Chapter 4
Semantic Segmentation for Autonomous Driving

Jingwei Yang, Sicen Guo, Mohammud Junaid Bocus, Qijun Chen, and Rui Fan

Abstract The task of semantic segmentation involves labeling each pixel in an image with its corresponding object class, which is achieved by clustering regions belonging to the same category using artificial intelligence. This is an important step from image processing to image analysis and has numerous applications in areas such as automatic driving, image enhancement, and 3D map reconstruction. With the emergence of deep learning, several sophisticated and efficient algorithms have been developed for this task. This chapter aims to review these methods, starting with a discussion of state-of-the-art semantic segmentation methods for both single modality and data fusion, emphasizing their contributions and significance in the field. Additionally, an overview of commonly used datasets is provided to assist researchers in selecting the appropriate dataset for their needs and goals. A comprehensive summary of evaluation metrics used to assess semantic segmentation results, along with corresponding benchmarks for a number of classic datasets, is also presented. Finally, practical applications of semantic segmentation in autonomous driving are explored, and conclusions are drawn on the current state of the art.

J. Yang · S. Guo · Q. Chen · R. Fan (✉)
The Robotics and Artificial Intelligence Laboratory (RAIL), State Key Laboratory of Intelligent Autonomous Systems, The Department of Control Science and Engineering and Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, People's Republic of China
e-mail: rfan@tongji.edu.cn

J. Yang
e-mail: jw_yang@tongji.edu.cn

S. Guo
e-mail: guosicen@tongji.edu.cn

Q. Chen
e-mail: qjchen@tongji.edu.cn

M. J. Bocus
University of Bristol, Bristol, United Kingdom
e-mail: junaid.bocus@bristol.ac.uk

R. Fan et al. (eds.), *Autonomous Driving Perception*, Advances in Computer Vision and Pattern Recognition, https://doi.org/10.1007/978-981-99-4287-9_4

4.1 Introduction

Semantic segmentation is a critical task in computer vision that aims to partition an image into visually significant regions for further visual comprehension and analysis [1]. This high-level task is essential for complete scene understanding, as it is used in 2D images, videos, and even 3D or volumetric data. The importance of scene comprehension in computer vision is highlighted by the growing number of applications that rely on extracting knowledge from images, such as autonomous driving [2, 3], precision agriculture [4], geological testing [5], and medical treatment [6].

Since the first successful demonstration in the 1980s [7], significant progress has been made in the field of autonomous driving. This technology has immense potential benefits, including reducing traffic congestion and improving driving safety. Nevertheless, it is an arduous task to guarantee the reliability of autonomous driving. An autonomous car must be able to accurately identify, assess, and make decisions about road conditions and plan routes. Even a minor error in any of these steps can have serious consequences. To address this challenge, perception systems in autonomous driving must be accurate, robust, and capable of real-time processing. Accuracy guarantees reliable environmental information, robustness ensures proper functioning even in adverse weather conditions or sensor degradation, while real-time processing is crucial for high-speed driving.

Semantic segmentation is a critical technology for autonomous driving, as it enables the understanding of driving scenes (as shown in Fig. 4.1). Historically, researchers have used a range of traditional computer vision and machine learning techniques to address this issue. Semantic segmentation is essential for numerous applications, including scene interpretation and robot perception [1, 8, 9]. Initially, researchers employed random forest and conditional random field (CRF) methods to classify various semantic elements. However, the current mainstream techniques for semantic segmentation have shifted towards deep learning, resulting in significant improvements in segmentation performance [10–12]. In this section, we aim to provide a comprehensive overview of both single-modal and data fusion-based semantic segmentation.

Significant progress has been made in the field of single-modal image segmentation using deep learning-based semantic segmentation models, with exceptional

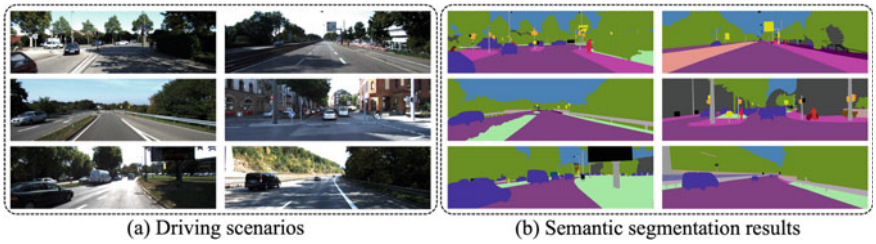

Fig. 4.1 Semantic segmentation used for autonomous driving

accuracy rates on widely used benchmarks. Early fusion involves combining conventional unimodal semantic segmentation networks and reusing existing models. For example, Gouprie et al. [13] adapted the RGB network in [14] to RGB-D semantic segmentation by concatenating input RGB and depth channels. Late fusion, on the other hand, aims to fuse multi-level features with different types through different streams [15–17]. For instance, [16] used two CNN models to extract features from RGB and depth images, and SVM classifiers to distinguish different semantic categories. [15] introduced a gated fusion layer that learns the respective contributions of high-level modality-specific features to enable more effective combination. Furthermore, hierarchical fusion allows the integration of multimodal features at different levels, as demonstrated in the works of [18, 19].

This chapter is structured as follows: Sect. 4.2 provides an overview of the latest segmentation algorithms and their characteristics, covering both single-modal and data fusion-based semantic segmentation. Section 4.3 provides a summary of the datasets and corresponding benchmarks utilized for semantic segmentation. In Sect. 4.4, evaluation metrics are discussed. In Sect. 4.5, various applications of semantic segmentation are explored. The chapter concludes with a discussion of current challenges in Sect. 4.6 and a summary in Sect. 4.7.

4.2 State of The Art

4.2.1 Single-Modal Networks

Pixel-level single-channel semantic segmentation is more challenging than image-level image classification and region-level object detection. Semantic segmentation is particularly crucial for a large number of real-world applications, such as autonomous driving [20, 21], surrounding sensing [22–24], and medical diagnosis [25, 26].

4.2.1.1 Single-Modal Convolutional Neural Networks

Fully Convolutional Networks

The introduction of the Fully Convolutional Network (FCN) [27] marked a significant advancement in end-to-end semantic segmentation. It has become the most advanced and robust framework for scenario analysis. In traditional Convolutional Neural Networks (CNNs), several fully connected layers are added after the final convolutional layer, which converts the feature map into a fixed-length feature vector. However, FCN can process input images of any size and upsample the final layer's feature map using a deconvolution layer. As a result, the output prediction for each pixel preserves the spatial information of the original input image, recovering its original size. This outperforms initial CNN methods such as R-CNN [28] and SDS [29]. However, the classification of individual pixels does not fully consider the

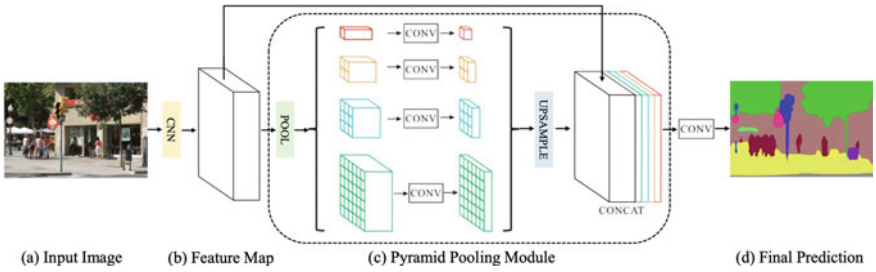

Fig. 4.2 PSPNet [32] architecture. Given an input image (**a**), it first uses CNN to get the feature map of the last convolutional layer (**b**), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (**c**). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (**d**)

relationship between pixels, resulting in a lack of spatial consistency. This approach also increases the memory footprint and computational complexity. To address this, FastFCN [30] formulates the task of extracting high-resolution feature maps into a joint upsampling problem, reducing computational complexity by more than three times without sacrificing performance.

Pyramid Models

Recent semantic segmentation approaches have made significant improvements by incorporating pyramid-based multi-resolution representations to increase the receptive fields. However, traditional fully convolutional network (FCN) [31] models often suffer from misclassification of similar categories due to inadequate utilization of global information and contextual relationships across different receptive fields. The spatial pyramid pooling (SPP) [31] model addresses this limitation by obtaining global image-level features to improve the ability of FCN-based models to utilize contextual information. Nevertheless, to reduce the loss of contextual information within different sub-regions, [32] proposes a hierarchical global prior, known as the pyramid pooling module (PPM), as illustrated in Fig. 4.2. The PPM module combines the features of four different pyramid scales to capture contextual information of varying scales and sub-regions. Finally, the resulting representation is passed through a convolutional layer to produce the final per-pixel prediction.

The DeepLab model also addresses this challenge by using Atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) modules. Compared to traditional convolution, dilated convolution can obtain a bigger receptive field without sacrificing spatial resolution. This architecture has evolved over several generations:

DeepLabV1: Uses Atrous Convolution and Fully Connected CRF to control the resolution at which image features are computed.
DeepLabV2: Uses ASPP to consider objects at different scales and segment with much improved accuracy.
DeepLabV3: Apart from using Atrous Convolution, DeepLabV3 uses an improved

ASPP module with batch normalization and image-level features. Unlike its predecessors, DeepLabV3 does not incorporate the CRF (Conditional Random Field) method. The final feature map obtained from the network captures abundant semantic information, however, due to the pooling and convolution operations with striding used in the network backbone, it lacks the finer details of object boundaries.
DeepLabv3+ [33]: It improves upon DeepLabv3 by adding a decoder module to the architecture. While the encoder module encodes multi-scale contextual information using atrous convolution at multiple scales, the decoder module refines the segmentation results along object boundaries in a simple yet effective manner.

Encoder-Decoder Architectures
UNet [34] aims to enhance a standard convolutional network by adding a series of layers that gradually upsample the previous layers. This allows for the combination of high resolution features from the encoder with upsampled feature maps to fuse different levels of information and learn context information. By using a u-shaped architecture, the decoder can propagate context information to higher resolution layers and allow for better segmentation results.

The network structure LinkNet [35] is based on UNet's encoder-decoder structure. There is also a jump connection between the Encoder and Decoder, which allows the network to forget certain information during encoding and revisit it during decoding. Because the amount of information required by the network to decode and generate the image is relatively low, this reduces the number of parameters required by the network. Various operations can be used to implement the jump connection. Another advantage of using jump connections is that you can easily implement reverse gradient flow with the same connection. Tiramisu [36], another semantic segmentation algorithm, connects the two and sends the hidden encoder output to the corresponding decoder input.

The novelty of SegNet [37] lies in the way the decoder upsamples its lower-resolution input feature maps. Specifically, the pooling index computed in the maximum pooling step of the corresponding encoder is used by the decoder to perform nonlinear upsampling. Inverse pooling upsamples the feature maps, maintaining the integrity of high-frequency information while ignoring neighboring information for feature maps with lower resolution.

Unlike SegNet's symmetric encoder-encoder architecture, the ENet [38] architecture consists of a large encoder and a small decoder. ENet believes that the encoder should be able to manipulate lower resolution data, allowing it to process information and filter ideas. In contrast, the decoder is primarily responsible for sampling the output of the encoder and only needs to fine-tune the details.

High-Resolution Network (HRNet) [39] proposes a different approach to previous methods where high-resolution feature maps are recovered from low-resolution representations. Instead, HRNet suggests keeping high-resolution representations during the entire feature extraction and fusion process, avoiding loss of essential shape and bounding details that may occur with a high-to-low encoder. HRNet achieves high performance by using multi-resolution sub-networks in parallel and progressively and repeatedly fusing multi-scale information.

Prior approaches such as pooling-based [40, 41] and dilated convolution-based [42, 43] extensions, make non-adaptive use of homogeneous contextual dependencies for all image regions, which overlooks the variations in local representation and contextual dependencies for different categories.

Attention

Attention mechanism has become a widely used technique in various deep learning tasks, including natural language processing and image recognition. It is a crucial technique that deserves a thorough understanding in the field of deep learning. The inspiration for studying the Attention Mechanism comes from the human brain's ability to selectively focus on a region of interest from a visual signal. Humans learn to focus their attention on specific regions of an image, based on previously observed information, rather than processing all pixels of the entire image at once and adjusting the focus over time. This focus of attention enables humans to quickly select relevant information and disregard irrelevant information, similar to how the Attention Mechanism works in deep learning.

