technologies

MDPI

Review

Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives

Giulia Rizzoli , Francesco Barbato and Pietro Zanuttigh *

Department of Information Engineering, University of Padova, Via Gradenigo 6/A, 35131 Padova, Italy; giulia.rizzoli@dei.unipd.it (G.R.); francesco.barbato@dei.unipd.it (F.B.)
* Correspondence: zanuttigh@dei.unipd.it

Abstract: The perception of the surrounding environment is a key requirement for autonomous driving systems, yet the computation of an accurate semantic representation of the scene starting from RGB information alone is very challenging. In particular, the lack of geometric information and the strong dependence on weather and illumination conditions introduce critical challenges for approaches tackling this task. For this reason, most autonomous cars exploit a variety of sensors, including color, depth or thermal cameras, LiDARs, and RADARs. How to efficiently combine all these sources of information to compute an accurate semantic description of the scene is still an unsolved task, leading to an active research field. In this survey, we start by presenting the most commonly employed acquisition setups and datasets. Then we review several different deep learning architectures for multimodal semantic segmentation. We will discuss the various techniques to combine color, depth, LiDAR, and other modalities of data at different stages of the learning architectures, and we will show how smart fusion strategies allow us to improve performances with respect to the exploitation of a single source of information.

Keywords: semantic segmentation; autonomous driving; multimodal; LiDAR; depth; modality fusion; deep learning

check for updates

Citation: Rizzoli, G.; Barbato, F.; Zanuttigh, P. Multimodal Semantic Segmentation in Autonomous Driving: A Review of Current Approaches and Future Perspectives. *Technologies* **2022**, *10*, 90. https://doi.org/10.3390/technologies10040090

Academic Editors: Manoj Gupta, Eugene Wong and Gwanggil Jeon

Received: 22 June 2022
Accepted: 19 July 2022
Published: 25 July 2022

1. Introduction

In recent years, the autonomous driving field has experienced an impressive development, gaining a huge interest and expanding into many sub-fields that cover all aspects of the self-driving vehicle [1,2]. Examples are vehicle-to-vehicle communications [3], energy-storage devices, sensors [4], safety devices [5], and more. Among them, a fundamental field is scene understanding, a challenging Computer Vision (CV) task which deals with the processing of raw environmental data to construct a representation of the scene in front of the car that allows for the subsequent interaction with the environment (e.g., route planning, safety breaks engagement, packet transmission optimizations, etc.).

Scene understanding is the process of perceiving, analysing, and elaborating on an interpretation of an observed scene through a network of sensors [6]. It involves several complex tasks, from image classification, to more advanced ones like object detection and Semantic Segmentation (SS). The first task deals with the assignment of a global label to an input image; however, it is of limited use in the autonomous driving scenario, given the need for localizing the various elements in the environment [1]. The second task provides a more detailed description, localizing all identified objects and providing classification information for them [7]. The third task is the most challenging one, requiring the assignment of a class to each pixel of an input image. Due to the accurate semantic description this problem provides, it requires complex machine learning architectures and can be identified as the basic goal for a scene understanding pre-processor. It will be the subject of this work.

Most approaches for semantic segmentation were originally developed by using as input a single RGB camera (see Section 1.1 for a brief review of the task). However, the development of self-driving vehicles provided with many onboard sensors requires the generalization toward different modalities. The joint employment of the various data streams coming from the sensors (RGB, LiDAR, RADAR, stereo setups, etc.) allows a much more in-depth understanding of the environment.

The importance of multimodal data for autonomous driving applications came under the spotlight for the first time in the DARPA's Grand Challenge in 2007. All three teams on the podium underlined the necessity of such an approach, especially focusing on LiDAR perception systems.

In later years, LiDARs found many applications in the development of large-scale datasets for the training of deep architectures, e.g., the well-known KITTI [8] benchmark. Although such sensors provide very high accuracy, they come with a couple of major downsides, namely the high cost, the presence of delicate moving parts, and the fact that the depth map produced is sparse, rather than dense as the images from standard cameras. To tackle the first problem, more cost-effective, consumer-grade technologies have been used, such as stereo cameras, matricial Time-of-Flight or structured-light sensors [9,10]. On the other hand, these technologies are less accurate and suffer the effect of sunlight, claiming for approaches accounting for the unreliability of their data.

The investigation of approaches able to leverage multiple heterogeneous datastreams (like those produced by the aforementioned sensors) is the focus of this survey, wherein we investigate the various proposed approaches for multi-modal semantic segmentation in autonomous driving. In particular, we will focus on 2.5D scenes (RGB and depth, including stereo vision setups), 2D + 3D fusion (RGB and LiDAR), and also report some additional, specific setups (e.g., by also using thermal data).

1.1. Semantic Segmentation with Deep Learning

In this section, we will report the main approaches for semantic segmentation from a single data source, overviewing the task history and highlighting the landmarks of its evolution. A graphic example of a possible deployment of the task in autonomous driving scenarios is reported in Figure 1.

Figure 1. The car screen shows an example of semantic segmentation of the scene in front of the car.

Early approaches to semantic segmentation were based on the use of classifiers on small image patches [11–13], until the introduction of deep learning, which has enabled great improvements in this field as well.

The first approach to showcase the deep learning potential on this task is found in [14], which introduced an end-to-end convolutional model, the so-called Fully Convolutional Network (FCN) model, which is made of an encoder (or contraction segment) and a decoder (or expansion segment). The former maps the input into a low-resolution feature representation, which is then upsampled in the expansion block. The encoder (also called backbone) is typically a pretrained image classification network used as a feature extractor. Among these networks, popular choices are VGG [15], ResNet [16], or the more lightweight MobileNet [17].

Other remarkable architectures that followed FCN are ParseNet (Liu et al. [18]), which models global context directly rather than only relying on a larger receptive field, and DeconvNet (Noh et al. [19]) which proposes an architecture that contains overlapping deconvolution and unpooling layers to perform nonlinear upsampling, resulting in improving the performance at the cost of increasing the complexity of the training procedure.

A slightly different approach is proposed in the Feature Pyramid Network (FPN), developed by Lin et al. [20], where a bottom-up pathway, a top-down pathway, and lateral connections are used to join low-resolution and high-resolution features and to better propagate the low-level information into the network. Inspired by the FPN model, Chen et al. [21,22] proposes the DeepLab architecture, which adopts pyramid pooling modules wherein the feature maps are implicitly downsampled through the use of dilated convolutions of different rates. According to the authors, dilated convolutions allow for an exponential increase in the receptive field without a decrease in resolution or increase in parameters, as may happen in the traditional pooling or stride-based approaches. Chen et al. [22] further extended the work by employing depth-wise separable convolutions.

Nowadays the current objective in semantic segmentation consists of improving the multiscale feature learning while making a trade-off between keeping the inference time low and increasing the receptive field/upsampling capability.

One recent strategy is feature merging through attention-based methods. Recently, such techniques gained a lot of traction in Computer Vision, following its success in Natural Language Processing (NLP) tasks. The most famous approach of this class is the transformer architecture [23], introduced by Vaswani et al. in 2017 in an effort to reduce the dependence of NLP architectures on recurrent blocks, which have difficulty in handling long-time relationships between input data. This architecture has been adapted to the image understanding field in the Vision Tranformers (ViT) [24,25] work, which presents a convolution-free, transformer-based vision approach able to surpass previous state-of-the-art techniques in image classification (at the cost of much higher memory and training data requirements). Transformers have been used as well in semantic segmentation in numerous works [26–28].

Although semantic segmentation was originally tackled by RGB data, recently many researchers started investigating its application for LiDAR data [29–34]. The development of such approaches is supported by an ever-increasing number of datasets that provide labeled training samples, e.g., Semantic KITTI [35]. More in detail, PointNet [29,30] was one of the first general-purpose 3D pointcloud segmentation architectures, but although it achieved state-of-the-art results on indoor scenes, the sparse nature of LiDAR data led to a significant performance decrease in outdoor settings, limiting its applicability in autonomous driving scenarios. An evolution of this technique is developed in RandLANet [31], where an additional grid-based downsampling step is added as preprocessing, together with a feature aggregation based on random-centered KD-trees, to better handle the sparse nature of LiDAR samples. Other approaches are SqueezeSeg [33] and RangeNet [36], wherein the segmentation is performed through a CNN architecture. In particular, the LiDAR data is converted to a spherical coordinate representation allowing one to exploit 2D semantic segmentation techniques developed for images. The most recent and better-performing architecture is Cylinder3D [34], which exploits the prior knowledge of LiDAR topologies—in particular their cylindrical aspect—to better represent the data fed into the architecture. The underlying idea is that the density of points in each voxel is inversely dependent on the distance from the sensor; therefore the architecture samples the data according to a cylindrical grid, rather than a cuboid one, leading to a more uniform point density.

Given the recent growth in the availability of heterogeneous data, the exploitation of deep multimodal methods attracted great research interest (in Section 4, a detailed overview is reported). RGB data carries a wealth of visual and textual information, which in many cases has successfully been used to enable semantic segmentation. Nevertheless, depth measurements provide useful geometric cues, which help significantly in the discrimination of visual ambiguities, e.g., to distinguish between two objects with a similar appearance. Moreover, RGB cameras are sensitive to light and weather conditions which can lead to failures in outdoor environments [37]. Thermal cameras give temperature-based characteristics of the objects, which can better enhance the recognition of some objects, thereby improving the resilience of semantic scene understanding in challenging lighting conditions [38].

1.2. Outline

In this paper, we focus on analyzing and discussing deep learning based fusion methods in multimodal semantic segmentation. The survey is organized as follows: Section 2 describes the most common sensors and their arrangements in autonomous driving setups; in Section 3 the main datasets for this application are listed, pointing out their features with particular attention to data diversity; finally, Section 4 reports several methods to address data fusion. As a conclusion, in Section 5 the open challenges and future outlooks are remarked upon.

2. Multimodal Data Acquisition and Preprocessing

One of the key aspects of an autonomous driving system is the choice of the acquisition devices and the infrastructure which allows them to exchange information among themselves and to the central perception system. Over the years many setups have been proposed, introducing different cameras, LiDARs, RADAR sensors, GPS systems, and IMU units. In this section, we will report an overview of the most commonly employed sensors, their placement, and the post-processing steps needed to convert the provided data into a machine-friendly format. In Figure 2 we report an example of sensor setup. The vehicle shown was used during the generation of [39]. In the work, the authors remark how it was chosen to be close to real autonomous vehicles (such as TESLA https://www.tesla.com/autopilot (accessed on 21 July 2022), Waymo https://waymo.com/ (accessed on 21 July 2022), and Argo https://www.argoverse.org/ (accessed on 21 July 2022), ...).

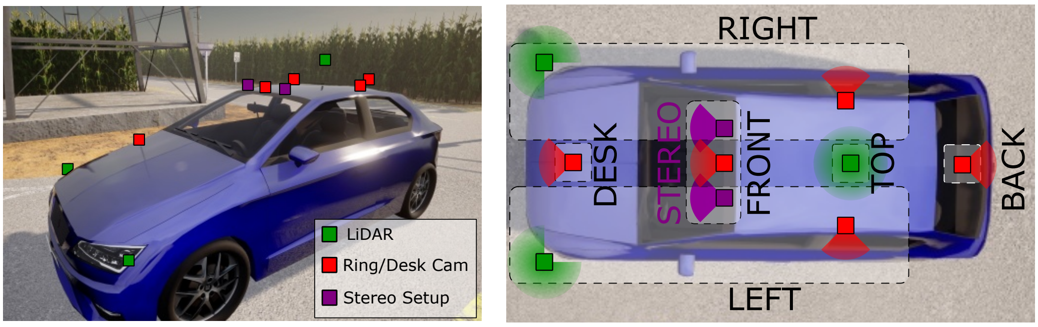

Figure 2. Figure (derived from the one in [39]) showcasing the multi-sensor setup used in the data collection.

2.1. RGB Cameras

Standard color cameras are employed in almost all setups (as underlined by the datasets reported in Section 3). Due to their limited cost, many systems rely on the combination of multiple cameras looking in different directions, both to improve the scene understanding and to allow a 360° Field-of-View (which may be helpful in the identification of obstacles/dangers coming from directions different than the heading one, but incurs additional processing costs related to the stitching and understanding of the bigger scene). Even if standard cameras provide an extremely useful representation of the scene, the data they provide suffers from some key limitations. First of all, they do not provide distance information, making it impossible to access precise information about the positions and sizes of the objects. Secondly, they are strongly affected by the illumination and weather conditions. Dark environments, direct sunlight, rain, or fog can strongly reduce the usefulness of the data provided by these devices [37]. This suggests that the combination of color cameras with other devices is a goal worth investigating, particularly with sensors resilient to the weaknesses of the cameras themselves.

2.2. Thermal Cameras

A thermal (or thermographic) camera is a special type of camera, which rather than acquiring information from the visible light spectrum (380∼750 nm) captures information in the near-infrared range (1∼14 µm) [38]. These wavelengths have the particular property of being the vector of irradiation heat, allowing to capture the heat sources in the scene (e.g., the heat produced by the vehicles).

