse) integrating the features from different layers. However, the projection between image, point cloud, and BEV may cause the information loss or deformation. So, the BiFNet [161] (bidirectional fusion network) introduces a dense space transformation (DST) module to solve the mutual relationship between the coordinates of the camera, point cloud, and BEV (shown in Figure 4a). Moreover, the proposed context-based feature fusion (CBF) module integrates the 2D and 3D feature by the adaptive weighted summation operation, which stimulates the useful feature and suppresses the useless feature. The weight matrix for summation is learned by the three-layer perceptron.

2. Cross-Level Fusion Strategy

The one-stage and multi-stage fusion strategies only fuse the 2D and 3D features in a certain layer of the two-stream network. However, one-stage and multi-stage fusion strategies fail to take the advantage of the hierarchy of commonly used segmentation networks. Therefore, a series of cross-level fusion strategies are proposed to integrate the 2D and 3D feature from each layer of hierarchy.

(1) RGBD

The network proposed by Yuan et al. [162] extracts the features of the RGB and HHA-encoded depth images, respectively, by the Deeplab. The features from each layer of the hierarchy are extracted by the atrous convolutions, and the depth feature learning is hierarchically inserted into the layer of RGB feature learning. Then, the depth features and RGB features are fused by an element-wise summation fusion strategy in the fusion layer to preserve the essential information from both branches. ACNET [163] (shown in Figure 5a) (attention complementary network) is a ResNet-based triple-branch network consisting of a RGB branch, a depth branch, a fusion branch, and a novel feature fusion operator. The author thinks that fusing the feature too early or late may hurt the original RGB and depth information. Therefore, the features from each layer of both RGB and depth branch are fused in the fusion layer hierarchically. The main contributions of this network are not only the fusion layer but also the novel cascaded feature fusion operator based on the attention complementary module (ACM). The features from both the RGB branch and depth branch are attention-based weighted summed up and the parameters of weight are learned by the ACM based on the feature map. Additionally, the undifferentiated concatenation appears ambiguous to learn the cross-level complementary feature. This method is shared with a good quality code at https://github.com/anheidelonghu/ACNet (accessed on 29 September 2021). Chen et al. [164] propose a cross-level distillation stream for the complementary feature learning. In addition, a channel-wise attention mechanism is proposed to adaptively select the complementary feature from each modality in each level. Moreover, TSNet [165] (shown in Figure 5d) is a three-stream self-attention network with a RGB stream with VGGNet16, a depth stream with ResNet34 and a novel cross-level distillation stream with self-attention which extracts the complementary feature in the bottom-up path. The ResNet is specially selected to preserve the edge contour information of depth image when the network goes deeper. Moreover, the features of different data in each layer of encoder are fused by the convolution layer and the ASPP (atrous spatial pyramid pooling).

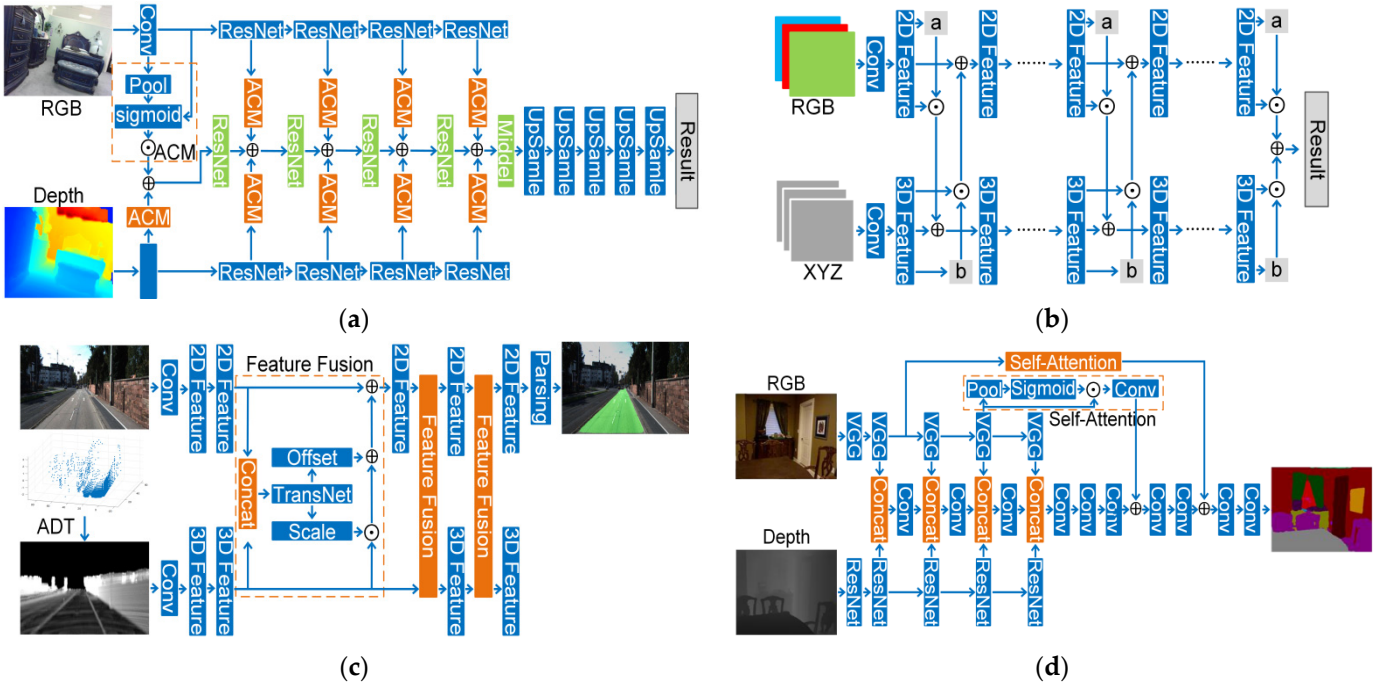


Figure 5. The representative segmentation methods using cross-level feature-based fusion strategy. (**a**) ACNet; (**b**) The method with cross fusion strategy; (**c**) PLARD; (**d**) TSNet.

(2) Point Cloud Registered with Images

FloorNet [166] is a triple-stream hybrid network with a PointNet branch, a floorplan branch, and an image branch to generate pixel-wise floorplan. The PointNet branch exploits PointNet to extract the features of 3D point clouds. The floorplan branch uses the FCN with the skip connections to learn the feature of point-density image generated from the top-down view. The image branch employs the dilated residual network (DRN) for the semantic features and utilizes the stacked hourglass CNN (HG) for the room layout features. The main innovation of this network is the mechanism of features fusion and the propagation method across different branches. The feature fusion process firstly applies the pooling module to integrate the features of disordered points and every 20 frames of video sequence. Secondly, the aforementioned features are projected to the cell of the top-down feature map in floorplan branch, then all the features in the same cell are summed up for fusion. Moreover, features from floorplan branch are also propagated to PointNet branch by reversing the pooling operation. This method is shared with a good quality code at https://github.com/art-programmer/FloorNet (accessed on 29 September 2021). Caltagirone et al. [167] propose a novel cross fusion strategy (shown in Figure 5b) based on the FCN. Both the image and point cloud feature vector are propagated and self-adaptive weighted summed up in each layer across different streams. Kim et al. [168] introduce a two-stream network whose image stream learns the 2D feature by the modified ENet and the point cloud stream consists of the 3D convolution layer, max-pooling layer, and up-sampling layer. The input of the point clouds stream is the voxel generated by the point clouds and encoded with the roughness and the porous feature. Moreover, the proposed project module projects the 3D feature to the 2D feature map according to the camera intrinsic parameters. Then, the image and point cloud feature fusion is achieved by the summation operation. The network proposed by Chiang et al. [169] takes the 3D mesh as input and extracts the 2D textural appearance, 3D local geometry, and 3D global context features for the holistic 3D point clouds semantic segmentation. The 2D appearance feature is learned by the 2D-CNN based on the images rendered by the 3D mesh and is then projected into the 3D local geometric feature map and 3D global context feature map for the feature fusion using the concatenation. Then, the two-stream encoder

is composed of the sub-volume encoder, global scene encoder, segmentation decoder, and skip-connection for further feature learning and holistic point cloud segmentation. This method is shared with a good quality code at https://github.com/ken012git/joint_point_based (accessed on 29 September 2021). However, the aforementioned methods fuse the 2D and 3D feature, which are simply projected together. The Progressive LiDAR Adaptation-aided Road Detection (PLARD) [170] (shown in Figure 5c) integrates the adapted point clouds features with image feature map by the cascaded summation. The proposed altitude difference-based transformation (ADT) transforms the feature map of point clouds into the feature map of image based on the scale and offset parameters learned by the TransNet. The existing representations, such as the range image, voxel, and point clouds, have their own advantages and disadvantages. The range image is regular and generally dense, but the data projection may cause the physical dimensions distortion. The voxel representation is regular but parse, and the computation grows cubicly when voxel resolution increase. The point cloud is geometrically accurate but disordered. To integrate the advantages of aforementioned three representations, the three-stream network RPVNet [171] (range-point-voxel fusion network) is proposed to take the advantages and alleviate the shortcoming of range image, voxel, and point clouds with the gated fusion module (GFM) for the self-adaptive features fusion. The self-adaptive mechanism is popular for filtering the useless features. Moreover, the gated fusion module is proposed to measure the importance of each feature based on the mature gating mechanism. Even though the network structure of RPVNet is complicated, the efficiency is guaranteed due to the RPV interaction mechanism using the hash mapping, the simple MLPs on point branch without local neighbors searching and taking a relatively lower resolution and sparse convolution in the voxel branch.

3. Others

There are some methods utilizing machine learning techniques that are worth including in this review.

(1) RGBD Images

Nakajima et al. [172] came up with a novel method that incrementally segmented not only known but also unknown objects using both color and geometric information, which achieved semi-real-time performance. The method includes the 3D segmentation map generated by the data-based fusion method and the incremental clustering based on the feature-based method. The 3D segmentation map is the crucial part of this method and will be updated incrementally when segmenting unknown objects. The process of generating 3D segmentation map includes: 3D reconstruction based on dense SLAM; superpixel segmentation using the modified SLIC based on the distance metric generated by CIELAB color, normal map and image coordinates; and agglomerative clustering the superpixel segmentation result. To cluster the superpixel for the object-level segmentation, an incremental clustering method is proposed by fusing and updating the geometric feature, deep feature, and entropy extracted by the depth image, RGB image, and 3D segmentation map. In addition, a weighted affinity is computed based on the similarity of geometric feature, the similarity of deep feature and entropy of the probability distribution of CNNs for incremental clustering improvement.

(2) Point Cloud Registered with Images

Multiple feature fusion is proved effective by Martinovic et al. [173] but sacrificing the efficiency. The descriptor for each 3D point includes the 2D features of image (mean RGB color, LAB value of mean RGB, spin-image (SI) descriptor) and 3D feature of point clouds (normal, height above ground plane, inverse height, depth from defined facade plane). The 132-dimensional descriptors of point cloud are clustered by the random forest classifier for the facade grouping based on the proposed 3D Weak Architectural Rules (3DWR) and the result is post-processed by CRF.

Table 8. Comparative summary of segmentation methods using feature-based fusion strategy (P: point cloud; Data 2: intermediate data).

