



**HAL**  
open science

# Real-time multimodal semantic scene understanding for autonomous UGV navigation

Yifei Zhang

► **To cite this version:**

Yifei Zhang. Real-time multimodal semantic scene understanding for autonomous UGV navigation. Image Processing [eess.IV]. Université Bourgogne Franche-Comté, 2021. English. NNT : 2021UBFCK002 . tel-03154783

**HAL Id: tel-03154783**

**<https://theses.hal.science/tel-03154783v1>**

Submitted on 1 Mar 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**THÈSE DE DOCTORAT**

**DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ**

**PRÉPARÉE À L'UNIVERSITÉ DE BOURGOGNE**

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Instrumentation et informatique d'image

par

**YIFEI ZHANG**

**Real-time multimodal semantic scene understanding for autonomous navigation**

Thèse présentée et soutenue à Le creusot, le 19 January 2021

Composition du Jury :

FREMONT VINCENT	Professeur à Ecole Centrale Nantes	Rapporteur
AINOUZ SAMIA	Professeur à INSA Rouen	Rapporteur
DUCOTTET CHRISTOPHE	Professeur à Université St-Etienne	Examineur/Président
MOREL OLIVIER	MCF à Univ. de Bourgogne Franche-Comté	Co-encadrant
MÉRIAUDEAU FABRICE	Professeur à Univ. de Bourgogne Franche-Comté	Co-directeur de thèse
SIDIBÉ DÉSIRÉ	Professeur à Université Paris-Saclay	Co-directeur de thèse



# ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisors Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau for their constant encouragement and guidance on my thesis. Without their patient instruction, insightful criticism, and expert guidance, the completion of experiments and papers would not have been possible. In preparing the doctoral thesis, they have spent much time reading through each draft and provided me with inspiring advice. I am very thankful to Désiré for his continuous support and valuable advice in all our discussions. Thank him for choosing me and giving me the chance to pursue my dream of scientific research. I am very grateful to Olivier for his insightful comments, which enable my thesis work further improved, and thanks for his encouragement, which makes me more confident in my presentation. I would also like to convey my sincere thanks to Fabrice for his patience, motivation, enthusiasm, and trust.

Then, my faithful appreciation also goes to the jury members for their kindness in participating in my Ph.D. defense committee. Many thanks to Prof. Vincent Fremont and Prof. Samia Ainouz for their precious time reading the thesis and their constructive remarks, and to Prof. Christophe Ducottet for his support as the president of the defense and valuable suggestions for the thesis work.

I also owe my sincere gratitude to Samia B. and Frederick for their friendly reception. During my visit to the IBISC laboratory, they have provided a lot of convenience for my life and work. During the epidemic, they provided me with the necessary conditions to prevent epidemics and tried their best to ensure my health and research progress.

I would also like to thank all my colleagues in the VIBOT team. During the two years in Le Creusot, they have given me generous support and kindness in my research work and teaching. To name a few: David F., Cédric, Ralph, Nathalie, Christophe S., Christophe L., Zawawi, Nathan P., Thomas, Thibault, Daniel... I am incredibly very thankful to Mojdeh and Cansen for their help in my early work of the thesis. They spent a lot of time teaching me how to do academic research and developing good research habits. Besides, I enjoyed the time working with Marc. I have fond memories of the days when we went to Prague to attend the conference. I also miss having dinner with friends and playing games at Mojdeh and Guillaume's house, climbing with David S., eating Girls' lunch with Ashvaany and Abir, training with Raphael and Nathan C. in Toulouse, with Eric and Fauvet taught in Nanjing together... Thank you very much for giving me many precious memories. Although I am far away from home, I still feel warm because of them.



Last, my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years. Also, I would like to express my love to my boyfriend Qifa, who gave me endless tenderness and time in listening to me and help me fight through the tough times. Their love and support make my dream come true.

# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Context and Motivation . . . . .	1
1.2	Background and Challenges . . . . .	3
1.2.1	Multimodal Scene Understanding . . . . .	3
1.2.2	Semantic Image Segmentation . . . . .	4
1.2.3	Deep Multimodal Semantic Segmentation . . . . .	5
1.2.4	Scenarios and Applications . . . . .	7
1.2.5	Few-shot Semantic Segmentation . . . . .	7
1.3	Contributions . . . . .	8
1.4	Organization . . . . .	9
<b>2</b>	<b>Background on Neural Networks</b>	<b>11</b>
2.1	Basic Concepts . . . . .	12
2.1.1	Convolution . . . . .	12
2.1.2	Convolutional Neural Networks . . . . .	13
2.1.3	Encoder-decoder Architecture . . . . .	14
2.2	Neural Network Layers . . . . .	15
2.2.1	Convolution layer . . . . .	15
2.2.2	Pooling Layer . . . . .	16
2.2.3	ReLu Layer . . . . .	17
2.2.4	Fully Connected Layer . . . . .	17
2.3	Optimization . . . . .	18
2.3.1	Batch Normalization . . . . .	18
2.3.2	Dropout . . . . .	18
2.4	Model Training . . . . .	19

2.4.1	Data Preprocessing . . . . .	19
2.4.2	Weight Initialization . . . . .	20
2.4.3	Loss Function . . . . .	20
2.4.4	Gradient Descent . . . . .	22
2.5	Evaluation Metrics . . . . .	23
2.6	Summary . . . . .	24
<b>3</b>	<b>Literature Review</b>	<b>25</b>
3.1	Fully-supervised Semantic Image Segmentation . . . . .	26
3.1.1	Taxonomy of Deep Multimodal Fusion . . . . .	26
3.1.1.1	Early Fusion . . . . .	27
3.1.1.2	Late Fusion . . . . .	28
3.1.1.3	Hybrid Fusion . . . . .	29
3.1.1.4	Statistical Fusion . . . . .	31
3.1.2	Discussion . . . . .	32
3.2	Datasets . . . . .	34
3.2.1	Popular Datasets for Image Segmentation Task . . . . .	34
3.2.2	Multimodal Datasets . . . . .	36
3.2.2.1	RGB-D Datasets . . . . .	36
3.2.2.2	Near-InfraRed Datasets . . . . .	37
3.2.2.3	Thermal Datasets . . . . .	38
3.2.2.4	Polarization Datasets . . . . .	38
3.2.2.5	Critical Challenges for Multimodal Data . . . . .	39
3.2.3	Comparative Analysis . . . . .	40
3.2.3.1	Accuracy . . . . .	40
3.2.3.2	Execution Time . . . . .	43
3.2.3.3	Memory Usage . . . . .	44
3.3	Summary . . . . .	44
<b>4</b>	<b>Deep Multimodal Fusion for Semantic Image Segmentation</b>	<b>47</b>
4.1	CMNet: Deep Multimodal Fusion for Road Scene Segmentation . . . . .	48

4.1.1	Introduction . . . . .	48
4.1.2	Baseline Architectures . . . . .	49
4.1.2.1	Early Fusion . . . . .	49
4.1.2.2	Late Fusion . . . . .	50
4.1.3	Proposed Method . . . . .	50
4.1.4	Dataset . . . . .	52
4.1.4.1	Polarization Formalism . . . . .	52
4.1.4.2	POLABOT Dataset . . . . .	54
4.1.5	Experiments . . . . .	55
4.1.5.1	Evaluation on Freiburg Forest Dataset . . . . .	55
4.1.5.2	Evaluation on POLABOT Dataset . . . . .	58
4.2	A Central Multimodal Fusion Framework . . . . .	59
4.2.1	Introduction . . . . .	59
4.2.2	Method . . . . .	60
4.2.2.1	Central Fusion . . . . .	60
4.2.2.2	Adaptive Central Fusion Network . . . . .	60
4.2.2.3	Statistical Prior Fusion . . . . .	61
4.2.3	Experiments . . . . .	63
4.2.3.1	Implementation Details . . . . .	63
4.2.3.2	Evaluation on POLABOT Dataset . . . . .	64
4.2.3.3	Evaluation on Cityscapes Dataset . . . . .	64
4.3	Summary . . . . .	67
<b>5</b>	<b>Few-shot Semantic Image Segmentation</b>	<b>69</b>
5.1	Introduction on Few-shot Segmentation . . . . .	70
5.2	MAPnet: A Multiscale Attention-Based Prototypical Network . . . . .	70
5.2.1	Introduction . . . . .	70
5.2.2	Problem Setting . . . . .	72
5.2.3	Method . . . . .	72
5.2.4	Experiments . . . . .	74

5.2.4.1	Setup . . . . .	74
5.2.4.2	Evaluation . . . . .	75
5.2.4.3	Test with weak annotations . . . . .	76
5.3	RDNet: Incorporating Depth Information into Few-shot Segmentation . . . .	78
5.3.1	Introduction . . . . .	78
5.3.2	Problem Setting . . . . .	80
5.3.3	Method . . . . .	81
5.3.4	Cityscapes-3i Dataset . . . . .	83
5.3.5	Experiments . . . . .	84
5.3.5.1	Setup . . . . .	84
5.3.5.2	Experimental Results . . . . .	84
5.3.5.3	Feature Visualization . . . . .	86
5.4	Summary . . . . .	86
<b>6</b>	<b>Conclusion and Future Work</b>	<b>89</b>
6.1	General Conclusion . . . . .	89
6.2	Future Perspectives . . . . .	90

# LIST OF FIGURES

1.1	Example of multimodal semantic segmentation from Cityscapes dataset. The prediction was generated by our CMFnet+BF2 framework. . . . .	2
1.2	Number of papers published per year. Statistical analysis is based on the work by Caesar [15]. Segmentation includes image/instance/panoptic segmentation and joint depth estimation. . . . .	4
1.3	An illustration of deep multimodal segmentation pipeline. . . . .	5
2.1	Example of a convolutional neural network with four inputs. . . . .	13
2.2	Example of the encoder-decoder architecture. . . . .	15
2.3	Example of max pooling operation. . . . .	16
2.4	Examples of popular activation functions. . . . .	17
3.1	An illustration of different fusion strategies for deep multimodal learning. . .	26
3.2	FuseNet architecture with RGB-D input. Figure reproduced from [69]. . . .	28
3.3	Convolved Mixture of Deep Experts framework. Figure extracted from [177].	29
3.4	Fusion architecture with self-supervised model sdaptation modules. Figure extracted from [177]. . . . .	30
3.5	Individual semantic segmentation experts are combined modularly using different statistical methods. Figure extracted from [13] . . . . .	31
3.6	Accumulated dataset importance. Statistical analysis is based on the work by [15]. . . . .	33
3.7	Real-time and accuracy performance. Performance of SSMA fusion method using different real-time backbones on the Cityscapes validation set (input image size: $768 \times 384$ , GPU: NVIDIA TITAN X). . . . .	43
3.8	Real-time and accuracy performance. Performance of different fusion methods on the Tokyo Multi-Spectral dataset.(input image size: $640 \times 480$ , GPU: NVIDIA 1080 Ti graphics card). . . . .	44
4.1	Typical early fusion and Late fusion architectures comparison. . . . .	49

4.2	Our proposed fusion architecture: CMnet for multimodal fusion based on late fusion architecture. . . . .	51
4.3	The electric and magnetic field of light as well as their continuous self-propagating. . . . .	53
4.4	Reflection influence on polarimetry. (a) and (b) represent a zoom on the non-polarized and polarized area, respectively. Figure extracted from [11]. .	54
4.5	(a) Mobile robot platform used for the acquisition of the POLABOT dataset. It is equipped with the IDS Ucam, PolarCam, Kinect 2 and a NIR camera. (b)(c)(d) Multimodal images in the POLABOT dataset. . . . .	55
4.6	Two segmented examples from Freiburg Forest dataset. RGB and/or EVI images were given as inputs. . . . .	57
4.7	Two segmented examples from POLABOT dataset. RGB and/or POLA images were given as inputs. . . . .	58
4.8	Typical fusion strategies with RGB and depth input. (a) Early fusion. (b) Late fusion. (c) Central fusion. As a comparison, the proposed fusion structure integrates the feature maps in a succession of layers into a central branch. . . . .	60
4.9	Overview of the central fusion network (CMFnet). The adaptive gating unit automatically produces the weights of modality-specific branch in each layer. GAP denotes to Global Average Pooling, "x" means multiplication, "C" means concatenation. . . . .	62
4.10	Training loss of models on the POLABOT dataset. . . . .	65
4.11	Improvement/error maps of the proposed CMFnet and CMFnet+BF2 in comparison to the RGB baseline. Note that the improved pixels and the misclassified pixels are recorded in green and red, respectively. . . . .	66
4.12	Segmentation results on the POLABOT dataset. The first row of examples contains the RGB image and the corresponding polarimetric image. The second row of examples shows, from left to right, the ground truth image, the segmentation outputs of the individual experts (RGB and Polar), and the fusion results of CMFnet-BF2 and CMFnet-BF3. . . . .	67
4.13	Two sets of segmentation results on Cityscapes dataset. The first row of examples contains the RGB image and the corresponding depth image. From left to right of the second row of examples: ground truth, the prediction of RGB input, average and LFC. From left to right of the second row of examples: the fusion results of CMoDE, CMFnet, CMFnet-BF2 and CMFnet-BF3, respectively. . . . .	68