This implies that they are able to work even in dark (in the usual sense) conditions because each object can be considered as a light source. Due to this property, these cameras can be very useful in night-time autonomous driving scenarios. A thermal camera output, in general, has two forms: the raw heatmap of the scenes (computed from the wavelength emitted by each object in the scene) or a color-coded post-processing. The second format is usually more meaningful than the first because the encoding uses special perceptive functions to map differences in temperature to differences in color [40].

2.3. Depth Cameras

Another approach to solving the problems affecting color cameras is to use depth sensors. As in the case of thermal cameras, the idea is to change the captured quantity from visible light to something more resilient to illumination/environmental changes. In the case of depth cameras, the acquired quantity is the distance-from-the-camera information for each pixel. Depth information cannot be directly inferred from a single standard image, and this has led to the development of multiple, complementary, active and passive techniques to acquire the depth information, e.g., stereo setups [9], matricial Time-of-Flight [10], RADARs [41], and LiDARs [42]. The last three actually belong to the same macro-class of techniques, which is ToF and differ in the way the time delay is computed (directly or indirectly) and on the medium used to extract the information (radio waves or light). In Table 1 we summarize in a qualitative manner the various sensors, classifying them depending on:

- the resilience to environmental conditions;
- the working range;
- the sparsity of the output depthmap; and
- the cost.

Table 1. Qualitative comparison between depth sensors. More details reported at [10,41,43].

Sensor	Range	Sparsity	Robustness	Direct Sun Perf.	Night Perf.	Cost
Passive Stereo	Far	Dense	Low	Medium	Low	Very Low
Active Stereo	Medium	Dense	Medium	Medium	Good	Low
Matricial ToF	Medium	Dense	High	Low	Good	Medium
LiDAR	Far	Sparse	High	Good	Good	High
RADAR	Far	Very Sparse	Medium	Good	Good	Low

2.3.1. Stereo Camera

Passive Stereo camera systems are one of the most common and cost-effective approaches for depth estimation. They employ two or more color cameras positioned at a known distance with respect to each other (commonly referred to as "baseline") to reconstruct a dense depthmap of the scene. The estimation procedure follows two main steps. The first is pixel matching between the two images (i.e., pixels representing the same location in the scene are found and coupled with each other). The second is the actual depth computation, wherein the distance between coupled pixels (disparity) is converted into the depthmap applying the well-known relation $d = {}^{bf}/_{p}$ pixel-wise, where d is the distance, b is the baseline, f is the camera focal length and p is the disparity. Clearly, the challenging part in the depth computation lies in the first step, the stereo matching, and many efficient algorithms were proposed to tackle the problem (from traditional computer vision algorithms like SGM [44] to recent deep learning-based strategies [45–47]).

In a similar fashion as for thermal data, alternative encodings for depthmaps exist. One example is HHA [48], which encodes in the three channels the horizontal disparity, the height above ground, and the angle that the pixel's local surface normal makes with the inferred gravity direction.

Active Stereo camera systems aid the stereo matching by adding a light projector to the stereo setup. This allows one to artificially increase the texture contrast, reducing the number of wrongly matched pixels. These systems, however, suffer in strong sunlight conditions, because the sunlight can overshadow or add noise to the projected light, thus strongly limiting the performance of the approach, that can instead be quite useful at night or in low light conditions.

2.3.2. Time-of-Flight

A matricial Time-of-Flight camera is a device able to calculate the distance between each scene point and the device [10]. This is done by measuring the round-trip time of the light traveling from the light transmitter, which illuminates the target to the photo-detector. ToF sensors are categorized into indirect (iToF) and direct (dToF) sensors. In iToF the distance is measured by calculating the shift in phase of the original emitted light signal, which is continuously modulated, and the received light signal. iToF sensors have demonstrated good spatial resolution with a greater ability to detect multiple objects over a wide (but still limited by the camera optics) field of view (FoV) [49]. However, such sensors come with a significant drawback, that being that their light source modulation frequency is directly proportional to the maximum range, but inversely proportional to the precision attainable, thereby constraining them to a short range of typically less than 30 m. This limitation makes them less suited for autonomous driving applications. In dToF, the depth information is collected by measuring the time the light pulse takes to hit the target and return to the sensor, which requires the pulsing laser and the camera acquisition to be synchronized. dToF are typically employed also in LiDARs due to their longer range and reliability.

2.3.3. LiDAR

A LiDAR is a long-range, omnidirectional depth sensor, which comes with high robustness in geometry acquisition at the expense of a higher cost [39]. It employs one or multiple focused laser beams whose ToF is measured to generate a 3D representation of the environment in the form of a point cloud. Generally speaking, a point cloud consists of the 3D location and the intensity of the incident light collected at every frame. LiDARs have different operating principles [50]. In the scanning type, a collimated laser beam illuminates a single point at a time, and the beam is raster-scanned to illuminate the field of view point-by-point. In the flash type, a wide diverging laser beam illuminates the whole field of view in a single pulse. In the latter approach, the acquired frames do not need to be patched together, and the device is not sensitive to platform motion, which allows for more precise imaging. Motion can produce “jitter” in scanning LiDAR due to the delay in time as the laser rasters over the area.

Due to the sparsity and uneven distribution of point clouds, LiDAR-only perception tasks are challenging [50]. Whereas images are dense tensors, 3D point clouds can be represented in a variety of ways, resulting in several families of preprocessing algorithms. Besides directly representing the 3D coordinates of the acquired points, projection methods are the most intuitive approaches to having a direct correspondence with RGB images. Common choices for multi-modal applications consist of:

- spherical projection;
- perspective projection; and
- bird’s-eye view.

In the first case, each 3D point is projected onto a sphere by using azimuth and zenith angles to create a spherical map. The result is a dense representation; however, it can differ in terms of size from the camera image. This does not happen in perspective projection where the 3D points are projected into the camera coordinate system; hence the depthmap has the same size. The main drawback of this method is that it leaves many pixels empty, and upsampling techniques are required to reconstruct the image. The latter approach, as the name suggests, directly provides the objects’ positions on the ground plane. Although it preserves the objects’ length and width, it loses height information and, as a result, some physical characteristics.

Point-based approaches utilize a raw pointcloud as input and provide point-by-point labeling as output. These algorithms can handle any unstructured pointcloud. As a direct consequence, the key challenge in processing raw pointclouds is extracting local contextual information. Several approaches were used to create an ordered feature sequence from unordered 3D LiDAR data, which was subsequently translated to 3D LiDAR data by by using convolutional deep networks [51].

- **Voxel-based** : convert 3D LiDAR data to voxels in order to represent structured data. These algorithms typically accept voxels as input and predict one semantic label for each voxel [32,34].
- **Graph-based**: create a graph by using 3D LiDAR data. A vertex generally represents a single point or a set of points, whereas edges indicate vertexes’ adjacency connections [52,53].
- **Point Convolution**: establish a similarity between points e.g., by sorting the K-nearest points according to their spatial distance from the centers [29–31].
- **Lattice Convolution**: provide a transformation between pointclouds and sparse permutohedral lattices so that convolutions can be performed efficiently [54,55].

Despite their high cost and moving components (in spindle-type lidars, whereas other technologies like solid-state lidars do not have this issue), LiDARs are being used as part of the vision systems of several high-level autonomous vehicles.

2.3.4. RADAR

RADAR (Radio Detection and Ranging) sensors can also give distance information; however, depth information coupled with RGB data is rarely produced by them. RADARs send out radio waves to be reflected by an obstacle, measure the signal runtime, and use the Doppler effect to estimate the object’s radial motion. They can withstand a variety of lighting and weather situations; however, due to their low resolution, semantic understanding with RADARs is difficult. Their application in driving is usually restricted to directional proximity sensors, usually to aid in cruise control or assistive parking. Nevertheless, some works [56,57] propose strategies that allow their use in semantic segmentation setups. An interesting approach to automatic RADAR samples labelling is presented in [58], wherein the authors exploit both image- and LiDAR-labeled samples to infer the correct RADAR-point classification.

2.4. Position and Navigation Systems

Many devices allow the absolute position and orientation of the vehicle to be established. Global Positioning System (GPS) receivers and Inertial Measurement Unit (IMU) are common examples of such devices. Global Navigation Satellite Systems (GNSS) were first utilized in cars as navigation tools in driver assistance features [59], but they are now also used in conjunction with HD Maps for autonomous vehicle path planning and autonomous vehicle self-localization. Internal vehicle information (i.e., "proprioceptive sensor") is provided by IMUs and odometers. IMUs measure the acceleration and rotational rates of cars and are currently employed in autonomous driving for accurate localization. These sensors can be leveraged to aid camera segmentation architectures in the creation of lane-level HD Maps [60]. On the other hand, it is possible to improve coarse GPS measurements through camera-vision systems [61].

3. Datasets

One of the biggest challenges involved in the use of deep learning-based architectures is the need for large amounts of labeled data, fundamental for their optimization [62]. This is reflected in a very active and diverse field [63,64] that deals with the generation (in case of synthetic datasets) or collection (in case of real-world datasets) and subsequent labeling of data suitable for training deep learning models. A fundamental task for autonomous driving that suffers greatly from the data availability problem is semantic segmentation. In this task, the action of producing a label coincides with assigning to each pixel in an image (or to each point in a pointcloud) a semantic class. The complexity of this task is the main reason for the huge time and cost involved in the collection of datasets for semantic segmentation. In Table 2 a high-level summary is reported for each of the datasets used in the methods described in Section 4 differentiating them by the type of scene content (e.g., indoor or outdoor).

In the following, we will focus on semantic segmentation datasets, with special attention to the current problems and challenges of the available datasets. For a comprehensive list of general datasets for autonomous driving applications one may refer to [63], and to [64] for RGB-D tasks. Very few large-scale (more than 25k labeled samples) semantic segmentation datasets are available for autonomous driving settings, and even fewer take care of the multimodal aspect of the sensors present in vehicles.

In Section 3, we will go over the most commonly used driving datasets that support this task, reporting their characteristics and classifying them according to the following criteria in Table 2:

- modalities provided (i.e., type of available sensors);
- tasks supported (i.e., provided labeling information);
- data variability offered (i.e., daytime, weather, season, location, etc.); and
- acquisition domain (i.e., real or synthetic).

For the dataset description, we will follow the order reported in Table 2, which summarizes the discussed datasets. Some of the dataset names were compressed into acronyms, the expanded name can be found at the end of the document in the abbreviations listing.

Table 2. Comparison between multi-modal datasets. Shorthand notation used: *Type* Real/Synthetic; *Cameras* Grayscale/Color/FishEye/Thermal/Polarization/Event/MultiSpectral/Depth; *Daytime* Morning/Day/Sunset/Night; *Location* City/Indoor/Outdoor/Region/Traffic (left/right-handed), † indicates that the cities/regions considered belong to the same state, ǂ indicates that single views of the 3D scene are labeled, ∗ indicates variability with no control or categorization. The table is color-coded to indicate the scenarios present in each dataset: □ Driving, ░ Exterior, ▒ In/Out, ▓ Interior.