	Name	Fusion Strategy	Data	Data 2	Data Set	Acc	Class Acc	AP	IoU	F1-Score	Scene	Code
[150]	_	One-stage	RGBD	_	NYUD2	_	35.1	32.5	_	_	Indoor	Available
[149]	_ _	One-stage	RGBD	Graph HHA	NYUD v2 SUN	_ _	57 52.5	_ _	43.6 40.2	_ _	Indoor	No
[151]	_	One-stage	RGBD	_	RGBD Synthia Cityscapes	_	_ _	_ _	70.7 57.8	_ _	Outdoor	No
[152]	3DMV	One-stage	RGBD	Voxel	ScanNet	71.2	_	_	_	_	Indoor	Available
[155]	SPLATNet	One-stage	RGB, P	Lattice Indexing	ShapeNet	_	_	_	83.7	_	Indoor	Available
[154]	_	One-stage	RGB, P	BEV	KITTI	_	_	88.5	_	91.4	Outdoor	No
[153]	TSF-FCN	One-stage	RGB, P	BEV	KITTI	_	_	95.4	_	95.42	Outdoor	No
[156]	MVP-Net	One-stage	RGB, P	_	ScanNet	_	_		64.1	_	Indoor	No
[161]	BiFNet	Multi-stage	RGB, P	BEV	KITTI	_	_	95.8	_	97.88	Outdoor	No
[157]	LSTM-CF	Multi-stage	RGBD	HHA	NYUD v2 SUN	–	49.4 48.1	_	_	_	Indoor	No
[158]	_	Multi-stage	RGBD	HHA	RGBD NYUD v2	70.3	51.7	_	41.2	54.2	Indoor	No
[159]	3D-SIS	Multi-stage	RGBD	Voxel	ScanNet	_	36.2	_	_	_	Indoor	No
[160]	_	Multi-stage	RGB, P	BEV	KITTI	_	—	96.9	_	95.98	Outdoor	No
[162]	_	Cross-level	RGBD	HHA	NYUD v2	_	49.9	_	37.4	51.2	Indoor	No
[163]	ACNET	Cross-level	RGBD	_	SUN RGBD NYUD v2	_ _	_ _	_ _	48.1 48.3	_	Indoor	Available
[164]	_	Cross-level	RGBD	HHA	NLPR	_	_	_	_	86.2	Outdoor	No
[165]	TSNet	Cross-level	RGBD	_	NYUD v2	73.5	59.6	_	_	46.1	Indoor	No
[166]	Floor-Net	Cross-level	RGB, P	Voxel	Self-collected	_	57.8	_	_		Outdoor	Available
[167]	_	Cross-level	RGB, P	_	KITTI	_	_	96.2	_	96.25	Outdoor	No
[168]	_	Cross-level	RGB, P	Voxel	Self-collected	_	_	74.7	_		Outdoor	No
[170]	PLARD	Cross-level	RGB, P	ADT	KITTI	_	_	_	_	97.77	Outdoor	No
[169]	_	Cross-level	3D Mesh	Voxel, Mesh	ScanNet	_	_	_	63.4	_	Indoor	Available
[171]	RPV-Net	Cross-level	RGB, P	Voxel	Semantic-KITTI	_	_	_	70.3	_	Outdoor	No
[172]	_	Others	RGBD	_	NYUD v2	_	_	_	46.1	_	Indoor	No
[173]	_	Others	RGB, P	_	RueMonge	_	61.4	_	_	_	Outdoor	No

4.2. Detection

Since the task of object detection aims to obtain the localization, bounding box, and the corresponding class, the structure of the network and fusion strategy contains great differences with the segmentation methods. In our opinions, the fusion strategies for the object detection can be grouped into data–result-based and feature-based. The category is defined based on how they achieve the 2D and 3D information fusion.

4.2.1. Data–Result-Based Fusion Strategy

The data–result-based fusion strategy integrates the 2D detection result with 3D raw data and vice versa. For example, the 2D detection results can be projected to 3D as the regions of interest for the further refinement based on the 3D geometric feature. The comparative summary of detection methods using data–result-based fusion strategy is shown in Table 9; their details will be introduced in the following paragraphs.

To improve the 2D detection, some researchers project the 3D proposals or bounding boxes onto the 2D images as the regions of interest (RoI). Then, the 2D feature is further learned for 2D bounding box regression based on the RoIs [3]. Arcos-García et al. [174] put forward a robust traffic sign detection method utilizing both the point clouds and the images collected by a vehicle equipped with LiDAR and RGB cameras. Firstly, the point cloud is pre-processed by removing the points whose distance from trajectory registered by MMS are further than 15 m and removing the ground by using a ground region growing method based on the voxelized point clouds. Since the traffic signs are planes made of

retro-reflective materials, the intensity information is useful for traffic signs segmentation. The unsupervised classification algorithm based on the gaussian mixture models (GMM) is selected for the coarse traffic signs segmentation by collecting the components with the largest mean intensity and then being refined by the DBSCAN algorithm and PCA. Secondly, they project the 3D clustering result to the proper image based on the cameras calibration parameters as 2D RoI for the further refinement using the visual information and the 2D ConvNet. Guan et al. [175] also propose a robust method in which the 3D traffic signs detection results based on point clouds are projected onto the RGB image as the 2D ROIs. The supervized Gaussian-Bernoulli deep Bolzmann machine model is utilized for the final detection. The method proposed by Barea et al. [176] fuses the 2D proposals generated by the 2D semantic segmentation result and the 2D proposals projected from the 3D box. Those two kinds of proposals are integrated depending on the overlap situation and geometric restrictions. With the development of deep learning techniques, this fusion strategy is generalized to the neural network. Guan et al. [177] introduced a two-stage convolutional capsule network for the traffic signs detection which exploits both mobile LiDAR point clouds and images. At the first stage, the traffic signs are detected by a supervoxel segmentation method based on the pole height, road width, intensity, geometrical structure, and traffic sign size. Then, the 3D segmentation patches are projected onto the image planes as 2D region proposals. Finally, the 2D region proposals are further refined by a novel convolutional capsule network consists of a convolutional CNNs for low-level feature extraction and a series of capsule layers encoded with the low-level feature and high-order vectorial capsule representation for more powerful and robust feature extraction.

Not only the 3D proposals or bounding boxes can be transformed as the 2D RoIs, but also the 2D proposal or bounding boxes can be projected into 3D as the frustums or 3D RoIs which will be refined based on 3D geometric feature. Lahoud et al. [178] take the full use of 2D information for 3D searching space reduction. Then, the 3D information is utilized to estimate the orientation, localization, and prediction of the bounding box of objects. The Faster R-CNN is used to estimate the initial location of objects based on the images. Then, the results of 2D detection are projected as the 3D frustums, which helps to improve the efficiency compared with the traditional approaches with the sliding window. Based on the interest region of 3D frustums, the orientation estimation is achieved by the Manhattan frame estimation (MFE). Moreover, the histograms with coordinate information of each point along each axis are used as the input of MLP for the geometric feature learning and the 3D bounding box regression. Finally, the LP-MAP (Linear Programed-Maximizing the Aposteriori) is utilized for label refinement based on the appearance information, geometric information, and relationship between labels by introducing the local marginal variables as unary term and binary term. The unary term represents the probability of assigning the box a label, which includes the geometric features of length, width, height, aspect ratios, volume, and deep learning features acquired by the Fast RCNN based on the projected 3D box in the image plane. The binary term indicates the probability of assigning one box a label when giving the label of another box, which reflects the co-occurrence and the spatial distribution of class in 3D scenes. According to the method proposed by Du et al. [179], both 2D and 3D information are used to segment the car in point cloud. Any 2D detection network can be used to generate the 3D proposal for the car dimensions estimation. Based on the result of car dimension estimation, the generalized models and score maps are generated by the 3D CAD dataset. Additionally, the 2D proposals are projected as 3D RoIs. Then, a model fitting method named 3DPVs is selected for points filtering based on the generalized models and score maps. Finally, a two-stage refinement CNN takes the filtered points as the input of the 3D detection. The 2D detection results (2D proposals) of Frustum PointNets [180] (shown in Figure 6a) are projected as the 3D frustums, which helps to reduce the research space. The 2D detector of this method is the FPN pretrained by the ImageNet and COCO and refined by the KITTI 2D. According to the SiFRNet [181] (shown in Figure 6c) proposed by Zhou et al., the result of 2D detector provides 3D frustum for the

point cloud. Then, the 2D feature vector and the 3D point coordinates in each frustum are fused by the concatenation and further learned by the proposed Point-UNet composed of the T-Net and the Point-SENet. The Frustum ConvNet [182] (shown in Figure 6d) utilizes the 2D proposals to generate the 3D frustums first. Then, the frustum-wise features of point clouds derived by the PointNet are reformed and concatenated for FCN to generate the 3D bounding box and classification result. This method is shared with a good quality code at https://github.com/zhixinwang/frustum-convnet (accessed on 29 September 2021). The RoarNet [183] includes both the 2D and 3D part. The 2D part generates the 2D bounding boxes and estimates their 3D pose for 3D region proposal generation. Based on the 3D feasible regions, the 3D part takes two simplified PointNet for better 3D proposals generation and bounding box regression.

Table 9. Comparative summary of detection methods using data–result-based fusion strategy (P: point cloud; Data 2: intermediate data).

	Name	Task	Data	Data 2	Dataset	IoU Threshold	AP	Scene	Code
[178]	_	3D Det	RGBD	Graph	SUN RGBD	0.25	45.12	Indoor	No
[180]	F-PointNet	3D Det	RGBD	_	KITTI	0.7 car	71.26	Outdoor	No
						0.5 ped.	45.44		
					SUN RGBD	0.5 cyc.	59.71	Indoor	
						0.25	54		
[174]	_	2D Det	RGB, P	Voxel	GTSRB	_	99.71	Outdoor	No
[184]	IPOD	2D Det	RGB, P	_	KITTI	0.7	88.96	Outdoor	No
		3D BEV Det					82.92		
		3D Det					72.88		
[175]	_	2D Det	RGB, P	Voxel	Self-collected		93.3	Outdoor	No
[176]	_	2D BEV Det	RGB, P	—	KITTI	0.7	80.64	Outdoor	No
[179]	PC-CNN	3D Det	RGB, P	BEV	KITTI	0.5	82.09	Outdoor	No
						0.7	53.59		
		3D BEV Det				0.5	83.89		
						0.7	76.86		
	MC-CNN	3D Det				0.5	84.65		
						0.7	54.32		
		3D BEV Det				0.50.7	87.86		
							77.91		
[177]	_	2D Det	RGB, P	Voxel	Self-collected	_	95.7	Outdoor	No
					SUN RGBD	0.25	58.4	Indoor	
[181]	SiFRNet	3D Det	RGB, P	BEV	KITTI	overall	66.99	Outdoor	No
						0.7 car	73.95		
						0.5 ped.	61.05		
[182]	Frustum ConvNet	3D Det	RGB, P	–	SUN RGBD	0.5 cyc.	65.97	Indoor	Available
						0.25	57.55		
				–	KITTI	0.7 car	76.82	Outdoor	
						0.5 ped.	46.49		
						0.5 cyc.	67.1		
[183]	RoarNet	3D Det	RGB, P	_	KITTI	0.7	72.77	Outdoor	No

However, there are some special methods that utilize the 2D segmentation result to improve the process of 3D detection. IPOD [184] (shown in Figure 6b) is a 3D detection network consisting of the background removing part, point-based proposal generation part, proposal feature generation module, and box prediction network. The 3D RoI generated by the images segmentation results are used as the 3D proposals. Vora et al. introduced a novel PointPainting [185] method that appends the class score of each point with the 2D segmentation result. According to this method, the point clouds are projected onto the image plane firstly. Once the point falls on a pixel, the relevant pixel feature vector will be concatenated with the point's coordinate. Then, the concatenated feature map can be imported to any point-cloud-only method for further feature learning.

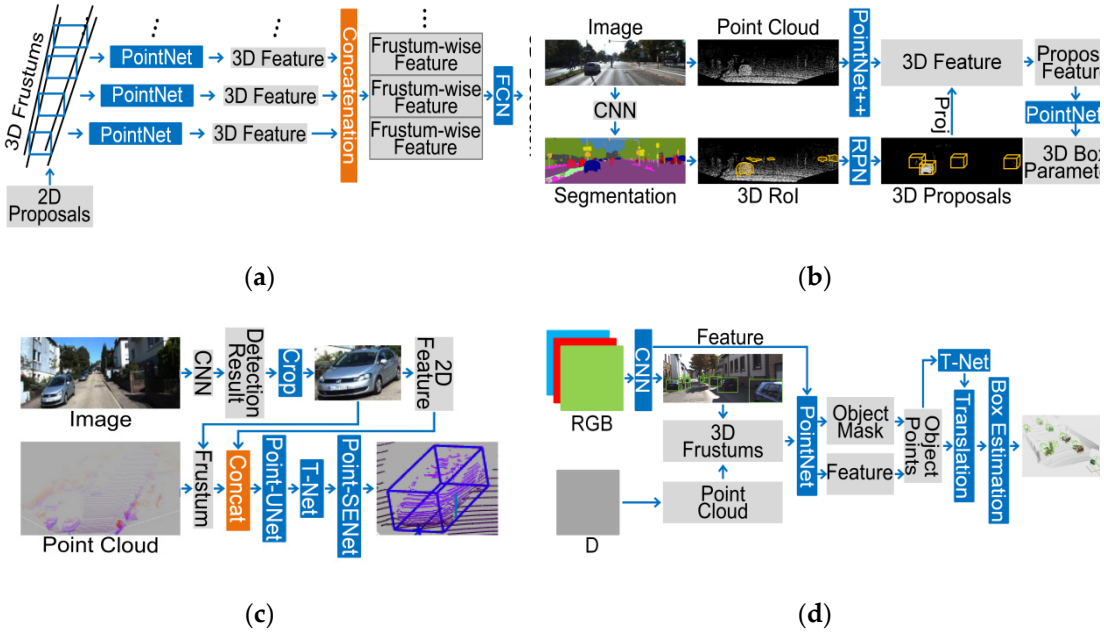


Figure 6. The representative detection methods using data–result-based fusion strategy. (**a**) Frustum PointNet; (**b**) IPOD; (**c**) SIFRNet; (**d**) Frustum ConvNet.