5.1	An overview of the proposed method (MAPnet). Given a query image of a new category, e.g., aeroplane, the goal of few-shot segmentation is to predict a mask of this category regarding only a few labeled samples. . . .	71
5.2	Illustration of the proposed method (MAPnet) for few-shot semantic segmentation. . . . .	73
5.3	Qualitative results of our method for 1-way 1-shot segmentation on the PASCAL-5 <sup>i</sup> dataset. . . . .	77
5.4	Training loss of models with and without attention-based gating (ABG) for 1-way 1-shot segmentation on PASCAL-5 <sup>0</sup> . . . . .	78
5.5	Qualitative results of our model using scribble and bounding box annotations for 1-way 5-shot setting. The chosen example in support images shows the annotation types. . . . .	79
5.6	Overview of the proposed RDNet approach. R and D indicate the RGB and depth image input, respectively. The abstract features of labeled support images are mapped into the corresponding embedding space (circles). Multiple prototypes (blue and yellow solid circles) are generated to perform semantic guidance (dashed lines) on the corresponding query features (rhombus). RDNet further produces the final prediction by combining the probability maps from RGB and depth stream. . . . .	80
5.7	Details of the proposed RDNet architecture. It includes two mirrored streams: an RGB stream and a depth stream. Each stream processes the corresponding input data, including a support set and a query set. The prototypes of support images are obtained by masked average pooling. Then the semantic guidance is performed on the query feature by computing the relative cosine distance. The results from these two streams are combined at the late stage. . . . .	82
5.8	Qualitative results of our method for 1-way 1-shot semantic segmentation on Cityscapes-3 <sup>i</sup> . . . . .	86
5.9	Visualization using t-SNE [179] for RGB and depth prototype representations in our RDNet. . . . .	87





# LIST OF TABLES

3.1	Typical early fusion methods reviewed in this chapter. . . . .	27
3.2	Typical late fusion methods reviewed in this chapter. . . . .	29
3.3	Typical hybrid fusion methods reviewed in this chapter. . . . .	30
3.4	Summary of popular datasets for image segmentation task. . . . .	35
3.5	Summary of popular 2D/2.5D multimodal datasets for scene understanding. . . . .	36
3.6	Examples of multimodal image datasets mentioned in Subsection 3.2.2. For each dataset, the top image shows two modal representations of the same scene. The bottom image is the corresponding groundtruth. . . . .	39
3.7	Performance results of deep multimodal fusion methods on SUN RGB-D dataset. . . . .	40
3.8	Performance results of deep multimodal fusion methods on NYU Depth v2 dataset. . . . .	41
3.9	Experimental results of deep multimodal fusion methods on Cityscapes dataset. Input images are uniformly resized to $768 \times 384$ . . . . .	41
3.10	Experimental results of deep multimodal fusion methods on Tokyo Multi-Spectral dataset. The image resolution in the dataset is $640 \times 480$ . . . . .	42
3.11	Parameters and inference time performance. The reported results on the Cityscapes dataset are collected from [178]. . . . .	44
4.1	Performance of segmentation models on Freiburg Multispectral Forest dataset. EF, LF refer to early fusion and late fusion respectively. We report pixel accuracy (PA), mean accuracy (MA), mean intersection over union (MIoU), frequency weighted IoU (FWIoU) as metric to evaluate the performance. . . . .	56
4.2	Comparison of deep unimodal and multimodal fusion approaches by class. We report MIoU as metric to evaluate the performance. . . . .	56
4.3	Segmentation performance on POLABOT dataset . . . . .	59

4.4	Performance comparison of our method with baseline models on the PO-LABOT dataset. Note that SegNet is used as the unimodal baseline and the backbone of multimodal fusion methods. . . . .	65
4.5	Ablation study of our method. Per class performance of our proposed framework in comparison to individual modalities with ENet baseline on Cityscapes dataset. . . . .	66
4.6	Performance of fusion models with ENet backbone on Cityscapes dataset.	66
5.1	Training and evaluation on PASCAL-5 <sup>i</sup> dataset using 4-fold cross-validation, where <i>i</i> denotes the number of subsets. . . . .	75
5.2	Results of 1-way 1-shot and 1-way 5-shot semantic segmentation on PASCAL-5 <sup>i</sup> using mean-IoU(%) metric. The results of 1-NN and LogReg are reported by [158]. . . . .	76
5.3	Results of 1-way 1-shot and 1-way 5-shot segmentation on PASCAL-5 <sup>i</sup> using binary-IoU(%) metric. $\Delta$ denotes the difference between 1-shot and 5-shot. . . . .	76
5.4	Evaluation results of using different types of annotations in mean-IoU(%) metric. . . . .	78
5.5	Training and evaluation on Cityscapes-3 <sup>i</sup> dataset using 3-fold cross-validation, where <i>i</i> denotes the number of subsets. . . . .	83
5.6	Results of 1-way 1-shot and 1-way 2-shot semantic segmentation on Cityscapes-5 <sup>i</sup> using mean-IoU(%) metric. . . . .	85
5.7	Per-class mean-IoU(%) comparison of ablation studies for 1-way 1-shot semantic segmentation . . . . .	85
5.8	Results of 1-way 1-shot semantic segmentation using binary IoU and the runtime. . . . .	86

# INTRODUCTION

## 1.1/ CONTEXT AND MOTIVATION

Since the 1960s, scientists are dedicated to creating machines that can see and understand the world like humans, which led to the emergence of computer vision. It has now become an active subfield of artificial intelligence and computer science for processing visual data. This thesis tackles the challenge of semantic segmentation in scene understanding, particularly multimodal image segmentation of the outdoor road scenes. Semantic segmentation, as a high-level task in the computer vision field, paves the way towards complete scene understanding. From a more technical perspective, semantic image segmentation refers to the task of assigning a semantic label to each pixel in the image [129, 54, 202]. This terminology was further distinguished from instance-level segmentation [38] that devotes to produce per-instance mask and class label. Recently, panoptic segmentation [91, 24] is getting popular which combines pixel-level and instance-level semantic segmentation. Although there are many traditional machine learning algorithms available to tackle these challenges, the rise of deep learning techniques [96, 59] gains unprecedented success and tops other approaches by a large margin. The various milestones in the evolution of deep learning significantly promote the advancement of semantic segmentation research.

Especially in recent years, deep multimodal fusion methods benefit from the massive amount of data and increased computing power. These fusion methods fully exploit hierarchical feature representations in an end-to-end manner. Multimodal information sources provide rich but redundant scene information, which also accompanied by uncertainty. Researchers engage in designing compact neural networks to extract valuable features, thus enhancing the perception of intelligent systems. The underlying motivation for deep multimodal image segmentation is to learn the optimal joint representation from rich and complementary features of the same scene. Moreover, the availability of multiple sensing modalities has encouraged the development of multimodal fusion, such as 3D LiDARs, RGB-D cameras, thermal cameras, etc. These modalities are usually used as

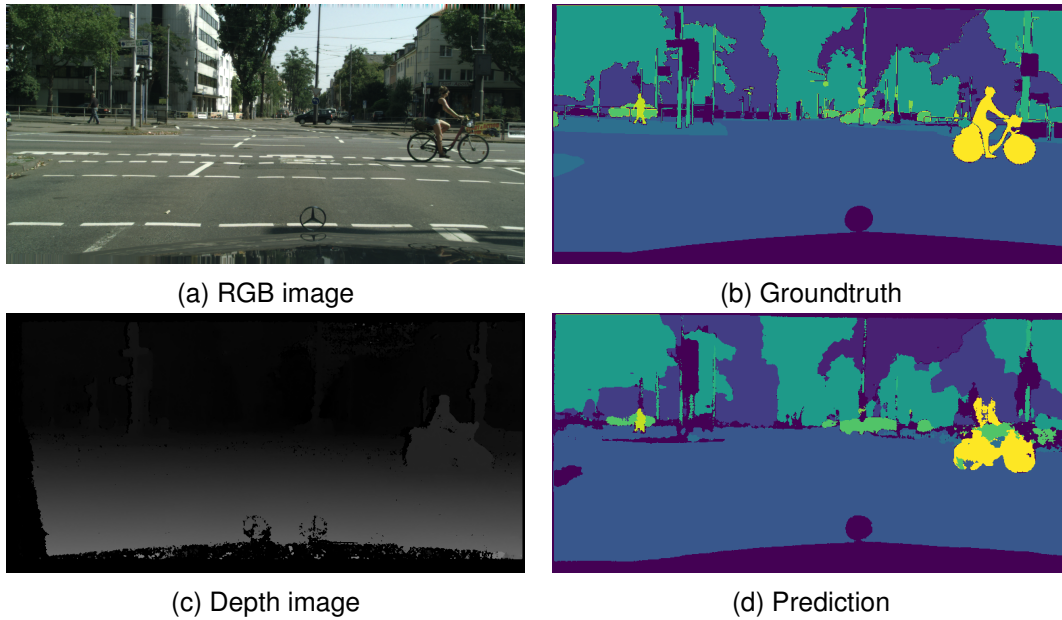


Figure 1.1: Example of multimodal semantic segmentation from Cityscapes dataset. The prediction was generated by our CMFnet+BF2 framework.

complementary sensors in complex scenarios, reducing the uncertainty of scene information. For example, visual cameras perform advanced information processing in lighting conditions, while LiDARs are robust to challenging weather conditions such as rain, snow, or fog. Thermal cameras work well in the nighttime as they are more sensitive to infrared radiation emitted by all objects with a temperature above absolute zero [50]. Arguably, the captured multimodal data provide more spatial and contextual information for robust and accurate scene understanding. Compared to using a single modality, multi-modalities significantly improve the performance of learning models [40, 113, 197, 2, 199]. Figure 1.1 illustrates an example of semantic image segmentation with RGB and depth input.

Besides, much research dedicates to exploring advanced technologies under limited supervision. Recently, few-shot learning has emerged as a hot topic in the computer vision community. Deep learning-based image understanding techniques require a large number of labeled images for training. Few-shot semantic segmentation, on the contrary, aims at generalizing the segmentation ability of the model to new categories given a few samples. Namely, the trained neural network predicts pixel-level mask of new categories on the query image, given only a few labeled support images. In this thesis, we explore the attention mechanism-based method for few-shot segmentation task, aiming to improve the semantic feature representation and generalization capabilities of the models.

However, the generalization and discrimination abilities of existing unimodal few-shot segmentation methods still remain to be improved, especially for complex scenes. The semantic understanding of outdoor road scenes is usually affected by environmental changes, such as occlusion of objects and variable lighting conditions, which makes

learning and prediction of the few-shot network difficult. In order to obtain complete scene information, we are also committed to extending the RGB-centric approach to take advantage of complementary depth information. The original intention of our work is to incorporate supplementary multimodal image information into a few-shot segmentation model. These multimodal data provide rich color and geometric information of scenes, leading to more accurate segmentation performance.

## 1.2/ BACKGROUND AND CHALLENGES

### 1.2.1/ MULTIMODAL SCENE UNDERSTANDING

Humans live in a complex multi-source environment. Whether it is video, image, text or voice, each form of information can be called a modality. We adopt the definition of **modality** from [95], which refers to each detector acquiring information about the same scene. The range of modalities is wider than our perception ability. In addition to the information obtained by vision, we can also collect multimodal information by various sensors such as radar and infrared cameras. Multimodal fusion systems work like the human brain, which synthesizes multiple sources of information for semantic perception and further decision making. Ideally, we would like to have an all-in-one sensor to capture all the information, but for most complex scenarios, it is hard for a single modality to provide complete knowledge. Consequently, the primary motivation for multimodal scene understanding is to obtain rich characteristics of the scenes by integrating multiple sensory modalities. Our work focuses on deep learning-based multimodal fusion technology, which also involves multimodal collaborative learning, multimodal feature representation, multimodal alignment, etc.

As a multi-disciplinary research, the meaning of multi-modality varies in different fields. For example, in medical image analysis, the principal modalities involve Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Single-Photon Emission Computed Tomography (SPECT) [112], to name a few. Benefiting from the complementary and functional information about a target (e.g., organ), multimodal fusion models can achieve a precise diagnosis and treatment [122, 116, 217]. In multimedia analysis, multimodal data collected from audio, video as well as text modalities [3, 126, 6] are used to tackle semantic concept detection, including human-vehicle interaction [41], biometric identification [49, 167, 28]. In remote sensing applications, multimodal fusion leverages the high-resolution optical data, synthetic aperture radar, and 3D point cloud [4, 206]. In this thesis, we clarify the definition of **modality** for semantic segmentation tasks as a single image sensor. Relevant sensory modalities reviewed and experimented include RGB-D cameras, Near-Infrared cameras, thermal cameras, and

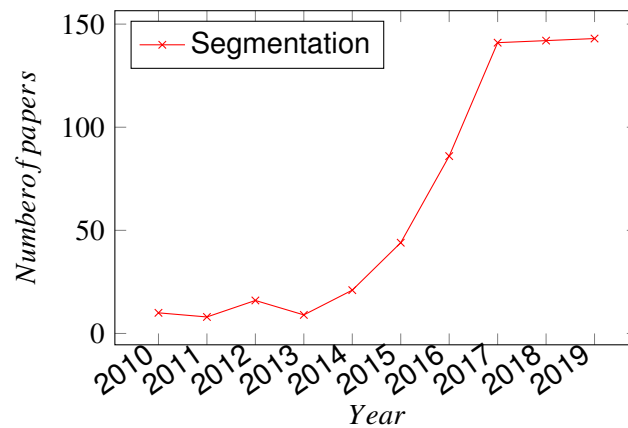


Figure 1.2: Number of papers published per year. Statistical analysis is based on the work by Caesar [15]. Segmentation includes image/instance/panoptic segmentation and joint depth estimation.

polarization cameras.