CNNs are limited in their ability to understand complex scenes due to the physical design of their convolutional kernel, which constrains feature information flow to a fixed-size local region. In order to address this limitation, Point-wise Spatial Attention Network (PSANet) [44] (Fig. 4.3) introduces adaptive attention masks to allow each location on the feature map to associate with other locations, thereby moderating the local neighborhood constraint. PSANet also employs bidirectional information propagation paths, which allow each location to aggregate information from all other locations to aid in its own prediction, while simultaneously allowing information from each location to be distributed globally to assist in the prediction of all other locations.

With the frequent use of attention mechanism in neural networks, non-local neural networks [45] obtain the attention mask by calculating the correlation matrix between each spatial point in the feature map. Figure 4.4a depicts the Non-Local structure. The block first computes the similarity between all positions. When the size of the input feature map is $C \times H \times W$, the computational complexity is $O(C \times H^2 \times W^2)$. Then matrix multiplication is performed to collect the influence of all locations on themselves, and the computational complexity of this step is also $O(C \times H^2 \times W^2)$. The high complexity brought by this matrix multiplication results in prohibitive com-

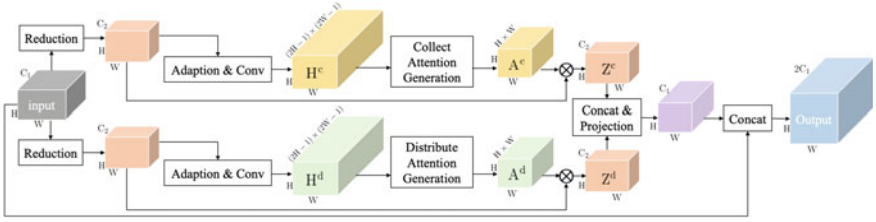

Fig. 4.3 Architecture of PSANet [44], consisting of two parallel branches

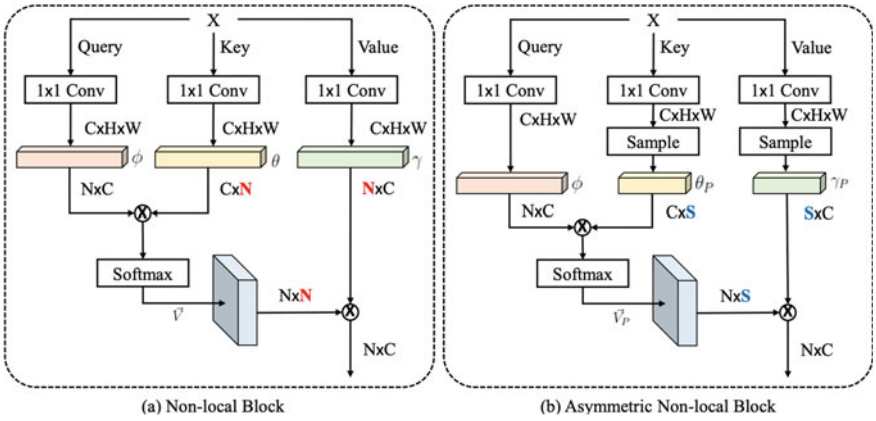

Fig. 4.4 **a** Architecture of a standard non-local block [45]; **b** the asymmetric non-local block [46]

putational costs and a massive GPU memory occupation. To address this issue, Asymmetric Non-local Neural Network (ANN) [46] selectively samples a small number of representative points from feature maps. Figure 4.4b shows the architecture of the ANN block, and it can be seen that the output size of the non-local blocks remains the same as long as the output of the key and value branches remains the same size. Therefore, the ANN block efficiently reduces the computational complexity by selecting a small number of representative points from the key and value branches, while maintaining comparable performance.

Non-local Net [45] computes attention using a whitened pairwise term and a unary term, which are tightly coupled, making it difficult to learn them separately. To address this limitation, Disentangled non-local block (DNLNet) [47] decouples these terms to facilitate their independent learning. Another method, Dual Attention Network (DANet) [48], encodes global context using a self-attention mechanism that incorporates both spatial and channel attention. Current methods such as PSANet [44] and DANet [48] use adaptive weights to calculate pair-wise similarity or learn pixel-wise attention maps. However, these approaches overlook the importance of global guidance from the global information extractor. In contrast to these works, Adaptive Pyramid Context Network (APCNet) [49] takes global-guided local affinity into account to estimate the degree of subregion contribution from local and global representations and exploits multi-scale representation with a feature pyramid.

Despite the superior performance of attention mechanisms over other methods, such as ASPP, PPM, large convolutional kernels, and stacked convolutional layers, the additional computational and GPU memory usage is often unaffordable. Therefore, some networks focus on reducing computational effort. Criss-Cross Network (CCNet) [50] aims to reduce the large computation budget introduced by full spatial attention. Another method, Interlaced Sparse Self-Attention Network (ISANet) [51], factors the dense affinity matrix as the product of two sparse affinity matrices, signif-

icantly reducing computation and memory complexity, particularly when processing high-resolution feature maps.

The attention-based methods have high computational complexity and require a large amount of GPU memory due to the generation of a large attention map. The generation and use of the attention map are computed with respect to all positions, which is a bottleneck. To address these issues, Expectation Maximization Attention (EMA) [52] proposes a novel attention-based method that uses the expectation maximization (EM) algorithm [53] to find a more compact basis set, significantly reducing computational complexity. Unlike previous methods that treat all pixels as reconstruction bases [44, 45], EMA finds a more efficient solution. EncNet (Context Encoding Module) [54] adds a class-dependent feature map to provide prior information about the scene, which "simplifies" the problem for the network. This approach differs from the standard attention mechanism.

Transformer Models

Transformers [56] rely on self-attention mechanisms and capture long-range dependencies among tokens in a sentence. In addition, transformers are highly parallelizable and suitable for training on large datasets, which has led to their success in natural language processing (NLP). This success has inspired the development of several computer vision methods that combine CNNs with self-attention mechanisms to tackle semantic segmentation problems [55].

The transformer-based OCR [57] approach uses the representation of the corresponding object class to characterize a pixel, which strengthens the pixel representation. Swin transformer [58] uses a variant of ViT [59], and proposes a hierarchical transformer whose representation is computed with shifted window. This hierarchical architecture has the flexibility to model at various scales and has different linear computational complexity with regard to image size. Segmenter [55] (Fig. 4.5) is also a transformer encoder-decoder architecture for semantic image segmentation. It relies on a ViT backbone and introduces a masked decoder inspired by

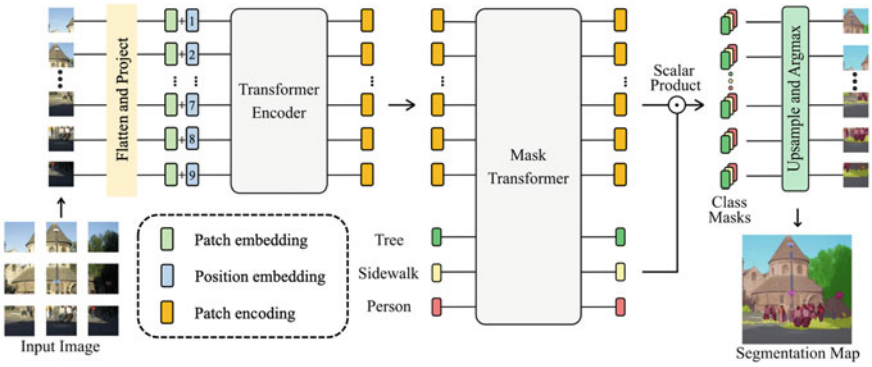

Fig. 4.5 Segmenter model [55]. The encoder module receives image patches, and the decoder module takes the input to a mask transformer to generate segmentation masks

DEtection TRansformer (DETR) [60]. Its architecture does not use convolutions but captures global context at every layer of the model during the encoding and decoding stages. SegFormer [61], a simple, efficient yet powerful semantic segmentation framework, unifies Transformers with lightweight multilayer perceptron (MLP) decoders. Twins [62] presents two powerful vision transformer backbones and dub them twin transformers: Twins-PCPVT and Twins-SVT. They find that interleaving local and global attention can produce impressive results, yet it comes with higher throughputs. ResNeSt [63] combines channel-wise attention with multi-path representation into a single unified Split-Attention block, which improves learned feature representations widely.

4.2.2 Data-Fusion Networks

Semantic segmentation is a popular computer vision task that involves partitioning an image into semantically meaningful regions. Typically, photometric data such as RGB is used for this task. However, structural information refers to the spatial relationships and arrangements of objects in the image, and it can enhance the performance and robustness of the segmentation algorithm due to its geometric property. To make the most of structural information in semantic segmentation tasks, data-fusion networks can combine different input visual information to create a richer representation of the image.

4.2.2.1 Structured Visual Information

- For humans to achieve stereo vision, perceiving depth information is crucial. In computer vision, images are acquired through cameras. However, in our everyday life, objects have three dimensions, whereas camera-captured images only contain two dimensions, resulting in the loss of one-dimensional information, which represents the relative distance of objects in space. Depth information in computer vision usually refers to the distance of each point in space relative to the camera, and the relative distance between the points can be easily calculated once the depth information is known.
- Transformed disparity information is utilized to differentiate potholes in the road. To accomplish this, a disparity map is subjected to a transformation process, whereby an energy function is minimized with respect to both the camera roll angle and the road disparity projection model [64].
- Thermal information is a visual representation that records the heat or temperature radiated by the object itself or its surroundings. An infrared (IR) camera is utilized to capture this data by using an infrared detector and an optical imaging lens to collect the infrared radiation energy distribution of the object in question. The resulting thermal image is then projected onto the photosensitive element of the IR detector, providing a detailed portrayal of the object's thermal profile.

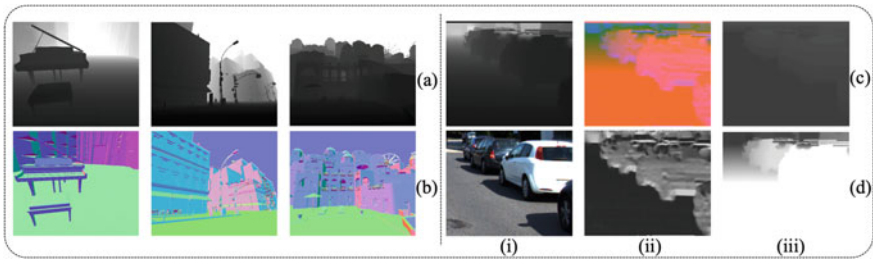

Fig. 4.6 Examples of structured visual information: **a** depth images, **b** surface normal images, (i)–(iii) in row **c** show depth image, HHA image, and height image, respectively. (i)–(iii) in row **d** show RGB image, surface normal angle image, and horizontal disparity image, respectively

- HHA information representation uses three parameters to capture depth images, namely the difference in horizontal position, the height above ground of each pixel, and the angle between the local surface normal and the estimated gravity direction. However, the HHA encoding method has limitations since it mainly emphasizes the interdependence between these parameters, while neglecting the unique contribution of each parameter to the overall representation (Fig. 4.6c, d).
- Surface normal information is a 3D spatial representation that represents the vector perpendicular to the given object. It is an important visual feature and is used in many computer vision applications [65, 66] (Fig. 4.6b).

4.2.2.2 RGB+X Data Fusion

Existing fusion methods fuse the structural visual information described in Sect. 4.2.2.1 with RGB information. Structural visual information captures geometric details such as distance, shape, and structure, which are not present in RGB images. These details provide a more precise understanding of the location and boundary of objects, aiding in semantic segmentation.

Recently, the combination of RGB images with depth information has been highly successful in the field of semantic segmentation, and FuseNet [18] is a typical example of an encoder-decoder architecture that fuses these inputs (as shown in Fig. 4.7). It uses two encoders, each with a VGG-16 [67] backbone, to extract features from the RGB and depth images separately. As the network delves deeper, the depth encoder's feature maps become incorporated into the RGB encoder. RedNet [68] takes this concept a step further by integrating multi-level features even more extensively along the top-down pathway.

