Multimodal Segmentation Techniques in Autonomous Driving

Subjects: Computer Science, Artificial Intelligence

Contributor: Giulia Rizzoli, Francesco Barbato, Pietro Zanuttigh

Semantic Segmentation has become one of the key steps toward scene understanding, especially in autonomous driving scenarios. In the standard formulation, Semantic Segmentation uses only data from color cameras, which suffer significantly in dim lighting or adverse weather conditions. A solution to this problem is the use of multiple heterogeneous sensors (e.g., depth and thermal cameras or LiDARs) as the input to machine learning approaches tackling this task, allowing to cover for the shortcomings of color cameras and to extract a more resilient representation of the scene.

Keywords: semantic segmentation ; autonomous driving ; multimodal ; LiDAR ; survey

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 that 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.

2. Semantic Segmentation with Deep Learning

A graphic example of a possible deployment of the task in autonomous driving scenarios is reported in **Figure 1**.

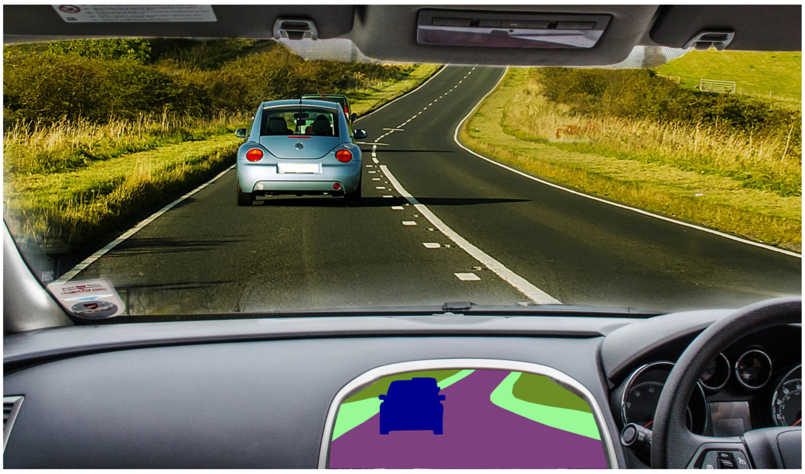

Figure 1. The car screen shows an example of semantic segmentation of the scene in front of the car.

Early approaches to semantic segmentation were based on the use of classifiers on small image patches [8][9][10], 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 [11], 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 [12], ResNet [13], or the more lightweight MobileNet [14].

Other remarkable architectures that followed FCN are ParseNet (Liu et al. [15]), which models global context directly rather than only relying on a larger receptive field, and DeconvNet (Noh et al. [16]) 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. [17], 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. [18][19] 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 researchers, 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. [19] 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 [20], 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) [21][22] 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 [23][24][25].

Although semantic segmentation was originally tackled by RGB data, recently many researchers started investigating its application for LiDAR data [26][27][28][29][30][31]. The development of such approaches is supported by an ever-increasing number of datasets that provide labeled training samples, e.g., Semantic KITTI [32]. More in detail, PointNet [26][27] 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 [28], 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 [30] and RangeNet [33], 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 [31], 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.

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 [34]. 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 [35].

3. Multimodal Segmentation Techniques in Autonomous Driving

Table 1 shows a summarized version of the methods, 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.).