4.2.2. Feature-Based Fusion Strategy

Similar to the segmentation network which fuses the 2D and 3D feature, most of the detection network are designed with a two-stream structure. The feature-based fusion strategies are also divided into three groups: one-stage, multi-stage, and cross-level fusion strategy. The comparative summary of detection methods using the feature-based fusion strategy is shown in Table 10, and their details will be introduced in the following paragraphs.

PRN [186] (3D Region Proposal Network) is the first joint object recognition network (shown in Figure 7a) which jointly learns the 2D color and 3D geometric feature. The author firstly puts forward the deep sliding shapes with 3D ConvNets for the 3D proposal feature learning based on the volumetric representation generated by the RGBD images and utilizes the 2D VGG to extract the color features at the same time. To achieve the feature fusion, a concatenation operation followed by a fully connected layer is selected for the voxel-wise feature fusion. Deng et al. [187] aims to predicting the 3D locations, physical sizes, and orientation simultaneously by the RGBD images (shown in Figure 7b). The 2D ROI proposal is generated by the MCG algorithm based on the RGB and depth values. Based on the 2D RoIs, the appearance feature of RGB image and geometric feature of depth image are extracted, respectively, and then fused by the concatenation. Rather than directly back projecting the 2D segmented pixel to the 3D space, this method generates the 3D box proposals derived from the corresponding 2D segment proposals. Then, the 3D bounding boxes are refined by learning the offsets of a seven-element vector for the 3D proposals. Furthermore, a multi-task loss is utilized to the jointly train the classification and the bounding box regression process. This method is shared with a good quality code at https://github.com/phoenixnn/Amodal3Det (accessed on 29 September 2021).

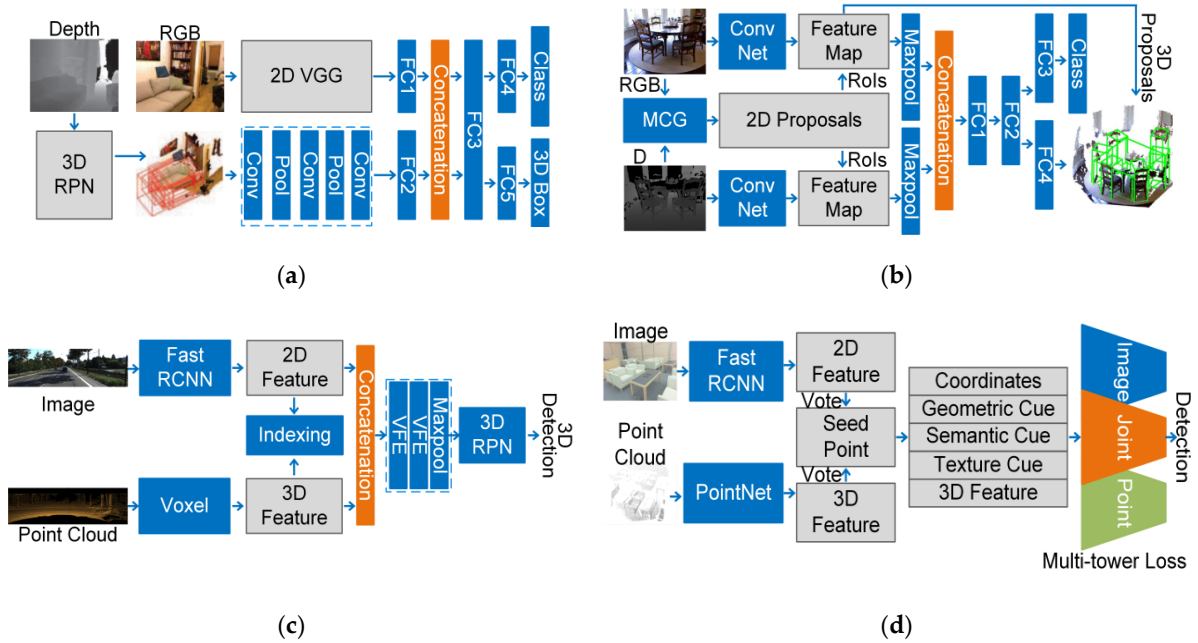


Figure 7. The representative detection methods using one-stage feature-based fusion strategy. (**a**) The joint object recognition network; (**b**) The structure of the method; (**c**) MVX-Net; (**d**) ImVoteNet.

(2) Point Cloud Registered with Images

The BEV and voxel are also commonly used in the task of detection which take both 2D and 3D feature into consideration. The following two methods integrate the 2D and 3D information based on the BEV representation. Wang et al. [188] put forward a novel sparse non-homogeneous pooling layer for the feature fusion, which is a cross-bridge between the MSCNN backbone and VoxelNet backbone for the 2D and 3D feature fusion. HDNet [189] is a single-stage detector which takes both the geometric and semantic features from HD maps to improve the 3D detection. To achieve the features fusion, the point clouds need to be transformed into the BEV representation. The features from each layer of HD map feature learning network are fused with the 3D feature map acquired by the average pool down-sampling, coping, and bilinear up-sampling operation.

The voxel is useful for the feature fusion as the intermediate data. However, the MVX-Net [190] (shown in Figure 7c) compares the point-based and voxel-based fusion method. The result shows that the point-based fusion method performs slightly better than the voxel-based fusion method. The PointFusion aggregates the dense 2D context information for each 3D point using the point-wise concatenation operation. The VoxelFusion fuses the 2D feature and the 3D feature learned by stacked VFE using the voxel-wise concatenation operation.

To solve the problem that BEV and voxel may cause the geometric information loss due to the projection and the resolution of the grid, the networks achieving the pointwise 2D and 3D feature fusion based on raw unordered point cloud are worthy of further research. ImVoteNet [191] (shown in Figure 7d) is a two-stream 3D detection network consisting of the modified VoteNet, 2D vote generation module, 3D vote generation module, and novel multi-tower formulation. In the point cloud stream, K seeds are selected from the point clouds. Then, each seed point is encoded with the concatenated coordinate, 2D votes and 3D votes. The 2D votes includes the geometric cue, semantic cues and texture cues based on the 2D detection results generated by Faster R-CNN. The 3D vote contains the 3D coordinates and local point clouds feature extracted by the PointNet. For better training and testing, a novel multi-tower formulation that including the image tower, point tower and joint tower is proposed. The image tower and point tower are only used for the training process; the joint tower is used for both training and testing process.

There is a special method which takes the raw point cloud, BEV voxel, and perspective view images generated by point cloud as inputs. Zhou et al. propose an end-to-end multi-view fusion (MVF) [192] algorithm which consists of the dynamic voxelization (DV) and the multi-view feature fusion operation. Compared with the traditional hard voxelization method, which may result in information loss, non-deterministic voxel embeddings, and unnecessary computation, the dynamic voxelization overcomes those drawbacks by preserving the complete mapping between points. Moreover, the number of voxels and the number of points inside each voxel are dynamic, so it eliminates the information loss caused by the stochastic dropout of points. Additionally, this method takes the advantages of raw unordered points, BEV voxelization, and perspective view representation. The point-wise feature extracted from the raw points, voxels and perspective view images are concatenated together for the feature fusion. Although this method only takes the 3D point cloud as input, it still provides an idea of integrating the feature from multiple forms of data.

2. Multi-stage Fusion Strategy

(1) RGBD Images

Xu et al. [193] introduced a multi-task network trained by a weighted multi-task loss for the 2D detection. The 2D and 3D features are fused three times in this method. Firstly, the depth information and the 3D coordinates are generated based on the camera intrinsic parameters and disparity information estimated by the monocular images. Then, the front view feature maps encoded with depth information are concatenated with the original monocular images for data enforcement. After generating the 2D region proposals based on a series of convolutional layers, the 3D point feature generated by mean pooling and the 2D features generated by max pooling of the same RoI are integrated by the concatenation operators for the multi-class classifier, 2D box regression, 3D orientation regression, 3D orientation regression, and 3D dimension regression. Finally, the final jointly learned features from 2D stream and 3D stream are fused by the summation operation for accurate 3D location estimation. This method is shared with a good quality code at https://github.com/mrharicot/monodepth (accessed on 29 September 2021).

(2) Point Cloud Registered with Images

The following four methods take the BEV as intermediate data for the 2D and 3D feature fusion. MV3D [194] takes the front view and bird's eye view representation generated by the point cloud and the RGB images as input. The 3D proposal part generates the 3D candidate boxes based on the bird's eye view representation. Then, the 3D proposals are projected on the front view representation and the image for the region-wise feature extraction. Finally, the region-based fusion part integrates all the region-wise features by the proposed hierarchical deep fusion strategy which consists of the element-wise mean operation. Liang et al. [195] came up with an end-to-end deep sliding network with camera stream for the image feature learning, BEV stream for the LiDAR feature extraction, and a proposed novel feature fusion structure with the continuous fusion layer for 3D object detection. The feature fusion layer consists of the multi-layer perceptrons and the weighted summation. The image features from each layer of ResNet block are fused into the multi-scale feature map and then will be further fused with the feature maps from the BEV stream. AVOD [162] (Aggregate View Object Detection Network) is a two-stream network (shown in Figure 8a) aiming at oriented 3D bounding box regression and category classification. There are two stages of feature fusion in the network. The first fusion stage fuses the image feature and BEV feature via the element-wise mean operation for the better 3D proposal. The second fusion stage fuses the 2D and 3D features of top K proposals with the same fusion operation for the final 3D bounding box regression. The SCANet [196] includes three contributions: introducing the spatial-channel attention (SCA) module for multi-scale and global context feature fusion by the spatial and channel-wise attention, providing a ESU (extension spatial upsample) module which achieves the multi-scale features fusion to recover the lost spatial information, and proposing a novel multi-level feature fusion scheme with the concatenation and element-wise mean operation for BEV and RGB feature

fusion. The feature fusion strategy is used twice. The first point-wise fused feature is utilized for 3D proposals. Secondly, the ROI-wise feature of the BEV and RGB are fused for the 3D bounding box refinement.

Figure 8. The representative detection methods using multi-stage feature-based fusion strategy. (**a**) AVOD; (**b**) PointFusion.

Moreover, the following two methods fuse both 2D and 3D information for the disordered point cloud. PointFusion [197] is a 3D object detection network (shown in Figure 8b) with a novel feature fusion part to integrate the image features and point clouds features that extracted by the ResNet and PointNet, respectively. The feature fusion part includes a vanilla global fusion network and a novel dense fusion network. During the dense fusion process, the point-wise features and the global feature of point clouds and the image feature are fused by the concatenation operation and further used to predict the point-wise offsets of each corner. Moreover, the global fusion network fuses the image feature and point clouds features using concatenation to achieve the 3D proposal generation. The multi-task and multi-sensor fusion method created by Liang et al. [198] is an end-to-end network which takes the advantages of both point-wise and ROI-wise feature fusion. Additionally, an integrated ground estimation module is used to extract the geometric information of the ground. Moreover, the depth completion and the pseudo-LiDAR points generated based on the RGB-D images help for the denser point-wise feature fusion. The point-wise feature fusion strategy fuses the RGB-D feature and the voxel-based BEV feature maps by the element-wise concatenation. Then, the point-wise fused feature is utilized for the 3D proposals generating by using the NMS (non-maximum suppression) and score threshold. Finally, the proposals are further refined by the ROI-wise fused feature which concatenates the features of images ROI and BEV ROI projected by 3D detection results.