### 1.2.2/ SEMANTIC IMAGE SEGMENTATION

In recent years, there have been many studies addressing semantic image segmentation with deep learning techniques [54, 56]. As a core problem of computer vision, scene understanding relies heavily on semantic segmentation technology to obtain semantic information and infer knowledge from imagery. Looking back at the history of semantic image segmentation, Fully Convolutional Network (FCN) [117] was first proposed for effective pixel-level classification. In FCN, the last fully connected layer is substituted by convolutional layers. DeconvNet [128], which is composed of deconvolution and unpooling layers, was proposed in the same year. Badrinarayanan et al. [5] introduced a typical encoder-decoder architecture with forwarding pooling indices, mentioned as SegNet. Another typical segmentation network with multi-scale features concatenation, U-Net [150], was initially proposed for biomedical image segmentation. In particular, U-Net employs skip connections to combine deep semantic features from the decoder with low-level fine-grained feature maps of the encoder. Then a compact network called ENet [131] was presented for real-time segmentation. In the work of PixelNet, Bansal et al. [7] explore the spatial correlations between pixels to improve the efficiency and performance of segmentation models. It is worth noting that Dilated Convolution was introduced in DeepLab [17] and DilatedNet [201], which helps to keep the resolution of output feature maps learn with large receptive fields. Besides, a series of Deeplab models also achieves excellent success on semantic image segmentation [20, 19, 21].

Furthermore, Peng et al. [135] dedicated to employing larger kernels to address both the classification and localization issues for semantic segmentation. RefineNet [108] ex-

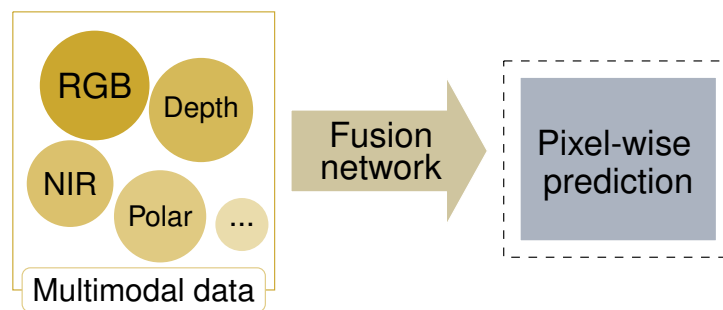


Figure 1.3: An illustration of deep multimodal segmentation pipeline.

Explicitly exploits multi-level features for high-resolution prediction using long-range residual connections. Zhao et al. [214] presented an image cascade network, known as ICNet, that incorporates multi-resolution branches under proper label guidance. In more recent years, semantic segmentation for adverse weather conditions [154, 137] and nighttime [61, 171] has also been addressed to perform the generalization capacity and robustness of deep learning models. Figure 1.2 shows the number of papers about segmentation published in the past decade. The statistical results include well-known computer vision conferences such as CVPR, ECCV, ICCV, BMVC, etc., as well as top journals such as IJCV, PAMI, etc.

In addition to the aforementioned networks, many practical deep learning techniques (e.g., Spatial Pyramid Pooling [73], CRF-RNN [215], Batch Normalization [79], Dropout [52]) were proposed for improving the effectiveness of learning models. Notably, multi-scale feature aggregation was frequently used in semantic segmentation [212, 98, 100, 173, 99]. These learning models experimentally achieve significant performance improvement. Lin et al. [110] introduced the Global Average Pooling (GAP) that replaces the traditional fully connected layers in CNN models. GAP computes the mean value for each feature map without additional training of model parameters. Thus it minimizes overfitting and makes the network more robust. Related applications in multimodal fusion networks can be found in [219, 178, 76]. Also, the  $1 \times 1$  convolution layer is commonly used to allow complex and learnable interaction across modalities and channels [78, 178]. Besides, attention mechanism has become a powerful tool for image recognition [18, 148, 103]. The attention distribution enables the model to selectively pick valuable information [181], achieving more robust feature representation and more accurate prediction.

### 1.2.3/ DEEP MULTIMODAL SEMANTIC SEGMENTATION

Before the tremendous success of deep learning, researchers expressed an interest in combining data captured from multiple information sources into a low-dimensional space, known as *early fusion* or *data fusion* [89]. Machine learning techniques used for such fusion include Principal Component Analysis (PCA), Independent Components Analy-



sis (ICA), and Canonical Correlation Analysis (CCA) [145]. As the discriminative classifiers [94] become increasingly popular (e.g. SVM [47] and Random Forest [58]), a growing body of research focus on integrating multimodal features at the late stage, such fusion strategy was called *late fusion* or *decision fusion*. These fusion strategies had become mainstream for a long time until the popularity of convolutional neural networks.

Compared to conventional machine learning algorithms, deep learning-based methods have competitive advantages in high-level performance and learning ability. In many cases, deep multimodal fusion methods extend the unimodal algorithms with an effective fusion strategy. Namely, these fusion methods do not exist independently but derive from existing unimodal methods. The representative unimodal neural networks, such as VGGNet [164] and ResNet [70], are chosen as the backbone network for processing data in a holistic or separated manner. The initial attempt of deep multimodal fusion for image segmentation is to train the concatenated multimodal data on a single neural network [30]. We will present a detailed literature review of recent achievements in Chapter 3, covering various existing fusion methodologies and multimodal image datasets. Next, we point out three core challenges of deep multimodal fusion:

**Accuracy** As one of the most critical metrics, accuracy is commonly used to evaluate the performance of a learning system. Arguably, the architectural design and the quality of multimodal data have a significant influence on accuracy. How to optimally explore the complementary and mutually enriching information from multiple modalities is the first fundamental challenge.

**Robustness** Generally, we assume that deep multimodal models are trained under the premise of extensive and high-quality multimodal data input. However, multimodal data not only brings sufficient information but also brings redundancy and uncertainty. Ensuring network convergence can be a significant challenge with the use of redundant multimodal data. Moreover, sensors may behave differently or even in reverse during information collection. The poor performance of individual modality and the absence of modalities should be seriously considered.

**Real-time** In practical applications, multimodal fusion models need to satisfy specific requirements, including the simplicity of implementation, scalability, etc. These factors have a vital impact on the efficiency of the autonomous navigation system.

#### 1.2.4/ SCENARIOS AND APPLICATIONS

As one of the major challenges in scene understanding, deep multimodal fusion for semantic segmentation task cover a wider variety of scenarios. For instance, Hazirbas et al. [69] address the problem of pixel-level prediction of indoor scenes using color and depth data. Schneider et al. [156] present a mid-level fusion architecture for urban scene segmentation. Similar works in both indoor and outdoor scene segmentation can be found in [178]. Furthermore, the work by Valada et al. [176] led to a new research topic in scene understanding of unstructured forested environments. Considering non-optimal weather conditions, Pfeuffer and Dietmayer [137] investigated a robust fusion approach for foggy scene segmentation. As an illustration, Figure 1.3 shows the pipeline of deep multimodal image segmentation.

Besides the image segmentation task mentioned above, there are many other scene understanding tasks that benefit from multimodal fusion, such as object detection [40, 87, 10], human detection [120, 113, 185, 62], salient object detection [16, 189], trip hazard detection [119] and object tracking [219]. Especially for autonomous systems, LiDAR is always employed to provide highly accurate three-dimensional point cloud information [205, 81]. Patel et al. [134] demonstrated the utility of fusing RGB and 3D LiDAR data for autonomous navigation in the indoor environment. Moreover, many works adopting point cloud maps reported in recent years have focused on 3D object detection (e.g., [23, 136, 195]). It is reasonably foreseeable that deep multimodal fusion of homogeneous and heterogeneous information sources can be a strong emphasis for intelligent mobility [45, 194] in the near future.

#### 1.2.5/ FEW-SHOT SEMANTIC SEGMENTATION

Few-shot segmentation presents a significant challenge for semantic scene understanding under limited supervision. Namely, this task targets at generalizing the segmentation ability of the model to new categories given a few samples. Existing methods generally address this problem by learning a set of parameters or prototypes from **support** images and guiding the pixel-wise segmentation on the **query** image. However, few-shot learning comes with the problem of data imbalance. Small-scale data may lead to overfitting and insufficient model expression ability. Therefore in this thesis, we are committed to exploring effective few-shot segmentation models with advanced deep learning techniques such as multiscale feature aggregation and attention mechanism. We expect that the discriminative power, the generalizability, and the training efficiency of the segmentation model can be significantly improved. In addition, distinguishing objects with similar characteristics, especially when they are placed in an overlapping manner, is one of the most challenging tasks for few-shot segmentation. The model's ability to recognize irregular

objects and their boundary contours is also crucial.

Moreover, multimodal data such as depth maps are frequently used to provide rich geometric information of the scenes in fully-supervised semantic segmentation. Deep neural networks usually exploit the depth maps as an additional image channel or point cloud in 3D space. Arguably, the integration of supplementary scene information leads to significant performance improvement. However, existing methods focus on the unimodal few-shot segmentation. In our work, we take inspiration from existing RGB-centric methods for few-shot semantic segmentation and explore the effective use of depth information and fusion architecture for few-shot segmentation.

### 1.3/ CONTRIBUTIONS

Our contribution mainly consists of two parts, deep multimodal fusion for fully-supervised semantic image segmentation and semi-supervised semantic scene understanding. Regarding the former case, our contributions can be summarized as follows:

- I We propose a late fusion-based neural network for outdoor scene understanding. In particular, we introduce the first-of-its-kind dataset for multimodal image segmentation, which contains aligned raw RGB images and polarimetric images.
- II We present a novel multimodal fusion framework for semantic segmentation. The fusion model adaptively learns the joint feature representations of both low-level and high-level modality-specific feature via a central neural network and statistical post-processing.
- III We provide a systematic review of 2D/2.5D deep multimodal image segmentation in fusion methodology and dataset. We gather quantitative experimental results of multimodal fusion methods on different benchmark datasets, including their accuracy, runtime, and memory footprint.

In the case of semi-supervised semantic scene understanding, our contributions are:

- I We propose a novel few-shot segmentation method based on the prototypical network. The proposed network provides effective semantic guidance on the query feature by a multiscale feature enhancement module. The attention mechanism is employed to fuse the similarity-guided probability maps.
- II We present a two-stream deep neural network based on metric learning, which incorporates depth information into few-shot semantic segmentation. We also build a novel benchmark dataset, known as Cityscapes-3<sup>i</sup>, to evaluate the multimodal few-shot semantic image segmentation.

The different contributions have been published in the following papers:

- **Journal papers**

1. Yifei Zhang, Olivier Morel, Ralph Seulin, Fabrice Mériaudeau, Désiré Sidibé. "A Central Multimodal Fusion Framework For Outdoor Scene Image Segmentation", submitted to Multimedia Tools and Applications, 2020.
2. Yifei Zhang, Désiré Sidibé, Olivier Morel, Fabrice Mériaudeau. "Deep Multimodal Fusion for Semantic Image Segmentation: A Survey", Image and Vision Computing (2020): 104042.

- **Conference papers**

1. Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, Désiré Sidibé. "Exploration of Deep Learning-based Multimodal Fusion for Semantic Road Scene Segmentation". 14th International Conference on Computer Vision Theory and Applications (VISAPP 2019), Feb 2019, Prague, Czech Republic.
2. Yifei Zhang, Désiré Sidibé, Olivier Morel, Fabrice Mériaudeau. "Multiscale Attention-Based Prototypical Network For Few-Shot Semantic Segmentation", 25th International Conference on Pattern Recognition (ICPR 2020), Jan 2021, Milan, Italy.
3. Yifei Zhang, Désiré Sidibé, Olivier Morel, Fabrice Mériaudeau. "Incorporating Depth Information into Few-Shot Semantic Segmentation", 25th International Conference on Pattern Recognition (ICPR 2020), Jan 2021, Milan, Italy.

Other works and publications:

1. Marc Blanchon, Olivier Morel, Yifei Zhang, Ralph Seulin, Nathan Crombez, Désiré Sidibé. "Outdoor Scenes Pixel-Wise Semantic Segmentation using Polarimetry and Fully Convolutional Network". 14th International Conference on Computer Vision Theory and Applications (VISAPP 2019), Feb 2019, Prague, Czech Republic.

## 1.4/ ORGANIZATION

This thesis is divided into six chapters as follows:

- Chapter 2 introduces the fundamental knowledge in deep neural networks, including the basic network architecture, layers, optimization, model training, and evaluation metrics.

- Chapter 3 comprehensively studies the related works on multimodal image datasets for segmentation task and fully-supervised image segmentation methods.
- The proposed CMnet network and CMFnet+BF2 framework for outdoor scene image segmentation are presented in Chapter 4.
- Chapter 5 introduces the background knowledge on few-shot segmentation and presents MAPnet method for unimodal few-shot image segmentation as well as RDNet for multimodal outdoor scene few-shot segmentation.
- Finally, chapter 6 draws conclusions about our work and summarizes future perspectives.

# 2

## BACKGROUND ON NEURAL NETWORKS

*“Our intelligence is what makes us human, and AI is an extension of that quality.”*

– Yann Le Cun , *AI Scientist at Facebook*

**T**his chapter presents the necessary background knowledge of deep neural networks. The essential network architecture and its components are stated. As a branch of machine learning, deep learning is based on artificial neural networks, and it can also be seen as an imitation of the human brain. The increased processing power afforded by graphical processing units, the enormous amount of available data, and the development of more advanced algorithms has led to the rise of deep learning, which has made significant progress in the field of computer vision.