Metadata				Sensors						Diversity					Labels			Size	
Name	Created	Update	Type	Cameras	LiDARs	Stereo	GT Depths	RADARs	IMU	Daytime	Seasons	Location	Weather	Env. Control	Sem Seg	Bboxes	Opt. Flow	Sequences	Labeled Sampl.
KITTI [8,65–67]	2012	2015	R	2G/2C	1	2	-	-	+	-	-	-	-	-	-	+	-	1(6h)	200
Cityscapes [68]	2016	2016	R	2C	-	1	-	-	+	-	-	27C †	-	-	+	-	-	-	5000
Lost and found [69]	2016	2016	R	2C	-	1	-	-	-	-	-	-	-	-	+	-	-	112	2104
Synthia [70–72]	2016	2019	S	1C	-	-	1	-	-	DS	+	-	2	-	+	-	-	-	9400
Virtual KITTI [73,74]	2016	2020	S	2C	-	1	1	-	+	MDS	-	-	4	+	+	+	+	35	17k
MSSSD/MF [75]	2017	2017	R	1C/1T	-	-	-	-	-	DN	-	-	-	-	+	-	-	-	1569
RoadScene-Seg [76]	2018	2018	R	1C/1T	-	-	-	-	-	DN	-	-	-	-	-	-	-	-	221
AtUlm [77]	2019	2019	R	1G	4	-	-	-	-	-	-	-	-	-	+	-	-	-	1446
nuScenes [78]	2019	2020	R	6C	1	1	-	5	+	-	-	T	-	-	+	+	-	-	40k
SemanticKITTI [35]	2019	2021	R	-	1	-	-	-	-	-	-	-	-	-	+	-	-	22	43,552
ZJU [79]	2019	2019	R	2C/1FE/1P	-	1	-	-	-	DN	-	-	-	-	-	-	-	-	3400
A2D2 [80]	2020	2020	R	6C	5	-	-	-	+	-	-	-	-	-	+	+	-	-	41,280
ApolloScape [81]	2020	2020	R	6C	2	1	-	-	+	*	-	4R †	*	-	+	+	-	-	140k
DDAD [82]	2020	2020	R	6C	4	-	-	-	-	-	-	2R	-	-	-	-	-	-	16,600
KITTI 360 [83]	2021	2021	R	2C/2FE	1	1	-	-	+	-	-	-	-	-	+	+	-	-	78k
WoodScape [84]	2021	2021	R	4FE	1	-	-	-	+	-	-	10C	-	-	+	+	-	-	10k
EventScape [85]	2021	2021	S	1C/1E	-	-	1	-	+	-	-	4C	-	-	+	+	-	743(2 h)	-
SELMA [39]	2022	2022	S	8C	3	3	7	-	-	DSN	-	7C	9	+	+	+	-	-	31k×27
Freiburg Forest [86]	2016	2016	R	2C/1MS	-	1	-	-	-	-	-	-	-	-	+	-	-	-	336
POLABOT [87]	2019	2019	R	2C/1P/1MS	-	1	-	-	-	-	-	-	-	-	+	-	-	-	175
SRM [88]	2021	2021	R	1C/1T	-	-	-	-	-	-	-	-	2	-	+	-	-	-	2458
SSW [88]	2021	2021	R	1C/1T	-	-	-	-	-	-	-	-	2	-	+	-	-	-	1571
MVSEC [89]	2018	2018	R	2G/2E	1	1	-	-	+	DN	-	IO	-	-	-	-	-	14(1h)	-
PST900 [90]	2019	2019	R	2C/1T	-	1	-	-	-	-	-	IO	-	-	+	-	-	-	4316
NYU-depth-v2 [91]	2012	2012	R	1C + 1D	-	-	1	-	-	-	-	-	-	-	+	-	-	-	1449 ǂ
SUN-RGBD [92]	2015	2015	R	1C + 1D	-	-	1	-	-	-	-	-	-	-	+	-	-	-	10k ǂ
2D-3D-S [93]	2017	2017	R	1C + 1D	-	-	1	-	-	-	-	-	-	-	+	-	-	-	270
ScanNet [94]	2017	2018	R	1C + 1D	-	-	1	-	-	-	-	-	-	-	+	-	-	-	1513
Taskonomy [95]	2018	2018	R	1C + 1D	-	-	1	-	-	-	-	-	-	-	∼	-	-	-	4 m ǂ

Summary

KITTI [8,65–67] was the first large-scale dataset to tackle the important issue of multimodal data in autonomous vehicles. The KITTI vision benchmark was introduced in 2012 and contains a real-world 6-h-long sequence recorded using a LiDAR, an IMU, and two stereo setups (with one grayscale and one color camera each). Although the complete suite is very extensive (especially for depth estimation and object detection), the authors did not focus much on the semantic labeling process, opting to label only 200 training samples for semantic (and instance) segmentation and for optical flow.

Cityscapes [68] became one of the most common semantic segmentation datasets for autonomous driving benchmarks. It is a real-world dataset containing 5000 finely labeled, high-definition (2048 × 1024) images captured in multiple German cities. Additionally, the authors provide 25,000 coarsely labeled samples—polygons rather than object borders, with many unlabeled areas (see Figure 3)—to improve deep architectures' performance through data variability. The data was captured with a calibrated and rectified stereo setup in high-visibility conditions, allowing the authors to provide high-quality estimated depthmaps for each of the 30,000 samples. Given its popularity in semantic segmentation settings, this dataset is also one of the most used for monocular depth estimation or 2.5D segmentation tasks.

Lost and Found [69] is an interesting road-scene dataset that tackles lost cargo scenarios, it includes pixel-level segmentation of the road and of the extraneous objects present on the surface. It was introduced in 2016 and includes around 2000 samples. The dataset comprises 112 stereo video sequences with 2104 annotated frames in a real-world scenario.

Synthia [70–72] is one of the oldest multimodal synthetic datasets providing labeled semantic segmentation samples. First introduced in 2016, it provides color, depth, and semantic information generated from the homonym simulator. The authors tackled data diversity by simulating the four seasons, by rendering the dataset samples from multiple PoVs, not only from road-view, but also from building height, and by considering day/night times. The dataset provides multiple versions, but only one supports (partially) the Cityscapes dataset label-set; it contains 9400 total samples.

Virtual KITTI [73,74] is an extension of the KITTI dataset. It is a synthetic dataset produced in Unity (https://unity.com/ (accessed on 21 July 2022)) which contains scenes modeled after the ones present in the original KITTI dataset. The synthetic nature of the dataset allowed the authors to produce a much greater number of labeled samples than those present in KITTI, while also maintaining a higher precision (due to the automatic labeling process). Unfortunately, the dataset does not provide labels for the LiDAR pointclouds.

MSSSD/MF [75] is a real-world dataset and one of the few that provides multispectral (thermal + color) information. It is of relatively small size, with only 1.5k samples, recorded in day and night scenes. Regardless, it represents an important benchmark for real-world applications, because thermal cameras are one of the few dense sensors resilient to low-visibility conditions such as fog or rain for which consumer-grade options exist.

RoadScene-Seg [76] is real-world dataset that provides 200 unlabeled road-scene images captured with an aligned color + infrared setup. Given the absence of labels, the only validation metric supported for architectures in this dataset is a qualitative evaluation by humans.

AtUlm [77] is a non-publicly available real-world dataset developed by Ulm University in 2019. It has been acquired with a grayscale camera and 4 LiDARs. In total the dataset contains 1446 finely annotated samples (grayscale images).

nuScenes [78] is a real-world dataset and one of the very few providing RADAR information. It is the standard for architectures aiming to use such sensor modality. The number of sensors provided is very impressive, as the dataset contains samples recorded from six top ring-cameras (two of which form a stereo setup), one top-central LiDAR, five ring RADARs placed at headlight level, and an IMU. The labeled samples are keyframes extracted with a frequency of 2 Hz from the recorded sequences, totaling 40k samples. The environmental variability lies in the recording location. The cities of Boston and Singapore were chosen as they offer different traffic handedness (Boston right-handed, Singapore left-handed).

Semantic KITTI [35] is an extension to the KITTI dataset. Here the authors took on the challenge of labeling in a point-wise manner all the LiDAR sequences recorded in the original set. It has rapidly become one of the most common benchmarks for LiDAR semantic segmentation, especially thanks to the significant number of samples made available.

ZJU [79] is a real-world dataset and the only among the one listed supporting the light polarization modality. It was introduced in 2019 and features 3400 labeled samples provided with color, (stereo) depth, light polarization, and an additional fish-eye camera view to cover the whole scene.

A2D2 [80] is another real-world dataset which focuses highly on the multimodal aspect of the data provided. It was recorded by a research team from the AUDI car manufacturer and provides five ring LiDARs, six ring cameras (two of which form a stereo setup) and an IMU. The semantic segmentation labels refer to both 2D images and LiDAR pointclouds, for a total of 41k samples. The daytime variability is very limited, offering only high-visibility day samples. The weather variability is slightly better, as it was changing throughout the recorded sequences, but no control over the conditions is offered.

ApolloScape [81] is a large-scale real-world dataset that supports a multitude of different tasks (semantic segmentation, lane segmentation, trajectory estimation, depth estimation, and more). As usual, we focus on the semantic segmentation task, for which ApolloScape provides ∼150k labeled RGB images. Together with the color samples, the dataset also provides depth information. Unfortunately the depth maps contain only static objects, and all information about vehicles or other road occupants is missing. This precludes the possibility of directly exploiting the dataset in multimodal settings because a deployed agent wouldn't have access to such static maps.

DDAD [82] is a real-world dataset developed by the Toyota Research Institute, whose main focus is on monocular depth estimation. The sensors provided include six ring cameras and four ring LiDARs. The data was recorded in seven cities across two states: San Francisco, the Bay Area, Cambridge, Detroit, and Ann Arbor in the USA, and Tokyo and Odaiba in Japan. The dataset provides semantic segmentation labels only for the validation and test (non-public) sets, significantly restricting its use-case.

KITTI 360 [83] is a real-world dataset first released in 2020, which provides many different modalities (Stereo Color, LiDAR, Spherical, and IMU) and labeled segmentation samples for them. The labeling is performed in the 3D space, and the 2D labels are extracted by re-projection. In total, the dataset contains 78K labeled samples. Like KITTI, the dataset is organized in temporal sequences, recorded from a synchronized sensor setup mounted on a vehicle. As such, it offers very limited environmental variability.

WoodScape [84] is another real-world dataset providing color and LiDAR information. As opposed to its competitors, its 2D information is extracted only by using fish-eye cameras. In particular, the dataset provides information coming from four fish-eye ring cameras and a single top-LiDAR (360° coverage), recorded from more than ten cities in five different states. In total, the dataset contains 10k 2D semantic segmentation samples.

EventScape [85] is a very recent (2021) synthetic dataset developed by using the CARLA simulator [96], providing color, (ground truth) depth, event camera, semantic segmentation, bounding boxes, and IMU information for 743 sequences for a total of 2 h of video across four cities.

SELMA [39] is a very recent (2022) synthetic dataset developed in a modified CARLA simulator [96] whose goal is to provide multimodal data in a multitude of environmental conditions, while also allowing a researcher to control such conditions. It is heavily focused on semantic segmentation, providing labels for all of the sensors offered (seven co-placed RGB/depth cameras, and three LiDARs). The environmental variability takes the form of three daytimes (day, sunset, night), nine weather conditions (clear, cloudy, wet road, wet road and cloudy, soft/mid-level/heavy rain, mid-level/heavy fog), and 8 synthetic towns. The dataset contains 31k unique scenes recorded in all 27 environmental conditions, resulting in 800k samples for each sensor.

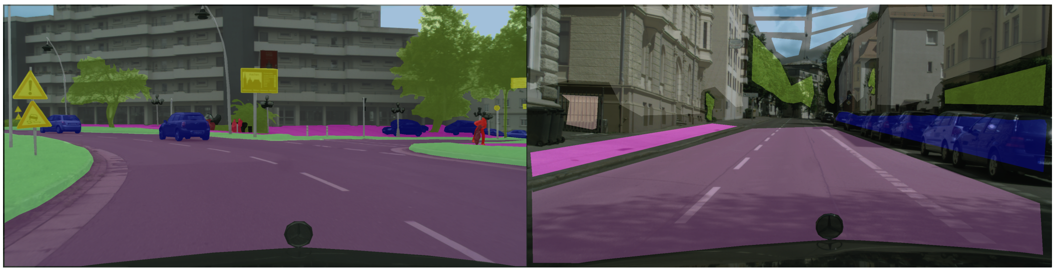

Figure 3. Example of finely (**left**) and coarsely (**right**) labeled Cityscapes [68] samples.

4. Multimodal Segmentation Techniques in Autonomous Driving

This section is the core of this work, wherein we present a detailed review of recent and well-performing approaches for multi-modal semantic segmentation.
We will start with a brief overview of the field and of the most common design choices, before moving to an in-depth description of the works, starting with RGB and depth data fusion in Section 4.1 (the most common choice). Then, we will discuss approaches combining RGB with LiDAR data in Section 4.2. Finally, approaches exploiting less conventional data sources (e.g., RADAR, event or thermal cameras) will be discussed in Section 4.4. Table 3 shows a summarized version of the methods discussed in the following sections, comparing them according to

- modalities used for the fusion;
- datasets used for training and validation;
- approach to feature fusion (e.g., sum, concatenation, attention, etc.); and
- fusion network location (e.g., encoder, decoder, specific modality branch, etc.).

On the other hand, in Table 4, we report the numerical score (mIoU) attained by the methods in three benchmark datasets, respectively: Cityscapes [68] for 2.5D SS in Table 4a, KITTI [8] for 2D + 3D SS in Table 4b and MSSSD/MF [75] for RGB + Thermal SS in Table 4c.

Table 3. Summary of recent multimodal semantic segmentation architectures. Modality shorthand: Dm, raw depth map; Dh, depth HHA; De, depth estimated internally; E, event camera; T, thermal; Lp, light polarization; Li, LiDAR; Ls, LiDAR spherical; F, optical flow. Location: D, decoder; E, encoder. Direction: D, decoder; C, color; B, bi-directional; M, other modality.