3. Cross-Level Fusion Strategy

Since the hierarchy is not necessary for detection network, there are few network take cross-level fusion strategies. However, the EPNet is an exception because it is a multi-task network which achieves both the semantic segmentation and the object detection. So, it needs the point-wise feature extraction and the cross-level fusion strategy. EPNet [199] (shown in Figure 9) introduces a novel LI-Fusion module which is a point-wise fusion manner enhancing the point clouds feature map with the image feature. The LI-Fusion includes a point-wise correspondence generation part and a LiDAR-guided fusion part. The point-wise correspondence generation part projects the point clouds onto the image plane. The bilinear interpolation is utilized to solve the problem that the projected points may fall between the adjacent pixels. Based on the correspondence between points and pixels, the point-wise image feature representation is fused with the point-wise feature vector by the LiDAR-guided weighted concatenation operation. This method is shared with a good quality code at https://github.com/happinesslz/EPNet (accessed on 29 September 2021).

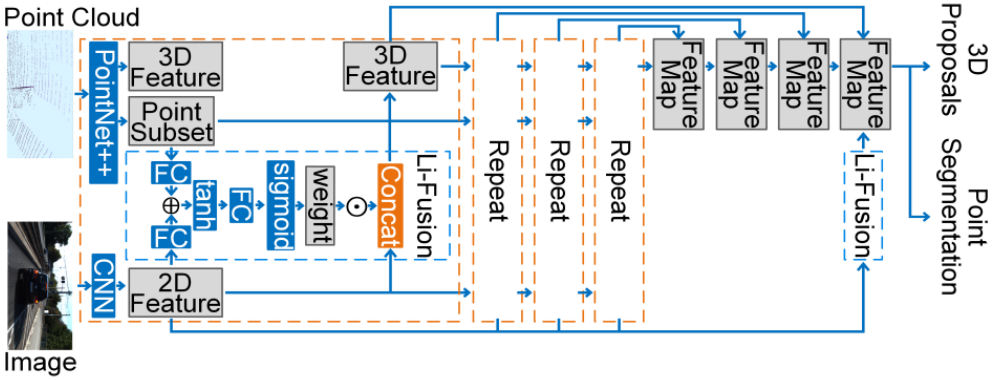


Figure 9. The representative detection methods using cross-level feature-based fusion strategy.

Table 10. Comparative summary of detection methods using feature-based fusion strategy (P: point cloud; Data 2: intermediate data).

	Name	Task	Fusion Strategy	Data	Data 2	Dataset	IoU Threshold	AP	Scene	Code
[186]	PRN	3D Det	One-stage	RGBD	_	SUN RGBD	0.25	26.9	Indoor	No
[187]	_	3D Det	One-stage	RGBD	_	NYUD v2	0.25	40.9	Indoor	Available
[188]	_	3D Det	One-stage	RGB, P	Voxel	KITTI	0.5	26	Outdoor	No
[189]	HDNet	3D Det	One-stage	RGB, P	BEV	KITTI	0.7	84.7	Outdoor	No
[190]	MVX-Net	3D Det 3D BEV Det	One-stage	RGB, P	Voxel	KITTI	0.7	73.7 84.4	Outdoor	No
[192]	MVF	3D Det 3D BEV Det	One-stage	P	PV,BEV Voxel	Waymo	0.7 car 0.5 ped. 0.7 car 0.5 ped.	62.9 65.3 80.4 74.4	Outdoor	No
[191]	ImVoteNet	3D Det	One-stage	RGB, P		SUN RGBD	0.25	63.4	Indoor	No
[193]	–	2D Det Orientation 3D BEV Det 3D Det	Multi-stage	RGB	P,FV	KITTI	0.7	84.9 84.6 10.5 5.7	Outdoor	Available
[194]	MV3D	3D Det	Multi-stage	RGB, P	BEV,FV	KITTI	0.25 0.5 0.7	91.7 91.2 63.5	Outdoor	No
[195]	–	2D Det 3D Det 2D BEV Det	Multi-stage	RGB, P	BEV	KITTI	0.7	85.4 70.9 84.0	Outdoor	No
[200]	AVOD	3D Det 3D BEV Det	Multi-stage	RGB, P	BEV	KITTI	0.7 car 0.5 ped. 0.5 cyc. 0.7 car 0.5 ped. 0.5 cyc.	73.4 44.8 54.3 83.4 52.5 58.8	Outdoor	Available
[197]	PointFusion	3D Det	Multi-stage	RGB, P	–	SUN RGBD KITTI	0.25 0.7 car 0.5 ped. 0.5 cyc.	45.4 64.7 28.3 35.3	Indoor Outdoor	No
[198]	–	2D Det 3D Det 3D BEV Det	Multi-stage	RGB, P	BEV Voxel	KITTI	0.7 car	90.2 77.3 85.4	Outdoor	No
[196]	SCANet	3D Det	Multi-stage	RGB, P	BEV	KITTI	0.7 car	67.0	Outdoor	No
[199]	EPNet	3D Det 3D BEV Det	Cross-level	RGB, P	Grid	SUN RGBD KITTI	0.25 0.7 car	81.2 88.8	Indoor Outdoor	Available

1. One-stage Fusion Strategy
(1) RGBD Images

4.3. Further Discussion

Since the feature extraction is one of the most important parts of segmentation and detection, the existing fusion strategies can be also categorized as the feature-based and non-feature-based fusion strategies. Although the general structure or pipeline differ a lot between the methods of segmentation and detection, there are still some similarities between the methods that take the same fusion strategy. We provide a further discussion to help the researchers to understand those fusion strategies better and help them to choose the proper one for their own method.

4.3.1. Non-Feature-Based Fusion Strategy

The non-feature-based fusion strategy includes the data-based, result-based, and data–result-based fusion strategy. Those types of strategies can be regarded as the pre-processing or post-processing of segmentation and detection. So, the 2D and 3D information are integrated at the input or output stage. According to our analysis of existing methods, the data-based and result-based fusion strategies are more suitable for the task of segmentation, and the data–result-based fusion strategy is more suitable for the task of detection. In addition, we abstract the general pattern of each kind of fusion strategy and depict them in Figure 10.

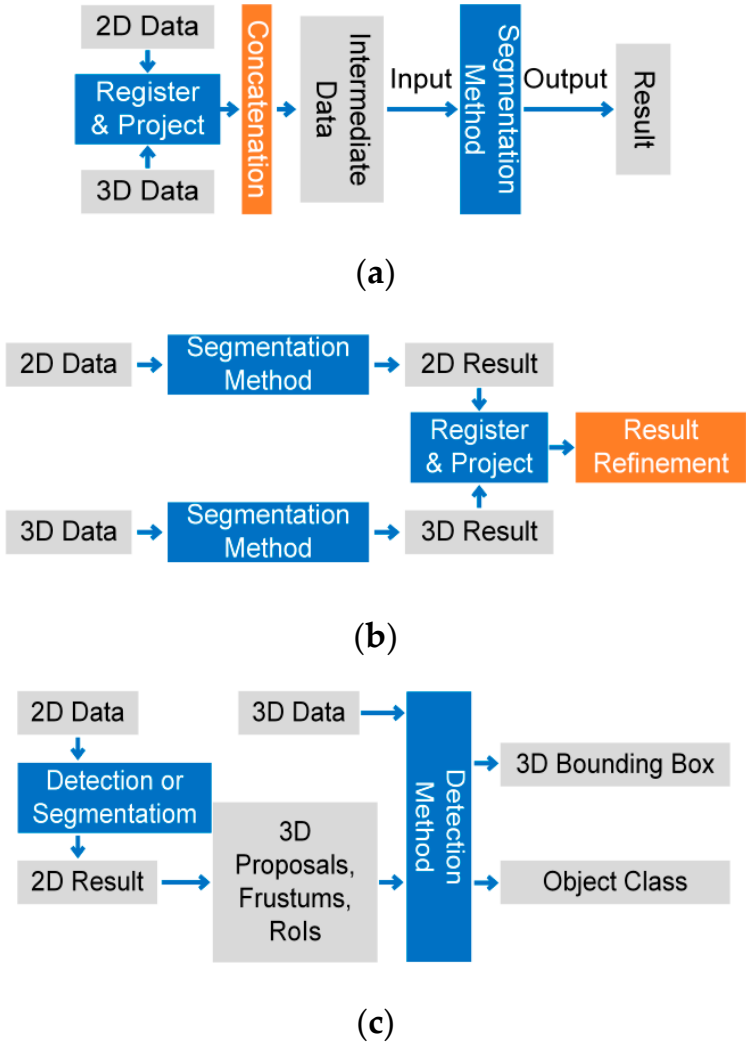


Figure 10. The general structure of non-feature-based fusion strategy. (**a**) Data-based fusion strategy; (**b**) result-based fusion strategy; (**c**) data–result-based fusion strategy.

1. Data-Based Fusion Strategy

The data-based fusion strategy is a brief strategy which integrates the 2D information of images and 3D information of point clouds. The RGBD image contains four channels of values in which the RGB value represents the 2D information and the depth value represents the 3D information. Inspired by the RGBD images, some intermediate data are proposed for the 2D and 3D information integration. For example, Erkent et al. projected the 2D segmentation results and point cloud to the BEV occupancy grid and then concatenated them into a multi-channel image as the input of the VGG-16 network for segmentation. Moreover, the spherical images are also introduced as the intermediate data for data fusion which achieves the accuracy improvement of segmentation.

2. Result-Based Fusion Strategy

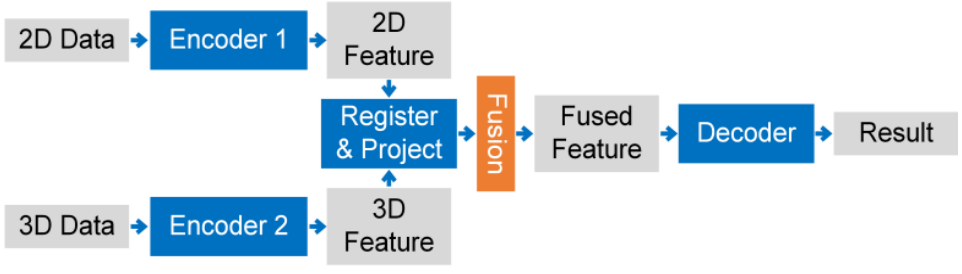
The post-processing of segmentation helps a lot to improve the accuracy of segmentation by integrating the different conditions and constraints, in which the conditional random field (CRF) plays an important role. Inspired by the existing post-processing methods, the result-based fusion strategy is proposed by integrating the 2D and 3D segmentation result based on their mutual relationship. The common idea is projecting the 2D and 3D segmentation result together for the result refinement. The CRF is widely used in this strategy. For example, the method proposed by Gu et al. [29] utilizes the CRF to penalize the inconsistency of 2D and 3D segmentation.

3. Data–Result-Based Fusion Strategy

Since the proposed-based method is important for the task of detection, the data–result-based fusion strategy is proposed to integrate the 2D and 3D information by projecting the 2D detection or segmentation results as the 3D proposals, frustums or RoIs and vice versa. The proposals, frustums, or RoIs help to lessen the processing area and lead to bounding box generation. Therefore, we name this kind of fusion strategy a data–result-based fusion strategy because the proposal, frustum, or RoIs connect the detection or segmentation result of one kind of data with another kind of data. Thanks to this strategy, both 2D and 3D data make a contribution for the bounding box generation in the task of detection.

4.3.2. Feature-Based Fusion Strategy

Most of the feature-based fusion strategy extracts the 2D and 3D features, respectively, first and then integrates their feature maps. Both 2D and 3D branches are composed of suitable networks for the 2D and 3D feature learning. For example, the classical 2D convolutional networks are applied on RGB and depth images, such as ResNet, ENet, NasNet, etc. Moreover, PointNet and PointNet++ are the most popular for the point cloud branch, according to how many times the feature-based fusion strategy based on deep learning can be grouped into one-stage, multi-stage, and cross-level based on how they achieve the feature fusion in the network. The general pattern of different fusion strategies is shown in Figure 11.