Table 1. 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 [36]	**2021**	[37][38][39]	**DmDe**	**+**	**-**	**+**	**-**	**+**	**+**	**D**	**D/C**	**2**	**+**	**+**
SSMA [40]	**2019**	[37][39][41][42][43]	**DmDhT**	**-**	**+**	**+**	**+**	**+**	**-**	**E**	**D**	**2**	**+**	**+**
CMX [44]	**2022**	[37][38][39][43][45][46][47][48]	**EDhLpT**	**+**	**+**	**-**	**+**	**-**	**-**	**E**	**D/B**	**2**	**+**	**+**
AsymFusion [49]	**2021**	[37][38][50]	**Dm**	**+**	**-**	**-**	**+**	**-**	**-**	**E**	**B**	**2**	**-**	**+**
SA-Gate [51]	**2020**	[37][38]	**Dh**	**+**	**-**	**+**	**+**	**-**	**-**	**E**	**B**	**2**	**+**	**+**
ESANet [52]	**2021**	[37][38][39]	**Dm**	**+**	**-**	**-**	**-**	**-**	**-**	**E**	**C**	**2**	**+**	**+**
DA-Gate [53]	**2018**	[37][38][39][48]	**DmDe**	**-**	**-**	**-**	**-**	**+**	**-**	**N/A**	**N/A**	**1**	**-**	**-**
RFBNet [54]	**2019**	[37][43]	**Dh**	**+**	**+**	**+**	**+**	**-**	**-**	**E**	**B**	**2**	**-**	**+**
MMSFB-snow [55]	**2021**	[37][41][55]	**DmT**	**-**	**-**	**+**	**+**	**-**	**-**	**E**	**D**	**2**	**+**	**+**
AdapNet [56]	**2017**	[37][41][42]	**DmT**	**+**	**+**	**-**	**+**	**-**	**-**	**D**	**D**	**2**	**-**	**-**
RFNet [57]	**2020**	[37][58]	**Dm**	**+**	**-**	**-**	**+**	**-**	**-**	**E**	**C**	**2**	**+**	**+**
RSSAWC [59]	**2019**	[37][59]	**DmLi**	**+**	**-**	**+**	**-**	**-**	**-**	**E**	**D**	**2**	**-**	**-**
PMF [60]	**2021**	[32][61]	**Li**	**+**	**+**	**+**	**+**	**+**	**-**	**E**	**M**	**2**	**+**	**+**
MDASS [62]	**2019**	[37][63]	**DmF**	**+**	**-**	**-**	**-**	**-**	**-**	**E**	**D**	**2/3**	**+**	**+**
CMFnet [64]	**2021**	[37][65]	**DmLp**	**-**	**+**	**+**	**-**	**-**	**-**	**E**	**D/B**	**3+**	**-**	**+**
CCAFFMNet [66]	**2021**	[45][67]	**T**	**-**	**-**	**+**	**+**	**-**	**-**	**E**	**C**	**2**	**+**	**+**
DooDLeNet [68]	**2022**	[45]	**T**	**-**	**+**	**+**	**-**	**-**	**-**	**E**	**D**	**2**	**+**	**+**
GMNet [69]	**2021**	[45][70]	**T**	**+**	**+**	**-**	**+**	**-**	**+**	**E**	**D**	**2**	**+**	**+**
FEANet [71]	**2021**	[45]	**T**	**+**	**+**	**-**	**-**	**-**	**-**	**E**	**C**	**2**	**-**	**+**
EGFNet [72]	**2021**	[45][70]	**T**	**+**	**+**	**+**	**+**	**-**	**-**	**E**	**D**	**2**	**-**	**+**
ABMDRNet [73]	**2021**	[45]	**T**	**+**	**+**	**+**	**+**	**+**	**+**	**E**	**D**	**2**	**-**	**+**
AFNet [74]	**2021**	[45]	**T**	**+**	**+**	**-**	**+**	**-**	**-**	**E**	**D**	**2**	**-**	**-**
FuseSeg-Thermal [75]	**2021**	[45]	**T**	**+**	**-**	**+**	**-**	**-**	**-**	**E**	**C**	**2**	**+**	**+**

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
RTFNet [57]	2019	[45]	T	+	-	-	-	-	-	E	C	2	-	+
FuseSeg-LiDAR [76]	2020	[77]	LsLi	-	-	+	-	-	-	E	M	2	+	+
RaLF3D [78]	2019	[77]	LsLi	+	-	+	-	-	-	E	D	2	+	+
DACNN [79]	2018	[38][39][48]	DmDh	+	-	-	-	-	-	E	D	2	-	-
xMUDA [80]	2020	[32][61][81]	Li	-	-	+	-	+	+	D	D	2	-	+

On the other hand, in **Table 2**, researchers report the numerical score (mIoU) attained by the methods in three benchmark datasets, respectively: Cityscapes [37] for 2.5D SS in **Table 2**a, KITTI [77] for 2D + 3D SS in **Table 2**b and MSSSD/MF [45] for RGB + Thermal SS in **Table 2**c.