We start with the concept of convolution and introduce Convolutional Neural Networks (CNNs) as well as the encoder-decoder architecture for semantic image segmentation. Such a network structure is the basis for our design of deep multimodal fusion methods. We then show how a deep neural network comprises a series of consecutive layers and explains how these neural network layers function in detail. Several optimization methods, such as batch normalization and dropout, are also presented. Besides, we describe the common model training techniques, including data preprocessing, weight initialization, gradient descent optimization algorithms, and various loss functions. The mentioned background knowledge about deep learning technology provides theoretical support for our research methods. Finally, we discuss the evaluation metrics for evaluation and comparison of image segmentation algorithms.

## 2.1/ BASIC CONCEPTS

### 2.1.1/ CONVOLUTION

In the field of mathematics, **convolution** is a kind of operation, similar to addition and multiplication. In digital signal processing, we usually employ the convolution technique to combine two signals to form a third signal. Although the convolution operation used in convolutional neural networks is not exactly the same as the definition in mathematics, we first define what convolution is then explain how to use the convolution operation in convolutional neural networks.

In general, convolution is a mathematical operator that generates a third function  $s$  from two functions  $f$  and  $g$ . Formally,

$$s(t) = (f * g)(t) \quad (2.1)$$

In continuous space, convolution can be defined as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)d\tau \quad (2.2)$$

In discrete space, it can be written as:

$$(f * g)(n) = \sum_{k=-\infty}^{\infty} f(k)g(n - k) \quad (2.3)$$

According to Equations (2.2) and (2.3), we can see that convolution is the accumulation of the persistent consequences of instantaneous action. Therefore convolution is used as an effective method of mixing information, which is frequently applied in various fields such as signal analysis and image processing (e.g., image blur, edge enhancement).

In deep learning-based image processing, the function  $f$  is usually called **Input**, while the function  $g$  is called **Kernel function**. The output  $s$  is sometimes referred to as **Feature map**. The inputs are usually multidimensional arrays of data, and kernel functions are the parameters of multidimensional arrays optimized by learning algorithms. Suppose that we take a two-dimensional image  $I$  as input, and the two-dimensional kernel is  $G$ , then

$$S(i, j) = (I * G)(i, j) = \sum_m \sum_n I(m, n)G(i - m, j - n) \quad (2.4)$$

Because of the commutativity of convolution, Equation (2.4) can also be written as:

$$S(i, j) = (G * I)(i, j) = \sum_m \sum_n I(i - m, j - n)G(m, n) \quad (2.5)$$

At this point,  $G$  is equivalent to the learned filter, while the feature map  $S$  can be computed

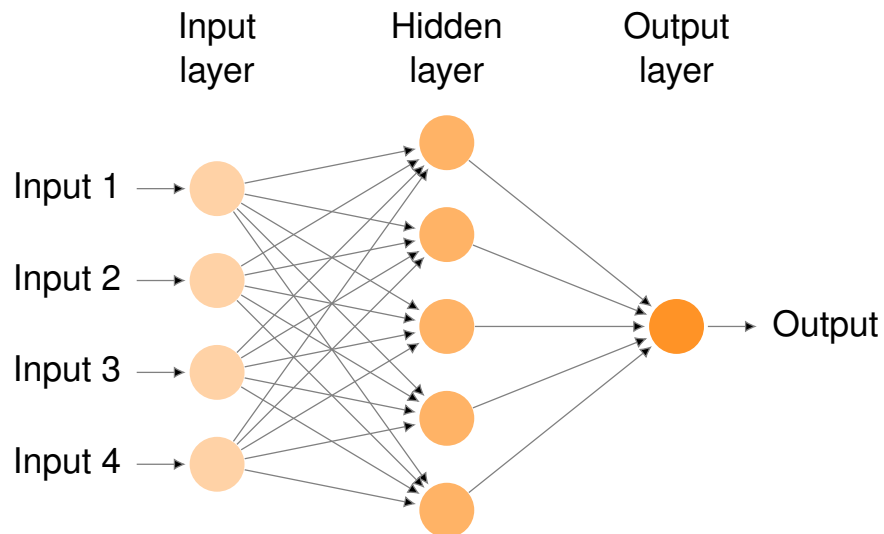


Figure 2.1: Example of a convolutional neural network with four inputs.

by matrix multiplication. These filters can filter out unwanted noise data and be sensitive to specific features. After training the learning algorithm, convolution kernels can filter out valuable information on the input image. This process is the key to the significant image processing capability of convolutional neural networks.

### 2.1.2/ CONVOLUTIONAL NEURAL NETWORKS

In recent years, benefiting from the rapid development of deep learning technology, the computer vision field has achieved unprecedented success. As one of the most widely used neural networks, **Convolutional neural networks (CNNs)** are the core learning algorithms for visual pattern recognition. They were developed from perceptrons, vector-mapping algorithms inspired by associative learning of the brain, and the idea of “integrate and fire” neurons. Researchers have employed CNNs and their variants (e.g., ResNet) to tackle a variety of challenges such as image classification, object recognition, action recognition, pose estimation, neural style transfer, etc. Previous studies have shown that they outperform humans in some recognition tasks.

Figure 2.1 illustrates a typical architecture of convolutional neural networks. In detail, CNNs are composed of multiple neural units, which can be generally divided into three types, namely, the **input layer**, the **hidden layer** and the **output layer**. The input layer of a convolutional neural network is mainly used to obtain input information, which can process multidimensional data. For example, one-dimensional data is usually time or sampled spectrum, while two-dimensional data may contain multiple channels, such as a three-channel color image, and three-dimensional data may be 3D images such as CR and MRI image. In particular, we focus on 2D/2.5D multimodal images in this thesis.



The output layer of a convolutional neural network usually employs a logical function or softmax function to output the classification labels. The practical application varies according to the type of task. For instance, the output layer can be the central coordinates, size, and classification of objects in object detection. In semantic image segmentation, the output layer directly outputs the classification label of each pixel.

In general, each neural unit in the input layer directly connects to the original data, and provides feature information to the hidden layer. Each neural unit in the hidden layer represents different weights for different neural units in the input layer, so it tends to be sensitive to a certain recognition pattern. The values in the output layer vary according to the activation degree of hidden layers, which is the final recognition result of the model. Compared with the input layer and output layer, the hidden layer is more complex because it is designed for abstract feature extraction. It usually includes the convolutional layer, pooling layer and fully connected layer. In Section 2.2, we introduce the common layers in CNNs.

### 2.1.3/ ENCODER-DECODER ARCHITECTURE

A convolutional encoder-decoder network is a standard architecture used for tasks requiring dense pixel-wise predictions. Such neural network design pattern is frequently used in semantic image segmentation, which is partitioned into two parts, the **encoder** and the **decoder**. Figure 2.2 shows a typical example of the encoder-decoder architecture.

In general, an encoder takes the image input and progressively computes higher-level abstract features. The role of the encoder is to encode the low-level image features into a high-dimensional feature vector. The spatial resolution of the feature maps is reduced progressively via the down-sampling operation. Multiple common backbone networks such as VGG, Inception, and ResNet, can be employed for abstract feature extraction in the encoder. The extracted semantic information is then passed into the decoder to compute feature maps of progressively increasing resolution via un-pooling or up-sampling. The decoder restores the learned valuable feature representation into a pixel-level segmentation mask, which is one of the reasons why the encoder-decoder architecture can be effectively used for image segmentation tasks.

Besides, different variations of the encoder-decoder architecture have been explored to improve the segmentation performance. To name a few, skip connections [150] have been used to recover the fine spatial details during decoding which get lost due to successive down-sampling operations involved in the encoder. Moreover, larger context information using image-level features [115], recurrent connections [139, 215], and larger convolutional kernels [135] has also significantly improved the accuracy of semantic segmentation.

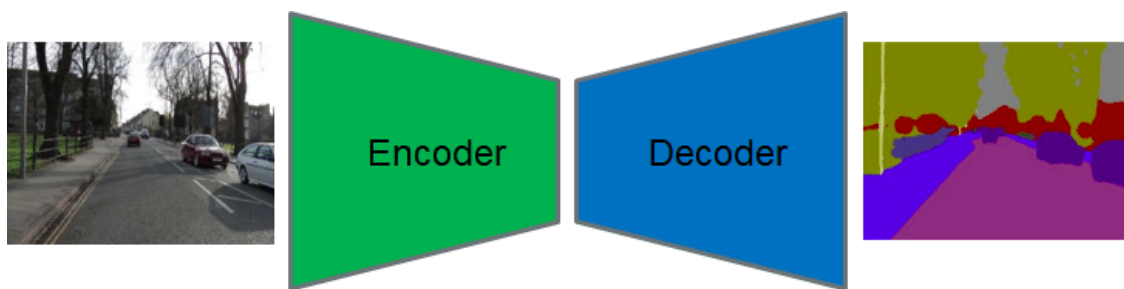


Figure 2.2: Example of the encoder-decoder architecture.

## 2.2/ NEURAL NETWORK LAYERS

### 2.2.1/ CONVOLUTION LAYER

Convolutional layers are the primary building blocks used in convolutional neural networks. The convolution as a filter enables the neural network to extract effective high-level features. The feature map, also called the activation map, can be generated by repeatedly applying the same filter, which indicates the locations and strength of detected features in the input image. The filter contains the weights that must be learned during the training of the layer. Moreover, the filter size or kernel size will significantly affect the shape of the output feature map. It is worth noting that the interaction of the filter with the border of the image may lead to border effects, especially for the small size input image and very deep network. Usually, we can fix the border effect problem by adding extra pixels to the edge of the image, which is called **padding**. Besides, the amount of movement between the filter applications to the input image is referred to as the **stride**, and it is almost always symmetrical in height and width dimensions. For example, the stride (2, 2) means moving the filter two pixels right for each horizontal movement of the filter and two pixels down for each vertical movement of the filter when creating the feature map. The stride of the filter on the input image can be seen as the downsampling of the output feature map.

Next, we introduce three essential properties of the convolutional layers: sparse interactions, parameter sharing, and equivariant representations.

**Sparse interactions** Convolutional neural networks have sparse interactions by making the kernel smaller than the input. When processing an image with thousands of pixels, we can detect small valuable features such as the edges of the image by taking only tens to hundreds of pixels. This not only reduces the storage requirements of the model but also improves its overall computation efficiency.

**Parameter sharing** Parameter sharing is the sharing of weights by all neurons in a particular feature map. Each neuron is connected only to a subset of the input image, which

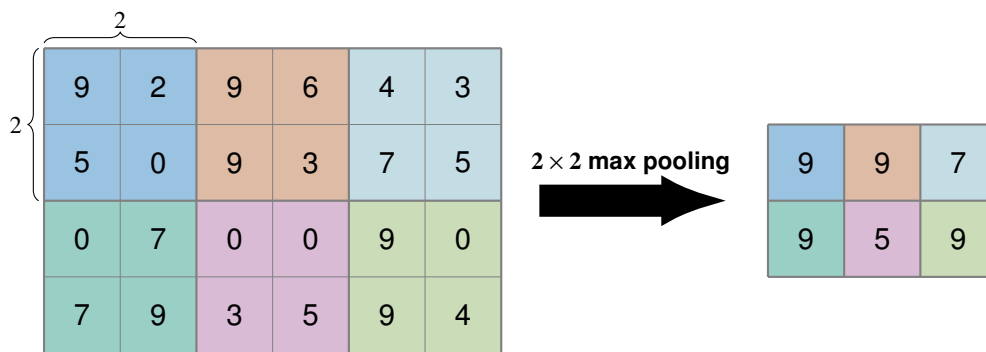


Figure 2.3: Example of max pooling operation.

is also called local connectivity. This property helps to reduce the number of parameters in the whole system and makes the computation more efficient.

**Equivariant representations** For convolution operation, parameter sharing makes the neural network layers have an equivariant representation. Namely, if the input is slightly shifted, the result of the convolution operation is the same. Note that the convolution is not naturally equivalent to some other transformations such as image scaling or rotation transformation.

### 2.2.2/ POOLING LAYER

Pooling operation plays a vital role in the structure of convolutional neural networks. First of all, pooling layers improve the spatial invariance to some extent, such as translation invariance, scale invariance, and deformation invariance. Namely, even if the image input is transformed slightly, the pooling layer can still produce similar pooling features, making the learning system more robust. Secondly, pooling operation is equivalent to feature downsampling, which increases the receptive field size. For some visual tasks, a large receptive field helps learn long-range spatial relationships and implicit spatial models. In addition, pooling operation greatly reduces the model parameters, which leads to a lower risk of overfitting. Suppose that the dimension of the image input is  $c \times w \times h$ , where  $c$  is the number of channels, and  $w$  and  $h$  are the width and height, respectively. If the stride of the pooling layer is set to 2, the dimension of the output image will be  $c \times w/2 \times h/2$ . In this case, both the computational cost and memory consumption will be reduced by a factor of 4.