(**a**)

Figure 11. *Cont.*

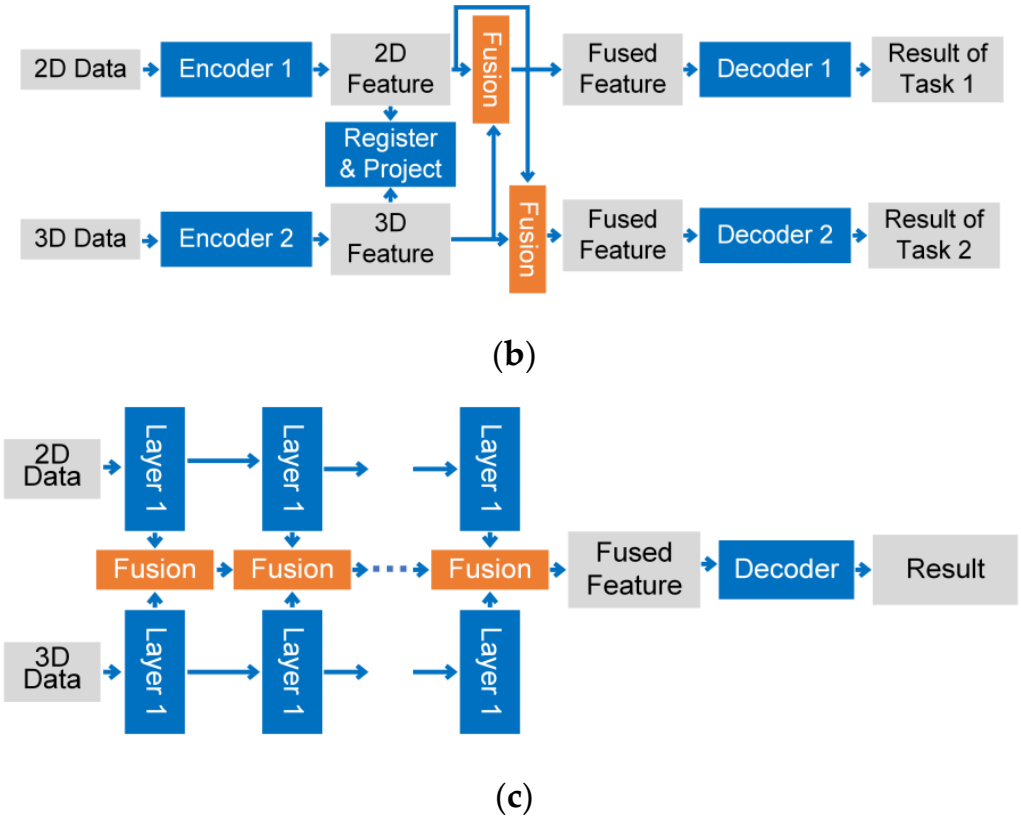


Figure 11. The general structure of feature-based fusion strategy. (**a**) One-stage feature-based fusion strategy; (**b**) multi-stage fusion strategy; (**c**) cross-level fusion strategy.

1. One-Stage Fusion Strategy

The one-stage fusion strategy is a brief strategy when it comes to feature fusion in network. The feature map from 2D and 3D branches are integrated by the summation, concatenation or pooling operation once in the network, so the position selection is important. For example, the 3DMV integrates the 2D feature of multi-view RGB images with the voxels encoded with the depth information by maxpooling, the SPLANet adopts the summation operation to integrate the feature maps of images and point cloud, NVX-Net and ImVoteNet take the concatenation operation based on the mutual relationship between image and point cloud. Moreover, the position between encoder and decoder is the most common selected position for the feature fusion because it ensures the sufficient feature learning for both data and the fused feature.

2. Multi-Stage Fusion Strategy

Although the one-stage fusion strategy achieves the feature fusion successfully, the fusion effect is limited by the one-stage fusion operation and the fusion position. To overcome this problem, the multi-stage fusion strategy is proposed for the better feature fusion. Before the final stage of bounding box regression or semantic score generation, the feature from 2D and 3D branches are integrated more than once at different positions in a network. Moreover, the multi-stage fusion strategy is suitable for multi-task networks such as instance segmentation and object detection because both of them need to generate multiple forms of result. The detection method aims to know the location and the bounding box of objects, and one of the popular instance segmentation methods is obtaining the instance-wise detection box and instance mask at the same time. For example, the 3D-SIS network takes two concatenation operations at early and late stages for the bounding box and instance mask generation, respectively. AVOD takes two pixel-wise mean pool operations for the proposal generation and bounding box regression in an end-to-end

network. Additionally, the PointFusion selects two concatenation operations at similar stages for the localization and bounding box offset generation and class recognition.

3. Cross-Level Fusion Strategy

Due to the hierarchy of the semantic segmentation network, cross-level fusion strategies are proposed to take the full use of 2D and 3D features from each layer and take the advantages of hierarchy. This fusion strategy breaks the bottlenecks when the one-stage and multi-stage fusion strategies may hurt the feature because of the improper fusion position. The ACNet and TSNet are triple-stream networks which integrate the features from RGB and depth streams at fusion stream for feature fusion, further feature learning, and resolution recovering. Moreover, the self-attention mechanism is used to filter out the useless information. However, unlike the one-stage and multi-stage fusion strategies, which are suitable for both segmentation and detection, the cross-level fusion strategy is barely used in the network for the detection due to the lack of hierarchical structure for the feature extraction. However, there is an exception named EPNet, which achieves semantic segmentation and object detection in an end-to-end network.

5. Trend, Open Challenges, and Promising Direction

Since the tasks of segmentation and detection have become important parts of real-world scenes, understanding the accuracy, reliability, and automation of segmentation and detection methods needs to be improved. Therefore, a series of deep learning techniques based on different kinds of data have been booming in recent years due to the high degree of automation and acceptable accuracy. However, the shortcomings of processing single kinds of data emerge. For example, the image fails to represent the geometric information but contains clear appearance information, point cloud represents the 3D geometric information precisely but struggles with the lack of appearance information due to the irregularity and sparsity. Therefore, series of segmentation and detection methods are constantly promoted to integrate the 2D and 3D information together for the feature complementation and accuracy improvement. Consequently, we are going to summarize some general trends, opening challenges, and promising directions in this section based on the methods reviewed above.

5.1. Trend

1. Since the 3D results are more suitable for applications in the real world, more methods tend to achieve segmentation and detection in the 3D domain.
2. The feature-based fusion strategy has gained more attention in current research based on the deep learning techniques because the feature fusion may improve the feature extraction and take the better use of both 2D and 3D feature in the deep-learning-based methods. Moreover, the multi-stage and cross-level feature achieved better fusion and accuracy, which may be the trends of future research also.
3. There is more and more research aiming to create the methods to achieve multiple tasks at the same time. For example, the combination of detection and segmentation lead to instance-level segmentation. Moreover, the depth or shape completion may be needed along with segmentation and detection for more elaborate results.
4. The interpretability of the network has become an area of growing interest and definitively will become the important research point.

5.2. Open Challenges

Based on the research trends and the methods reviewed above, there are several open challenges summarized in the following lines:

1. Dependency on Dataset

Since most of the methods are supervised, the quality, quantity, and diversity of the dataset influences the deep-learning-based methods a lot. However, it is impossible for a dataset to include every class in the real world and a perfect amount of data. Moreover,

creating a useful dataset is a labor-exhausted, time-consuming, and expensive task. So, creating a dataset that includes the registered 2D and 3D data takes more workload than creating a regular 2D or 3D dataset. Therefore, the dataset may limit the research.

2. Uncertain Correspondence between 2D and 3D Domain

Since the commonly used 2D and 3D data are organized by different data structures in different coordinates, every fusion strategy needs to project the 2D and 3D data, result, or feature map into the same coordinate. Traditionally, the correspondence is represented by six DoF parameters, which is only useful for the naive projection. Even though the naive projection is suitable for data-based and result-based fusion strategy, the feature-based fusion strategy may need more elaborate correspondence. For example, the uncertain mutual relationship between 2D and 3D neighboring regions may affect the feature fusion effect.

3. Open Challenges of fusion strategies

(1) Lack of the Suitable Intermediate Data for Data-based Fusion Strategy

Since the data-based fusion strategy is brief, efficient, and achieves acceptable accuracy in some situation, this kind of fusion strategy is worthy of further research. The intermediate data is the key point due to the great difference between 2D and 3D data. However, the popular voxel and BEV representation struggle with information loss and the balance between the efficiency and resolution. Therefore, a lack of suitable intermediate data becomes the main bottleneck of the data-based fusion strategy.

(2) Uncertain Relationship of 2D and 3D Results for Result-based Fusion Strategy

The result-based fusion strategy is a kind of post-processing method for the segmentation result refinement. After projecting the 2D and 3D results to the same coordinates, the CRF is commonly used for the result refinement based on an energy function which consists of the 2D unary potential, 3D unary potential, and 2D–3D joint relationship. However, the quality of both 2D and 3D result are unclear. Therefore, the reliability of this fusion strategy is also unclear.

(3) Dependency on the quality of RoIs/Frustums/Proposals for Data–result-based Fusion Strategy

The data–result-based fusion strategy integrates the 2D and 3D information by projecting the detection result of one kind of data to another kind of data as RoIs/frustums/proposals for further feature learning and bounding box regression. However, this kind of fusion strategy relies on the quality of the first stage of detection.

(4) Choosing Suitable Fusion Operation for Feature-based Fusion Strategy

The feature-based fusion strategy slightly outperforms compared with the non-feature-based fusion strategy because it helps to take the better use of both 2D and 3D feature and achieves better fusion for the deep-learning-based methods. After projecting the 2D and 3D feature maps into the same feature space, a suitable feature vector fusion operation need to be selected. There are some popular operations, for example: concatenation, pooling, and summation. However, the fusion operation is often selected based on the intuition of researchers. Moreover, there is no research study about how different feature fusion operations affect the performance of the methods.

(5) Unknown Effect of Different Fusion Position

The fusion effect of the one-stage and multi-stage fusion strategy relies on the fusion position. According to the existing method, there is no evidence showing which fusion position is the most suitable for the 2D and 3D information integration. The early fusion achieves a high degree of feature fusion but fails to take the full use of 2D and 3D information independently, and the late fusion vice versa. Moreover, the middle fusion balances the 2D feature, 3D feature, and joint feature learning process. Thanks to the symmetrical structure of the popular segmentation network, which consists of the encoder and decoder, it is easy to select the middle stage of network between the encoder and decoder for feature

fusion. However, the suitable fusion position is still uncertain for the detection network. So, the unknown effect of different fusion position may confuse researchers.

(6) Combination of Different Fusion Strategies

Both the feature-based and non-feature-based fusion strategy have realized series of achievements; both of them are suitable for different situations. There is currently no research that integrates the multiple fusion strategies in one method.

4. Balance the Accuracy and Efficiency

Since the deep-learning-based method integrating both 2D and 3D information has become a hot research topic, more and more methods are constantly proposed with better and better accuracy. However, most of them sacrifice the efficiency due to the increasing number of parameters need to be memorized during the training, which may hurt the efficiency of the application. Therefore, the balance between accuracy and efficiency remains an open challenge for almost every deep-learning-based method.

5. The interpretability of the models

Although the existing deep-learning-based method has achieved outstanding accuracy, a lack of interpretability limits the promotion of existing models. Figuring out how the deep-learning-based techniques work may effectively help to learn more useful features.

5.3. Promising Direction

According to the open challenges above, the promising directions are further summarized in the following line, which may inspire the researchers in the future.

1. Unsupervised or Weakly supervised Method

Since the supervised deep-learning-based methods rely on the expensive dataset, the unsupervised and weakly supervised methods are worthy of further research.

2. Adaptive Transform or Adaption Method Before Projection

Rather than naively projecting the 2D and 3D data into the same coordinate based on the six DoFs of the sensor, a suitable adaptive transform or adaption before projection may help to improve the performance. For example, only projecting the pixels with useful features may help to improve the efficiency and robustness. Moreover, appropriate adaptive transformation of the point cloud may help to ensure the permutation invariance of the rigid objects.

3. The Promising Directions of Fusion Strategies

(1) Better Intermediate Data

A suitable intermediate data need to be created for the data-based fusion strategy to alleviate the problem of information loss. For example, the BEV and voxel representation struggle with the balance of resolution and computation load. Therefore, the hierarchical intermediate data with the pyramidal structure may help to break through this bottleneck and may be the promising direction.

(2) Elaborate Correspondence of 2D and 3D Feature Map

Since the neighboring region selection is important for the local feature learning, the correspondence between the neighboring regions of 2D and 3D feature map may need more information than the six DoF parameters. For example, the pointwise and Frustum-wise feature fusion need image and point cloud feature extracted from different neighboring regions. Therefore, more elaborate correspondence between 2D and 3D neighboring regions should be discovered for different fusion strategies in the future.

(3) Integration of 2D and 3D Proposals/RoIs/Frustums

The data–result-based fusion strategy relies on the quality of the proposals/RoIs/Frustums. However, the misdetection may happen because of the low quality of proposals caused by

the shape missing of the point cloud. Therefore, the shape complement of point cloud may be a promising direction.