Table 2. Architectures Performance Comparison.

Name	Backbone	mIoU
(a) Cityscapes dataset (2.5D SS).		
LWM [36]	ResNet101 [13]	83.4
SSMA [40]	ResNet50 [13]	83.29
CMX [44]	MiT-B4 [24]	82.6
AsymFusion [49]	Xception65 [82]	82.1
SA-Gate [51]	ResNet101 [13]	81.7
ESANet [52]	ResNet34 [13]	80.09
DA-Gate [53]	ResNet101 [13]	75.3
RFBNet [54]	ResNet50 [13]	74.8
MMSFB-snow [55]	ResNet50 [13]	73.8
AdapNet [56]	AdapNet [56]	71.72
RFNet [57]	ResNet18 [13]	69.37
RSSAWC [59]	ICNet [83]	65.09
MDASS [62]	VGG16 [12]	63.13
CMFnet [64]	VGG16 [12]	58.97
(b) KITTI dataset (2D + 3D SS).		
PMF [60]	ResNet34 [13]	63.9
FuseSeg-LiDAR [76]	SqueezeNet [84]	52.1
RaLF3D [78]	SqueezeSeg [30]	37.8
xMUDA [80]	SparseConvNet3D [85] ResNet34 [13]	49.1
(c) MSSSD/MF dataset (RGB + Thermal SS).		
CMX [44]	MiT-B4 [24]	59.7
CCAFFMNet [66]	ResNeXt50 [86]	58.2
DooDLeNet [68]	ResNet101 [13]	57.3
GMNet [69]	ResNet50 [13]	57.3
FEANet [71]	ResNet101 [13]	55.3
EGFNet [72]	ResNet152 [13]	54.8
ABMDRNet [73]	ResNet50 [13]	54.8
AFNet [74]	ResNet50 [13]	54.6
FuseSeg-Thermal [75]	DenseNet161 [87]	54.5
RTFNet [57]	ResNet152 [13]	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 [88][89]. This simple early fusion combination strategy is not too effective because it struggles to capture the different types 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 into 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 [40][73][76].

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 [40][44][51]. This hierarchical fusion strategy leverages multilevel features via progressive feature merging and generates 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 [52] 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 [40].

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 [80].

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.
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.
3. Wang, J.; Liu, J.; Kato, N. Networking and Communications in Autonomous Driving: A Survey. IEEE Commun. Surv. Tutor. 2019, 21, 1243–1274.
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.
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.
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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
13. 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.

14. 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.

15. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. arXiv 2015, arXiv:1506.04579.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. Gade, R.; Moeslund, T.B. Thermal cameras and applications: A survey. Mach. Vis. Appl. 2014, 25, 245–262.

36. Gu, Z.; Niu, L.; Zhao, H.; Zhang, L. Hard pixel mining for depth privileged semantic segmentation. IEEE Trans. Multimed. 2020, 23, 3738–3751.

37. 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.

38. 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.

39. 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.

40. Valada, A.; Mohan, R.; Burgard, W. Self-supervised model adaptation for multimodal semantic segmentation. Int. J. Comput. Vis. 2020, 128, 1239–1285.

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. Xiang, K.; Yang, K.; Wang, K. Polarization-driven semantic segmentation via efficient attention-bridged fusion. Opt. Express 2021, 29, 4802–4820.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.

73. 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.
74. Xu, J.; Lu, K.; Wang, H. Attention fusion network for multi-spectral semantic segmentation. Pattern Recognit. Lett. 2021, 146, 179–184.
75. 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.
76. Krispel, G.; Opitz, M.; Waltner, G.; Possegger, H.; Bischof, H. Fuseseg: Lidar point cloud segmentation fusing multi-modal data. arXiv 2020, arXiv:1912.08487.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. Couprie, C.; Farabet, C.; Najman, L.; LeCun, Y. Indoor semantic segmentation using depth information. arXiv 2013, arXiv:1301.3572.
89. 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.

Retrieved from https://encyclopedia.pub/entry/history/show/63309