Common pooling methods include **average pooling** and **maximum pooling**. Maximum pooling calculates the maximum value of the target patch, which retains more texture information of the image input, whereas average pooling keeps more background information and tends to transfer the comprehensive information in the architecture of convo-

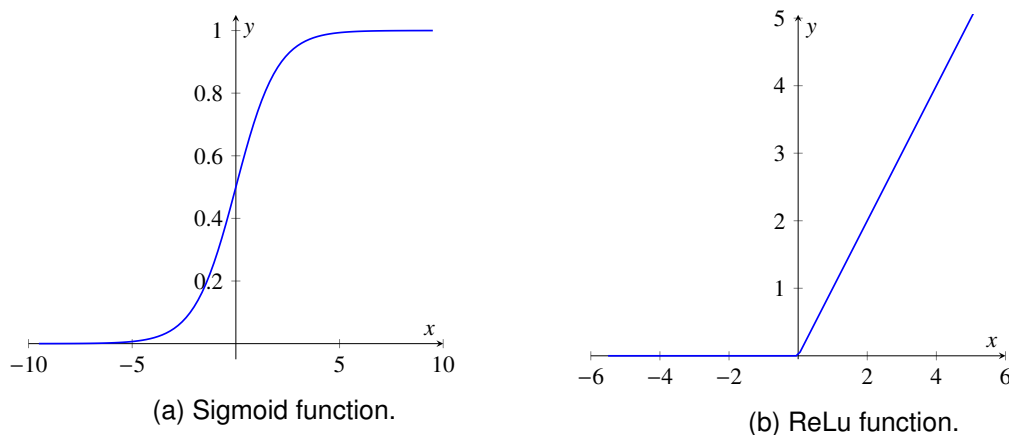


Figure 2.4: Examples of popular activation functions.

lutional neural networks. Figure 2.3 shows an example of  $2 \times 2$  maximum pooling.

### 2.2.3/ RELU LAYER

In neural networks, the activation function performs the nonlinear transformation to the input, making it capable of learning and performing more complex tasks. In order to increase the nonlinearity of neural networks, some nonlinear functions are introduced. Obviously, the accumulation of multiple linear functions is still linear, while linear functions have limited expression. The use of nonlinear functions makes the network more expressive and thus better fits the target function. Two common nonlinear functions used in convolutional neural networks are the **sigmoid function** and the **rectified linear unit (ReLU)**.

As shown in Figure 2.4, we can observe that the ReLU activation function has more advantages than the sigmoid function. ReLU can carry out negative suppression so as to be more sparsely active. More importantly, ReLU activation functions suffer less from the vanishing gradient problem. The derivative of the sigmoid function has good activation only when it is near zero. The gradient in the positive and negative saturation region is close to zero. Also, the derivative of the ReLU function is easy to calculate, which can accelerate the model training to some extent.

### 2.2.4/ FULLY CONNECTED LAYER

The fully connected layer is usually used as a classifier to connect the hidden layer and the final output. In the architecture of convolutional neural networks, adding several fully connected layers after the convolution layers can map the generated feature map into a fixed-length feature vector. The final output represents the numerical description of the input image. This structural property is conducive to the realization of image-level

classification and regression tasks.

Although multiple fully connected layers can significantly improve the nonlinear expression ability of learning models, a large number of neurons increase the model complexity. Plenty of model parameters will reduce the efficiency of the learning algorithm and even lead to overfitting. Therefore, the trade-off between accuracy and efficiency has been deeply explored in deep learning technology research. For the segmentation task, however, spatial information should be stored to make a pixel-wise classification. Hence, the fully connected layer is usually substituted by another convolution layer with a large receptive field.

## 2.3/ OPTIMIZATION

### 2.3.1/ BATCH NORMALIZATION

Training deep neural networks with multiple hidden layers is quite challenging. One reason is that the model is updated layer-by-layer backward from the output to the input using an error estimate that assumes the weights in the layers prior to the current layer are fixed. This slows down the training by requiring lower learning rates and careful parameter initialization and makes it notoriously hard to train models with saturating nonlinearities [79]. Therefore batch normalization, as an effective optimization technique, is proposed to standardize the inputs to a layer for each mini-batch while training very deep neural networks. It stabilizes the learning process and dramatically reduces the number of training epochs required to train deep networks.

Batch normalization can be implemented during training by calculating the mean and standard deviation of each input variable to a layer per mini-batch and using these statistics to perform the standardization. Alternatively, a running average of mean and standard deviation can be maintained across mini-batches but may result in unstable training. For example, He et al. [72] used batch normalization after the convolutional layers in their very deep model, referred to as ResNet. The reported results achieved state-of-the-art in the image classification task. In our work of model design, we usually add batch normalization transformation before nonlinearity.

### 2.3.2/ DROPOUT

Deep neural networks are likely to get overfitting while training with few examples. As early as 2012, Hinton et al. [74] has proposed the concept of dropout, which is now widely used in advanced neural networks. Probabilistically dropping out nodes in the network is a simple and effective regularization method [169]. In each iteration, some nodes are

randomly deleted, and only the remaining nodes are trained. This optimization method reduces the correlation between nodes and the complexity of the model to achieve the effect of regularization.

Generally, dropout only needs to set a hyper-parameter that is the proportion of nodes randomly preserved in each layer. Namely, the parameter matrix of this layer is calculated with the binary matrix generated by the hyper-parameter via the point by point product. Suppose that the hyper-parameter is set to 0.7, then 30% of the nodes will be randomly ignored and produce no outputs from the layer. A good value for dropout in a hidden layer is between 0.5 and 0.8. Input layers use a larger dropout rate, such as of 0.8.

## 2.4/ MODEL TRAINING

The process of training neural networks is the most challenging part of using deep learning techniques and is by far the most time consuming, both in terms of effort required for configuration and computational complexity required for execution. In the following, we summarize the commonly used techniques in model training, including data preprocessing, weight initialization, loss function and gradient descent optimization.

### 2.4.1/ DATA PREPROCESSING

Generally, training deep learning models requires a lot of data because of the huge number of parameters needed to be tuned by the learning algorithm. Data preprocessing is a prerequisite guarantee for effectively model training, and we need to be careful to prepare the training data to achieve the best prediction results. For example, many deep learning models have normalized input processing, namely whitening operation, which changes the average pixel value of the image to zero and the variance of the image to unit variance. In detail, the mean and variance of the original image are first calculated, then each pixel value of the original image is transformed. This operation enables the convergence of the neural network faster. Common data preprocessing methods also include data quality assessment, feature aggregation, feature sampling, dimensionality reduction, and feature encoding.

Besides, data augmentation [160] is frequently used in model training, which increases the amount of input data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. In the real world scenarios, we may have a small-scale dataset of images taken in a limited set of conditions. In the case of limited data, data augmentation can increase the diversity of training samples, so as to improve the robustness of the model and avoid overfitting. Typical operations include flipping, rotation, shift, resize, random scale, random crop, color jittering, contrast, noise,

fancy PCA, GAN, etc.

### 2.4.2/ WEIGHT INITIALIZATION

Training a deep learning model means learning good values for all the weights and the bias from labeled examples. In particular, the bias allows to shift the activation function by adding a constant. In order to consistently update the weights, the models require each parameter to have the corresponding initial value. For convolutional neural networks, the nonlinear function is superimposed by multiple layers, and how to select the initial value of parameters becomes a problem worthy of discussion.

In general, the purpose of weight initialization [57, 121] is to prevent the layer activation output from exploding or disappearing in the forward transfer process of deep neural networks. In either case, the loss gradient is either too large or too small to flow backward advantageously. The learning model will then take a longer time to converge. Also, it is notable that initializing all the weights with zeros leads the neurons to learn the same features during training. The model can not get the update of parameters correctly. For example, assume we initialize all the biases to zero and the weights with some constant  $\alpha$ . If we forward propagate an input  $(x_1, x_2)$  in the network, the output of hidden layers will be  $\text{relu}(\alpha x_1 + \alpha x_2)$ . Namely, the hidden layers will have an identical influence on the cost, which will result in identical gradients.

In practice, researchers usually employ the Xavier initialization [57] to keep the variance the same across every layer. Another common initialization is He initialization [71], in which the weights are initialized by multiplying by two the variance of the Xavier initialization.

### 2.4.3/ LOSS FUNCTION

The neural networks are usually trained using the gradient descent optimization algorithm. Generally, the optimization problem involves an objective function that indicates the direction of optimization. During the optimization process, the network tries to find a candidate solution to maximize or minimize the objective function. Under constraint conditions, we calculate and minimize the model error via a loss function or cost function, which evaluates the fit of the learning model. This series of constraints is a regularization term that helps to prevent overfitting. The weights are updated using the backpropagation of error algorithm. Therefore we need to choose a suitable loss function when designing and configuring the model.

Suppose that there is a series of training samples  $\{(x_i, y_i)\}_{i=1, \dots, N}$  in supervised learning. The model learns the mapping relation of  $x \rightarrow y$ , so that given a  $x$ , even if the  $x$  is not

in the training samples, it can get the output  $\hat{y}$  as close to the real  $y$  as possible. The loss function is a key component to indicate the direction of model optimization, which is used to measure the difference between the output  $\hat{y}$  of the model and the real output  $y$ , namely,  $L = f(y_i, \hat{y}_i)$ .

In the following, we introduce several loss functions commonly used for classification and regression in deep learning, including mean squared error loss, mean absolute error loss and cross-entropy loss.

\* **Mean Squared Error Loss**

Mean Squared Error Loss (MSE), also known as Quadratic Loss or  $L_2$  Loss, is the most commonly used loss function in machine learning and deep learning regression tasks. Formally,

$$J_{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (2.6)$$

Under the assumption that the error between the output of the model and the real value follows a Gaussian distribution, the minimum mean square error loss function and the maximum likelihood estimate are essentially consistent. Therefore, in the scenario where the assumption can be satisfied (such as regression), the mean square error loss is a good choice for the loss function. In scenarios where this assumption is not satisfied (such as classification), other losses have to be considered.

\* **Mean Absolute Error Loss**

Mean Absolute Error Loss (MAE), also known as  $L_1$  Loss, is another common loss function. MSE loss generally converges faster than MAE loss, however the latter is more robust to outlier, which can be defined as:

$$J_{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (2.7)$$

\* **Cross-Entropy Loss**

Previous loss functions introduced above are applicable to regression problems. For classification tasks, especially for semantic image segmentation, the most commonly used loss function is the cross-entropy loss that increases as the predicted probability diverges from the actual label. For the multi-class cross-entropy loss, we can obtain:

$$J_{CE} = - \sum_{i=1}^N y_i^{c_i} \log(\hat{y}_i^{c_i}) \quad (2.8)$$

Here,  $y_i^{c_i}$  can be 0 or 1, indicating whether class label  $c_i$  is the correct classification.



### 2.4.4/ GRADIENT DESCENT

A neural network is merely a complicated function, consisting of millions of parameters, that represents a mathematical solution to a problem. By optimizing various parameters of the network, the trained model is finally matched to the learning task under a certain parameter set. The gradient descent method is one of the most commonly used methods to achieve such learning process. In brief, the gradient descent method, also referred to as steepest descent method, is a method to find the minimum of the objective function.

From a mathematical point of view, the gradient indicates the direction of the fastest growth of the function. In other words, the opposite direction of the gradient is the direction of the fastest function decrease. Once we have the direction to move in, we need to decide the size of the step we take. The size of this step is called the learning rate. It is worth noting that if the value of the learning rate is too large, we might overshoot the minima, and keep bouncing along the ridges of the "valley" without ever reaching the minima. And small learning rate might lead to painfully slow convergence, even get stuck in a minima. The closer the gradient descent method is to the target value, the smaller the step size is and the slower the progress is.

There are three variants of gradient descent, including batch gradient descent, stochastic gradient descent, and mini-batch gradient descent. These variants differ in how much data we use to compute the gradient of the objective function. In detail, batch gradient descent computes the gradient of the cost function for the entire training dataset. Stochastic gradient descent in contrast performs a weight update for each training example and the corresponding label. As for mini-batch gradient descent, it performs an update for every mini-batch of  $n$  training examples. Taking batch gradient descent as an example, the weight update process can be defined as:

$$\theta := \theta - \alpha \nabla_{\theta} J(\theta) \quad (2.9)$$

where  $J(\theta)$  is the objective function parameterized by the weights  $\theta$  of the learning model. The parameters is updated in the opposite direction of the gradient of the objective function  $\nabla_{\theta} J(\theta)$ . The learning rate  $\alpha$  determines the size of the steps we take to reach a minimum.

Moreover, several gradient descent optimizers have been proposed, including Momentum SGD, AdaGrad, RMSProp, Adam, etc. For example, Momentum SGD [142] is a method that helps accelerate stochastic gradient descent in the relevant direction and dampens oscillations. It adds a fraction  $\gamma$  of the update vector of the past time step to the current update vector.

$$v_t = \gamma v_{t-1} + \alpha \nabla_{\theta} J(\theta) \quad (2.10)$$

Then the weights can be updated by  $\theta := \theta - v_t$ . The momentum term  $\gamma$  increases for dimensions whose gradients point in the same directions and reduces updates for dimensions whose gradients change directions. As a result, we gain faster convergence and reduced oscillation. Besides, AdaGrad [36] is an alternative algorithm for gradient-based optimization. It adapts the learning rate to the parameters, performs smaller updates for parameters associated with frequently occurring features, and larger updates for infrequent features. Adaptive Moment Estimation (Adam) [90] is a method that computes adaptive learning rates for each parameter. In addition to storing an exponentially decaying average of past squared gradients, Adam also keeps an exponentially decaying average of past gradients, similar to momentum. Each algorithm has its advantages and disadvantages, we can choose the appropriate optimizer according to the input data and training requirements.