RFNet [69] follows a similar approach to FuseNet by using two separate branches to extract features from RGB and depth images. The RGB branch is considered the primary branch in RFNet, while the depth branch is treated as a secondary branch. The two branches are combined using the attention feature complementary (AFC) module (as shown in Fig. 4.8). Furthermore, a spatial pyramid pooling module is implemented

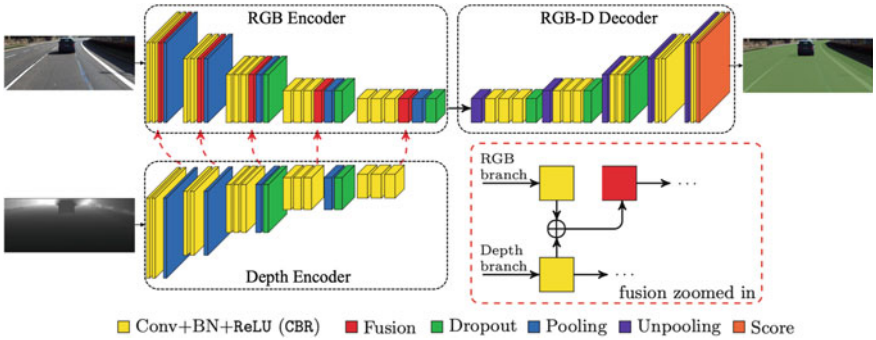

Fig. 4.7 Architecture of FuseNet [18], containing two branches to extract RGB and depth features, respectively

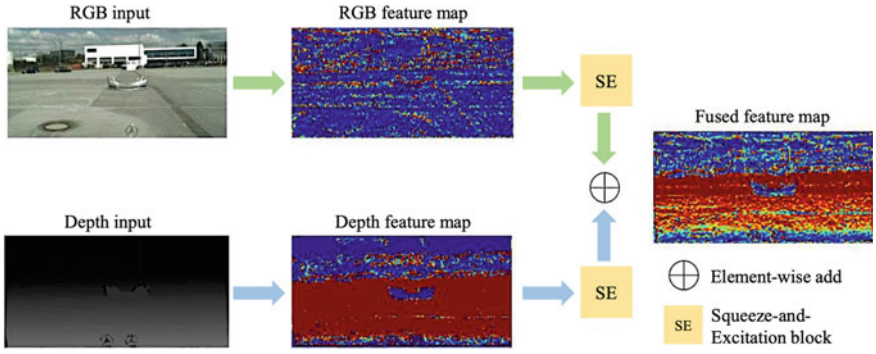

Fig. 4.8 Data fusion method in RFNet [69]

to aggregate the merged RGB-D features and produce feature maps incorporating information at multiple scales. The inclusion of a squeeze-and-excitation block in the AFC module, as proposed in [70], facilitates the acquisition of global knowledge to highlight significant channels and suppress less relevant ones. As a result, the AFC module can efficiently leverage informative features from both branches.

RFBNet [71] adds a novel stream to RGB and depth streams. These three streams are combined using approaches such as summation, concatenation, and self-supervised model adaptation (SSMA) block [72]. The resulting features are then passed through a decoder to generate the final predictions (as shown in Fig. 4.9). Specifically, RFBs are utilized in the upper layers to adeptly capture the interplay between the three streams. SSMA suggests a fusion approach that merges streams and multi-level features specific to different modalities.

Depth-aware CNN [73] utilizes depth-aware convolution and depth-aware average pooling, which can be integrated into the traditional segmentation framework to make full use of the depth information at low cost. Furthermore, in both the encoder and

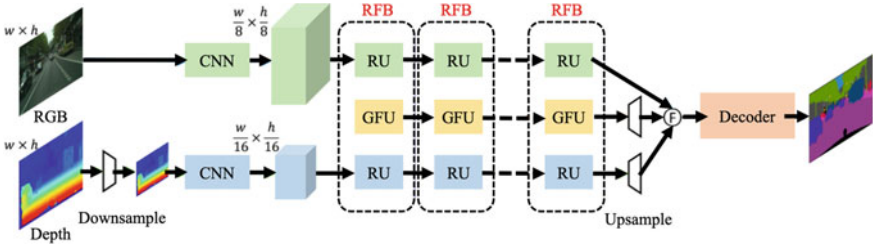

Fig. 4.9 Architecture of RFBNet [71]. Three streams: RGB stream (green), depth stream (blue), and interaction stream (yellow). F denotes the fused method

decoder, RGB and spectral image feature maps are processed separately by the depth-aware CNNs.

In addition to depth information, significant efforts have been directed towards integrating thermal and RGB images. The two types of visual information complement each other, as thermal images are proficient at representing diverse environments with varying illumination and encoding the structural details of the scene [74, 75]. The majority of previous research has split the encoder component of the semantic segmentation model into two branches of networks, extracting features from both RGB and thermal images, and then progressively merging the thermal characteristics with the RGB features based on the number of layers in the network. The following describes the current studies on fusion of RGB and thermal images.

MFNet [76] is an encoder-decoder structure data fusion network, featuring two encoders for extracting features from both RGB and thermal images (as shown in Fig. 4.10). To incorporate contextual information and extract features from each encoder, mini-inception modules utilize parallel convolutional and hole convolutional layers. In the fusion stage, the short-cut module concatenates the RGB and thermal features from each encoder at every scale before they are added to the output of the previous decoder layer, yielding the ultimate fused result. MFNet maximizes the complementarity between thermal image and RGB information for enhanced semantic segmentation in autonomous driving scenarios during the night.

RTFNet [77] is composed of three main components: an RGB encoder, a thermal encoder, and a decoder (as shown in Fig. 4.11). To extract the necessary features from the RGB and thermal images, the ResNet model is utilized as a feature extractor. The fusion process involves adding the extracted RGB and thermal features on a per-pixel basis to obtain the fused output. This fused output is then fed into the decoder to recover the feature map resolution and ultimately obtain the semantic segmentation result.

FuseSeg [78] (as shown in Fig. 4.12) also consists of two encoders and a decoder. The two encoders use DenseNet [79] to extract RGB and thermal features respectively. The RGB encoder incorporates the corresponding thermal features with the RGB features using element summation. These fused feature maps are then further

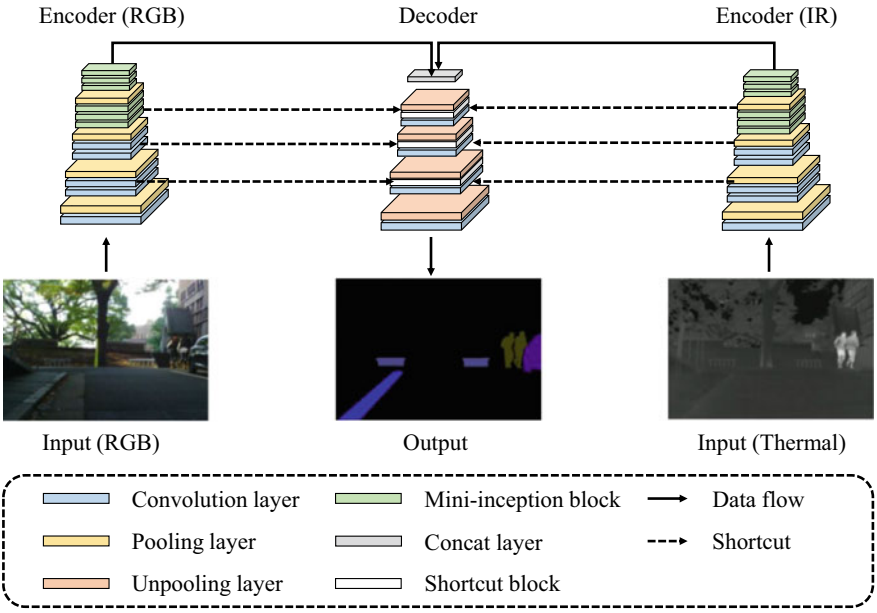

Fig. 4.10 Architecture of MFNet [76]. The output of RGB and IR encoders are fused to feed into the decoder

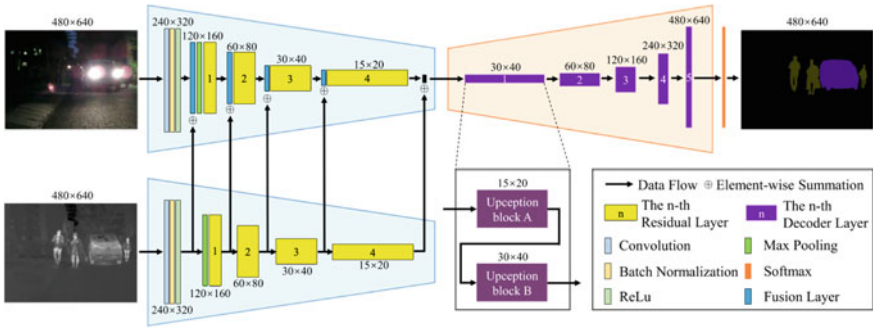

Fig. 4.11 Architecture of RTFNet [77], consisting of an RGB encoder, a thermal encoder (blue) and a decoder (orange)

combined with the decoder's corresponding feature maps using the tensor cascade method. Additionally, the bottom feature maps are directly copied to the decoder.

FEANet [80] comprises two components and aims to improve the comprehension of spatial information (as shown in Fig. 4.13). The first component of the network utilizes two encoders to extract RGB and thermal features separately. The output from each encoder layer is then fed into the feature-enhanced attention module (FEAM), which blends and refines the two features through weightage and fusion. This process enhances the features, allowing for effective capture and utilization of the comple-

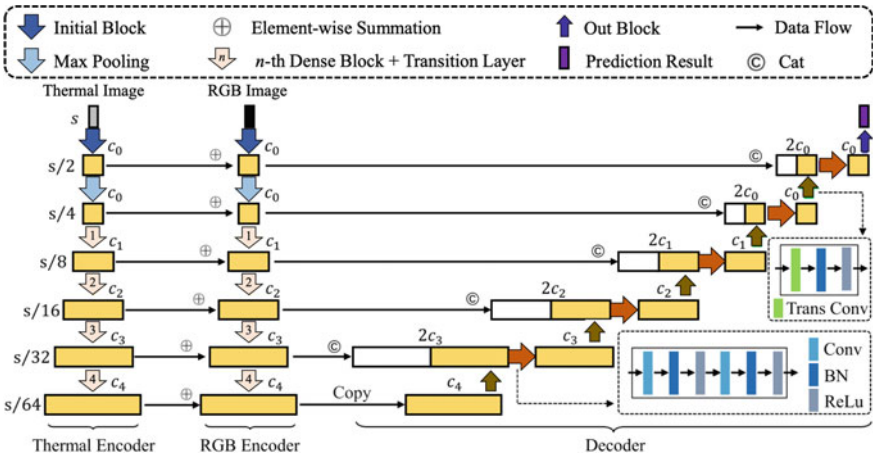

Fig. 4.12 Architecture of FuseSeg [78], consisting of an RGB encoder, a thermal encoder, and a decoder. It uses DenseNet as the backbone of the encoders

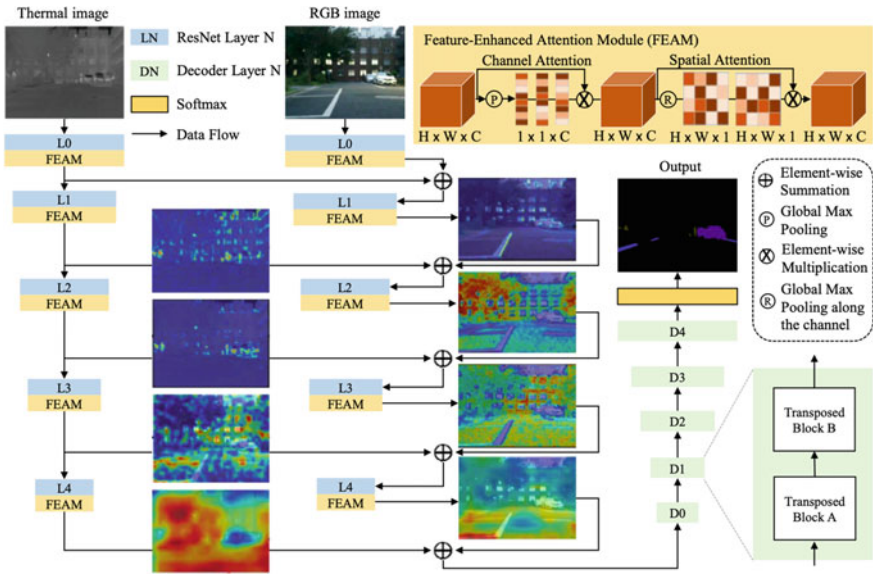

Fig. 4.13 Architecture of FEANet [80], consisting of a thermal stream, an RGB stream, and an output stream

mentary relationship between the two distinct features. The second component of the network uses a decoder to recover the resolution, which takes the enhanced fused features as input.