Metadata				Fusion Approach						Fusion Architecture				
Name	**Year**	**Dataset(s)**	**Modality(ies)**	+	×	⊙	**Ad-Hoc Block**	**Ad-Hoc Loss**	**Multi-Task**	**Location**	**Direction**	**Parallel Branches**	**Skip Connections**	**Multi-Level Fusion**
LWM [97]	2021	[68,91,92]	DmDe	+	-	+	-	+	+	D	D/C	2	+	+
SSMA [98]	2019	[68,70,86,92,94]	DmDhT	-	+	+	+	+	-	E	D	2	+	+
CMX [99]	2022	[68,75,79,85,91–94]	EDhLpT	+	+	-	+	-	-	E	D/B	2	+	+
AsymFusion [100]	2021	[68,91,95]	Dm	+	-	-	+	-	-	E	B	2	-	+
SA-Gate [101]	2020	[68,91]	Dh	+	-	+	+	-	-	E	B	2	+	+
ESANet [102]	2021	[68,91,92]	Dm	+	-	-	-	-	-	E	C	2	+	+
DA-Gate [103]	2018	[68,91–93]	DmDe	-	-	-	-	+	-	N/A	N/A	1	-	-
RFBNet [104]	2019	[68,94]	Dh	+	+	+	+	-	-	E	B	2	-	+
MMSFB-snow [88]	2021	[68,70,88]	DmT	-	-	+	+	-	-	E	D	2	+	+
AdapNet [105]	2017	[68,70,86]	DmT	+	+	-	+	-	-	D	D	2	-	-
RFNet [106]	2020	[68,69]	Dm	+	-	-	+	-	-	E	C	2	+	+
RSSAWC [77]	2019	[68,77]	DmLi	+	-	+	-	-	-	E	D	2	-	-
PMF [107]	2021	[35,78]	Li	+	+	+	+	+	-	E	M	2	+	+
MDASS [108]	2019	[68,73]	DmF	+	-	-	-	-	-	E	D	2/3	+	+
CMFnet [109]	2021	[68,87]	DmLp	-	+	+	-	-	-	E	D/B	3+	-	+
CCAFFMNet [110]	2021	[75,76]	T	-	-	+	+	-	-	E	C	2	+	+
DooDLeNet [111]	2022	[75]	T	-	+	+	-	-	-	E	D	2	+	+
GMNet [112]	2021	[75,90]	T	+	+	-	+	-	+	E	D	2	+	+
FEANet [113]	2021	[75]	T	+	+	-	-	-	-	E	C	2	-	+
EGFNet [114]	2021	[75,90]	T	+	+	+	+	-	-	E	D	2	-	+
ABMDRNet [115]	2021	[75]	T	+	+	+	+	+	+	E	D	2	-	+
AFNet [116]	2021	[75]	T	+	+	-	+	-	-	E	D	2	-	-
FuseSeg-Thermal [117]	2021	[75]	T	+	-	+	-	-	-	E	C	2	+	+
RTFNet [106]	2019	[75]	T	+	-	-	-	-	-	E	C	2	-	+
FuseSeg-LiDAR [118]	2020	[8]	LsLi	-	-	+	-	-	-	E	M	2	+	+
RaLF3D [119]	2019	[8]	LsLi	+	-	+	-	-	-	E	D	2	+	+
DACNN [120]	2018	[91–93]	DmDh	+	-	-	-	-	-	E	D	2	-	-
xMUDA [121]	2020	[35,78,80]	Li	-	-	+	-	+	+	D	D	2	-	+

Table 4. Architectures Performance Comparison.

Name	Backbone	mIoU
(a) Cityscapes dataset (2.5D SS).		
LWM [97]	ResNet101 [16]	83.4
SSMA [98]	ResNet50 [16]	83.29
CMX [99]	MiT-B4 [27]	82.6
AsymFusion [100]	Xception65 [122]	82.1
SA-Gate [101]	ResNet101 [16]	81.7
ESANet [102]	ResNet34 [16]	80.09
DA-Gate [103]	ResNet101 [16]	75.3
RFBNet [104]	ResNet50 [16]	74.8
MMSFB-snow [88]	ResNet50 [16]	73.8
AdapNet [105]	AdapNet [105]	71.72
RFNet [106]	ResNet18 [16]	69.37
RSSAWC [77]	ICNet [123]	65.09
MDASS [108]	VGG16 [15]	63.13
CMFnet [109]	VGG16 [15]	58.97
(b) KITTI dataset (2D + 3D SS).		
PMF [107]	ResNet34 [16]	63.9
FuseSeg-LiDAR [118]	SqueezeNet [124]	52.1
RaLF3D [119]	SqueezeSeg [33]	37.8
xMUDA [121]	SparseConvNet3D [125] ResNet34 [16]	49.1
(c) MSSSD/MF dataset (RGB + Thermal SS).		
CMX [99]	MiT-B4 [27]	59.7
CCAFFMNet [110]	ResNeXt50 [126]	58.2
DooDLeNet [111]	ResNet101 [16]	57.3
GMNet [112]	ResNet50 [16]	57.3
FEANet [113]	ResNet101 [16]	55.3
EGFNet [114]	ResNet152 [16]	54.8
ABMDRNet [115]	ResNet50 [16]	54.8
AFNet [116]	ResNet50 [16]	54.6
FuseSeg-Thermal [117]	DenseNet161 [127]	54.5
RTFNet [106]	ResNet152 [16]	53.2

Early attempts of multimodal semantic segmentation approaches combine RGB data and other modalities into multi-channel representations that were then fed into classical semantic segmentation networks based on the encoder–decoder framework [128,129]. This simple early fusion combination strategy is not too effective because it struggles to capture the different type of information carried by the different modalities (e.g., RGB images contain color and texture, whereas the other modalities typically better represent the spatial relations among objects). Within this reasoning, feature-level and late-fusion approaches have been developed. Fusion strategies have typically been categorized in early, feature and late-fusion strategies, depending on the fact that the fusion happens at the input level, in some intermediate stage or at the end of the understanding process. However, most recent approaches try to get the best of the three modalities by performing multiple fusion operations at different stages of the deep network [98,115,118].

A very common architectural choice is to adopt a multi-stream architecture for the encoder with a network branch processing each modality (e.g., a two-stream architecture for RGB and depth) and additional network modules connecting the different branches that combine modality-specific features into fused ones and/or carry information across the branches [98,99,101]. This hierarchical fusion strategy leverages multilevel features via progressive feature merging and generate a refined feature map. It entails fusing features at various levels rather than at early or late stages.

The feature fusion can take place through simple operations e.g., concatenation, element-wise addition, multiplication, etc., or a mixture of these, which is typically addressed as a fusion block, attention, or gate module. In this fashion, multi-level features can be fed from one modality to another, e.g., in [102] where depth cues are fed to the RGB branch, or mutually between modalities. The fused content can either reach the next layer or the decoder directly through skip connections [98].

The segmentation map is typically computed by a decoder taking in input the fused features and/or the output of some of the branches. Multiple decoders can also be used but it is a less common choice [121]. We also remark that both symmetrical approaches (by using the same architecture for all modalities) and asymmetrical ones (setting a main modality from which the output is computed and by using the others as side information) have been proposed. Finally, the loss function can be just the cross-entropy, or any other loss for semantic segmentation on the output maps. Furthermore multi-task strategies employing different losses on the estimate of some of the modalities from others have also been proposed as further described in the following sub-sections [97,103].

4.1. Semantic Segmentation from RGB and Depth Data

Wang et al. [100] claim that typical methods relying on fusing the multimodal features into one branch in a hierarchical manner are still lacking rich feature interactions. They design a bidirectional fusion scheme (AsymFusion) wherein they maintain the two branches with shared weights and promote the propagation of informative features at later fusion layers by making use of an asymmetric fusion block (see Figure 4). In their architecture, the encoders of the two modalities are sharing convolutional parameters (except for the batch normalization layers which are modality-specific) and at each layer a mutual fusion is performed introducing two operations: channel shuffle and pixel shift. The authors hold that features fused by symmetrical fusion methods at both branches tend to learn similar representations, therefore asymmetric operations might be significant. To avoid bringing redundant information at both the encoder branches, channel shuffle fuses two features by exchanging features corresponding to a portion of channels, whereas pixel shift constantly shifts one pixel on a feature map introducing zero padding.

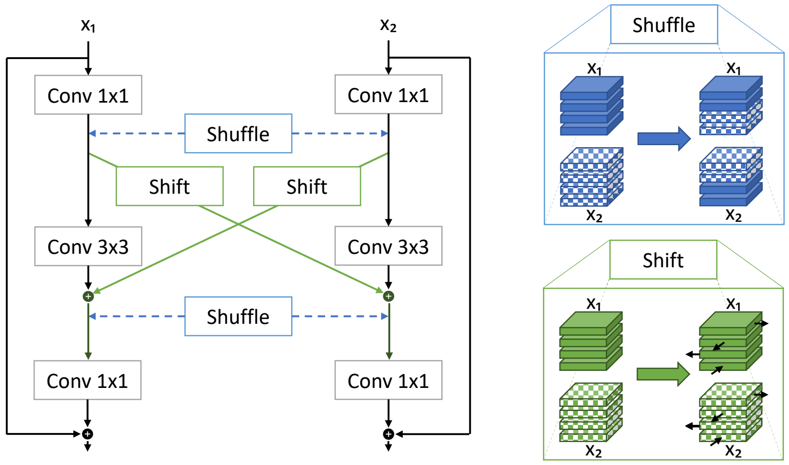

Figure 4. Asymmetric fusion block of [100].

Chen et al. [101] propose a unified and efficient cross-modality guided encoder whose architecture is depicted in Figure 5. It not only effectively re-calibrates RGB feature responses, but also takes into account the noise of the depth and accurately distills its information via multiple stages, alternately aggregating the two re-calibrated representations. The separation-and-aggregation gate (SA-Gate) is designed with two operations to ensure informative feature propagation between modalities. Formerly, feature re-calibration is performed for each individual modality. It is then followed by feature aggregation across modality boundaries. The operations are classified as feature separation and feature combination. The first consists of a global average pooling along the channel-wise dimensions of two modalities, which is followed by concatenation and a MLP operation to obtain an attention vector. This operation finds its motivation in filtering out exceptional depth activations that may overshadow confident RGB responses, reducing the probability of misleading information propagation. The same principle is implemented as a re-calibration step in a symmetric and bi-directional manner. Feature combination generates spatial-wise gates for both modalities to control information flow of each modality feature map with a soft attention mechanism. At each layer, the normalized output of the SA-Gate is added to each modality branch; thus the refined result will be passed on to the encoder's next layer, resulting in more precise and efficient encoding of the two modalities.

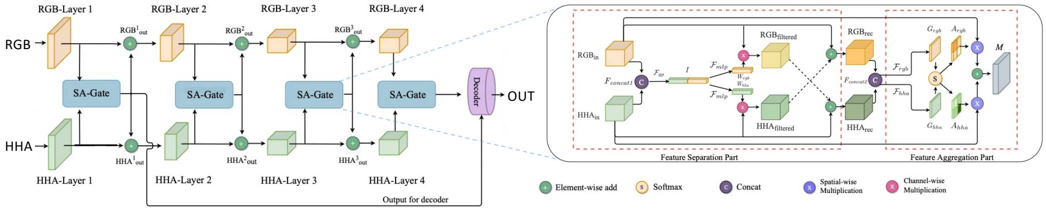

Figure 5. Figure from [101] showing its cross-modality feature propagation scheme. Adapted with authors' permission from [101]. Copyright 2020, Springer Nature Switzerland AG.

Valada et al. [98] present a multimodal fusion framework that incorporates an attention mechanism for effectively correlating multimodal features at mid- and high levels, and for better object boundary refinement (see Figure 6). Each modality is individually fed into a computationally efficient unimodal semantic segmentation architecture, AdapNet++ [105], that includes a strong encoder with skip refinement phases, as well as an efficient atrous spatial pyramid module and a decoder with multiscale residual units. By using the proposed Self-Supervised Model Adaptation (SSMA) block, the encoder uses a late fusion approach to join feature maps from modality-specific streams. In the SSMA block, the features are concatenated and re-weighted through a bottleneck which is used for dimensionality reduction and to improve the representational capacity.

Vachmanus et al. [88] adapt the SSMA architecture with the addition of another parallel bottleneck, with the aim of better capturing the temperature feature in snowy environments. To this end, they introduced two thermal datasets, SRM and SSW (see Table 2), while still testing their network on depth data.

A similar approach is presented in the work by **Zhang et al. [109]**, wherein the modalities are mixed together in a central branch through cross-attention mechanisms. Differently from SSMA, the weighting is performed in each branch separately and the features mixed correspond to the re-weighted outputs. Moreover, the final prediction is performed exploiting a statistics-aware module, able to extract more meaningful information from the concatenated multi-resolution features.

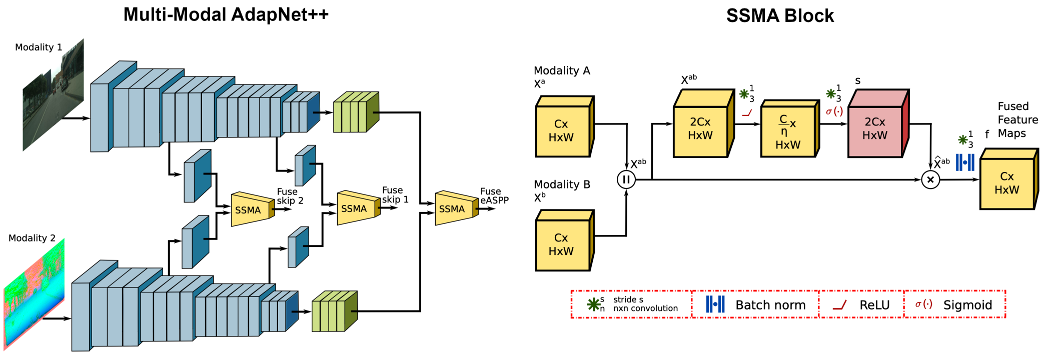

Figure 6. Figure from [98] that explains the work's multimodal semantic segmentation scheme. Reprinted with permission from the authors of [98]. Copyright 2019, Springer Nature Switzerland AG.