(4) Ablation Study about Different Fusion Operation

A series of ablation studies about different fusion operation of feature-based fusion strategy are necessary. The suitable situation and the fusion effect of different feature vector fusion operation need to be figured out. We advise researchers to execute the comparative analysis of concatenation, summary, max/min/average pooling, and adaptive weighting in the following research.

(5) Ablation Study about Different Fusion Position

A series of ablation studies about different fusion positions for one-stage and multi-stage feature-based fusion strategy are necessary. To the best of our knowledge, few researchers have considered the ablation study about different fusion positions to prove that the fusion position they choose is relatively perfect for their methods. There is no evident clue showing which fusion position outperforms the others. Therefore, further ablation studies about fusion position are needed.

(6) Integrating the Feature-based and Non-feature-based fusion strategy

Both feature-based and non-feature-based fusion strategy achieve the 2D and 3D information fusion and accuracy improvement successfully. Therefore, integrating multiple fusion strategies in one network might be a promising direction for accuracy improvement. For example, the SPLATNet achieves the 2D, 3D, and joint segmentation in one end-to-end network using the feature-based fusion strategy. Additionally, it utilizes the result-based method as post-processing to integrate the 2D, 3D, and joint segmentation results. Therefore, whether the integration of feature-based and result-based fusion strategy works is worth further verification.

4. More Efficient and Concise Feature Learning Architecture

More research on improving both the accuracy and efficiency is necessary for better application. The number of parameters and local regions selection should be researched further to find the cause of the main computation load. Removing the redundant 2D and 3D data may help to simplify the process of feature learning and fusion. Moreover, better local region selection methods need to be created for both better local feature and more efficient neighbor researching.

5. The Significance of Layers or Feature Vectors

Analyzing the significance of each layer and feature of the deep-learning-based model may help to explain their functions. Captum is a useful pytorch toolset to visualize which layer or feature plays an important role in the model. Therefore, the significance analysis of each layer and feature may help to increase the interpretability by using toolsets such as Captum.

6. Conclusions

To the best of our knowledge, this is the first review that aims to conclude the fusion strategies of integrating the 2D and 3D information for the task of segmentation and detection in both indoor and outdoor scenes based on deep learning techniques. Compared with existing papers, this paper focuses more on summarizing the similarities between different methods and put forward a taxonomy of grouping these fusion strategies into six categories (data-based, result-based, data–result-based, one-stage feature-based, multi-stage feature-based, and cross-level feature-based), which makes it clearer for researchers to understand the current research situation. Moreover, we abstract the general pattern of each fusion strategy possessing good generalization, which may inspire researchers to create their own methods. Since the methods based on the deep learning techniques rely on the quality and data quantity of the dataset, we also introduce some representative datasets in Section 3, which includes the 2.5D RGBD datasets, the datasets with the registered 2D and 3D data, and the datasets of fine-grained 3D models. In addition, this paper includes

lots of methods and suitable datasets, being as comprehensive as possible. There are 38 datasets and 61 methods included, and their important details are introduced in this paper. To help the researcher pick suitable datasets and a fusion strategy for their own research, we deliver the comparative analysis of datasets and methods in a series of tabular forms. Furthermore, the fusion of 2D and 3D information has been proven effective for the accuracy improvement and the advantages complementation.

Author Contributions: This work was carried out by collaboration between all of the authors. Conceptualization, J.Z. and Y.W.; methods collecting, Y.W., X.D., X.N. and Y.C. (Yuanyuan Cui); Writing—original draft, J.Z. and Y.W.; opinions and modification, X.H., M.G., Y.C. (Yuee Cao) and R.Z.; grammar and spelling check, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: State Key Laboratory of Geo-Information Engineering [grant number SKLGIE2019-Z-3-1]; Fundamental Research Funds of Beijing University of Civil Engineering and Architecture [grant number X18063]; National Natural Science Foundation of China [grant number 41601409]; Beijing Natural Science Foundation [grant number 8172016]; Open Research Fund Program of LIESMARS [grant number 19E01]; National Key Research and Development Program of China [grant number 2018YFC0807806]; BUCEA Post Graduate Innovation Project [grant number 31081021004]; National Natural Science Foundation of China [grant number 41971350]; Beijing Advanced Innovation Centre for Future Urban Design Project [grant number UDC2019031724]; Teacher Support Program for Pyramid Talent Training Project of Beijing University of Civil Engineering and Architecture [grant number JDJQ20200307]; Open Research Fund Program of Key Laboratory of Digital Mapping and Land Information Application, Minisitry of Natural Resources [ZRZYBWD202102].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable. No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dong, S.; Wang, P.; Abbas, K. A survey on deep learning and its applications. *Comput. Sci. Rev.* **2021**, *40*, 100379. [CrossRef]
2. Bello, S.A.; Yu, S.; Wang, C.; Adam, J.M.; Li, J. Review: Deep learning on 3D point clouds. *Remote. Sens.* **2020**, *12*, 1729. [CrossRef]
3. Guo, Z.; Huang, Y.; Hu, X.; Wei, H.; Zhao, B. A survey on deep learning based approaches for scene understanding in autonomous driving. *Electronics* **2021**, *10*, 471. [CrossRef]
4. Arshad, S.; Kim, G.-W. Role of deep learning in loop closure detection for visual and lidar SLAM: A survey. *Sensors* **2021**, *21*, 1243. [CrossRef]
5. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
6. Wang, J.; Zhu, H.; Wang, S.-H.; Zhang, Y.-D. A review of deep learning on medical image analysis. *Mob. Netw. Appl.* **2021**, *26*, 351–380. [CrossRef]
7. Liu, X.; Song, L.; Liu, S.; Zhang, Y. A review of deep-learning-based medical image segmentation methods. *Sustainability* **2021**, *13*, 1224. [CrossRef]
8. Amanullah, M.A.; Habeeb, R.A.A.; Nasaruddin, F.H.; Gani, A.; Ahmed, E.; Nainar, A.S.M.; Akim, N.M.; Imran, M. Deep learning and big data technologies for IoT security. *Comput. Commun.* **2020**, *151*, 495–517. [CrossRef]
9. Xie, Y.; Tian, J.; Zhu, X.X. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geosci. Remote. Sens. Mag.* **2020**, *8*, 38–59. [CrossRef]
10. Minaee, S.; Boykov, Y.Y.; Porikli, F.; Plaza, A.J.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef] [PubMed]
11. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
12. Wu, Y.; Wang, Y.; Zhang, S.; Ogai, H. Deep 3D object detection networks using LiDAR data: A review. *IEEE Sens. J.* **2021**, *21*, 1152–1171. [CrossRef]
13. Ming, Y.; Meng, X.; Fan, C.; Yu, H. Deep learning for monocular depth estimation: A review. *Neurocomputing* **2021**, *438*, 14–33. [CrossRef]
14. Yao, G.; Yilmaz, A.; Meng, F.; Zhang, L. Review of wide-baseline stereo image matching based on deep learning. *Remote Sens.* **2021**, *13*, 3247. [CrossRef]
15. Raj, T.; Hashim, F.H.; Huddin, A.B.; Ibrahim, M.F.; Hussain, A. A survey on LiDAR scanning mechanisms. *Electronics* **2020**, *9*, 741. [CrossRef]
16. Bi, S.; Yuan, C.; Liu, C.; Cheng, J.; Wang, W.; Cai, Y. A survey of low-cost 3D laser scanning technology. *Appl. Sci.* **2021**, *11*, 3938. [CrossRef]

17. Zhang, J.; Lin, X. Advances in fusion of optical imagery and LiDAR point cloud applied to photogrammetry and remote sensing. *Int. J. Image Data Fusion* **2017**, *8*, 1–31. [CrossRef]
18. Wang, Z.; Wu, Y.; Niu, Q. Multi-sensor fusion in automated driving: A survey. *IEEE Access* **2019**, *8*, 2847–2868. [CrossRef]
19. Debeunne, C.; Vivet, D. A review of visual-LiDAR fusion based simultaneous localization and mapping. *Sensors* **2020**, *20*, 2068. [CrossRef]
20. Fayyad, J.; Jaradat, M.A.; Gruyer, D.; Najjaran, H. Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors* **2020**, *20*, 4220. [CrossRef]
21. 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**, 1–18. [CrossRef]
22. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
23. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Sist.* **2012**, *2*, 1097–1105. [CrossRef]
24. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. *arXiv* **2014**, arXiv:1409.4842.
25. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
26. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
27. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
28. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
29. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
30. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via. region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, UK, 2016; pp. 379–387.
31. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
32. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
33. Yoo, D.; Park, S.; Lee, J.-Y.; Paek, A.S.; Kweon, I.S. AttentionNet: Aggregating weak directions for accurate object detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2015; pp. 2659–2667.
34. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
35. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
36. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
37. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
38. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
39. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
40. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 109–117.
41. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
42. Pinheiro, P.O.; Lin, T.-Y.; Collobert, R.; Dollár, P. Learning to refine object segments. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 75–91.
43. Zeiler, M.D.; Taylor, G.W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In Proceedings of the 2011 IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2018–2025.
44. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014, ECCV 2014, Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2014. [CrossRef]
45. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
46. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germaby, 5–9 October 2015; pp. 234–241.
47. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
48. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.

49. Guo, Y.; Wang, H.; Hu, Q.; Liu, H.; Liu, L.; Bennamoun, M. Deep learning for 3D point clouds: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]
50. Liu, W.; Sun, J.; Li, W.; Hu, T.; Wang, P. Deep learning on point clouds and its application: A survey. *Sensors* **2019**, *19*, 4188. [CrossRef]
51. Zhang, J.; Zhao, X.; Chen, Z.; Lu, Z. A review of deep learning-based semantic segmentation for point cloud. *IEEE Access* **2019**, *7*, 179118–179133. [CrossRef]
52. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 945–953.
53. Yang, Z.; Wang, L. Learning relationships for multi-view 3D object recognition. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 7505–7514.
54. Wei, X.; Yu, R.; Sun, J. View-GCN: View-based graph convolutional network for 3D shape analysis. In Proceedings of the CVPR 2020: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–18 June 2020; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2020; pp. 1847–1856.
55. Maturana, D.; Scherer, S. Voxnet: A 3D convolutional neural network for real-time object recognition. In Proceedings of the IROS 2015—IEEE/RSJ International Conference on Intelligent Robots and Systems, Hamburg, Germany, 28 September–2 October 2015; pp. 922–928.
56. Riegler, G.; Ulusoy, A.O.; Geiger, A. Octnet: Learning deep 3D representations at high resolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nonolulu, HI, USA, 21–26 July 2017; pp. 6620–6629.
57. Han, X.-F.; Laga, H.; Bennamoun, M. Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1578–1604. [CrossRef]
58. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3D classification and segmentation. *arXiv* **2017**, arXiv:1612.00593v2.
59. Wu, W.; Qi, Z.; Fuxin, L. Pointconv: Deep convolutional networks on 3D point clouds. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Longh Beach, CA, USA, 15–21 June 2019; pp. 9613–9622.
60. Xu, Y.; Fan, T.; Xu, M.; Zeng, L.; Qiao, Y. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Computer Science Logic*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 90–105.
61. Simonovsky, M.; Komodakis, N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 29–38.
62. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J. Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph.* **2019**, *38*, 1–12. [CrossRef]
63. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3D object proposal generation and detection from point cloud. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Longh Beach, CA, USA, 15–21 June 2019; pp. 770–779.
64. Zarzar, J.; Giancola, S.; Ghanem, B. Pointrgcn: Graph convolution networks for 3D vehicles detection refinement. *arXiv* **2019**, arXiv:1911.12236.
65. Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; Jia, J. STD: Sparse-to-dense 3D object detector for point cloud. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1951–1960.
66. Lehner, J.; Mitterecker, A.; Adler, T.; Hofmarcher, M.; Nessler, B.; Hochreiter, S. Patch refinement-localized 3D object detection. *arXiv* **2019**, arXiv:1910.04093.
67. Qi, C.R.; Litany, O.; He, K.; Guibas, L. Deep hough voting for 3D object detection in point clouds. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9277–9286.
68. Li, B.; Zhang, T.; Xia, T. Vehicle detection from 3D lidar using fully convolutional network. *arXiv* **2016**, arXiv:1608.07916.
69. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3DSSD: Point-based 3D single stage object detector. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 11037–11045.
70. Lawin, F.J.; Danelljan, M.; Tosteberg, P.; Bhat, G.; Khan, F.S.; Felsberg, M. Deep projective 3D semantic segmentation. In *Programming Languages and Systems*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2017; pp. 95–107.
71. Wu, B.; Wan, A.; Yue, X.; Keutzer, K. Squeezeseg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D lidar point cloud. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Queensland, AU, 21–25 May 2018; pp. 1887–1893. [CrossRef]
72. Tchapmi, L.; Choy, C.; Armeni, I.; Gwak, J.; Savarese, S. SEGCloud: Semantic segmentation of 3D point clouds. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 537–547.
73. Rethage, D.; Wald, J.; Sturm, J.; Navab, N.; Tombari, F. Fully-convolutional point networks for large-scale point clouds. In *Advances in Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 625–640.
74. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hi-erarchical. feature learning on point sets in a metric space. Advances in Neural Information Processing Systems. *arXiv* **2017**, arXiv:1706.02413.
75. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. Pointcnn: Convolution on X-transformed points. *arXiv* **2018**, arXiv:1801.07791.
76. Wu, B.; Liu, Y.; Lang, B.; Huang, L. DGCNN: Disordered graph convolutional neural network based on the Gaussian mixture model. *Neurocomputing* **2018**, *321*, 346–356. [CrossRef]
77. Landrieu, L.; Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4558–4567.