MFFENet [81] consists of two encoders, a feature fusion layer, and a multi-label supervision layer. It concatenates the multi-scale features with the features that

contain global semantic information. To extract features, two parallel encoders based on DenseNet are situated in the top section. Feature fusion employs elementwise summation, with features at various levels being combined at the encoders' peak. The spatial attention mechanism module is used to improve fused features by focusing on foreground objects. The upgraded features are supplied to a multi-label supervision layer, which is responsible for semantic segmentation.

GMNet [82] classifies the extracted features into three groups that correspond to specific aspects of the visual content: fine-grained visual details, semantic information, and intermediate-level features. During the fusion process, GMNet integrates the low-level and high-level features separately. Specifically, for the low-level features, GMNet employs spatial and channel attention mechanisms to combine information across different feature maps. As for the high-level features, GMNet leverages multi-scale null convolution to capture fine-grained details while preserving the semantic richness of the information.

ABMDRNet [83] considers the differences between two types of visual information by mitigating their differences before fusing. Specifically, a bidirectional translation technique is employed to reduce feature differences between the images, followed by a weighting scheme to dynamically select the discriminative features for semantic segmentation.

CCAFFMNet [84] consists of two encoders and a decoder. The chief function of these encoders is to extract features from both RGB and thermal images, respectively. An innovative channel-coordinate attention feature-fusion module (CCAFFM) is deployed within the encoder to effectuate this. Specifically, the aforementioned module is responsible for extracting the RGB features and subsequently gauging the spatial correlation between each RGB and thermal feature. It then proceeds to fuse the feature maps, thereby generating fused features which are ultimately fed into the fusion layer of the RGB encoder. Finally, the fused features produced by each fusion layer are input into the decoder to enable the recovery of the resolution.

SNE-RoadSeg [86] proposed a data-fusion CNN architecture RoadSeg, which can extract and fuse features from both RGB images and the inferred surface normal information for accurate freespace detection. Based on it, SNE-RoadSeg+ [85] (Fig. 4.14) consists of SNE+ module for more accurate surface normal estimation, and a data-fusion DCNN (RoadSeg+) that can greatly minimize the trade-off between accuracy and efficiency with the use of deep supervision.

RDFNet [87] presents a fusion network that receives RGB and HHA [88] information as inputs and enhances the diverse visual information as well as the features across various levels. The HHA information is obtained via depth map coding.

These methods of fusing RGB and other visual information (Table 4.1) can combine the spatial structure features in the scene, thus boosting the performance of semantic segmentation. However, in the majority of the aforementioned approaches, the RGB and other visual features are first extracted separately before being fused, thereby limiting their ability to distinguish the differences and complementarities between different types of visual information.

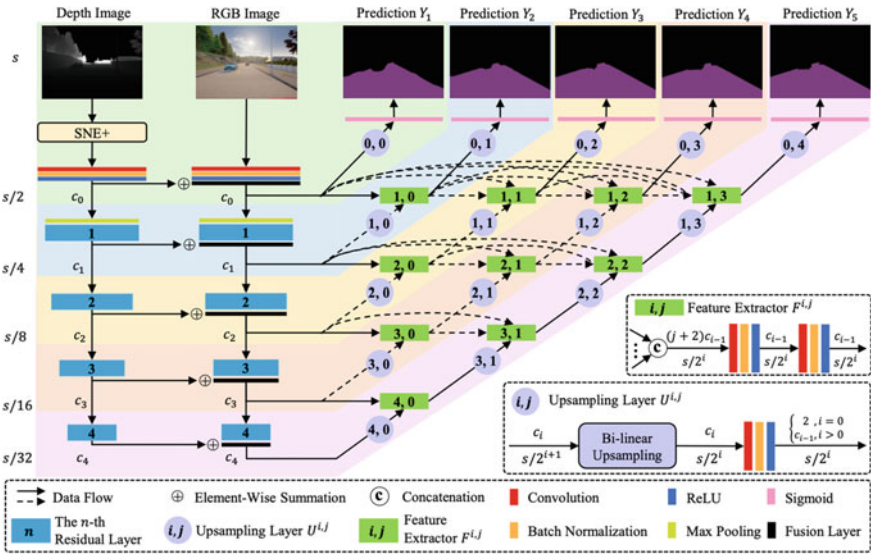

Fig. 4.14 Architecture of SNE-RoadSeg+ [85]. It consists of SNE+ (a module for achieving surface normal information), and RoadSeg+ (a data-fusion network)

Table 4.1 The results of semantic segmentation methods based on data fusion

Method	Year	Visual information	Dataset
FuseNet [18]	2017	RGB+Depth	SUNRGBD [89]
RedNet [68]	2018	RGB+Depth	SUNRGBD [89]
RFNet [69]	2020	RGB+Depth	Cityscapes [90], Lost and found [91]
RFBNet [71]	2019	RGB+Depth	ScanNet [92], Cityscapes [90]
Depth-aware CNN [73]	2018	RGB+Depth	NYU V2 [93], SUNRGBD [89], SID [94]
MFNet [76]	2017	RGB+Thermal	MFNet [76]
RTFNet [77]	2019	RGB+Thermal	MFNet [76]
FuseSeg [78]	2020	RGB+Thermal	MFNet [76]
FEANet [80]	2021	RGB+Thermal	MFNet [76]
MFFENet [81]	2021	RGB+Thermal	PST900 [95], MFNet [76]
GMNet [82]	2021	RGB+Thermal	PST900 [95], MFNet [76]
ABMDRNet [83]	2021	RGB+Thermal	MFNet [76]
CCAFFMNet [84]	2022	RGB+Thermal	RoadScene [96], MFNet [76],
SNE-RoadSeg [86]	2020	RGB+Surface normal	R2D Road [86], KITTI [97], SYNTHIA [98]
SNE-RoadSeg+ [85]	2021	RGB+Surface normal	KITTI [97], R2D road [86]
RDFNet [87]	2017	RGB+HHA	NYU V2 [93], SUNRGBD [89]

4.3 Public Datasets and Benchmarks

4.3.1 Public Datasets

In this section, we provide a summary of the datasets that have been used for training and testing in the context of semantic segmentation tasks. The datasets can be divided into two categories based on their data types: 2D datasets and 2.5/3D datasets. At the same time, for autonomous driving applications, models are often trained using synthetic data. Therefore, we also count the scenario types of the dataset in this section, including real scenarios and synthetic scenarios. Table 4.2 includes some classical datasets for semantic segmentation tasks.

4.3.1.1 2D Dataset

PASCAl VOC[1] [99] is a dataset for generic scenarios in semantic segmentation, consisting of 17125 images with a total of 20 categories (excluding background). For the semantic segmentation task, it provides 2913 images, of which 1464 were used for training with 3507 objects and 1449 were used for validation with 3422 objects. The test dataset is not publicly available.

PASCAL Context[2] [100] adds new annotated categories to PASCAL VOC 2010, increasing the number of categories to 540, with 59 commonly used categories and the others being re-labeled as background. The dataset includes 10103 images for training.

PASCAL Part[3] [101] provides additional annotations in PASCAL VOC 2010, and the components in each object are also annotated. This dataset includes the training dataset and validation dataset from PASCAL VOC, as well as 9637 labels for testing.

Stanford background dataset[4] [102] includes 715 images selected from the public datasets LabelMe [103], MSRC [104], PASCAL VOC 2007 [105], and Geometric Context [106]. The dataset is primarily targeted at outdoor scenarios with an image size of 320×240 pixels and includes at least one foreground object. The semantic and geometric labels were obtained through Amazon's Mechanical Turk (AMT) [107].

Semantic Boundaries Dataset (SBD)[5] [108] is a semantic boundary dataset that aggregates unannotated images from PASCAL VOC 2011 [109] and annotates 11355 images for semantic segmentation, 8498 of which were used for training and 2857 for testing. These annotations include the boundaries of each category and are based on category-level as well as instance-level information.

---

[1] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/.

[2] https://cs.stanford.edu/~roozbeh/pascal-context/.

[3] http://roozbehm.info/pascal-parts/pascal-parts.html.

[4] http://dags.stanford.edu/projects/scenedataset.html.

[5] http://home.bharathh.info/pubs/codes/SBD/download.html.

Table 4.2 Classic datasets for semantic segmentation tasks

Dataset	Year	Classes	Synthetic/Real	Data	Samples	scenario
Stanford background [102]	2009	8	R	2D	715	Outdoor
SBD [108]	2011	21	R	2D	11355	Generic
PASCAL VOC [99]	2012	20	R	2D	≈3 k	Generic
PASCAL Context [100]	2014	540	R	2D	≈20 k	Generic
PASCAL Part [101]	2014	20	R	2D	≈20 k	Body part
COCO [110]	2014	≥ 80	R	2D	204721	Generic
NYU Depth V2 [93]	2012	894	R	2.5D	≈1.5 k	Indoor
SUN3D [127]	2013	–	R	2.5D	19640	Indoor
SUNRGBD [89]	2015	37	R	2.5D	10335	Indoor
ScanNet [92]	2017	21	R	2.5D/3D	1513	Generic
CamVid [111]	2008	32	R	2D	701	Driving
KITTI (road) [97]	2013	10	R	2D	579	Driving
KITTI (semantics) [114]	2018	11	R	2D	400	Driving
Virtual KITTI [115]	2016	–	S	2D	17 k	Driving
SYNTHIA [98]	2016	11	S	2D	13407	Driving
Cityscapes [90]	2016	33	R	2D	25 k	City
ADE20K [116, 117]	2017	150	R	2D	≥25 k	Generic
MVD [120]	2017	66	R	2D	25 k	City
ApolloScape [121]	2018	25	R	2D	140 k	Driving
Highway driving [122]	2019	10	R	2D	1200	Driving
WoodScape [123]	2019	40	R	2D	10 k	Driving
IDD [124]	2019	34	R	2D	10003	Driving
A2D2 [125]	2020	38	R	2D	≈41 k	Driving
IDDA [126]	2020	–	S	2D	1000 k	Driving

Microsoft Common Objects in Context (COCO)[6] [110] is a dataset built by Microsoft that includes tasks such as detection, key points, and segmentation. The dataset for the segmentation task is currently the largest semantic segmentation dataset in computer vision, containing over 80 objects, and consists of two separate releases in 2014, with 82783 training images, 40504 validation images, and 40775 test images. In the 2015 version, there are 165482 training images, 81208 validation images, and 81434 test images.

The Cambridge-driving Labeled Video Dataset (CamVid)[7] [111] is a semantic segmentation dataset for driving scenarios, which consists of urban road scenarios, including videos of five driving scenarios. There are more than 700 annotated images with a resolution of 960×720 and 32 objects. The dataset includes 368 images for training, 100 for validation, and 233 for testing.

KITTI[8] [112, 113] is a dataset for the evaluation of various computer vision tasks in autonomous driving scenarios, including real-world scenarios such as urban, rural, and highway. Semantic segmentation and instance segmentation tasks [114] consists of 200 training images with semantic segmentation annotations and 200 test images corresponding to stereo2015 and flow2015. Road semantic segmentation task [97] includes 289 images for training and 290 images for testing.

Virtual KITTI[9] [115] uses real-to-virtual world cloning to create a synthetic video dataset consisting of five videos obtained from KITTI clones. It comprises a total of 21260 high-resolution frames, and they are generated from five different virtual worlds of urban scenarios under different imaging and weather conditions. Virtual KITTI2 adds a stereo camera that makes the scenarios more realistic.