Deng et al. [104] adapt the SSMA model and propose an interactive fusion structure to compute the inter-dependencies between the two modality-specific streams and to propagate them through the network. Their residual fusion block (RFB) is composed of two residual units and a gating function unit which adaptively aggregates the features and generates complementary ones. These are fed to the residual units as well as the next layer. In this way, the gating unit exploit the complementary relationship in a soft-attention manner (see Figure 7).

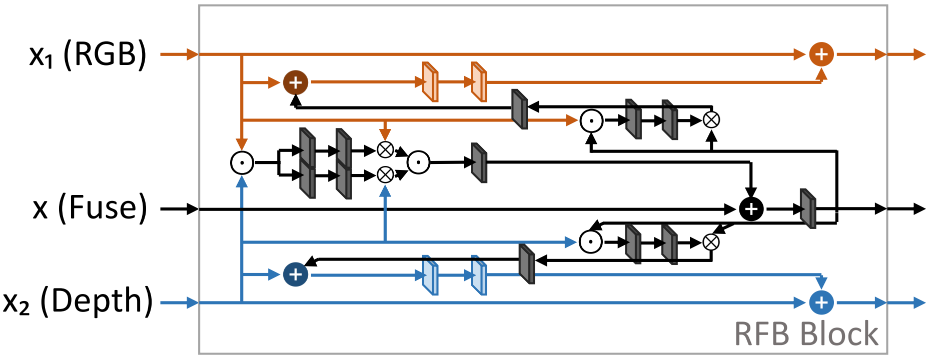

Figure 7. Architecture of the modified version of SSMA proposed by [104].

Seichter et al.'s [102] contribution, although mainly intended for indoor scenes, achieves good segmentation performance in outdoor settings as well. They target an efficient segmentation for embedded hardware, rather than by using high-end GPUs, meaning that their two branches encoder (depicted in Figure 8) is optimized to enable much faster inference than a single deep unimodal encoder. The depth encoder provides geometric information to the RGB one at several stages by using an attention mechanism. The latter aims for understanding which modality to focus on and which to suppress. It consists in an addition between the features reweighted through a squeeze-and-excitation (SA) module [130].

A similar approach is presented by Sun et al. [106], wherein the SA blocks and concatenation are used to merge the features into the RGB branch at multiple levels.

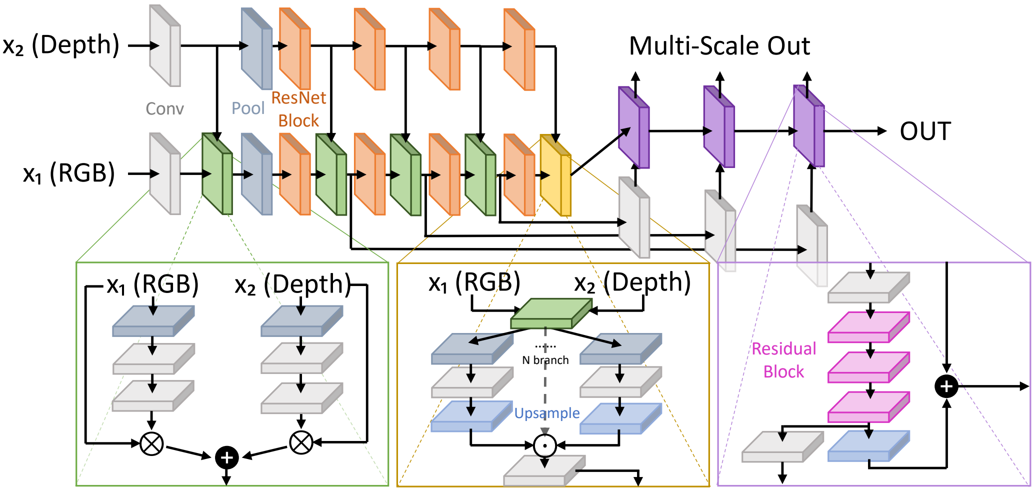

Figure 8. Two branches encoder architecture proposed in [102].

Kong et al. [103], differently from the common multi-scale approaches, exploit the benefit of processing the input image at a single fixed scale, but performing pooling at multiple convolutional dilate rates. Semantic segmentation is carried out by combining a CNN, used as a feature extractor, and a recurrent convolutional neural network, that includes a depth-aware gate. The gating module selects the size over which the features must be pooled, following the idea that larger depth values should have a smaller pooling field to precisely segment small objects. The module works with either estimated depth ("raw" measurements) or directly from monocular cues. A graphic representation of the fusion architecture may be found in Figure 9.

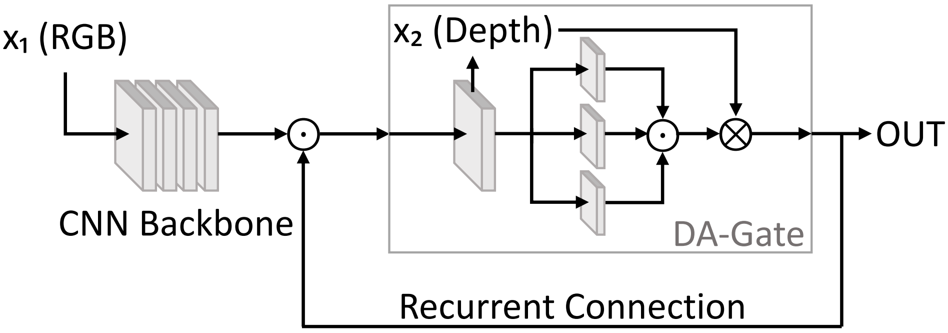

Figure 9. Fusion architecture proposed in [103].

Gu et al. [97] take a similar approach in the self-estimation of depth, noting how such information is not always available in real case scenarios. Therefore in their network (LWM) they establish a depth-privileged paradigm in which depth is provided only during the training process (Figure 10). They pay special attention to hard pixels, which are defined as pixels with a high probability of being misclassified. For this reason, they employ at different multi-scale outputs a loss weight module whose aim is to generate a loss weight map by additively fusing two metrics: depth prediction error and depth-aware segmentation error. The latter have the objective of measuring the "hardness" of a pixel. In the first case, for example, when the depth of two adjacent objects with a considerable distance gap is mispredicted, the delineation of the depth boundary between them may fail, resulting in the segmentation error. In the other, a local region of similar depth becomes a hard region when the categories of distinct subregions are confused due to similar visual appearance. Their network is based on a multi-task learning framework, which has one shared encoder branch and two distinct decoder branches for the segmentation and depth prediction branches. The final output, as well as four side outputs of the segmentation decoder branch, are feeded to the loss weight module.

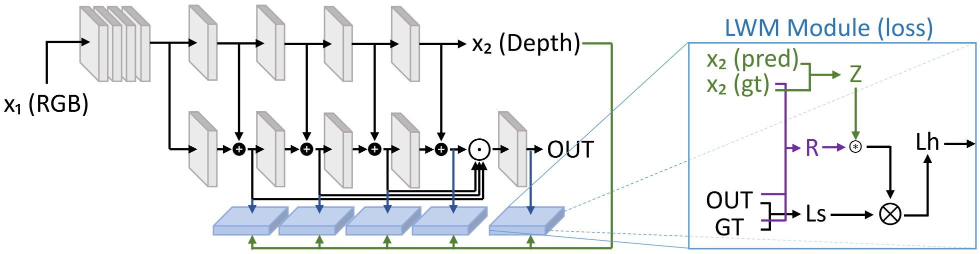

Figure 10. Architecture of [97], exploiting the LWM module.

Rashed et al. [108] focus on sensor fusion for an autonomous driving scenario wherein the dense depth map and the optical flow are considered. They establish a mid-fusion network (MDASS) that performs feature extraction for each modality separately and combines the modality cues at feature-level by using skip connections. In their experiments, they try to fuse at different stages by using a combination of two or three modalities. In addition, they analyzed the effect of using the ground truth measurement or a monocular depth estimate. A graphic representation of the architecture is available in Figure 11.

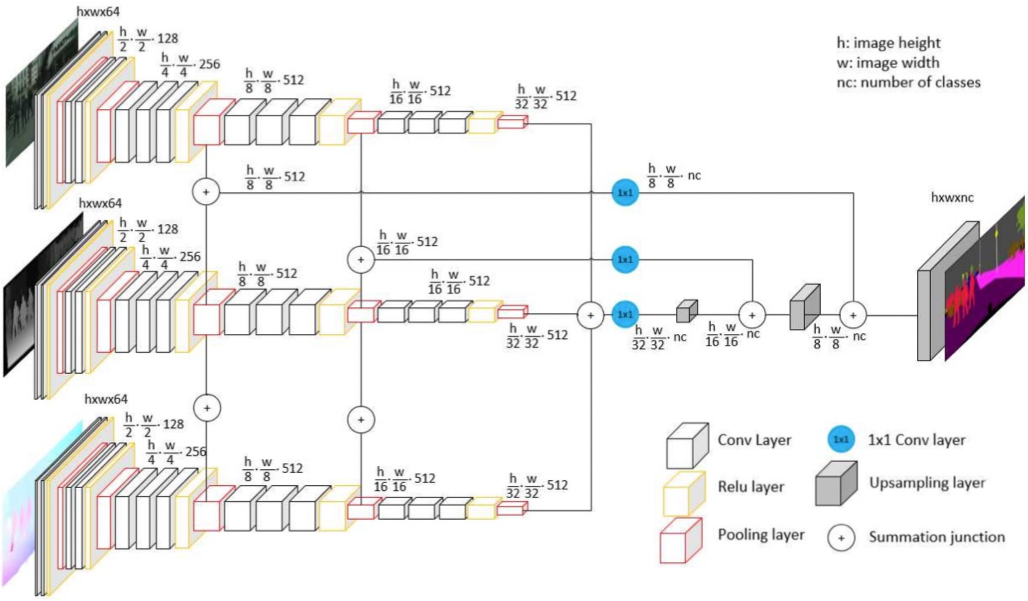

Figure 11. Architecture of [108] where the parallel, multimodal architecture is reported. Reprinted with permission from [108]. Copyright 2019, IEEE.

Liu et al. [99] propose an architecture (CMX, Figure 12) whose fundamental goal is to achieve enough flexibility to generalize across various multi-modal combinations (their approach is not limited to the fusion of RGB and depth data). They do so by exploiting a two-stream network (RGB and X-modality) with two ad-hoc modules for feature interaction and fusion: the cross-modal feature rectification module leverages the spatial and channel correlations to filter noise and calibrate the modalities, and the fusion module merges the rectified features by using a cross-attention mechanism. The latter finds its motivation behind the success of vision transformers and it is modeled into two stages. In the first stage, a cross-modal global reasoning is performed via a symmetric dual-path structure, and in the second stage a mixed channel embedding is applied to produce enhanced output features. The authors achieved remarkable results not just in fusing depth with RGB color, but also in fusing thermal data with color information.

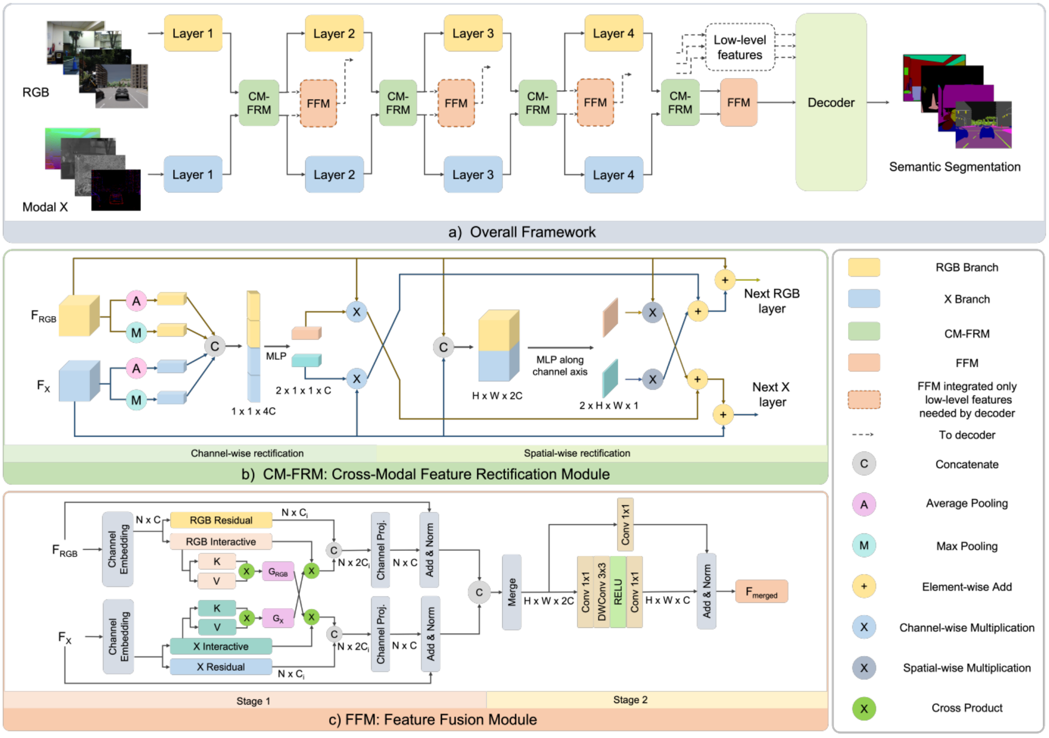

Figure 12. Figure from [99] where the CMX architecture and its modules are shown. Reprinted with permission from the authors of [99]. Copyright 2022, H. Liu.