## 2.5/ EVALUATION METRICS

The segmentation performance is generally affected by many factors, such as the pre-processing of data, fusion strategy, the choice of the backbone network, the practice of state-of-the-art deep learning technologies, etc. Therefore how to evaluate and compare the performance of segmentation algorithms is a critical issue. The validity and usefulness of a learning system can be measured in many aspects, such as **execution time**, **memory footprint**, and **accuracy**. It is notable that existing large-scale benchmark datasets promote the standardization of comparison metrics, providing a fair comparison of the state-of-the-art methods. In Chapter 3.2.3, we provide a comparative analysis of existing deep multimodal segmentation methods in terms of common metrics.

In general, accuracy is the most common evaluation criteria to measure the performance of pixel-level prediction [46]. For multimodal image segmentation, the most popular metrics are not different from those used in unimodal approaches, including Pixel Accuracy (PA), Mean Accuracy (MA), Mean Intersection over Union (MIoU), and Frequency Weighted Intersection over Union (FWIoU), which are first employed in [117]. In our work, we mainly report a series of segmentation results in Intersection over Union (IoU), also known as the Jaccard Index. The IoU for each category is defined as  $IoU = TP / (TP + FP + FN)$ , where TP, FP and FN denote true positives, false positives and false negatives, respectively.

For the sake of explanation, we denote  $n_{ij}$  as the number of pixels belonging to class  $i$  which are classified into class  $j$ , and we consider that there are  $n_{cl}$  classes, and  $t_i = \sum_j n_{ij}$  is the numbers of pixel in class  $i$ . Therefore we can define these accuracy metrics as follows:

- Pixel Accuracy

$$\sum_i n_{ii} / \sum_i t_i$$

- Mean Accuracy

$$(1/n_{cl}) \sum_i n_{ii} / t_i$$

- Mean Intersection over Union

$$(1/n_{cl}) \sum_i n_{ii} / (t_i + \sum_j n_{ji} - n_{ii})$$

- Frequency Weighted Intersection over Union

$$(\sum_k t_k)^{-1} \sum_i t_i n_{ii} / (t_i + \sum_j n_{ji} - n_{ii})$$

Besides, researchers usually evaluate the real-time performance of autonomous navigation systems by measuring the execution time and memory consumption. Although these indicators are usually closely related to hardware settings, they are also essential for model optimization.

## 2.6/ SUMMARY

This chapter reviewed some general concepts in deep learning, including convolution operations and essential neural network layers. This background knowledge laid the foundations for building a compact and efficient convolutional neural network. Especially for the semantic segmentation task, we introduced the encoder-decoder architecture for dense pixel-wise predictions. Besides, several optimization techniques, such as batch normalization and dropout, were stated in detail. Then we presented several standard techniques for neural network training, including data preprocessing, weight initialization, loss function, and gradient descent. Advanced model training strategies were summarized in the corresponding section. We highlighted their advantages and disadvantages, which offers the design choice and facilitates the model design in our experiments. In the last section, we provided a brief overview of image segmentation evaluation metrics, followed by their mathematical formulations.

## LITERATURE REVIEW

**D**uring the long history of computer vision, semantic image segmentation is one of the grand challenges, which involves labeling each pixel of the image into a predefined set of classes. Recent advances in deep learning technologies, especially deep convolutional neural networks (CNNs), have led to a significant improvement over traditional semantic segmentation methods. However, in some complex environments or under challenging conditions, it is necessary to employ multiple modalities that provide complementary information on the same scene. A variety of studies have demonstrated that deep multimodal fusion for semantic image segmentation achieves great performance gains. These fusion approaches take the benefits of multiple information sources and generate an optimal joint prediction automatically.

This chapter provides a systematic review of deep multimodal fusion methodologies, with the highlight of their contributions to model design. In detail, existing fusion methods are summarized according to a common taxonomy: early fusion, late fusion, and hybrid fusion. We also conduct a comprehensive survey of current semantic segmentation datasets, as well as the potential multimodal datasets. Multiple image data types are analyzed, such as depth images, Near-InfraRed images, thermal images, and polarization images. Moreover, we gather quantitative experimental results of multimodal fusion methods on different benchmark datasets, including their accuracy, runtime, and memory footprint. Based on their performance, we analyze the strengths and weaknesses of different fusion strategies. Current challenges and design choices are discussed, aiming to provide the reader with a comprehensive and heuristic view of deep multimodal image segmentation. Besides, we review existing few-shot semantic image segmentation methods as the preliminary work of Chapter 5.

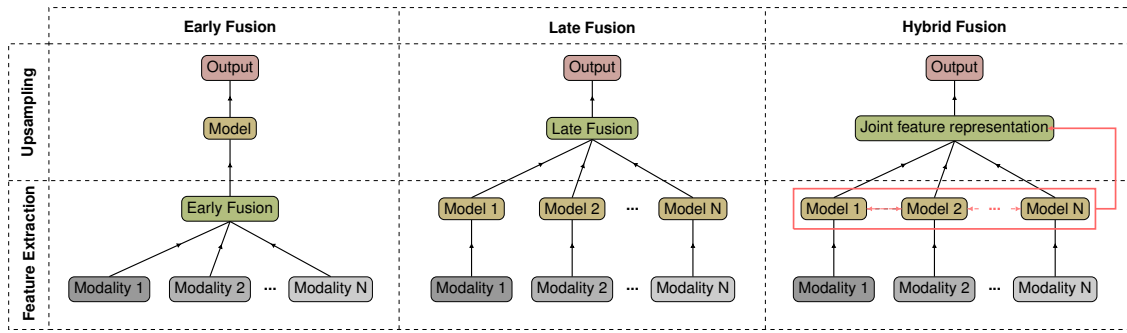


Figure 3.1: An illustration of different fusion strategies for deep multimodal learning.

### 3.1/ FULLY-SUPERVISED SEMANTIC IMAGE SEGMENTATION

In this subsection, we provide a comprehensive review of deep multimodal fusion methods according to our taxonomy. We highlight their benefits and drawbacks, providing interested readers with a complete overview of deep fusion strategies.

#### 3.1.1/ TAXONOMY OF DEEP MULTIMODAL FUSION

In the early works [3, 205, 114], the classification of multimodal fusion strategies involves various taxonomic methods, including data fusion, early fusion, late fusion, intermediate fusion, and hybrid fusion. In this review, we explicitly divide the deep multimodal fusion methods into early fusion, late fusion, and hybrid fusion, according to the fusion stage and motivation (see Figure 3.1).

Early fusion methods involve raw data-level fusion and feature-level fusion. The initial attempt of early fusion is to concatenate the raw data from different modalities into multiple channels. The learning model can be trained end-to-end using an individual segmentation network. Almost all the state-of-the-art segmentation networks are adaptable for such fusion strategy. Moreover, cross-modal interactions throughout the encoding stage, namely feature-level fusion, is also a distinctive manifestation of early fusion. For the sake of explanation, we denote the single segmentation network as  $I$ ,  $(x_1, x_2, \dots, x_n)$  is a set of  $n$  modalities as input, then the final prediction  $y$  can be defined as:

$$y = I(x_1, x_2, \dots, x_n). \quad (3.1)$$

On the contrary, late fusion methods aim to integrate multimodal feature maps at decision-level. More precisely, late fusion separately processes the multimodal data in different branches. During the decoding stage, all the feature maps computed by branches are mapped into a common feature space via fusion operations (e.g., concatenation, addition, averaging, weighted voting, etc.) [45], followed by a series of convolutional layers.

Table 3.1: Typical early fusion methods reviewed in this chapter.

Ref.	Method	Backbone	Contribution(s)	Year	Source Code
[30]	Coupric <sup>7</sup>	-	Initial attempt	2013	Available
[69]	FuseNet	VGG-16	Dense fusion/Sparse fusion	2016	Available
[118]	MVCNet	VGG-16	Multi-view consistency	2017	-
[78]	LDFNet	VGG-16	D&Y Encoder	2018	Available
[32]	RFBNet	AdapNet++	Bottom-up interactive fusion structure	2019	-
[76]	ACNet	ResNet-50	Multi-branch attention based network	2019	Available
[172]	RTFNet	ResNet-152	RGB-Thermal fusion with Upception blocks	2019	Available

Besides, we consider the common feature learned by the transformation network as a further refinement of decoding and prediction, some conventional intermediate fusion approaches (e.g., [187]) are therefore categorized into late fusion strategy in this review. Suppose that the segmentation networks  $(I_1, \dots, I_n)$  are used to process the multimodal data  $(x_1, x_2, \dots, x_n)$  from different modalities, and  $P$  is the fusion operation as well as the following convolutional layers, the final output  $y$  can be formulated as:

$$y = P(I_1(x_1), I_2(x_2), \dots, I_n(x_n)). \quad (3.2)$$

Hybrid fusion methods are elaborately designed to combine the strengths of both early and late fusion strategies. Generally, the segmentation network accesses the data through the corresponding branch. Then more than one extra module is employed to compute the class-wise or modality-wise weights and bridge the encoder and decoder with skip connections. Therefore the hybrid fusion networks can adaptively generate a joint feature representation over multiple modalities, yielding a better performance in terms of accuracy and robustness.

Based on such common taxonomy of fusion strategy, we systematically review the existing deep multimodal fusion networks for semantic image segmentation in the following sections.

### 3.1.1.1/ EARLY FUSION

The first attempt at deep multimodal fusion was made by Coupric et al. [30] in 2013. This work presents an early fusion strategy via a simple concatenation of RGB and depth channels before feeding into a segmentation network. In the case of similar depth appearance and location, this method shows positive results for indoor scene recognition. However, the simple concatenation of images provides limited help in multimodal feature extraction. The high variability of depth maps, to a certain extent, increase the uncertainty of feature learning.

To further explore semantic labeling on RGB-D data, FuseNet [69] was proposed in 2016

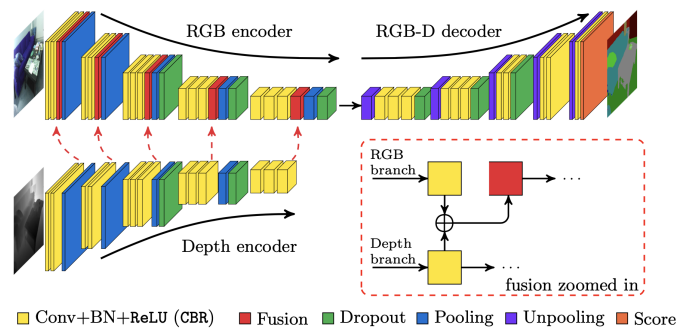


Figure 3.2: FuseNet architecture with RGB-D input. Figure reproduced from [69].

(see Figure 3.2). FuseNet is a clear example of incorporating the auxiliary depth information into an encoder-decoder segmentation framework. The abstract features obtained from the depth encoder are simultaneously fused to the RGB branch as the network goes deeper. Motivated by FuseNet, Ma et al. [118] proposed MVCNet to predict multi-view consistent semantics. Then Hung et al. [78] presented LDFNet that contains a well-designed encoder for the non-RGB branch, aiming to fully make use of luminance, depth, and color information. Recently, RFBNet [32] was proposed with an efficient fusion mechanism that explores the interdependence between the encoders. The Residual Fusion Block (RFB), which consists of two modality-specific residual units (RUs) and one gated fusion unit (GFU), was employed as the basic module to achieve the interactive fusion in a bottom-up way. Hu et al. [76] proposed a novel early fusion architecture based on attention mechanism, known as ACNet, which selectively gathers valuable features from RGB and depth branches. Besides, RTFNet [172] was particularly designed to fuse both RGB and thermal images by element-wise summation. Notably, average pooling and the fully connected layers in the backbone network was removed to avoid the excessive loss of spatial information.

### 3.1.1.2/ LATE FUSION

As early as 2014, Gupta et al. [63] proposed a geocentric embedding for object detection and segmentation. The authors employed two convolutional neural network streams to extract RGB and depth features, respectively. The feature maps obtained from these two streams are combined by SVM classifier at the late stage. Then the work by Li et al. [104] addresses semantic labeling of RGB-D scenes by developing a Long Short-Term Memorized Context Fusion (LSTM-CF) model. This network captures photometric and depth information in parallel, facilitating deep integration of contextual information. The global contexts and the last convolutional features of the RGB stream are fused by simple tensor concatenation.

Besides, Wang et al. [187] proposed a feature transformation network for learning the

Table 3.2: Typical late fusion methods reviewed in this chapter.

Ref.	Method	Backbone	Contribution(s)	Year	Source Code
[63]	Gupta'	-	CNN+SVM	2014	Available
[104]	LSTM-CF	Deeplab	LSTM-based context fusion	2016	Available
[176]	LFC	VGG-16	Late-fused convolution	2016	Available
[187]	Wang'	VGG-16	Feature transformation network	2016	-
[177]	CMoDE	AdapNet	Class-wise adaptive gating network	2017	Available
[25]	LSD-GF	VGG-16	Locality-sensitive DeconvNet with gated fusion	2017	-

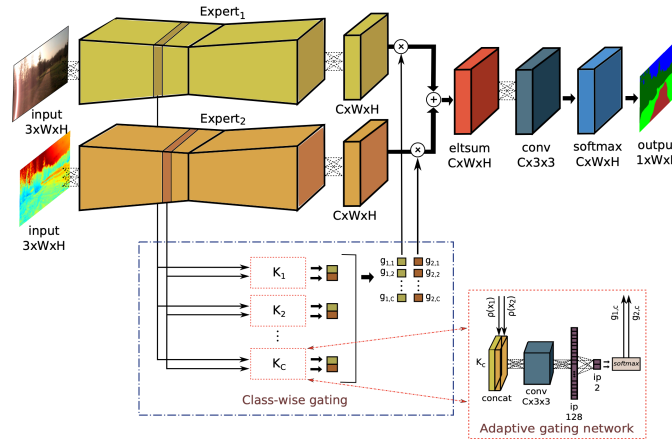


Figure 3.3: Convolved Mixture of Deep Experts framework. Figure extracted from [177].

common features between RGB and depth branches. This fusion structure bridges the convolutional networks with the deconvolutional networks by sharing feature representation. Another typical late fusion network, mentioned as LFC, was presented by Valada et al. [176]. This fusion architecture separately extracts multimodal features on the corresponding branch. The computed feature maps are summed up for joint representation, followed by a series of convolutional layers. Afterward, the authors extended the LFC method with a convolved mixture of deep expert units, referred to as CMoDE [177]. This deep fusion framework was inspired by the work [80, 39], in which multimodal features are mapped to a particular subspace. An adaptive gating subnetwork is employed to produce class-wise probability distribution over the experts (see Figure 3.3). In the work of LSD-GF, Cheng et al. [25] proposed a gated fusion module to adaptively merge RGB and depth score maps according to their weighted contributions.