SYNTHetic Collection of Imagery and Annotations (SYNTHIA)[10] [98] is used to address the problem of semantic segmentation for scene understanding in driving scenarios and consists of a photorealistic framework of virtual cities rendered out. It includes town, city, and highway scenarios in different seasons and weather with 11 pixel-level labeling categories (road, sidewalk, void, sky, building, fence, car, sign, vegetation, pole, pedestrian, and cyclist). 13407 images are extracted from the video sequences for training.

Cityscapes[11] [90] is a dataset for pixel-level and instance-level semantic segmentation, captured from 50 city streets and includes scenarios from the spring, summer, and autumn seasons. It uses a stereo camera to acquire stereo visual video sequences, of which 5000 images are annotated with high-quality pixel-level annotations, and 20000 images are coarsely annotated. The scenarios in this dataset are more complex than previous datasets.

---

[6] https://cocodataset.org/#home.

[7] http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/.

[8] https://www.cvlibs.net/datasets/kitti/.

[9] https://europe.naverlabs.com/research/computer-vision.

[10] https://synthia-dataset.net/downloads/.

[11] https://www.cityscapes-dataset.com/.

ADE20K[12] [116, 117] consists of 27000 images with 3688 objects from the dataset SUN [118] and Places [119], all images are anonymized. 25574 images are labeled for training and 2000 images for validation.

The Mapillary Vistas (MVD)[13] [120] is a large-scene streetscape dataset consisting of 25000 high-resolution images with 66 objects, of which 18000 are used for training, 2000 for validation, and 5000 for testing. The images in the dataset are taken in different weather, season, and daytime conditions.

ApolloScapes[14] [121] publishes over 140000 frames of pixel-level semantically annotated images, with each image semantically annotated for each pixel. It captures hundreds of moving objects for scenarios in various traffic conditions, providing 8900 instance-level annotations for moving objects. Meanwhile, pose information is also annotated with centimeter-level accuracy. The annotation accuracy and richness of this dataset are much higher than those of KITTI and Cityscapes.

Highway Driving[15] [122] is a densely annotated dataset with ten objects for the semantic video segmentation task, which takes into account the relationship between neighbouring frames for each frame annotation. It consists of 20 sequences of 60 frames at a frame rate of 30Hz, 15 of which are for training and five for testing.

WoodScape[16] [123] is the first publicly available fisheye dataset for autonomous driving, with image acquisition by the four fisheye cameras on the vehicle. The dataset includes a total of 10000 images with semantic annotations for 40 categories and over 100000 annotated images for other tasks.

IDD[17] [124] is a new dataset for understanding road scenes in unstructured environments. The images in the dataset are collected by front-facing cameras in cars, it includes 10003 images labeled with 34 categories and 182 driving sequences on Indian roads. The images are mostly in 1080p resolution, with some images in 720p and other resolutions. The categories annotated add unique categories such as animals and autorickshaws compared to other datasets. Conditions such as weather and lighting are highly variable in this dataset.

Audi Autonomous Driving (A2D2)[18] [125] is a large autonomous driving dataset released by Audi, which contains RGB images and corresponding 3D point cloud data. The dataset uses six cameras and five LiDAR sensors to capture the surrounding environment of vehicles, collecting scenarios from highways, rural roads, and cities in southern Germany under different weather conditions. The dataset is annotated with 41227 frames of non-sequential data, all of which include semantic segmentation annotations and point cloud annotations, including 12497 frames of 3D bounding box annotations for objects in the field of view of the front camera. In addition, 392556 consecutive frames of unannotated sensor data are included.

[12] https://groups.csail.mit.edu/vision/datasets/ADE20K/index.html.

[13] https://research.mapillary.com/publication/iccv17a/.

[14] http://apolloscape.auto.

[15] https://sites.google.com/site/highwaydrivingdataset/.

[16] https://drive.google.com/drive/folders/1X5JOMEfVlaXfdNy24P8VA-jMs0yzf_HR.

[17] http://idd.insaan.iiit.ac.in/accounts/login/?next=/dataset/download/.

[18] https://www.a2d2.audi/a2d2/en.html.

ItalDesign Dataset (IDDA)[19] [126] is a large-scale synthetic dataset for semantic segmentation, consisting of over 100 different source visual domains. The dataset is used to solve the problem of domain transfer between training, and test data in various weather and viewpoint conditions in seven different types of cities.

4.3.1.2 2.5D/3D Dataset

NYU Depth V2[20] [93] consists of video sequences of indoor scenarios captured by RGB and depth cameras and includes 464 scenarios from three cities. In total, there are 1449 annotated RGB images and depth images containing 26 scenario categories. In addition, 407024 images are not annotated.

SUN3D[21] [127] is similar to NYU Depth V2 in that it consists of large-size RGB-D video data with eight annotated sequences, with scenarios in the dataset captured at different times of the day. Each frame in the sequence is accompanied by a semantic annotation and camera pose information. The dataset consists of 415 video sequences, captured in 41 different buildings in 254 different scenarios.

SUNRGBD[22] [89] is a collection of 10000 images from four RGB-D sensors, including images from NYU depth v2, Berkeley B3DO [128], and SUN3D. The dataset is densely annotated, including 146617 2D polygons and 58657 3D bounding boxes with accurate object orientations, as well as 3D room layouts and categories of the scenario.

ScanNet[23] [92] is an RGB-D video dataset containing 2.5 million views in more than 1.5 k scans. It annotates with 3D camera poses, surface reconstructions, and instance-level semantic segmentations. It consists of a total of 1513 images with 21 categories of objects, of which 1201 scenes are used for training and 312 scenes for testing.

4.3.2 Online Benchmarks

In this section, we list the benchmark results corresponding to two classical datasets for the semantic segmentation task. The evaluation metrics for semantic segmentation are based on the generic IoU (as shown in Sect. 4.4) and give four related metrics: IoU class, iIoU class, IoU category, and iIoU category. IoU class considers the accuracy of the segmentation results for the global scenario and evaluates the segmentation accuracy for the whole scenario. However, the global IoU measure is biased towards object instances covering large areas of the image and is not appli-

[19] https://idda-dataset.github.io/home/.

[20] https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html.

[21] http://sun3d.cs.princeton.edu/.

[22] http://rgbd.cs.princeton.edu/challenge.html.

[23] https://github.com/ScanNet/ScanNet.

cable to scenarios with strong scale variation. Therefore, iIoU class builds on the IoU class by segmenting the image scenario at the instance level and predicting the result. This metric focuses more on the segmentation accuracy of the algorithm across instances, with each instance in the scenario being segmented. Different instances in the same category are evaluated separately to obtain the final result. IoU is divided into 'Class' and 'Category' according to the granularity of the segmentation, with the subdivision class being divided into the coarse category for coarse granularity. The IoU category and iIoU category are evaluated according to the granularity of the segmentation.

4.3.2.1 Cityscapes

Table 4.3 shows semantic segmentation results in Cityscapes benchmark. This task involves predicting a per-pixel semantic labeling of the image without considering higher-level object instance or boundary information. It introduces a loss term that is aware of the boundaries for semantic segmentation, using an inverse-transformation network that effectively learns the extent of parametric transformations between predicted and target boundaries [90]. It achieves the highest IoU class, iIoU class, and iIoU category, while HMS Attention [129] method achieves the highest IoU category.

Table 4.3 Semantic segmentation results in Cityscapes benchmark, where the best results are in bold type

Method	IoU class (%)	iIoU class (%)	IoU category (%)	iIoU category (%)
InverseFormNet [130]	**85.8**	**72**	93.1	**85.6**
HMS Attention [129]	85.4	70.4	**93.2**	85.4
Naive-Student [131]	85.2	68.8	92.9	82.0
ViT-Adapter [132]	85.2	68.3	92.8	83.4
Wide-ResNets [133]	85.1	71.2	93.0	85.1
OCR-Seg [134]	84.5	65.9	92.7	83.9
Panoptic-DeepLab [135]	84.5	68.7	92.9	82.8
DCNAS [136]	84.3	68.5	92.7	84.6
EfficientPS [137]	84.2	65.2	92.5	83.5
Axial-DeepLab [138]	84.1	66.0	92.6	79.7

Table 4.4 Semantic segmentation results in KITTI benchmark, where the best results are in bold type

Method	IoU class (%)	iIoU class (%)	IoU category (%)	iIoU category (%)
WRP [139]	**76.44**	**50.92**	**89.63**	73.69
UJS_model [140]	75.11	47.71	89.53	**75.75**
VideoProp [141]	72.82	48.68	88.99	75.26
SN_DN161 [142]	68.89	40.45	87.06	67.93
MSeg1080_RVC [143]	62.64	31.62	86.59	68.05
Chroma UDA [144]	60.36	31.70	80.73	61.91
IfN-DomAdap [145]	59.50	30.28	81.57	61.91
SegStereo [146]	59.10	28.00	81.31	60.26
SGDepth [147]	53.04	24.36	78.65	55.95
SDNet [148]	51.14	17.74	79.62	50.45
APMoE_ROB [149]	47.96	17.86	78.11	49.17

4.3.2.2 KITTI

Table 4.4 shows the KITTI semantic segmentation benchmark. It consists of 200 semantically annotated trains as well as 200 test images corresponding to the KITTI Stereo and Flow Benchmark 2015. WRP [139] method learns to refine geometrically-warped labels and infuse them with learned semantic priors in a semi-supervised setting by leveraging cycle-consistency across time. It achieves the highest IoU class, iIoU class, and IoU category, while UJS_model [140] method achieves the highest iIoU category.

4.4 Evaluation Metrics

To ensure the usefulness and significance of semantic segmentation in this field, strict performance evaluation is necessary. Moreover, it is imperative that the evaluation of a system is carried out using established and widely recognized metrics that ensure equitable comparisons with existing methods. Moreover, various aspects of the system must be scrutinized to establish its soundness and usefulness, including execution time, memory usage, and accuracy. Depending on the system's purpose or context, certain metrics may carry greater significance than others; for instance, in a real-time application, execution speed may take precedence over accuracy to a certain extent. However, in the interest of scientific rigor, it is crucial to provide a comprehensive set of metrics for any proposed method.

Many evaluation criteria have been proposed and are frequently used to assess the accuracy of any kind of technique for semantic segmentation. The essence of semantic segmentation is the classification of pixels, the evaluation metrics of classification methods include confusion matrix, accuracy, precision, recall, and F1-score. The confusion matrix is composed of n_{TP} (true positive), n_{FP} (false positive), n_{FN} (false negative), and n_{TN} (true negative). The evaluation of semantic segmentation performance involves computing metrics such as pixel accuracy and Intersection over Union (IoU). We remark on the following notation details.

We denote the total number of all classes as m, where i represents the ith class.