4.2. Semantic Segmentation from RGB and LiDAR Data

LiDAR acquisitions offer an accurate spatial representation of the physical world. However, the pointclouds from these sensors are relatively sparse and lack color information, which results in a significant classification error in fine-grained segmentation [42]. Due of the sparsity and irregular structure of LiDAR data, the combination with standard camera data for multimodal sensor fusion remains a challenging problem. A possible workaround is to obtain a dense pointcloud by merging multiple LiDAR sensors as in the work by Pfeuffer et al. [77] (unfortunately the employed dataset is not public). However, most of the existing approaches use a projection of the original pointcloud over the color frame and try to find an alignment that can be exploited for the fusion between the cross-modality features. Pointcloud data processing has been tackled in Section 2, whereas the main fusion strategies for LiDAR data are now described.

Zhuang et al. [107] present an approach (PMF) whereby RGB data and LiDAR's projected data (using a perspective projection model) are fed to a two-stream architecture with residual-based fusion modules toward the LiDAR branch (see Figure 13). The modules are designed to learn the complementary features of color and LiDAR data (i.e., the appearance information from color data and the spatial information from pointclouds). The output of the network are two distinct semantic predictions that are used for the optimization through several losses. Among them, a perception-aware loss, based on the predictions and on the perceptual confidence, is introduced to be able to measure the difference between the two modalities. A similar approach is proposed by Madawi et al. [119], wherein RGB images and LiDAR data are converted to a polar-grid mapping representation to be fed into an hybrid early and mid-level fusion architecture. The first is achieved by establishing a mapping between the LiDAR scan points and the RGB pixels. The network is composed of two branches. The first uses the LiDAR measurements, whereas in the second the RGB images are concatenated with the depth and intensity map from LiDAR. The features from the two streams are then fused additively at different levels of the upsampling by using skip connections.

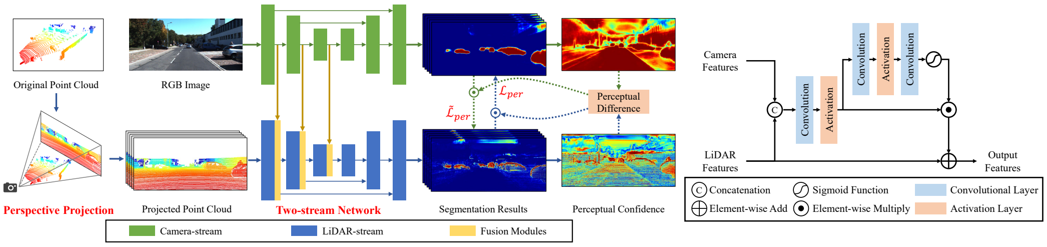

Figure 13. Figure from [107] showing the perception-aware multi-sensor fusion (PMF) architecture and fusion module. Reprinted with permission of the authors from [107]. Copyright 2021, IEEE.

Krispel et al. [118], in the architecture we refer to as FuseSeg-LiDAR, adopt a multi-layer concatenation of the features from the color information in a network for LiDAR data segmentation as depicted in Figure 14. The LiDAR data is spherically projected; hence alignment is required to enable a RGBD representation. Each RGB feature is the bilinear interpolation from the pixels adjacent to a non-discrete position computed as the alignment to the LiDAR range image by using a first-order polyharmonic spline interpolation.

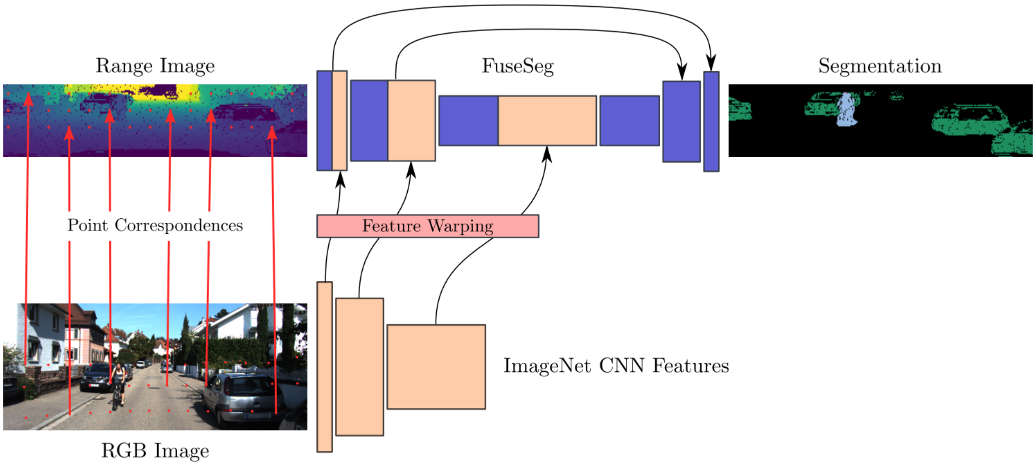

Figure 14. Figure from [118] that explains the FuseSeg-LiDAR architecture. Reprinted with permission from the authors of [118]. Copyright 2020, IEEE.

4.3. Pointcloud Semantic Segmentation from RGB and LiDAR Data

An alternative to the computation of a semantic map in the image space is to produce a semantically labeled pointcloud of the surrounding environment [51]. This approach is particularly well suited for LiDAR data, which typically have this structure.

Early works following this strategy aimed at 3D classification problems, where 3D representations were obtained by applying CNNs to 2D rendering pictures and combining multi-view features [131]. Then the attention moved to 3D semantic segmentation for indoor scenarios. Cheng et al. [132] proposed a method in which they back-project 2D image features into 3D coordinates. Then the network learns both 2D textural appearance and 3D structural features in a unified framework. The work of Jaritz et al. [133] instead aggregates 2D multi-view image features into 3D pointclouds, and then uses a point-based network to fuse the features in 3D canonical space to predict 3D semantic labels.

Jaritz et al. [121] provide a complex pipeline (xMUDA, see Figure 15) that can exchange 2D and 3D information to achieve an unsupervised domain adaptation for 3D semantic segmentation, leveraging the fact that LiDAR is robust to day-to-night domain shifts, and RGB camera images are deeply impacted by it. The architecture consists of a 2D and 3D network inspired by the U-Net model [134] that produces a feature vector of length equal to the number of points in the pointcloud. To obtain such a representation for the RGB image, the 3D points are projected to sample the 2D features at the corresponding pixel location. Each vector is fed to two classifiers to produce the segmentation prediction of the modality and the complementary one, obtaining four distinct segmentation outputs. With the aim of establishing a link between the 2D and 3D, they introduce a "mimicry" loss between the output probabilities. Each modality should be able to predict the output of the other. The final prediction is computed on the concatenated feature vectors of the two modalities.

Similarly to the previous approach, **Liu et al. [135]** adopt a 2D and 3D network called AUDA. Nevertheless, they believe that instead of sampling sparse 2D points in the source domain, the domain adaptation may benefit from using the entire 2D picture. The semantic prediction for the RGB image is achieved directly in this manner, and the calculated loss is used as supervision for the 3D prediction. They also offer an adaptive threshold-moving post-processing phase for boosting the recall rate for uncommon classes, as well as a cost-sensitive loss function to mitigate class imbalance.

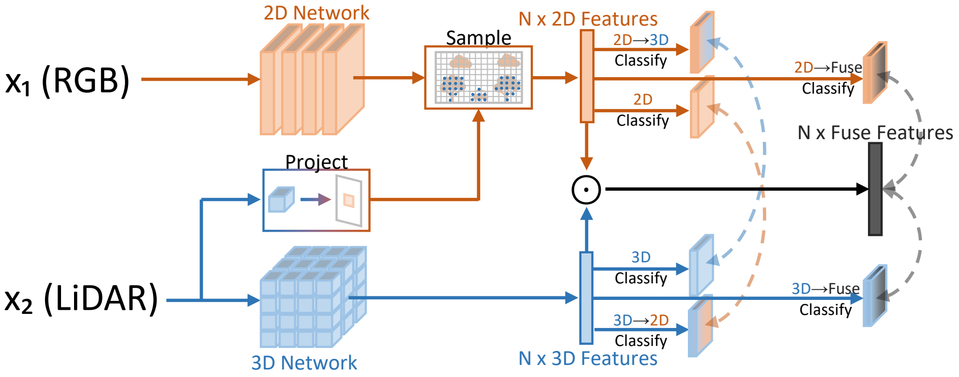

Figure 15. xMUDA [121] 2D/3D architecture.

4.4. Semantic Segmentation from Other Modalities

Even if color and 3D data are the two key sources of information for semantic understanding, other imaging techniques have also been exploited in combination with them. Some recent works combine color and 3D data with thermal imaging, radar acquisitions, and other sources of information.

Zhang et al. [115] employs a bi-directional image-to-image translation to reduce modality differences between RGB and thermal features (ABMDRNet, depicted in Figure 16). The RGB image is first fed to a feature extractor, then is upsampled and fed to a translation network, which is an encoder–decoder architecture, to obtain the corresponding thermal image. The same is done for the thermal image. The difference between the real and the pseudoimages is used as supervision to another decoder which takes as input the cross-modality features at multiple layers and fuses them. In their fusion strategy, the complementary information is exploited by re-weighting the importance of the single-modality features in a channel-dependent way, rather than in a spatial position-dependent way. Additionally, two modules are designed to exploit the multi-scale contextual information of the fused features.

Deng et al. [113] also addresses the fusion of RGB and thermal images by designing an encoder with a two-stream architecture, wherein each convolutional layer is followed by an attention module to re-weight the features. The idea is to enhance the difference between modalities, given that an object at night may be invisible in RGB maps but clearly visible in thermal maps. The information from the thermal branch is additively fused at each layer in the RGB one.

In **Zhou et al.'s GMNet [112]** the multi-layer RGB and thermal features are integrated by using two different fusion modules accounting the fact that deep-layer features provide richer contextual information. For the latter case, they design a densely connected structure to transmit global contextual inception data and a residual module to preserve original information. As opposed to other similar strategies, their decoder has multiple streams wherein different level features are joined. The semantic prediction is decoupled in the foreground, background, and boundary maps which all contribute to the optimization of the model.

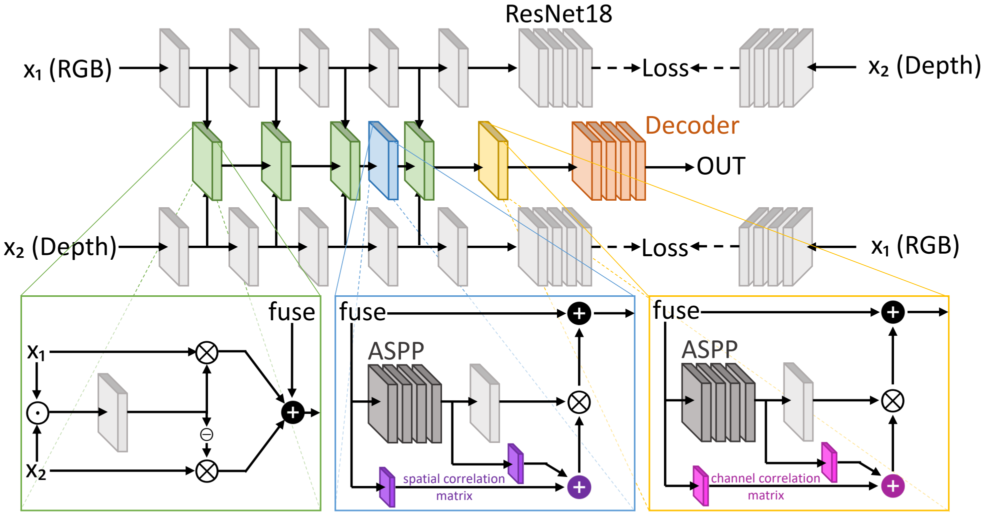

Figure 16. Architecture of the approach of Zhang et al. [115].