78. Yi, L.; Zhao, W.; Wang, H.; Sung, M.; Guibas, L.J. GSPN: Generative shape proposal network for 3D instance segmentation in point cloud. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3942–3951.
79. Yang, B.; Wang, J.; Clark, R.; Hu, Q.; Wang, S.; Markham, A.; Trigoni, N. Learning object bounding boxes for 3D instance segmentation on point clouds. *arXiv* **2019**, arXiv:1906.01140.
80. Wang, W.; Yu, R.; Huang, Q.; Neumann, U. SGPN: Similarity group proposal network for 3D point cloud instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, CA, USA, 18–23 June 2018; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2018; pp. 2569–2578.
81. Wang, X.; Liu, S.; Shen, X.; Shen, C.; Jia, J. Associatively segmenting instances and semantics in point clouds. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4091–4100.
82. Lai, K.; Bo, L.; Ren, X.; Fox, D. A large-scale hierarchical multi-view RGB-D object dataset. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 1817–1824.
83. Koppula, H.S.; Anand, A.; Joachims, T.; Saxena, A. Semantic labeling of 3D point clouds for indoor scenes. In Proceedings of the Neural Information Processing Systems, Granada, Spain, 12–17 December 2011; p. 6.
84. Janoch, A.; Karayev, S.; Jia, Y.; Barron, J.T.; Fritz, M.; Saenko, K.; Darrell, T. A category-level 3D object dataset: Putting the kinect to work. In *RGB-D Image Analysis and Processing*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2013; pp. 141–165.
85. Susanto, W.; Rohrbach, M.; Schiele, B. 3D object detection with multiple kinects. In *Programming Languages and Systems*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2012; pp. 93–102.
86. Silberman, N.; Fergus, R. Indoor scene segmentation using a structured light sensor. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–11 November 2011; pp. 601–608.
87. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 746–760.
88. Zhang, Q.; Song, X.; Shao, X.; Shibasaki, R.; Zhao, H. Category modeling from just a single labeling: Use depth information to guide the learning of 2D models. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2013; pp. 193–200.
89. Xiao, J.; Owens, A.; Torralba, A. SUN3D: A database of big spaces reconstructed using SfM and object labels. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2013; pp. 1625–1632.
90. Lai, K.; Bo, L.; Fox, D. Unsupervised feature learning for 3D scene labeling. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–5 June 2014; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2014; pp. 3050–3057.
91. Song, S.; Lichtenberg, S.P.; Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2015; pp. 567–576.
92. Martinez-Gomez, J.; García-Varea, I.; Cazorla, M.; Morell, V. Vidrilo: The visual and depth robot indoor localization with objects information dataset. *Int. J. Robot. Res.* **2015**, *34*, 1681–1687. [CrossRef]
93. Hua, B.-S.; Pham, Q.-H.; Nguyen, D.T.; Tran, M.-K.; Yu, L.-F.; Yeung, S.-K. Scenenn: A scene meshes dataset with annotations. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2016; pp. 92–101.
94. Handa, A.; Patraucean, V.; Badrinarayanan, V.; Stent, S.; Cipolla, R. Scenenet: Understanding real world indoor scenes with synthetic data. *arXiv* **2015**, arXiv:1511.07041.
95. McCormac, J.; Handa, A.; Leutenegger, S.; Davison, A.J. Scenenet rgb-d: 5 M photorealistic images of synthetic indoor trajectories with ground truth. *arXiv* **2016**, arXiv:1612.05079.
96. Georgakis, G.; Reza, M.A.; Mousavian, A.; Le, P.-H.; Košecká, J. Multiview RGB-D dataset for object instance detection. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 426–434.
97. Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niebner, M.; Savva, M.; Song, S.; Zeng, A.; Zhang, Y. Matterport3D: Learning from RGB-D data in indoor environments. In Proceedings of the International Conference 3D Vision 2017, Qingdao, China, 10–12 October 2017; pp. 667–676. [CrossRef]
98. Tombari, F.; Di Stefano, L.; Giardino, S. Online learning for automatic segmentation of 3D data. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 4857–4864.
99. Vasiljevic, I.; Kolkin, N.; Zhang, S.; Luo, R.; Wang, H.; Dai, F.Z.; Daniele, A.F.; Mostajabi, M.; Basart, S.; Walter, M.R. Diode: A dense indoor and outdoor depth dataset. *arXiv* **2019**, arXiv:1908.00463.
100. Armeni, I.; Sener, O.; Zamir, A.R.; Jiang, H.; Brilakis, I.; Fischer, M.; Savarese, S. 3D semantic parsing of large-scale indoor spaces. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1534–1543.
101. Armeni, I.; Sax, S.; Zamir, A.R.; Savarese, S. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv* **2017**, arXiv:1702.01105.

102. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Niessner, M. Scannet: Richly-annotated 3D reconstructions of indoor scenes. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017; pp. 2432–2443.
103. Sun, X.; Xie, Y.; Luo, P.; Wang, L. A Dataset for Benchmarking Image-Based Localization. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2017; pp. 5641–5649.
104. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H. Shapenet: An information-rich 3D model repository. *arXiv* **2015**, arXiv:1512.03012.
105. Uy, M.A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; Yeung, S.-K. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019.
106. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361. [CrossRef]
107. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
108. Ros, G.; Ramos, S.; Granados, M.; Bakhtiary, A.; Vazquez, D.; López, A. Vision-based offline-online perception paradigm for autonomous driving. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2015; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2015; pp. 231–238.
109. Zhang, R.; Candra, S.A.; Vetter, K.; Zakhor, A. Sensor fusion for semantic segmentation of urban scenes. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 25–30 May 2015; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2015; pp. 1850–1857.
110. 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 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 9296–9306.
111. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11618–11628. [CrossRef]
112. Can, G.; Mantegazza, D.; Abbate, G.; Chappuis, S.; Giusti, A. Semantic segmentation on Swiss3DCities: A benchmark study on aerial photogrammetric 3D pointcloud dataset. *Pattern Recognit. Lett.* **2021**, *150*, 108–114. [CrossRef]
113. Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A.S.; Hauswald, L.; Pham, V.H.; Mühlegg, M.; Dorn, S. A2D2: Audi autonomous driving dataset. *arXiv* **2020**, arXiv:2004.06320.
114. Tan, W.; Qin, N.; Ma, L.; Li, Y.; Du, J.; Cai, G.; Yang, K.; Li, J. Toronto-3D: A Large-scale Mobile LiDAR dataset for semantic segmentation of urban roadways. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 797–806.
115. Hackel, T.; Savinov, N.; Ladicky, L.; Wegner, J.D.; Schindler, K.; Pollefeys, M. Semantic3d.Net: A new large-scale point cloud classification benchmark. *arXiv* **2017**, arXiv:1704.03847. [CrossRef]
116. Tong, G.; Li, Y.; Chen, D.; Sun, Q.; Cao, W.; Xiang, G. CSPC-Dataset: New lidar point cloud dataset and benchmark for large-scale scene semantic segmentation. *IEEE Access* **2020**, *8*, 87695–87718. [CrossRef]
117. Weng, X.; Man, Y.; Cheng, D.; Park, J.; O'Toole, M.; Kitani, K.; Wang, J.; Held, D. All-in-One Drive: A Large-Scale Comprehensive Perception Dataset with High-Density Long-Range Point Clouds. 2020. Available online: https://www.researchgate.net/publication/347112693_All-In-One_Drive_A_Large-Scale_Comprehensive_Perception_Dataset_with_High-Density_Long-Range_Point_Clouds (accessed on 18 May 2021).
118. Chang, M.-F.; Ramanan, D.; Hays, J.; Lambert, J.; Sangkloy, P.; Singh, J.; Bak, S.; Hartnett, A.; Wang, D.; Carr, P.; et al. Argoverse: 3D tracking and forecasting with rich maps. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 8740–8749.
119. Huang, X.; Wang, P.; Cheng, X.; Zhou, D.; Geng, Q.; Yang, R. The Apolloscape open dataset for autonomous driving and its application. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2702–2719. [CrossRef]
120. Gaidon, A.; Wang, Q.; Cabon, Y.; Vig, E. Virtualworlds as proxy for multi-object tracking analysis. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4340–4349.
121. Fang, J.; Yan, F.; Zhao, T.; Zhang, F.; Zhou, D.; Yang, R.; Ma, Y.; Wang, L. Simulating lidar point cloud for autonomous driving using real-world scenes and traffic flows. *arXiv* **2018**, arXiv:1811.07112.
122. Yi, L.; Shao, L.; Savva, M.; Huang, H.; Zhou, Y.; Wang, Q.; Graham, B.; Engelcke, M.; Klokov, R.; Lempitsky, V. Large-scale 3D shape reconstruction and segmentation from shapenet core55. *arXiv* **2017**, arXiv:1710.06104.
123. Mo, K.; Zhu, S.; Chang, A.X.; Yi, L.; Tripathi, S.; Guibas, L.J.; Su, H. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–21 June 2019; pp. 909–918.
124. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3D Shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.