- Pixel Accuracy (PA). PA is used to evaluate the percentage of correctly classified pixels. It is a simple metric that computes the ratio between the number of properly classified pixels and the total number of pixels. The formula for calculating PA is:

$$\mathrm{PA} = \frac{n_{\mathrm{TP}} + n_{\mathrm{TN}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}} + n_{\mathrm{FN}} + n_{\mathrm{TN}}}. \tag{4.1}$$

- Class Pixel Accuracy (CPA). CPA represents the percentage of pixels that are actually classified as category i among all pixels predicted as category i:

$$\mathrm{CPA} = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}}}. \tag{4.2}$$

- Mean Pixel Accuracy (MPA). MPA is calculated by dividing the sum of the correctly predicted pixels for each class by the sum of the total predicted pixels for that class. It is an improvement over PA and takes into account the accuracy for each individual class rather than just the overall accuracy. The formula for MPA is:

$$\mathrm{MPA} = \frac{\sum_{i=1}^{m} \mathrm{CPA}_i}{m}. \tag{4.3}$$

- Intersection over Union (IoU). IoU is a standard metric used to measure the degree of coincidence between pixel predictions and the ground truth, and to determine whether the pixel prediction is a positive or negative sample by comparing it with a threshold. It calculates the ratio between the intersection and the union of two sets, namely the ground truth and the predicted segmentation. This ratio can be expressed as the number of true positives (intersection) over the sum of true positives, false negatives, and false positives (union). IoU is calculated on a per-class basis:

$$\mathrm{IoU} = \frac{n_{\mathrm{TP}}}{n_{\mathrm{TP}} + n_{\mathrm{FP}} + n_{\mathrm{FN}}} \tag{4.4}$$

- Mean intersection over Union (MIoU). MIoU is a commonly used evaluation metric in semantic segmentation. It measures the average intersection over union (IoU) for all classes in a dataset. It is calculated by first computing IoU for each class and then taking the mean of all class IoUs. The formula for MIoU is as follows:

$$\text{MIoU} = \frac{\sum_{i=1}^{m} \text{IoU}_i}{m}. \tag{4.5}$$

- Frequency Weighted Intersection over Union (FWIoU). It is an improved version of MIoU, which takes into account the class imbalance issue that often occurs in semantic segmentation datasets. It assigns different weights to each class based on its appearance frequency, so that the evaluation metric is not dominated by the performance of the majority classes. FWIoU is calculated as follows:

$$\text{FWIoU} = \frac{n_{\text{TP}} + n_{\text{FN}}}{n_{\text{TP}} + n_{\text{FP}} + n_{\text{TN}} + n_{\text{FN}}} \times \frac{n_{\text{TP}}}{n_{\text{TP}} + n_{\text{FP}} + n_{\text{FN}}}. \tag{4.6}$$

- F1 score (also known as the Dice score). It is a metric used to evaluate the accuracy of binary classification models. It is calculated by balancing the precision and recall of the test sample, and is given by the following formula:

$$F_\beta = \frac{n_{\text{TP}}^2(1+\beta^2)}{(1+\beta^2)n_{\text{TP}}^2 + n_{\text{TP}}(\beta^2 n_{\text{FN}} + n_{\text{FP}})} \tag{4.7}$$

  Generally, recall is considered as important as precision, so F1 score sets β to 1:

$$F_1 = \frac{2n_{\text{TP}}}{2n_{\text{TP}} + n_{\text{FP}} + n_{\text{FN}}}. \tag{4.8}$$

4.5 Specific Autonomous Driving Tasks

Semantic segmentation is utilized as a valuable tool for accomplishing specific tasks related to autonomous driving.

4.5.1 Freespace Detection

Freespace detection plays a crucial role in the visual perception of autonomous driving [150, 151]. It is essential to determine the drivable area for a vehicle to move safely. Freespace detection provides vital information for subsequent path planning, mainly to achieve road path planning and obstacle avoidance. However, environmental factors such as light conditions and extreme weather can affect the accuracy of freespace detection results obtained from cameras, LiDAR, and other sensors. Additionally, in congested road sections, surrounding vehicles may obstruct the view and affect the accuracy of freespace detection. Freespace detection is essentially a semantic segmentation problem that can be addressed through data fusion of different visual inputs. Figure 4.15 [86] shows the results of freespace detection.

Fig. 4.15 Examples on the KITTI road benchmark, where rows **a**–**f** show the freespace detection results obtained by RBNet [152], TVFNet [153], LC-CRF [154], LidCamNet [155], RBANet [156] and SNE-RoadSeg, respectively. The true positive, false negative and false positive pixels are shown in green, red and blue, respectively

4.5.2 Road Defect Detection

Road pothole detection systems have become critical not only for smart urban road maintenance but also for autonomous driving [64, 157–160], with the rapid development of computers. While today's autonomous vehicles primarily focus on large target information such as pedestrians, traffic signs, and other vehicles, road defects also play a crucial role in ensuring ride quality, vehicle handling, fuel consumption, and tire wear. Sensing information about the size and shape of potholes can help self-driving cars drive smoothly, thereby improving driving comfort while protecting the vehicle [161, 162]. In recent years, semantic segmentation has been widely used for road defect detection [163]. Figure 4.16 [164] shows the results of semantic segmentation applications.

4.5.3 Road Anomaly Detection

The ability to detect road anomalies is crucial for ensuring the safety of autonomous vehicles [170]. Road anomalies refer to areas where there is a height difference from the surrounding freespace area, and detecting them allows the vehicle to make timely adjustments and avoid potential risks. Recent advances in deep learning technology have led to the development of several semantic segmentation methods that have proven effective in detecting road anomalies. Figure 4.17 [171] shows the results of semantic segmentation based on data fusion applied in road anomaly detection.

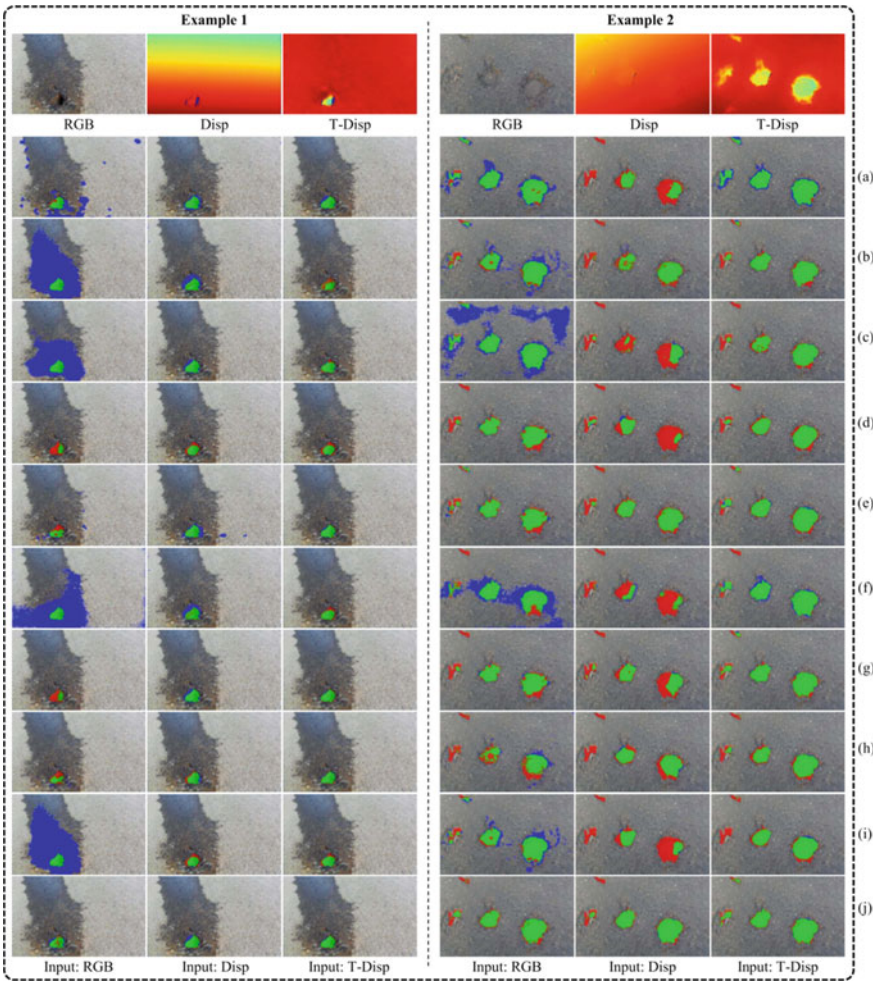

Fig. 4.16 Examples of the experimental results of SoTA CNNs: **a** FCN [27]; **b** U-Net [34]; **c** DenseASPP [165]; **d** DUpsampling [166]; **e** GSCNN [167]; **f** SegNet [37]; **g** DeepLabv3+ [33]; **h** PAN [168]; **i** ESPNet [169]; **j** GAL-DeepLabv3+ [164], where the true-positive, false-positive, and false-negative pixels are shown in green, blue and red, respectively

4.6 Existing Challenges

Multi-visual information fusion is particularly challenging for semantic segmentation tasks in autonomous driving because of its high requirements for accuracy, robustness, and real-time performance. Efficient processing of input information is a prerequisite for accurate perception of complex environments by driverless vehicles. In Sect. 4.2, we summarized the multi-visual information fusion network for semantic

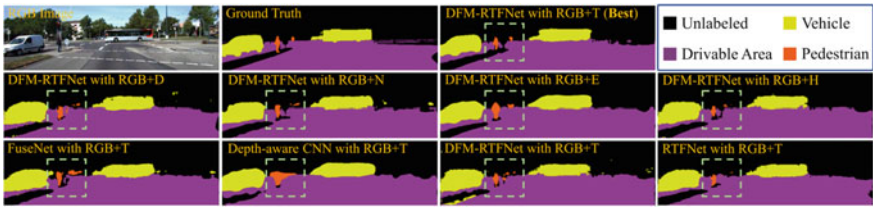

Fig. 4.17 Example of the experimental results on the KITTI semantic segmentation dataset. FuseNet [18], MFNet [76], Depth-aware CNN [73], RTFNet [77], and DFM-RTFNet [171] are all data-fusion networks. Significantly improved regions are marked with green dashed boxes

segmentation. Existing data fusion methods commonly combine RGB images with either thermal or depth images. Thermal images are particularly useful for detecting high-temperature regions, contours of targets, points or lines of sudden temperature changes, trends in temperature, and other related features. On the other hand, depth images are effective in capturing the physical structure of each object in the scene. However, these types of visual information are applicable only to specific scenes and scenarios. Currently, the fusion of RGB images with other visual information has received limited research attention. Therefore, selecting and acquiring the appropriate visual information based on the characteristics of the scenario remains a challenge. Due to its unique characteristics, such as color and texture, RGB data requires different network architectures and fusion methods compared to other types of visual information. While most existing fusion methods for RGB images involve overlaying and stitching, these approaches often fail to account for the spatial and temporal relationships between different types of visual information. As a result, fusing RGB images with other visual information is a significant challenge that requires the development of novel fusion methods. Furthermore, the computational costs and memory requirements associated with these fusion methods must be carefully considered to ensure that they are suitable for use in real-world applications.

4.7 Conclusion

In this chapter, we extensively explored semantic segmentation techniques for autonomous driving, covering both single-modal and data fusion approaches. We began by providing a comprehensive overview of semantic segmentation methods and then delved into the latest techniques for both scenarios, highlighting their significance and contributions to the field. We also presented common datasets that researchers can use to achieve their objectives and satisfy their requirements, along with evaluation metrics for assessing semantic segmentation performance and corresponding benchmarks for several classic datasets. Finally, we discussed various applications of autonomous driving and shared our perspective on the current state-of-the-art of semantic segmentation.

Acknowledgements This work was supported by the National Key R&D Program of China under Grant 2020AAA0108100, the National under Grant 62233013, and the Science and Technology Commission of Shanghai Municipal under Grant 22511104500.