Sun et al. [106] propose RTFNet, whereby the encoder and the decoder are asymmetrically designed. The features are extracted through a large encoder for each modality whereas the upsampling is made by a small decoder. The modalities are combined into the RGB branch at multiple levels of the encoder.

Sun et al. [117] propose a two-branch architecture, FuseSeg-Thermal (Figure 17), in which the thermal feature maps are hierarchically added to the RGB feature maps in the RGB encoder in the first step of a two-stage fusion. The fused feature maps, except for the bottom one, are then fused again in the second stage with the matching feature maps in the decoder by tensor concatenation, which is inspired by the U-Net design [134].

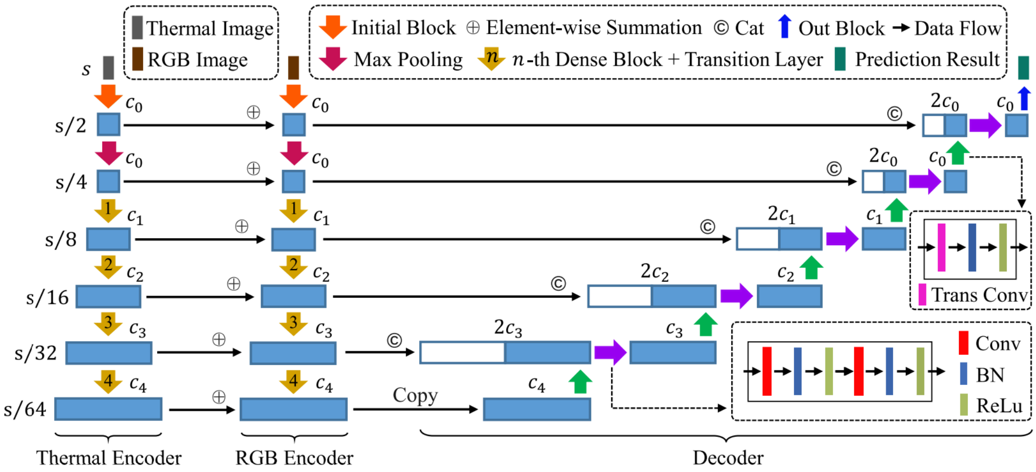

Figure 17. Figure from [117] showcasing the U-net-like architecture presented in the work. Reprinted with permission from the authors [117]. Copyright 2021, IEEE.

Another similar approach, which exploits the coarse-to-fine U-Net architecture, is the one presented by **Yi et al. [110]**, wherein thermal and color modalities are mixed through weights computed from multi-level attention blocks.

Similarly to previous approaches, in **Xu et al. [116]** a fusion module is used on the features extracted from a two-stream encoder to feed a single decoder. The modalities are scaled via cosine similarity, obtaining a channel-wise normalized product, and then the attention map is multiplied with the features that are then summed.

5. Conclusions and Outlooks

In this work, we overviewed the current approaches for multimodal road-scenes segmentation, with particular attention to the imaging modalities and datasets used. Several different approaches have been discussed and compared, showing how the combination of multiple inputs allows for improving the performance with respect to each modality when used alone. Even if there is a variety of different solutions, it is possible to notice a quite common design strategy based on having one network branch for each modality and some additional modules moving the information across them or merging the extracted features.

During our investigation, we were able to recognize some important issues that may be worth tackling by the research community. First of all, as is common when employing deep learning, data availability (and in particular labeled samples for supervised training) is a big bottleneck. This is particularly critical for semantic segmentation wherein labeling is extremely costly and the task itself is notably data-hungry. Therefore many—real and synthetic—datasets are required for optimization. Many of them have been introduced, but they are still far from being able to represent all the situations that can appear in a real-world driving scenario. In particular, the shortage is more critical for thermal data, where no "standard" large-scale dataset is currently available, precluding thorough training and evaluation, and leaving open the question of whether the availability of more data could make the exploitation of these sensors more effective (both alone or combined with standard cameras). On the other hand, a field where data is abundant but that is still mostly unexplored (due to the significant modality difference) is RGB+LiDAR fusion, especially when exploiting the LiDAR samples as raw pointclouds and not after projection. In fact, working in a fully three-dimensional environment can bring some additional understanding capabilities with respect to the 2D projection given by images. Also, the fusion of radar data with other approaches is still quite unexplored.

For the time being, there is no indication that one fusion scheme is preferable to the others. The search for an optimal fusion architecture is often driven by empirical results. In turn, current metrics compare the networks' accuracy on the semantic prediction directly rather than considering multi-modal resilience. The formulation of a metric for assessing multi-modal network robustness could help future improvements.

Author Contributions: Conceptualization, G.R., F.B. and P.Z.; investigation, G.R. and F.B.; resources, G.R., F.B. and P.Z.; writing—original draft preparation, G.R. and F.B.; writing—review and editing, G.R., F.B. and P.Z.; visualization, G.R. and F.B.; supervision, P.Z.; project administration, P.Z.; funding acquisition, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by the SID project "Semantic Segmentation in the Wild".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in the works analyzed in this survey can be found in Section 3 "Datasets", together with a reference to their presentation paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

A2D2	Audi Autonomous Driving Dataset.
CV	Computer Vision.
DARPA	Defense Advanced Research Projects Agency.
DDAD	Dense Depth for Autonomous Driving.
dToF	Direct Time-of-Flight.
FCN	Fully Convolutional Network.
FoV	Field-of-View.
FPN	Feature Pyramid Network.
GNSS	Global Navigation Satellite Systems.
GPS	Global Positioning System.
IMU	Inertial Measurement Unit.
iToF	Indirect Time-of-Flight.
LiDAR	Light Detection and Ranging.
mIoU	mean Intersection over Union.
MLP	Multi-Layer Perceptron.
MSSSD	Multi-Spectral Semantic Segmentation Dataset.
MVSEC	MultiVehicle Stereo Event Camera.
NLP	Natural Language Processing.
PoV	Point of View.
RADAR	Radio Detection and Ranging.
SELMA	SEmantic Large-scale Multimodal Acquisitions.
SGM	Semi-Global Matching.
SRM	Snow Removal Machine.
SS	Semantic Segmentation.
SSMA	Self-Supervised Model Adaptation.
SSW	Snowy SideWalk.
ToF	Time-of-Flight.
VGG	Visual Geometry Group.
ViT	Vision Tranformers.

References

1. Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access* **2020**, *8*, 58443–58469. [CrossRef]
2. Liu, L.; Lu, S.; Zhong, R.; Wu, B.; Yao, Y.; Zhang, Q.; Shi, W. Computing Systems for Autonomous Driving: State of the Art and Challenges. *IEEE Internet Things J.* **2021**, *8*, 6469–6486. [CrossRef]
3. Wang, J.; Liu, J.; Kato, N. Networking and Communications in Autonomous Driving: A Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 1243–1274. [CrossRef]
4. Broggi, A.; Buzzoni, M.; Debattisti, S.; Grisleri, P.; Laghi, M.C.; Medici, P.; Versari, P. Extensive Tests of Autonomous Driving Technologies. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 1403–1415. [CrossRef]
5. Okuda, R.; Kajiwara, Y.; Terashima, K. A survey of technical trend of ADAS and autonomous driving. In Proceedings of the Technical Papers of 2014 International Symposium on VLSI Design, Automation and Test, Hsinchu, Taiwan, 28–30 April 2014; pp. 1–4. [CrossRef]
6. Bremond, F. Scene Understanding: Perception, Multi-Sensor Fusion, Spatio-Temporal Reasoning and Activity Recognition. Ph.D. Thesis, Université Nice Sophia Antipolis, Nice, France, 2007.
7. Gu, Y.; Wang, Y.; Li, Y. A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection. *Appl. Sci.* **2019**, *9*, 2110. [CrossRef]
8. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012.
9. Fan, R.; Wang, L.; Bocus, M.J.; Pitas, I. Computer stereo vision for autonomous driving. *arXiv* **2020**, arXiv:2012.03194.
10. Zanuttigh, P.; Marin, G.; Dal Mutto, C.; Dominio, F.; Minto, L.; Cortelazzo, G.M. *Time-of-Flight and Structured Light Depth Cameras*; Springer International Publishing: Cham, Switzerland, 2016.
11. Brostow, G.J.; Shotton, J.; Fauqueur, J.; Cipolla, R. Segmentation and recognition using structure from motion point clouds. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 44–57.
12. Sturgess, P.; Alahari, K.; Ladicky, L.; Torr, P.H. Combining appearance and structure from motion features for road scene understanding. In Proceedings of the BMVC-British Machine Vision Conference, London, UK, 7–10 September 2009.

13. Zhang, C.; Wang, L.; Yang, R. Semantic segmentation of urban scenes using dense depth maps. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 708–721.
14. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
15. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 770–778.
17. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
18. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
19. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 1520–1528.
20. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
21. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
22. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
25. Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houlsby, N.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
26. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
27. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
28. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 7262–7272.
29. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
30. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
31. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. Randla-net: Efficient semantic segmentation of large-scale point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11108–11117.
32. Tchapmi, L.; Choy, C.; Armeni, I.; Gwak, J.; Savarese, S. SEGCloud: Semantic Segmentation of 3D Point Clouds. In Proceedings of the 2017 International Conference on 3D Vision (3DV), Qingdao, China, 10–12 October 2017; pp. 537–547. [CrossRef]
33. Wu, B.; Wan, A.; Yue, X.; Keutzer, K. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 1887–1893.
34. Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; Lin, D. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9939–9948.
35. Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; Gall, J. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9297–9307.
36. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. Rangenet++: Fast and accurate lidar semantic segmentation. In Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macao, China, 3–8 November 2019; pp. 4213–4220.
37. Secci, F.; Ceccarelli, A. On failures of RGB cameras and their effects in autonomous driving applications. In Proceedings of the IEEE 31st International Symposium on Software Reliability Engineering (ISSRE), Coimbra, Portugal, 12–15 October 2020.
38. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. *Mach. Vis. Appl.* **2014**, *25*, 245–262. [CrossRef]

39. Testolina, P.; Barbato, F.; Michieli, U.; Giordani, M.; Zanuttigh, P.; Zorzi, M. SELMA: SEmantic Large-scale Multimodal Acquisitions in Variable Weather, Daytime and Viewpoints. *arXiv* **2022**, arXiv:2204.09788.
40. Moreland, K. Why we use bad color maps and what you can do about it. *Electron. Imaging* **2016**, *2016*, 1–6. [CrossRef]
41. Zhou, Y.; Liu, L.; Zhao, H.; López-Benítez, M.; Yu, L.; Yue, Y. Towards Deep Radar Perception for Autonomous Driving: Datasets, Methods, and Challenges. *Sensors* **2022**, *22*, 4208. [CrossRef]
42. Gao, B.; Pan, Y.; Li, C.; Geng, S.; Zhao, H. Are We Hungry for 3D LiDAR Data for Semantic Segmentation? A Survey of Datasets and Methods. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 6063–6081. [CrossRef]
43. Jang, M.; Yoon, H.; Lee, S.; Kang, J.; Lee, S. A Comparison and Evaluation of Stereo Matching on Active Stereo Images. *Sensors* **2022**, *22*, 3332. [CrossRef]
44. Hirschmuller, H. Accurate and efficient stereo processing by semi-global matching and mutual information. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 807–814.
45. Zhou, K.; Meng, X.; Cheng, B. Review of stereo matching algorithms based on deep learning. *Comput. Intell. Neurosci.* **2020**, *2020*, 8562323. [CrossRef]
46. Li, J.; Wang, P.; Xiong, P.; Cai, T.; Yan, Z.; Yang, L.; Liu, J.; Fan, H.; Liu, S. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. In Proceedings of the 2022 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LO, USA, 19–24 June 2022.
47. Tonioni, A.; Tosi, F.; Poggi, M.; Mattoccia, S.; Stefano, L.D. Real-time self-adaptive deep stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 195–204.
48. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning rich features from RGB-D images for object detection and segmentation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 345–360.
49. Padmanabhan, P.; Zhang, C.; Charbon, E. Modeling and analysis of a direct time-of-flight sensor architecture for LiDAR applications. *Sensors* **2019**, *19*, 5464. [CrossRef]
50. Li, Y.; Ibanez-Guzman, J. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Process. Mag.* **2020**, *37*, 50–61. [CrossRef]
51. Camuffo, E.; Mari, D.; Milani, S. Recent Advancements in Learning Algorithms for Point Clouds: An Updated Overview. *Sensors* **2022**, *22*, 1357. [CrossRef] [PubMed]
52. Landrieu, L.; Simonovsky, M. Large-scale point cloud semantic segmentation with superpoint graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4558–4567.
53. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–12. [CrossRef]
54. Su, H.; Jampani, V.; Sun, D.; Maji, S.; Kalogerakis, E.; Yang, M.H.; Kautz, J. Splatnet: Sparse lattice networks for point cloud processing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2530–2539.
55. Rosu, R.A.; Schütt, P.; Quenzel, J.; Behnke, S. Latticenet: Fast point cloud segmentation using permutohedral lattices. *arXiv* **2019**, arXiv:1912.05905.
56. Prophet, R.; Deligiannis, A.; Fuentes-Michel, J.C.; Weber, I.; Vossiek, M. Semantic segmentation on 3D occupancy grids for automotive radar. *IEEE Access* **2020**, *8*, 197917–197930. [CrossRef]
57. Ouaknine, A.; Newson, A.; Pérez, P.; Tupin, F.; Rebut, J. Multi-view radar semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15671–15680.
58. Kaul, P.; De Martini, D.; Gadd, M.; Newman, P. RSS-Net: Weakly-supervised multi-class semantic segmentation with FMCW radar. In Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 19 October–13 November 2020; pp. 431–436.
59. Bengler, K.; Dietmayer, K.; Farber, B.; Maurer, M.; Stiller, C.; Winner, H. Three decades of driver assistance systems: Review and future perspectives. *IEEE Intell. Transp. Syst. Mag.* **2014**, *6*, 6–22. [CrossRef]
60. Zhou, Y.; Takeda, Y.; Tomizuka, M.; Zhan, W. Automatic Construction of Lane-level HD Maps for Urban Scenes. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 6649–6656. [CrossRef]
61. Guo, C.; Lin, M.; Guo, H.; Liang, P.; Cheng, E. Coarse-to-fine Semantic Localization with HD Map for Autonomous Driving in Structural Scenes. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 1146–1153. [CrossRef]
62. Aggarwal, C.C. *Neural Networks and Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2018; Volume 10, p. 978.
63. Yin, H.; Berger, C. When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 1–8.
64. Lopes, A.; Souza, R.; Pedrini, H. A Survey on RGB-D Datasets. *arXiv* **2022**, arXiv:2201.05761.
65. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets Robotics: The KITTI Dataset. *Int. J. Robot. Res. (IJRR)* **2013**, *32*, 1231–1237. [CrossRef]