### 3.1.1.3/ HYBRID FUSION

Previous studies have shown that simply concatenating multimodal features or fusing weighted feature maps at decision level may not be sufficient to meet the requirements of highly accurate and robust segmentation. The hybrid fusion strategy is proposed to combine the strengths of early fusion and late fusion as an alternative method.

In the early stages of hybrid fusion, Park et al. [130] extended the core idea of resid-



Table 3.3: Typical hybrid fusion methods reviewed in this chapter.

Ref.	Method	Backbone	Contribution(s)	Year	Source Code
[130]	RDFNet	ResNet-152	Extension of residual learning	2017	Available
[84]	DFCN-DCRF	VGG-16	Dense-sensitive FCN/ Dense-sensitive CRF	2017	Available
[102]	S-M Fusion	VGG-16	Semantics-guided Multi-level feature fusion	2017	-
[107]	CFN	RefineNet-152	Context-aware receptive field/ Cascaded structure	2017	-
[85]	RedNet	ResNet-50	Residual Encoder-Decoder structure	2018	Available
[178]	SSMA	AdapNet++	self-supervised model adaptation fusion mechanism	2019	Available

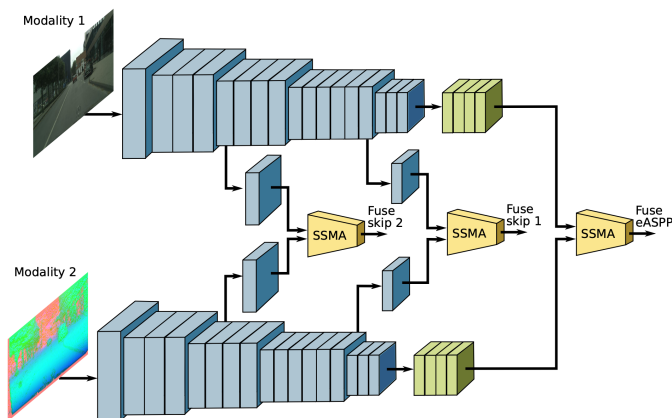


Figure 3.4: Fusion architecture with self-supervised model adaptation modules. Figure extracted from [177].

ual learning to deep multimodal fusion. This method, known as RDFNet, can effectively combine RGB-D features for high-resolution prediction through multimodal feature fusion blocks and multi-level feature refinement blocks. Afterward, Jiang et al. [84] introduced a fusion structure combining a fully convolutional neural network of RGB-D (DFCN) and a depth-sensitive fully-connected conditional random field (DCRF). The DFCN module can be considered as an extension of FuseNet, while the DCRF module is used to refine the preliminary prediction. CFN is a cascaded feature network introduced by Lin et al. [107]. The feature maps generated by the RGB branch are used to match the image regions to complementary branches. Experimentally, the use of context-aware receptive field (CaRF) enables the fusion network to achieve a competitive segmentation result. Additionally, semantics-guided multi-level fusion [102], referred to as S-M Fusion, was proposed to learn the feature representation in a bottom-up manner. This fusion strategy employed the cascaded Semantics-guided Fusion Block (SFB) to fuse lower-level features across modalities sequentially.

Moreover, Jiang et al. [85] described a residual encoder-decoder network for RGB-D semantic segmentation, named RedNet. The complementary features are fused into the RGB branch before upsampling. The skip-connection was used to bypass the spatial feature between the encoder and decoder. Instead of VGG, the residual module was applied as the basic building block. A more recent method addressed the issue of deep multimodal fusion using a Self-Supervised Model Adaptation module (SSMA) [178]. This

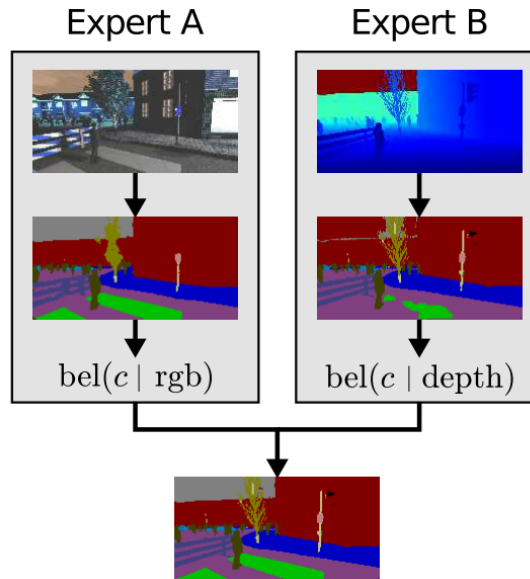


Figure 3.5: Individual semantic segmentation experts are combined modularly using different statistical methods. Figure extracted from [13]

fusion framework dynamically adapts the fusion of semantically mature multiscale representations. The latent joint representation generated from the SSMA block is integrated into decoder by two skip connections (see Figure 3.4). Arguably, the SSMA blocks enable the fusion model to exploit complementary cues from each modality-specific encoder, notably enhancing the discriminative power of feature representation.

#### 3.1.1.4/ STATISTICAL FUSION

As an alternative post-processing approach, statistical fusion is proposed to reduce the model uncertainty at the decision-level. Blum et al. [13] introduced a novel method to combine statistical fusion with deep learning-based segmentation prediction, including Bayes categorical fusion and Dirichlet fusion. The presented methods allow different training sets per expert (modality). Namely, individual baseline networks for every expert are completely independent of each other. The fusion can be applied based on the outputs of all the individual baselines. Without extra training on aligned data, only a small subset is needed for calibration of the statistical models. Therefore, we can produce the probability  $p(k|\text{all expert outputs})$  for every possible category  $k \in 1, \dots, K$ , given the outputs of all the unimodal experts. Then we can find the fused classification by choosing for every pixel the class with the highest probability, which is also called the belief of a class  $\text{bel}(k)$  (see Figure 3.5).

Formally,

$$\text{output class} = \arg \max_k p(k|\text{all expert outputs}) \quad (3.3)$$

Considering Bayes categorical fusion as an example, the unimodal classification output is used as index to the confusion matrix in order to produce the conditional class likelihoods. The final prediction is therefore the class with the highest joint likelihood.

Combining multiple classifiers in a statistical way is not a new concept [196], but this work leads to an interesting research direction in the combination of deep learning and statistics.

### 3.1.2/ DISCUSSION

Deep multimodal fusion for scene understanding is an extremely complex issue that involves several factors, including the spatial location of objects, the semantic context of the scenes, the effectiveness of fusion models, the physical properties of the modalities, etc. The fusion strategies mentioned above follow different design concepts to tackle this challenges. Early fusion methods make an effort to optimally integrate information from multimodal sources during feature extraction. Namely, the representative features from complementary modalities are automatically fused to the RGB branch or a gated branch at the early stage, while features are reconstructed via a common decoder. These works emphasize the importance of cross-modal information interaction. Late fusion methods generally map multimodal features into a common space at the decision level. In other words, the fusion model is trained to learn unimodal features separately. Thus, late fusion may offer more flexibility and scalability but lacks sufficient cross-modal correlation. Regarding hybrid fusion, such fusion strategy is elaborated to combine the strengths of early fusion and late fusion, achieving a more robust performance. However, the trade-off between accuracy and execution time should be carefully considered in architectural design.

This brings us to two main questions:

- **When to Fuse:** Many deep multimodal fusion methods are extended from existing unimodal methods, or derived from other typical neural networks. In the former case, multiple unimodal segmentation networks are integrated into a composite end-to-end training model in early, late, or multi-level stages. Early fusion strategy allows stronger cross-modal information interaction, while late fusion shows more flexibility and scalability for implementation. Extensive experiments demonstrate that both low-level and high-level features are valuable to the final prediction. Multi-level fusion is helpful for segmentation model to learn representative features. Fusing multimodal contextual information in multi-level stages represents the current trend. Moreover, semantic guiding across layers, such as skip connections, can be effectively used to bridge early feature extraction and late decision making. The state-of-the-art method SSMA shows a typical example.

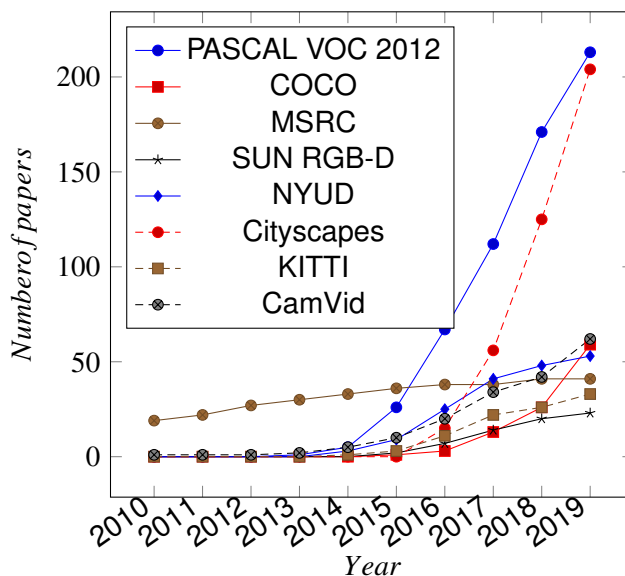


Figure 3.6: Accumulated dataset importance. Statistical analysis is based on the work by [15].

- How to Fuse:** Different from unimodal networks, deep multimodal fusion networks should consider multimodal information collaboration. Although deep learning-based methods learn representative features automatically, in many cases, multimodal input is likely to be imperfect. The redundancy, imbalance, uncertainty, and even contradiction of multimodal data may significantly affect the model's performance. Simple fusion operations, such as summation and concatenation, provide limited help to generate optimal joint feature representations. Experiments indicate that several adaptive fusion methods make remarkable progress in terms of accuracy, such as attention-based networks. One potential reason is that such a learning model takes into account the contribution of multimodal features at multiple stages. Such fusion methods usually contain specific gating units that assign class-wise or modality-wise weights. One extreme case that should be noted is modality missing. Most of the existing deep fusion models can not work effectively when the supplementary modality is unavailable. Piasco et al. [138] offers some ideas based on learning with the privileged information paradigm to tackle this challenge. Otherwise, the trade-offs between accuracy/speed [77, 75, 45] or memory/robustness should be carefully considered in the architectural design. In order to provide readers a more intuitive understanding, we show more detailed evaluations in Subsection 3.2.3.

## 3.2/ DATASETS

Over the last decade, a large number of datasets have been proposed to meet the needs of deep learning-based methods. The quantity and quality of training data significantly affect the performance of learning models. For this reason, many academic and research institutions have released several large-scale benchmark datasets for different scenarios. The creation of these well-annotated datasets actively promotes semantic scene understanding, which also facilitates the performance evaluation and inspires innovative approaches.

With the advent of multiple sensory modalities, numerous multimodal benchmark datasets have been released to the public successively. These datasets provide complementary properties of the same scene, such as geometric information, toward learning an improved feature representation. Figure 3.6 shows the accumulated dataset importance for image segmentation task since 2010. We observe that several large-scale datasets have emerged from 2015. Notably, PASCAL VOC 2012 [43] and Cityscapes [29] are two of the most popular datasets for semantic segmentation. As the representative RGB-D dataset, NYU-D [163] and SUN RGB-D [168] are frequently used for indoor scene understanding.

In the following parts, we provide a summary of current unimodal and multimodal datasets for semantic image segmentation. The aim is to grab the reader's interest in multimodal scene understanding and facilitate the preliminary experiments on deep multimodal segmentation.

### 3.2.1/ POPULAR DATASETS FOR IMAGE SEGMENTATION TASK

As one of the earliest pixel-wise labeled image databases, MSRC dataset [161] was released for full scene segmentation. It consists of 591 images and 23 object classes. However, along with the development of deep learning techniques, small-scale datasets can not meet the demands of model training. PASCAL VOC dataset [43] is one of the most popular object segmentation datasets, which derived from the early stage competition: PASCAL Visual Object Classes (VOC) challenge. It provides thousands of images with pixel-level labeling. Up to now, it has been augmented to several additional datasets with a set of extra annotations, such as PASCAL-Context [125], PASCAL-Part [22], SBDB [66]. Another similar large-scale dataset is Microsoft COCO dataset [111], which contains 81 categories of objects, including 21 categories of PASCAL VOC. It covers complex everyday scenes and their contextual information. PASCAL VOC and COCO dataset are not only the most popular benchmarks for fully supervised segmentation but also frequently-used in weakly supervised learning for object segmentation.