References

1. Wang X-F, Huang D-S, Xu H (2010) An efficient local chan-vese model for image segmentation. Pattern Recognit 43(3):603–618
2. Ess A, Müller T, Grabner H, Van Gool L (2009) Segmentation-based urban traffic scene understanding. In: British machine vision conference (BMVC). Citeseer, p 2
3. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 3354–3361
4. Oberweger M, Wohlhart P, Lepetit V (2015) Hands deep in deep learning for hand pose estimation. arXiv:1502.06807
5. Yoon Y, Jeon H-G, Yoo D, Lee J-Y, So Kweon I (2015) Learning a deep convolutional network for light-field image super-resolution. In: Proceedings of the IEEE international conference on computer vision workshops (ICCV Workshop), pp 24–32
6. Wan J, Wang D, Hoi SCH, Wu P, Zhu J, Zhang Y, Li J (2014) Deep learning for content-based image retrieval: a comprehensive study. In: Proceedings of the 22nd ACM international conference on multimedia, pp 157–166
7. Dickmanns ED, Mysliwetz BD (1992) Recursive 3-D road and relative ego-state recognition. IEEE Trans Pattern Anal Mach Intell 14(02):199–213
8. Wang Y, Zhou Q, Liu J, Xiong J, Gao G, Wu X, Latecki LJ (2019) Lednet: a lightweight encoder-decoder network for real-time semantic segmentation. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 1860–1864
9. Wang X, Huang D (2009) A novel density-based clustering framework by using level set method. IEEE Trans Knowl Data Eng 21(11):1515–1531
10. Huang D, Du J (2008) A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. IEEE Trans Neural Netw 19(12):2099–2115
11. Huang D (1999) Radial basis probabilistic neural networks: Model and application. Int J Pattern Recognit Artif Intell 13(07):1083–1101
12. Zhao Z-Q, Huang D-S, Sun B-Y (2004) Human face recognition based on multi-features using neural networks committee. Pattern Recognit Lett 25(12):1351–1358
13. Couprie C, Farabet C, Najman L, LeCun Y (2013) Indoor semantic segmentation using depth information: 1st international conference on learning representations, iclr 2013. In: 1st international conference on learning representations (ICLR)
14. Farabet C, Couprie C, Najman L, LeCun Y (2012) Learning hierarchical features for scene labeling. IEEE Trans Pattern Anal Mach Intell 35(8):1915–1929
15. Cheng Y, Cai R, Li Z, Zhao X, Huang K (2017) Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp 3029–3037
16. Gupta S, Girshick R, Arbeláez P, Malik J (2014) Learning rich features from rgb-d images for object detection and segmentation. In: European conference on computer vision (ECCV). Springer, pp 345–360
17. Wang J, Wang Z, Tao D, See S, Wang G (2016) Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In: European conference on computer vision (ECCV). Springer, pp 664–679
18. Hazirbas C, Ma L, Domokos C, Cremers D (2017) Fusenet: incorporating depth into semantic segmentation via fusion-based cnn architecture. In: Asian conference on computer vision (ACCV). Springer, pp 213–228

19. Song X, Herranz L, Jiang S (2017) Depth cnns for rgb-d scene recognition: learning from scratch better than transferring from rgb-cnns. In: Thirty-first AAAI conference on artificial intelligence
20. Sistu G, Leang I, Yogamani S (2019) Real-time joint object detection and semantic segmentation network for automated driving. arXiv:1901.03912
21. Siam M, Elkerdawy S, Jagersand M, Yogamani S (2017) Deep semantic segmentation for automated driving: taxonomy, roadmap and challenges. In: 2017 IEEE 20th international conference on intelligent transportation systems (ITSC). IEEE, pp 1–8
22. Pan B, Sun J, Leung HYT, Andonian A, Zhou B (2020) Cross-view semantic segmentation for sensing surroundings. IEEE Robot Autom Lett 5(3):4867–4873
23. Yang K, Hu X, Chen H, Xiang K, Wang K, Stiefelhagen R (2020) Ds-pass: detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing. In: 2020 IEEE intelligent vehicles symposium (IV). IEEE, pp 457–464
24. Liu C-W, Wang H, Guo S, Junaid Bocus M, Chen Q, Fan R (2023) Stereo matching: fundamentals, state-of-the-art, and existing challenges. Springer, submitted for publication
25. Khan MZ, Gajendran MK, Lee Y, Khan MA (2021) Deep neural architectures for medical image semantic segmentation. IEEE Access 9:83 002–83 024
26. Alalwan N, Abozeid A, ElHabshy AA, Alzahrani A (2021) Efficient 3d deep learning model for medical image semantic segmentation. Alex Eng J 60(1):1231–1239
27. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3431–3440
28. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on Computer vision and pattern recognition (CVPR), pp 580–587
29. Hariharan B, Arbeláez P, Girshick R, Malik J (2014) Simultaneous detection and segmentation. In: European conference on computer vision (ECCV). Springer, pp 297–312
30. Wu H, Zhang J, Huang K, Liang K, Yu Y (2019) Fastfcn: rethinking dilated convolution in the backbone for semantic segmentation. arXiv:1903.11816
31. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR), vol 2. IEEE, pp 2169–2178
32. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2881–2890
33. Chen L, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
34. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention (MICCAI). Springer, pp 234–241
35. Chaurasia A, Culurciello E (2017) Linknet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE visual communications and image processing (VCIP). IEEE, pp 1–4
36. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y (2017) The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPR Workshop), pp 11–19
37. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 39(12):2481–2495
38. Paszke A, Chaurasia A, Kim S, Culurciello E (2016) Enet: a deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147
39. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5693–5703

40. Florian L-C, Adam SH (2017) Rethinking atrous convolution for semantic image segmentation. In: Conference on computer vision and pattern recognition (CVPR), vol 6
41. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2881–2890
42. Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122
43. Liang-Chieh C, Papandreou G, Kokkinos I, Murphy K, Yuille A (2015) Semantic image segmentation with deep convolutional nets and fully connected crfs. In: International conference on learning representations (ICLR)
44. Zhao H, Zhang Y, Liu S, Shi J, Loy CC, Lin D, Jia J (2018) Psanet: point-wise spatial attention network for scene parsing. In: Proceedings of the European conference on computer vision (ECCV), pp 267–283
45. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7794–7803
46. Zhu Z, Xu M, Bai S, Huang T, Bai X (2019) Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 593–602
47. Yin M, Yao Z, Cao Y, Li X, Zhang Z, Lin S, Hu H (2020) Disentangled non-local neural networks. In: European conference on computer vision (ECCV). Springer, pp 191–207
48. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3146–3154
49. He J, Deng Z, Zhou L, Wang Y, Qiao Y (2019) Adaptive pyramid context network for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 7519–7528
50. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W (2019) Ccnet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 603–612
51. Huang L, Yuan Y, Guo J, Zhang C, Chen X, Wang J (2019) Interlaced sparse self-attention for semantic segmentation. arXiv:1907.12273
52. Li X, Zhong Z, Wu J, Yang Y, Lin Z, Liu H (2019) Expectation-maximization attention networks for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 9167–9176
53. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc: Ser B (Methodological) 39(1):1–22
54. Zhang H, Dana K, Shi J, Zhang Z, Wang X, Tyagi A, Agrawal A (2018) Context encoding for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7151–7160
55. Strudel R, Garcia R, Laptev I, Schmid C (2021) Segmenter: transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 7262–7272
56. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
57. Yuan Y, Chen X, Chen X, Wang J (2019) Segmentation transformer: object-contextual representations for semantic segmentation. arXiv:1909.11065
58. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 10 012–10 022
59. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. In: International conference on learning representations (ICLR)
60. NicolasCarion F, GabrielSynnaeve NU (2020) Alexanderkirillov, and sergeyzagoruyko. End-to-end object detection with transformers. In: Proceedings of the European conference on computer vision (ECCV)

61. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P (2021) Segformer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst 34:12 077–12 090
62. Chu X, Tian Z, Wang Y, Zhang B, Ren H, Wei X, Xia H, Shen C (2021) Twins: revisiting the design of spatial attention in vision transformers. Adv Neural Inf Process Syst 34:9355–9366
63. Zhang H, Wu C, Zhang Z, Zhu Y, Lin H, Zhang Z, Sun Y, He T, Mueller J, Manmatha R et al (2022) Resnest: split-attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2736–2746
64. Fan R, Ozgunalp U, Wang Y, Liu M, Pitas I (2022) Rethinking road surface 3-D reconstruction and pothole detection: from perspective transformation to disparity map segmentation. IEEE Trans Cybern 52(7):5799–5808
65. Ming N, Feng Y, Fan R (2022) SDA-SNE: spatial discontinuity-aware surface normal estimation via multi-directional dynamic programming. In: 2022 International conference on 3D vision (3DV), pp 486–494
66. Feng Y, Xue B, Liu M, Chen Q, Fan R (2023) D2NT: a high-performing depth-to-normal translator. In: 2023 IEEE international conference on robotics and automation (ICRA), pp 12360–12366
67. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
68. Jiang J, Zheng L, Luo F, Zhang Z (2018) Rednet: residual encoder-decoder network for indoor rgb-d semantic segmentation. arXiv:1806.01054
69. Sun L, Yang K, Hu X, Hu W, Wang K (2020) Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images. IEEE Robot Autom Lett 5(4):5558–5565
70. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 7132–7141
71. Deng L, Yang M, Li T, He Y, Wang C (2019) Rfbnet: deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation. arXiv:1907.00135
72. Valada A, Mohan R, Burgard W (2020) Self-supervised model adaptation for multimodal semantic segmentation. Int J Comput Vis 128(5):1239–1285
73. Wang W, Neumann U (2018) Depth-aware cnn for rgb-d segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 135–150
74. Lian Q, Lv F, Duan L, Gong B (2019) Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: a non-adversarial approach. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 6758–6767
75. Guo M, Wang Z, Yang N, Li Z, An T (2018) A multisensor multiclassifier hierarchical fusion model based on entropy weight for human activity recognition using wearable inertial sensors. IEEE Trans Hum-Mach Syst 49(1):105–111
76. Ha Q, Watanabe K, Karasawa T, Ushiku Y, Harada T (2017) Mfnet: towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 5108–5115
77. Sun Y, Zuo W, Liu M (2019) Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. IEEE Robot Autom Lett 4(3):2576–2583
78. Sun Y, Zuo W, Yun P, Wang H, Liu M (2020) Fuseseg: semantic segmentation of urban scenes based on rgb and thermal data fusion. IEEE Trans Autom Sci Eng 18(3):1000–1011
79. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4700–4708
80. Deng F, Feng H, Liang M, Wang H, Yang Y, Gao Y, Chen J, Hu J, Guo X, Lam TL (2021) Feanet: feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In: 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 4467–4473
81. Zhou W, Lin X, Lei J, Yu L, Hwang J-N (2021) Mffenet: multiscale feature fusion and enhancement network for rgb-thermal urban road scene parsing. IEEE Trans Multimed 24:2526–2538

82. Zhou W, Liu J, Lei J, Yu L, Hwang J-N (2021) Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. IEEE Trans Image Process 30:7790–7802
83. Zhang Q, Zhao S, Luo Y, Zhang D, Huang N, Han J (2021) Abmdrnet: adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2633–2642
84. Yi S, Li J, Liu X, Yuan X (2022) Ccaffmnet: dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module. Neurocomputing 482:236–251
85. Wang H, Fan R, Cai P, Liu M (2021) SNE-Roadseg+: rethinking depth-normal translation and deep supervision for freespace detection. In: 2021 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 1140–1145
86. Fan R, Wang H, Cai P, Liu M (2020) SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection. In: Proceedings of the European conference on computer vision (ECCV). Springer, pp 340–356
87. Park S-J, Hong K-S, Lee S (2017) Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 4980–4989
88. Gupta S, Arbelaez P, Malik J (2013) Perceptual organization and recognition of indoor scenes from rgb-d images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 564–571
89. Song S, Lichtenberg SP, Xiao J (2015) Sun rgb-d: a rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 567–576
90. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3213–3223
91. Pinggera P, Ramos S, Gehrig S, Franke U, Rother C, Mester R (2016) Lost and found: detecting small road hazards for self-driving vehicles. In: 2016 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 1099–1106
92. Dai A, Chang AX, Savva M, Halber M, Funkhouser T, Nießner M (2017) Scannet: richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 5828–5839
93. Silberman N, Hoiem D, Kohli P, Fergus R (2012) Indoor segmentation and support inference from rgbd images. In: European conference on computer vision (ECCV). Springer, pp 746–760
94. Armeni I, Sax S, Zamir AR, Savarese S (2017) Joint 2d-3d-semantic data for indoor scene understanding. arXiv:1702.01105
95. Shivakumar SS, Rodrigues N, Zhou A, Miller ID, Kumar V, Taylor CJ (2020) Pst900: Rgb-thermal calibration, dataset and segmentation network. In: 2020 IEEE international conference on robotics and automation (ICRA). IEEE, pp 9441–9447
96. Xu H, Ma J, Le Z, Jiang J, Guo X (2020) Fusiondn: a unified densely connected network for image fusion. In: Proceedings of the thirty-fourth AAAI conference on artificial intelligence
97. Fritsch J, Kuehnl T, Geiger A (2013) A new performance measure and evaluation benchmark for road detection algorithms. In: International conference on intelligent transportation systems (ITSC)
98. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM (2016) The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3234–3243
99. Everingham M, Winn J (2012) The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Anal Stat Model Comput Learn Tech Rep 2007:1–45
100. Mottaghi R, Chen X, Liu X, Cho N-G, Lee S-W, Fidler S, Urtasun R, Yuille A (2014) The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 891–898