66. Fritsch, J.; Kuehnl, T.; Geiger, A. A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms. In Proceedings of the International Conference on Intelligent Transportation Systems (ITSC), The Hague, The Netherlands, 6–9 October 2013.
67. Menze, M.; Geiger, A. Object Scene Flow for Autonomous Vehicles. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
68. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 3213–3223.
69. Pinggera, P.; Ramos, S.; Gehrig, S.; Franke, U.; Rother, C.; Mester, R. Lost and found: Detecting small road hazards for self-driving vehicles. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 1099–1106.
70. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. 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), Las Vegas, NV, USA, 27–30 June 2016.
71. Hernandez-Juarez, D.; Schneider, L.; Espinosa, A.; Vazquez, D.; Lopez, A.M.; Franke, U.; Pollefeys, M.; Moure, J.C. Slanted Stixels: Representing San Francisco's Steepest Streets. In Proceedings of the British Machine Vision Conference (BMVC), London, UK, 4–7 September 2017.
72. Zolfaghari Bengar, J.; Gonzalez-Garcia, A.; Villalonga, G.; Raducanu, B.; Aghdam, H.H.; Mozerov, M.; Lopez, A.M.; van de Weijer, J. Temporal Coherence for Active Learning in Videos. In Proceedings of the IEEE International Conference in Computer Vision, Workshops (ICCV Workshops), Seoul, Korea, 27 Octorber–2 November 2019.
73. Gaidon, A.; Wang, Q.; Cabon, Y.; Vig, E. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2016.
74. Cabon, Y.; Murray, N.; Humenberger, M. Virtual kitti 2. *arXiv* **2020**, arXiv:2001.10773.
75. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5108–5115. [CrossRef]
76. Xu, H.; Ma, J.; Le, Z.; Jiang, J.; Guo, X. FusionDN: A Unified Densely Connected Network for Image Fusion. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
77. Pfeuffer, A.; Dietmayer, K. Robust Semantic Segmentation in Adverse Weather Conditions by means of Sensor Data Fusion. In Proceedings of the 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–8.
78. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
79. Xiang, K.; Yang, K.; Wang, K. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Opt. Express* **2021**, *29*, 4802–4820. [CrossRef]
80. Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A.S.; Hauswald, L.; Pham, V.H.; Mühlegg, M.; Dorn, S.; et al. A2d2: Audi autonomous driving dataset. *arXiv* **2020**, arXiv:2004.06320.
81. Wang, P.; Huang, X.; Cheng, X.; Zhou, D.; Geng, Q.; Yang, R. The apolloscape open dataset for autonomous driving and its application. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2702–2719.
82. Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; Gaidon, A. 3D Packing for Self-Supervised Monocular Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
83. Liao, Y.; Xie, J.; Geiger, A. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *arXiv* **2021**, arXiv:2109.13410.
84. Yogamani, S.; Hughes, C.; Horgan, J.; Sistu, G.; Varley, P.; O'Dea, D.; Uricár, M.; Milz, S.; Simon, M.; Amende, K.; et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 Octorber–2 November 2019; pp. 9308–9318.
85. Gehrig, D.; Rüegg, M.; Gehrig, M.; Hidalgo-Carrió, J.; Scaramuzza, D. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robot. Autom. Lett.* **2021**, *6*, 2822–2829. [CrossRef]
86. Valada, A.; Oliveira, G.L.; Brox, T.; Burgard, W. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In Proceedings of the International Symposium on Experimental Robotics, Nagasaki, Japan, 3–8 October 2016; pp. 465–477.
87. Zhang, Y.; Morel, O.; Blanchon, M.; Seulin, R.; Rastgoo, M.; Sidibé, D. Exploration of Deep Learning-based Multimodal Fusion for Semantic Road Scene Segmentation. In Proceedings of the VISIGRAPP (5: VISAPP), Prague, Czech Republic, 25–27 February 2019; pp. 336–343.
88. Vachmanus, S.; Ravankar, A.A.; Emaru, T.; Kobayashi, Y. Multi-Modal Sensor Fusion-Based Semantic Segmentation for Snow Driving Scenarios. *IEEE Sens. J.* **2021**, *21*, 16839–16851. [CrossRef]
89. Zhu, A.Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; Daniilidis, K. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2032–2039. [CrossRef]

90. Shivakumar, S.S.; Rodrigues, N.; Zhou, A.; Miller, I.D.; Kumar, V.; Taylor, C.J. PST900: RGB-Thermal Calibration, Dataset and Segmentation Network. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), 31 May–31 August 2020; pp. 9441–9447.
91. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from rgbd images. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 746–760.
92. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
93. Armeni, I.; Sax, A.; Zamir, A.R.; Savarese, S. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *arXiv* **2017**, arXiv:cs.CV/1702.01105.
94. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Niessner, M. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
95. Zamir, A.R.; Sax, A.; Shen, W.; Guibas, L.J.; Malik, J.; Savarese, S. Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3712–3722.
96. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An Open Urban Driving Simulator. In Proceedings of the 1st Annual Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017; pp. 1–16.
97. Gu, Z.; Niu, L.; Zhao, H.; Zhang, L. Hard pixel mining for depth privileged semantic segmentation. *IEEE Trans. Multimed.* **2020**, *23*, 3738–3751. [CrossRef]
98. Valada, A.; Mohan, R.; Burgard, W. Self-supervised model adaptation for multimodal semantic segmentation. *Int. J. Comput. Vis.* **2020**, *128*, 1239–1285. [CrossRef]
99. Liu, H.; Zhang, J.; Yang, K.; Hu, X.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. *arXiv* **2022**, arXiv:2203.04838.
100. Wang, Y.; Sun, F.; Lu, M.; Yao, A. Learning deep multimodal feature representation with asymmetric multi-layer fusion. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 3902–3910.
101. Chen, X.; Lin, K.Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 561–577.
102. Seichter, D.; Köhler, M.; Lewandowski, B.; Wengefeld, T.; Gross, H.M. Efficient rgb-d semantic segmentation for indoor scene analysis. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13525–13531.
103. Kong, S.; Fowlkes, C.C. Recurrent scene parsing with perspective understanding in the loop. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 956–965.
104. Deng, L.; Yang, M.; Li, T.; He, Y.; Wang, C. RFBNet: Deep multimodal networks with residual fusion blocks for RGB-D semantic segmentation. *arXiv* **2019**, arXiv:1907.00135.
105. Valada, A.; Vertens, J.; Dhall, A.; Burgard, W. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 4644–4651.
106. Sun, Y.; Zuo, W.; Liu, M. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2576–2583. [CrossRef]
107. Zhuang, Z.; Li, R.; Jia, K.; Wang, Q.; Li, Y.; Tan, M. Perception-aware Multi-sensor Fusion for 3D LiDAR Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 16280–16290.
108. Rashed, H.; El Sallab, A.; Yogamani, S.; ElHelw, M. Motion and depth augmented semantic segmentation for autonomous navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019.
109. Zhang, Y.; Morel, O.; Seulin, R.; Mériaudeau, F.; Sidibé, D. A central multimodal fusion framework for outdoor scene image segmentation. *Multimed. Tools Appl.* **2022**, *81*, 12047–12060. [CrossRef]
110. Yi, S.; Li, J.; Liu, X.; Yuan, X. CCAFFMNet: Dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module. *Neurocomputing* **2022**, *482*, 236–251. [CrossRef]
111. Frigo, O.; Martin-Gaffé, L.; Wacongne, C. DooDLeNet: Double DeepLab Enhanced Feature Fusion for Thermal-color Semantic Segmentation. In Proceedings of the 2022 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LO, USA, 19–24 June 2022; pp. 3021–3029
112. Zhou, W.; Liu, J.; Lei, J.; Yu, L.; Hwang, J.N. GMNet: Graded-Feature Multilabel-Learning Network for RGB-Thermal Urban Scene Semantic Segmentation. *IEEE Trans. Image Process.* **2021**, *30*, 7790–7802. [CrossRef]
113. Deng, F.; Feng, H.; Liang, M.; Wang, H.; Yang, Y.; Gao, Y.; Chen, J.; Hu, J.; Guo, X.; Lam, T.L. FEANet: Feature-Enhanced Attention Network for RGB-Thermal Real-time Semantic Segmentation. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 4467–4473. [CrossRef]
114. Zhou, W.; Dong, S.; Xu, C.; Qian, Y. Edge-aware Guidance Fusion Network for RGB Thermal Scene Parsing. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2022; pp. 3571–3579.

115. Zhang, Q.; Zhao, S.; Luo, Y.; Zhang, D.; Huang, N.; Han, J. 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, Nashville, TN, USA, 20–25 June 2021; pp. 2633–2642.
116. Xu, J.; Lu, K.; Wang, H. Attention fusion network for multi-spectral semantic segmentation. *Pattern Recognit. Lett.* **2021**, *146*, 179–184. [CrossRef]
117. Sun, Y.; Zuo, W.; Yun, P.; Wang, H.; Liu, M. FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion. *IEEE Trans. Autom. Sci. Eng.* **2020**, *18*, 1000–1011. [CrossRef]
118. Krispel, G.; Opitz, M.; Waltner, G.; Possegger, H.; Bischof, H. Fuseseg: Lidar point cloud segmentation fusing multi-modal data. *arXiv* **2020**, arXiv:1912.08487. [CrossRef]
119. El Madawi, K.; Rashed, H.; El Sallab, A.; Nasr, O.; Kamel, H.; Yogamani, S. Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In Proceedings of the 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 27–30 October 2019; pp. 7–12.
120. Wang, W.; Neumann, U. Depth-aware CNN for RGB-D Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 135–150.
121. Jaritz, M.; Vu, T.H.; Charette, R.d.; Wirbel, E.; Pérez, P. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12605–12614.
122. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. [CrossRef]
123. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018.
124. Iandola, F.N.; Moskewicz, M.W.; Ashraf, K.; Han, S.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1 MB model size. *arXiv* **2016**, arXiv:1602.07360.
125. Graham, B.; Engelcke, M.; Maaten, L.V.D. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8–23 June 2018; pp. 9224–9232. [CrossRef]
126. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500. [CrossRef]
127. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. [CrossRef]
128. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. *arXiv* **2013**, arXiv:1301.3572.
129. Pagnutti, G.; Minto, L.; Zanuttigh, P. Segmentation and semantic labelling of RGBD data with convolutional neural networks and surface fitting. *IET Comput. Vis.* **2017**, *11*, 633–642. [CrossRef]
130. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8–23 June 2018; pp. 7132–7141. [CrossRef]
131. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 945–953.
132. Chiang, H.Y.; Lin, Y.L.; Liu, Y.C.; Hsu, W.H. A unified point-based framework for 3d segmentation. In Proceedings of the 2019 International Conference on 3D Vision (3DV), Québec City, QC, Canada, 16–19 September 2019; pp. 155–163.
133. Jaritz, M.; Gu, J.; Su, H. Multi-view pointnet for 3d scene understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
134. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
135. Liu, W.; Luo, Z.; Cai, Y.; Yu, Y.; Ke, Y.; Junior, J.M.; Gonçalves, W.N.; Li, J. Adversarial unsupervised domain adaptation for 3D semantic segmentation with multi-modal learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *176*, 211–221. [CrossRef]