125. Richtsfeld, A.; Morwald, T.; Prankl, J.; Zillich, M.; Vincze, M. Segmentation of unknown objects in indoor environments. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Algarve, Portugal, 7–12 October 2012; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2012; pp. 4791–4796.
126. Taghanaki, S.A.; Luo, J.; Zhang, R.; Wang, Y.; Jayaraman, P.K.; Jatavallabhula, K.M. Robust point set: A dataset for benchmarking robustness of point cloud classifiers. *arXiv* **2020**, arXiv:2011.11572.
127. De Deuge, M.; Quadros, A.; Hung, C.; Douillard, B. Unsupervised feature learning for classification of outdoor 3D scans. In Proceedings of the Australasian Conference on Robitics and Automation, Sydney, New South Wales, AU, 2–4 December 2013; pp. 1–27.
128. Serna, A.; Marcotegui, B.; Goulette, F.; Deschaud, J.-E. Paris-rue-madame database—A 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods. In Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods ICPRAM 2014, Angers, France, 6 March 2014; pp. 819–824.
129. Vallet, B.; Brédif, M.; Serna, A.; Marcotegui, B.; Paparoditis, N. Terra mobilita/iQmulus urban point cloud analysis benchmark. *Comput. Graph.* **2015**, *49*, 126–133. [CrossRef]
130. Roynard, X.; Deschaud, J.-E.; Goulette, F. Paris-lille-3D: A point cloud dataset for urban scene segmentation and classification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2108–21083.
131. Wang, Y.; Tan, X.; Yang, Y.; Liu, X.; Ding, E.; Zhou, F.; Davis, L.S. 3D pose estimation for fine-grained object categories. In *Transactions on Petri Nets and Other Models of Concurrency XV*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2019; pp. 619–632.
132. Ibrahim, M.; Akhtar, N.; Wise, M.; Mian, A. Annotation tool and urban dataset for 3D point cloud semantic segmentation. *IEEE Access* **2021**, *9*, 35984–35996. [CrossRef]
133. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D.; Breitkopf, U.; Jung, J. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction. *ISPRS J. Photogramm. Remote. Sens.* **2014**, *93*, 256–271. [CrossRef]
134. Zolanvari, S.; Ruano, S.; Rana, A.; Cummins, A.; da Silva, R.E.; Rahbar, M.; Smolic, A. Dublin city: Annotated lidar point cloud and its applications. *arXiv* **2019**, arXiv:1909.03613.
135. Hu, Q.; Yang, B.; Khalid, S.; Xiao, W.; Trigoni, N.; Markham, A. Towards semantic segmentation of urban-scale 3D point clouds: A dataset, benchmarks and challenges. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online Conference, 19–25 June 2021; pp. 4977–4987.
136. Varney, N.; Asari, V.K.; Graehling, Q. Dales: A large-scale aerial lidar data set for semantic segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Online Conference, 14–19 June 2020; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2020; pp. 717–726.
137. Ye, Z.; Xu, Y.; Huang, R.; Tong, X.; Li, X.; Liu, X.; Luan, K.; Hoegner, L.; Stilla, U. Lasdu: A large-scale aerial lidar dataset for semantic labeling in dense urban areas. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 450. [CrossRef]
138. Li, X.; Li, C.; Tong, Z.; Lim, A.; Yuan, J.; Wu, Y.; Tang, J.; Huang, R. Campus3d: A photogrammetry point cloud benchmark for hierarchical understanding of outdoor scene. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 238–246.
139. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Online Conference, 14–19 June 2020; pp. 2446–2454.
140. Wulff, F.; Schaufele, B.; Sawade, O.; Becker, D.; Henke, B.; Radusch, I. Early fusion of camera and lidar for robust road detection based on U-net fcn. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 30 June–1 July 2018; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA; pp. 1426–1431.
141. Erkent, O.; Wolf, C.; Laugier, C.; Gonzalez, D.S.; Cano, V.R. Semantic grid estimation with a hybrid bayesian and deep neural network approach. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 May 2018; pp. 888–895.
142. Zhou, K.; Ming, D.; Lv, X.; Fang, J.; Wang, M. CNN-based land cover classification combining stratified segmentation and fusion of point cloud and very high-spatial resolution remote sensing image Data. *Remote. Sens.* **2019**, *11*, 2065. [CrossRef]
143. Lee, J.-S.; Park, T.-H. Fast road detection by cnn-based camera–lidar fusion and spherical coordinate transformation. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 5802–5810. [CrossRef]
144. Gu, S.; Lu, T.; Zhang, Y.; Alvarez, J.M.; Yang, J.; Kong, H. 3-D LiDAR + monocular camera: An inverse-depth-induced fusion framework for urban road detection. *IEEE Trans. Intell. Veh.* **2018**, *3*, 351–360. [CrossRef]
145. Gu, S.; Zhang, Y.; Tang, J.; Yang, J.; Kong, H. Road detection through CRF based lidar-camera fusion. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 3832–3838.
146. Narita, G.; Seno, T.; Ishikawa, T.; Kaji, Y. Panoptic fusion: Online volumetric semantic mapping at the level of stuff and things. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019; pp. 4205–4212.
147. 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]
148. Riemenschneider, H.; Bódis-Szomorú, A.; Weissenberg, J.; Van Gool, L. Learning where to classify in multi-view semantic segmentation. In *Programming Languages and Systems*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2014; pp. 516–532.

149. Qi, X.; Liao, R.; Jia, J.; Fidler, S.; Urtasun, R. 3D graph neural networks for RGBD semantic segmentation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5199–5208.
150. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In *Programming Languages and Systems*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2014; pp. 345–360.
151. Jaritz, M.; De Charette, R.; Wirbel, E.; Perrotton, X.; Nashashibi, F. Sparse and dense data with CNNs: Depth completion and semantic segmentation. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 52–60.
152. Dai, A.; Nießner, M. 3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 458–474.
153. Lv, X.; Liu, Z.; Xin, J.; Zheng, N. A novel approach for detecting road based on two-stream fusion fully convolutional network. *IEEE Intell. Veh. Symp.* **2018**, 1464–1469. [CrossRef]
154. Yang, F.; Yang, J.; Jin, Z.; Wang, H. A Fusion model for road detection based on deep learning and fully connected CRF. In Proceedings of the 13th Annual Conference on System of Systems Engineering (SoSE), Paris, France, 19–22 June 2018; pp. 29–36.
155. Su, H.; Jampani, V.; Sun, D.; Maji, S.; Kalogerakis, E.; Yang, M.-H.; Kautz, J. Splatnet: Sparse lattice networks for point cloud processing. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2018; pp. 2530–2539.
156. Jaritz, M.; Gu, J.; Su, H. Multi-view pointnet for 3D scene understanding. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27 October–2 November 2019; pp. 3995–4003.
157. Li, Z.; Gan, Y.; Liang, X.; Yu, Y.; Cheng, H.; Lin, L. LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling. In *Machine Learning in Clinical Neuroimaging*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 541–557.
158. Liu, H.; Wu, W.; Wang, X.; Qian, Y. RGB-D joint modelling with scene geometric information for indoor semantic segmentation. *Multimed. Tools Appl.* **2018**, *77*, 22475–22488. [CrossRef]
159. Hou, J.; Dai, A.; NieBner, M. 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4416–4425.
160. Yu, D.; Xiong, H.; Xu, Q.; Wang, J.; Li, K. Multi-stage residual fusion network for lidar-camera road detection. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2019; pp. 2323–2328.
161. Li, H.; Chen, Y.; Zhang, Q.; Zhao, D. Bifnet: Bidirectional fusion network for road segmentation. *IEEE Trans. Cybern.* **2021**, 1–12. [CrossRef]
162. Yuan, J.; Zhang, K.; Xia, Y.; Qi, L. A fusion network for semantic segmentation using RGB-D data. In Proceedings of the Ninth International Conference on Graphic and Image Processing (ICGIP), Qingdao, China, 14–16 October 2018; p. 1061523.
163. Hu, X.; Yang, K.; Fei, L.; Wang, K. ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1440–1444.
164. Chen, H.; Li, Y. Three-stream attention-aware network for RGB-D salient object detection. *IEEE Trans. Image Process.* **2019**, *28*, 2825–2835. [CrossRef]
165. Zhou, W.; Yuan, J.; Lei, J.; Luo, T. TSNet: Three-stream self-attention network for RGB-D indoor semantic segmentation. *IEEE Intell. Syst.* **2021**, *36*, 73–78. [CrossRef]
166. Liu, C.; Wu, J.; Furukawa, Y. FloorNet: A unified framework for floorplan reconstruction from 3D scans. In *Medical Image Computing and Computer-Assisted Intervention*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2018; pp. 203–219.
167. Caltagirone, L.; Bellone, M.; Svensson, L.; Wahde, M. Lidar—camera fusion for road detection using fully convolutional neural networks. *Robot. Auton. Syst.* **2019**, *111*, 125–131. [CrossRef]
168. Kim, D.-K.; Maturana, D.; Uenoyama, M.; Scherer, S. Season-invariant semantic segmentation with a deep multimodal network. In *Experimental Robotics*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 255–270.
169. Chiang, H.-Y.; Lin, Y.-L.; Liu, Y.-C.; Hsu, W.H. A Unified point-based framework for 3D segmentation. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Montreal, QC, Canada, 16–19 September 2019; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2019; pp. 155–163.
170. Chen, Z.; Zhang, J.; Tao, D. Progressive lidar adaptation for road detection. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 693–702. [CrossRef]
171. Xu, J.; Zhang, R.; Dou, J.; Zhu, Y.; Sun, J.; Pu, S. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. *arXiv* **2021**, arXiv:2103.12978.
172. Nakajima, Y.; Kang, B.; Saito, H.; Kitani, K. Incremental class discovery for semantic segmentation with RGBD sensing. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 972–981.
173. Martinovic, A.; Knopp, J.; Riemenschneider, H.; Van Gool, L. 3D all the way: Semantic segmentation of urban scenes from start to end in 3D. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2015; pp. 4456–4465.
174. Arcos-García, Á.; Soilán, M.; Álvarez-García, J.A.; Riveiro, B. Exploiting synergies of mobile mapping sensors and deep learning for traffic sign recognition systems. *Expert Syst. Appl.* **2017**, *89*, 286–295. [CrossRef]
175. Guan, H.; Yan, W.; Yu, Y.; Zhong, L.; Li, D. Robust traffic-sign detection and classification using mobile lidar data with digital images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2018**, *11*, 1715–1724. [CrossRef]

176. Barea, R.; Perez, C.; Bergasa, L.M.; Lopez-Guillen, E.; Romera, E.; Molinos, E.; Ocana, M.; Lopez, J. Vehicle detection and localization using 3D lidar point cloud and image semantic segmentation. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Big Island, HI, USA, 4–7 November 2018; pp. 3481–3486.
177. Guan, H.; Yu, Y.; Peng, D.; Zang, Y.; Lu, J.; Li, A.; Li, J. A convolutional capsule network for traffic-sign recognition using mobile lidar data with digital images. *IEEE Geosci. Remote. Sens. Lett.* **2019**, *17*, 1067–1071. [CrossRef]
178. Lahoud, J.; Ghanem, B. 2D-driven 3D object detection in RGB-D images. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2017; pp. 4632–4640.
179. Du, X.; Ang, M.H.; Karaman, S.; Rus, D. A general pipeline for 3D detection of vehicles. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 3194–3200.
180. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum pointnets for 3D object detection from RGB-D data. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 918–927.
181. Zhao, X.; Liu, Z.; Hu, R.; Huang, K. 3D object detection using scale invariant and feature reweighting networks. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Association for the Advancement of Artificial Intelligence (AAAI): Palo Alto, CA, USA, 2019; Volume 33, pp. 9267–9274.
182. Wang, Z.; Jia, K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 4–8 November 2019; pp. 1742–1749.
183. Shin, K.; Kwon, Y.P.; Tomizuka, M. Roarnet: A robust 3D object detection based on region approximation refinement. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 2510–2515.
184. Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; Jia, J. Ipod: Intensive point-based object detector for point cloud. *arXiv* **2018**, arXiv:1812.05276.
185. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3D object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–18 June 2020; pp. 4603–4611.
186. Song, S.; Xiao, J. Deep sliding shapes for amodal 3D object detection in RGB-D images. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 808–816.
187. Deng, Z.; Latecki, L.J. Amodal detection of 3D objects: Inferring 3D bounding boxes from 2D ones in RGB-depth images. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 June 2017; pp. 398–406.
188. Wang, Z.; Zhan, W.; Tomizuka, M. Fusing bird's eye view lidar point cloud and front view camera image for 3D object detection. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 30 June–1 July 2018; pp. 1–6.
189. Yang, B.; Liang, M.; Urtasun, R. Hdnet: Exploiting hd maps for 3d object detection. In Proceedings of the Conference on Robot Learning, Zurich, Switzerland, 29–31 October 2018; pp. 146–155.
190. Sindagi, V.A.; Zhou, Y.; Tuzel, O. MVX-Net: Multimodal voxelnet for 3D object detection. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7276–7282.
191. Qi, C.R.; Chen, X.; Litany, O.; Guibas, L.J. Imvotenet: Boosting 3D object detection in point clouds with image votes. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online Conference, 14–19 June 2020; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2020; pp. 4403–4412.
192. Zhou, Y.; Sun, P.; Zhang, Y.; Anguelov, D.; Gao, J.; Ouyang, T.; Guo, J.; Ngiam, J.; Vasudevan, V. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In Proceedings of the Conference on Robot Learning, London, UK/Online Conference, 8–11 November 2020; pp. 923–932.
193. Xu, B.; Chen, Z. Multi-level fusion based 3D object detection from monocular images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2345–2353.
194. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3D object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6526–6534.
195. Liang, M.; Yang, B.; Wang, S.; Urtasun, R. Deep continuous fusion for multi-sensor 3D object detection. In *Lecture Notes in Computer Science*; Springer Science and Business Media: Berlin/Heidelberg, Germany, 2018; pp. 663–678.
196. Lu, H.; Chen, X.; Zhang, G.; Zhou, Q.; Ma, Y.; Zhao, Y. Scanet: Spatial-channel attention network for 3D object detection. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2019; pp. 1992–1996.
197. Xu, D.; Anguelov, D.; Jain, A. PointFusion: Deep sensor fusion for 3D bounding box estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 244–253.
198. Liang, M.; Yang, B.; Chen, Y.; Hu, R.; Urtasun, R. Multi-task multi-sensor fusion for 3D object detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7337–7345.
199. Huang, T.; Liu, Z.; Chen, X.; Bai, X. EPNet: Enhancing point features with image semantics for 3D object detection. In *Computer Vision—ECCV*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 35–52.
200. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3D proposal generation and object detection from view aggregation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 May 2018; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2018; pp. 1–8.