101. Chen X, Mottaghi R, Liu X, Fidler S, Urtasun R, Yuille A (2014) Detect what you can: detecting and representing objects using holistic models and body parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1971–1978
102. Gould S, Fulton R, Koller D (2009) Decomposing a scene into geometric and semantically consistent regions. In: 2009 IEEE 12th international conference on computer vision (ICCV). IEEE, pp 1–8
103. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) Labelme: a database and web-based tool for image annotation. Int J Comput Vis 77(1):157–173
104. Criminisi A et al (2004) Microsoft research cambridge object recognition image database. http://research.microsoft.com/vision/cambridge/recognition
105. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
106. Hoiem D, Efros AA, Hebert M (2007) Recovering surface layout from an image. Int J Comput Vis 75(1):151–172
107. Ipeirotis PG (2010) Analyzing the amazon mechanical turk marketplace. XRDS: Crossroads. ACM Mag Stud 17(2):16–21
108. Hariharan B, Arbeláez P, Bourdev L, Maji S, Malik J (2011) Semantic contours from inverse detectors. In: 2011 international conference on computer vision (ICCV). IEEE, pp 991–998
109. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2011) The PASCAL visual object classes challenge 2011 (VOC2011) Results. http://www.pascal-network.org/challenges/VOC/voc2011/workshop/index.html
110. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision (ECCV). Springer, pp 740–755
111. Brostow GJ, Fauqueur J, Cipolla R (2009) Semantic object classes in video: a high-definition ground truth database. Pattern Recognit Lett 30(2):88–97
112. Kuutti S, Bowden R, Jin Y, Barber P, Fallah S (2020) A survey of deep learning applications to autonomous vehicle control. IEEE Trans Intell Transp Syst 22(2):712–733
113. Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3061–3070
114. Alhaija H, Mustikovela S, Mescheder L, Geiger A, Rother C (2018) Augmented reality meets computer vision: efficient data generation for urban driving scenes. Int J Comput Vis
115. Gaidon A, Wang Q, Cabon Y, Vig E (2016) Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4340–4349
116. Zhou B, Zhao H, Puig X, Fidler S, Barriuso A, Torralba A (2017) Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 633–641
117. Zhou B, Zhao H, Puig X, Xiao T, Fidler S, Barriuso A, Torralba A (2019) Semantic understanding of scenes through the ade20k dataset. Int J Comput Vis 127(3):302–321
118. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) Sun database: large-scale scene recognition from abbey to zoo. In: 2010 IEEE computer society conference on computer vision and pattern recognition (CVPR). IEEE, pp 3485–3492
119. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A (2014) Learning deep features for scene recognition using places database. Adv Neural Inf Process Syst 27
120. Neuhold G, Ollmann T, Rota Bulo S, Kontschieder P (2017) The mapillary vistas dataset for semantic understanding of street scenes. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 4990–4999
121. Huang X, Cheng X, Geng Q, Cao B, Zhou D, Wang P, Lin Y, Yang R (2018) The apolloscape dataset for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPR), pp 954–960
122. Kim B, Yim J, Kim J (2020) Highway driving dataset for semantic video segmentation. arXiv:2011.00674

123. Yogamani S, Hughes C, Horgan J, Sistu G, Varley P, O'Dea D, Uricár M, Milz S, Simon M, Amende K et al (2019) Woodscape: a multi-task, multi-camera fisheye dataset for autonomous driving. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 9308–9318
124. Varma G, Subramanian A, Namboodiri A, Chandraker M, Jawahar C (2019) Idd: a dataset for exploring problems of autonomous navigation in unconstrained environments. In: 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1743–1751
125. Geyer J, Kassahun Y, Mahmudi M, Ricou X, Durgesh R, Chung AS, Hauswald L, Pham VH, Mühlegg M, Dorn S et al (2020) A2d2: audi autonomous driving dataset. arXiv:2004.06320
126. Alberti E, Tavera A, Masone C, Caputo B (2020) Idda: a large-scale multi-domain dataset for autonomous driving. IEEE Robot Autom Lett 5(4):5526–5533
127. Xiao J, Owens A, Torralba A (2013) Sun3d: a database of big spaces reconstructed using sfm and object labels. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1625–1632
128. Janoch A, Karayev S, Jia Y, Barron JT, Fritz M, Saenko K, Darrell T (2013) A category-level 3D object dataset: putting the kinect to work. In: Consumer depth cameras for computer vision. Springer, pp 141–165
129. Tao A, Sapra K, Catanzaro B (2020) Hierarchical multi-scale attention for semantic segmentation. arXiv:2005.10821
130. Borse S, Wang Y, Zhang Y, Porikli F (2021) Inverseform: a loss function for structured boundary-aware segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5901–5911
131. Chen L-C, Lopes RG, Cheng B, Collins MD, Cubuk ED, Zoph B, Adam H, Shlens J (2020) Naive-student: leveraging semi-supervised learning in video sequences for urban scene segmentation. In: Computer vision-ECCV, (2020) 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer, pp 695–714
132. Chen Z, Duan Y, Wang W, He J, Lu T, Dai J, Qiao Y (2022) Vision transformer adapter for dense predictions. arXiv:2205.08534
133. Chen L-C, Wang H, Qiao S (2020) Scaling wide residual networks for panoptic segmentation. arXiv:2011.11675
134. Yuan Y, Chen X, Wang J (2020) Object-contextual representations for semantic segmentation. In: Computer vision-ECCV, (2020) 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. Springer, pp 173–190
135. Cheng B, Collins MD, Zhu Y, Liu T, Huang TS, Adam H, Chen L-C (2020) Panoptic-deeplab: a simple, strong, and fast baseline for bottom-up panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12 475–12 485
136. Zhang X, Xu H, Mo H, Tan J, Yang C, Wang L, Ren W (2021) Dcnas: densely connected neural architecture search for semantic image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 13 956–13 967
137. Mohan R, Valada A (2021) Efficientps: efficient panoptic segmentation. Int J Comput Vis (IJCV) 129(5):1551–1579
138. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C (2020) Axial-deeplab: stand-alone axial-attention for panoptic segmentation. In: Computer vision-ECCV, (2020) 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. Springer, pp 108–126
139. Ganeshan A, Vallet A, Kudo Y, Maeda S-I, Kerola T, Ambrus R, Park D, Gaidon A (2021) Warp-refine propagation: semi-supervised auto-labeling via cycle-consistency. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 15 499–15 509
140. Cai Y, Dai L, Wang H, Li Z (2021) Multi-target pan-class intrinsic relevance driven model for improving semantic segmentation in autonomous driving. IEEE Trans Image Process 30:9069–9084
141. Zhu Y, Sapra K, Reda FA, Shih KJ, Newsam S, Tao A, Catanzaro B (2019) Improving semantic segmentation via video propagation and label relaxation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8856–8865

142. Bevandić P, Oršić M, Grubišić I, Šarić J, Šegvić S (2022) Multi-domain semantic segmentation with overlapping labels. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV), pp 2615–2624
143. Lambert J, Liu Z, Sener O, Hays J, Koltun V (2020) Mseg: a composite dataset for multi-domain semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2879–2888
144. Erkent Ö, Laugier C (2020) Semantic segmentation with unsupervised domain adaptation under varying weather conditions for autonomous vehicles. IEEE Robot Autom Lett 5(2):3580–3587
145. Bolte J-A, Kamp M, Breuer A, Homoceanu S, Schlicht P, Huger F, Lipinski D, Fingscheidt T (2019) Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, p 0
146. Yang G, Zhao H, Shi J, Deng Z, Jia J (2018) Segstereo: exploiting semantic information for disparity estimation. In: Proceedings of the European conference on computer vision (ECCV), pp 636–651
147. Klingner M, Termöhlen J-A, Mikolajczyk J, Fingscheidt T (2020) Self-supervised monocular depth estimation: solving the dynamic object problem by semantic guidance. In: European conference on computer vision (ECCV). Springer, pp 582–600
148. Ochs M, Kretz A, Mester R (2019) Sdnet: semantically guided depth estimation network. In: German conference on pattern recognition (GCPR). Springer, pp 288–302
149. Kong S, Fowlkes C (2018) Pixel-wise attentional gating for parsimonious pixel labeling. arXiv:1805.01556
150. Ozgunalp U, Fan R, Ai X, Dahnoun N (2017) Multiple lane detection algorithm based on novel dense vanishing point estimation. IEEE Trans Intell Transp Syst 18(3):621–632
151. Fan R, Wang H, Cai P, Wu J, Bocus MJ, Qiao L, Liu M (2022) Learning collision-free space detection from stereo images: homography matrix brings better data augmentation. IEEE/ASME Trans Mechatron 27(1):225–233
152. Chen Z, Chen Z (2017) Rbnet: a deep neural network for unified road and road boundary detection. In: International conference on neural information processing. Springer, pp 677–687
153. Gu S, Zhang Y, Yang J, Alvarez JM, Kong H (2019) Two-view fusion based convolutional neural network for urban road detection. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 6144–6149
154. Gu S, Zhang Y, Tang J, Yang J, Kong H (2019) Road detection through crf based LiDAR-camera fusion. In: 2019 international conference on robotics and automation (ICRA). IEEE, pp 3832–3838
155. Caltagirone L, Bellone M, Svensson L, Wahde M (2019) LiDAR-camera fusion for road detection using fully convolutional neural networks. Robot Auton Syst 111:125–131
156. Sun J-Y, Kim S-W, Lee S-W, Kim Y-W, Ko S-J (2019) Reverse and boundary attention network for road segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, p 0
157. Fan R, Ai X, Dahnoun N (2018) Road surface 3D reconstruction based on dense subpixel disparity map estimation. IEEE Trans Image Process 27(6):3025–3035
158. Ma N, Fan J, Wang W, Wu J, Jiang Y, Xie L, Fan R (2022) Computer vision for road imaging and pothole detection: a state-of-the-art review of systems and algorithms. Transp Safety Environ 4(4):tdac026
159. Fan R, Liu M (2020) Road damage detection based on unsupervised disparity map segmentation. IEEE Trans Intell Transp Syst 21(11):4906–4911
160. Guo S, Jiang Y, Li J, Zhou D, Su S, Junaid Bocus M, Zhu X, Chen Q, Fan R (2023) Road environment perception for safe and comfortable driving. Springer, submitted for publication
161. Fan R, Ozgunalp U, Hosking B, Liu M, Pitas I (2020) Pothole detection based on disparity transformation and road surface modeling. IEEE Trans Image Process 29:897–908

162. Fan J, Bocus MJ, Hosking B, Wu R, Liu Y, Vityazev S, Fan R (2021) Multi-scale feature fusion: learning better semantic segmentation for road pothole detection. In: 2021 IEEE international conference on autonomous systems (ICAS). IEEE, pp 1–5
163. Fan R, Wang H, Bocus MJ, Liu M (2020) We learn better road pothole detection: from attention aggregation to adversarial domain adaptation. In: Computer vision-ECCV, (2020) Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer, pp 285–300
164. Fan R, Wang H, Wang Y, Liu M, Pitas I (2021) Graph attention layer evolves semantic segmentation for road pothole detection: a benchmark and algorithms. IEEE Trans Image Process 30:8144–8154
165. Yang M, Yu K, Zhang C, Li Z, Yang K (2018) Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 3684–3692
166. Tian Z, He T, Shen C, Yan Y (2019) Decoders matter for semantic segmentation: data-dependent decoding enables flexible feature aggregation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3126–3135
167. Takikawa T, Acuna D, Jampani V, Fidler S (2019) Gated-scnn: gated shape cnns for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 5229–5238
168. Li H, Xiong P, An J, Wang L (2018) Pyramid attention network for semantic segmentation. arXiv:1805.10180
169. Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H (2018) Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 552–568
170. Wang H, Fan R, Sun Y, Liu M (2020) Applying surface normal information in drivable area and road anomaly detection for ground mobile robots. In: 2020 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 2706–2711
171. Wang H, Fan R, Sun Y, Liu M (2022) Dynamic fusion module evolves drivable area and road anomaly detection: a benchmark and algorithms. IEEE Trans Cybern 52(10):10 750–10 760