Table 3.4: Summary of popular datasets for image segmentation task.

Ref.	Dataset	Classes	Resolution	Images	Scene	Data	Year
[161]	MSRC	23	320x213	591	Outdoor	2D	2006
[60]	Stanford background	8	320x240	715	Outdoor	2D	2009
[14]	CamVid	32	960x720	701	Outdoor	2D	2009
[43]	PASCAL VOC	20	Variable	11K	Variable	2D	2012
[163]	NYU Depth v2	40	480x640	1449	Indoor	2.5D	2012
[111]	Microsoft COCO	80	Variable	330K	Variable	2D	2014
[55]	KITTI	11	Variable	400	Outdoor	2D/3D	2015
[29]	Cityscapes	30	2048x1024	5K	Outdoor	2.5D	2015
[151]	SYNTHIA	13	960x720	13K	Outdoor(synthetic)	2.5D	2016
[149]	GTA5	19	1914x1052	13K	Outdoor(synthetic)	2D	2016
[168]	SUN RGB-D	37	Variable	10K	Indoor	2.5D	2015
[216]	ADE20K	150	Variable	22K	Variable	2D	2017
[127]	Mapillary Vistas	66	1920x1080	25K	Outdoor	2D	2017
[203]	WildDash	28	Variable	1.8K	Outdoor	2D	2018

Furthermore, several outdoor road scene datasets are constantly emerging during the last decade, e.g. CamVid [14], KITTI [55], Cityscapes [29], Mapillary Vistas [127], toward promoting the commercialization and advancement of autonomous driving technology [200]. To be specific, CamVid database is the first collection of fully segmented videos, captured from a moving vehicle. It provides over 700 manually labeled images of naturally complex driving scenes sampling from the video sequences. After that, KITTI Vision Benchmark was published to tackle various real-world computer vision problems, such as stereo, optical flow, visual odometry/SLAM, and 3D object detection. It consists of around 400 semantically annotated images recorded by RGB cameras, grayscale stereo cameras, and a 3D laser scanner.

During the past few years, Cityscapes dataset has been a strong performer in outdoor scene semantic segmentation. This high-quality dataset contains around five thousand high-resolution images with pixel-level annotations, recording the street scenes from 50 different cities. Also, Cityscapes is a superior multimodal segmentation dataset, containing precomputed depth maps of the same scenes. Besides, Mapillary Vistas dataset [127] provides 25,000 high-resolution images of street scenes captured from all over the world at various conditions regarding weather, season, and daytime. The images were annotated into 66 object categories, aiming to support the development of state-of-the-art methods for road scene understanding. More recently, for the sake of robustness and performance evaluation, WildDash [203] was released to the research community. This new benchmark provides standard data of driving scenarios under real-world conditions for a fair comparison of semantic segmentation algorithms. It is worth noting that RailSem19 [204] is the first public outdoor scene dataset for semantic segmentation targeting the rail domain, which is useful for rail applications and road applications alike.

We present a summary of the reviewed segmentation datasets in Table 3.4. Further information are provided, including numbers of classes, size of the database, and the

Table 3.5: Summary of popular 2D/2.5D multimodal datasets for scene understanding.

Ref.	Dataset	Images	Scene	Multi-modal data	Year
[163]	NYUDv2	1449	Indoor	RGB/Depth	2012
[168]	SUN RGB-D	10K	Indoor	RGB/Depth	2015
[29]	Cityscapes	5K	Urban street	RGB/Depth	2015
[151]	SYNTHIA	13K	Urban street	RGB/Depth	2016
[176]	Freiburg Forest	5K	Forest	RGB/Depth/NIR	2016
[31]	ScanNet	19K	Indoor	RGB/Depth	2017
[64]	Tokyo Multi-Spectral	1569	Urban street	RGB/Thermal	2017
[174]	CATS 2	686	Variable	RGB/Depth/Thermal	2018
[26]	RANUS	40k	Urban street	RGB/NIR	2018
[210]	POLABOT	175	Outdoor	RGB/NIR/Polarization	2019
[159]	PST900	894	Subterranean	RGB/Thermal	2019
[92]	DISCOMAN	600K	Indoor	RGB/Depth	2019

type of scenes.

### 3.2.2/ MULTIMODAL DATASETS

Throughout the years, multimodal data are gaining the attention of researchers in various domains. The primary motivation for using multiple sensory modalities is to improve learning models' performance by enriching the feature representation. Table 3.5 lists numerous multimodal datasets reviewed in this survey, providing valuable information such as their application scenarios and data information. Next, we describe the potential multimodal datasets for image segmentation in detail, covering RGB-D datasets, Near InfraRed datasets, thermal datasets, and polarization datasets. Multiple samples can be found in Table 3.6.

#### 3.2.2.1/ RGB-D DATASETS

RGB-D cameras are widely used to augment the conventional color images with a depth map, which provides supplementary depth information about the distance of the object surface. Gupta et al. [63] proposed a method to encode horizontal disparity, height above ground, and the angle of the local surface normal into more efficient HHA images using raw depth images. Apart from semantic segmentation, depth information also makes significant contributions to other scene understanding tasks, such as object detection [63, 40] and pose estimation [157]. The first row in Table 3.6 illustrates RGB-D image examples sampling from the datasets reviewed in this part.

- **Indoor scenes:** One of the main difficulties for indoor scene segmentation is that object classes always come in various positions, shapes, and sizes. By taking ad-

vantage of RGB-D data, we can encode the pixel-level color and depth information of the same scene into a high-level feature representation. Such information fusion, to a certain extent, reduces the difficulty of indoor object recognition. NYUDv2 [163] is an early RGB-D database containing 795 training images and 654 testing images with pixel-wise labels for 40 semantic categories. A Microsoft Kinect camera captured all the RGB and depth image pairs with favorable frame synchronization. This dataset aims to inform a structured 3D interpretation of indoor scenes, having become one of the most popular multimodal benchmarks so far. Another standard benchmark for indoor scene recognition is SUN RGB-D [168]. It consists of around 10K RGB-D images with 37 indoor object classes. This dataset advances the state-of-the-art in all major scene understanding tasks and provides a fair comparison of deep multimodal fusion methods.

- **Outdoor scenes:** Unlike indoor scenes, the depth information of outdoor scenes is generally captured by stereo vision cameras or LiDAR due to Kinect's poor performance in sunlight. As one of the segmentation benchmark datasets, Cityscapes consists of thousands of high-quality depth images of the same scene. These depth maps overcome the lack of depth information of objects for road scene recognition. In order to simulate different seasons, weather, and illumination conditions, several synthetic RGB-D datasets (e.g., SYNTHIA [151]) are generated for driving scenes semantic segmentation.

### 3.2.2.2/ NEAR-INFRA-RED DATASETS

Infrared imaging captured from multi-spectral cameras shows high contrast of natural and artificial objects [182, 86]. In the computer vision field, multi-spectral images make up the data in the non-visible light spectrum and help better understand the scene characteristics. For example, Freiburg Forest dataset [176] was created to tackle the semantic segmentation problem in forested environments. It consists of 366 aligned color, depth, and near-infrared images with six classes pixel-wise annotation. Due to the abundant presence of vegetation in the unstructured forest environment, this dataset provides enhanced NIR images (e.g., Normalized Difference Vegetation Index images, Enhanced Vegetation Index images) to ensure border accuracy. Besides, RANUS dataset [26] has been released to the public in 2018. It consists of 40k spatially-aligned RGB-NIR pairs for real-world road scenes, and thousands of keyframes are annotated with ground truth masks for ten classes: sky, ground, water, mountain, road, construction, vegetation, object, vehicle, and pedestrian.

Apart from semantic segmentation, multi-spectral images are also used in other computer vision tasks, including pedestrian detection [27, 97], face recognition [44], image dehazing



[82, 37], video surveillance [8], to name a few.

### 3.2.2.3/ THERMAL DATASETS

Different from NIR images, thermal images are captured to recognize visible and invisible objects under various lighting conditions. The thermal imaging cameras are sensitive to all the objects that constantly emit thermal radiations [51]. The wavelength is generally detected up to  $14\mu m$ . In the early years, thermal imaging cameras were invented for military uses. With the cost of sensors decreasing, many scene understanding tasks can now benefit from thermal information [172].

Tokyo Multi-Spectral [64] is the first large-scale color-thermal dataset for urban scene segmentation. It contains both visible and thermal infrared images captured in daily and night conditions. There are 1569 images manually labeled to eight classes: car, person, bike, curve, car stop, guardrail, color cone, and bump. Then Shivakumar et al. [159] presented PST900, a dataset of 894 synchronized and calibrated RGB and thermal image pairs with pixel-level annotations across four distinct classes from the DARPA Subterranean Challenge. The long-wave infrared (LWIR) imagery was used as a supporting modality for semantic segmentation of subterranean scenes. These large-scale RGB-Thermal datasets broaden the research field of deep learning-based scene understanding, allowing for more in-depth exploration in poor visibility and adverse weather conditions.

### 3.2.2.4/ POLARIZATION DATASETS

As a universal phenomenon existing in natural scenes, polarimetric imaging is highly sensitive to the vibration pattern of the light [191]. In the natural environment, the polarization of light is generally obtained by reflection or scattering. The polarization images carry crucial information of reflection surface related to object shape and surface material properties. To tackle practical problems in computer vision, polarization images have been widely applied to object detection [193], image dehazing [155], depth estimation [165, 218].

Zhang et al. [210] released a small-scale segmentation dataset, known as POLABOT, that dedicates to the polarimetric imaging of outdoor scenes. Synchronized cameras collect hundreds of raw color, Near-InfraRed, and polarimetric images. All the images are manually labeled into eight classes according to the polarimetric characteristic of the scenes. For example, reflective areas such as windows and water are typically considered. More recently, Sun et al. [170] developed a multimodal vision system that integrates a stereo camera, a polarization camera, and a panoramic camera. The polarization camera is mainly used to detect specular materials such as glass and puddles, potentially

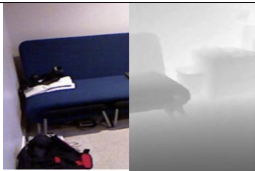

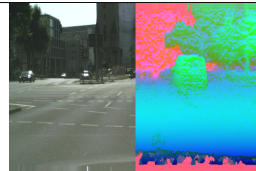
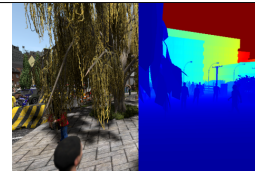
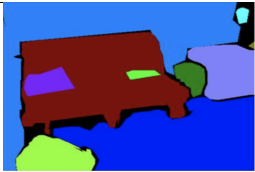


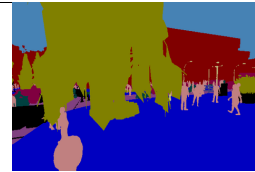
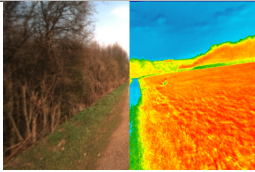





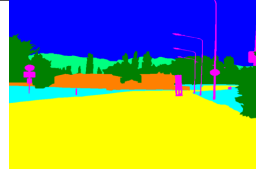

NYUDv2 (RGB+Depth)	SUN RGB-D (RGB+Depth)	Cityscapes (RGB+Depth)	SYNTHIA (RGB+Depth)
			
			
Freiburg Forest (RGB+NIR)	Tokyo Multi-Spectral (RGB+Thermal)	RANUS (RGB+NIR)	POLABOT (RGB+Polarization)
			
			

Table 3.6: Examples of multimodal image datasets mentioned in Subsection 3.2.2. For each dataset, the top image shows two modal representations of the same scene. The bottom image is the corresponding groundtruth.

dangerous for autonomous systems. Currently, the use of polarimetric data leads to new directions for deep multimodal fusion research. The polarimetric imaging offers great potential [143, 152] in scene understanding. For future perspectives, polarization cameras may be extremely valuable in autonomous driving [11, 170] and robotics [198, 146].

### 3.2.2.5/ CRITICAL CHALLENGES FOR MULTIMODAL DATA

Based on the review of multimodal image datasets, we summarized four critical challenges for multimodal data:

- **Data diversity:** different image sensors offer different representative features of the scene according to their physical properties. The accuracy and robustness of deep fusion models are closely related to the amount and variety of multimodal data. In addition to the multimodal types mentioned above, more data types are expected for complex tasks in computer vision.
- **Quantity and quality:** in order to meet the needs of deep learning model training,

Table 3.7: Performance results of deep multimodal fusion methods on SUN RGB-D dataset.

Method	Backbone	Input size	Modality	Fusion strategy	Mean Acc	Mean IoU
Bayesian SegNet [88]	VGG-16	-		-	45.9	30.7
Context [109]	VGG-16	-	RGB	-	53.4	42.3
RefineNet [108]	ResNet-152	-		-	58.5	45.9
LSTM-CF [104]	VGG-16	426x426		Late	48.1	-
FuseNet [69]	VGG-16	224x224		Early	48.30	37.3
DFCN-DCRF [84]	VGG-16	480x480		Early	50.6	39.3
S-M Fusion [102]	VGG-16	449x449		Hybrid	53.93	40.98
LSD-GF [25]	VGG-16	417x417	RGB-D	Late	58.0	-
SSMA [178]	ResNet-50	768x384		Hybrid	-	44.52
RDFNet [130]	ResNet-152	-		Early	60.1	47.7
RedNet [85]	ResNet-50	640x480		Hybrid	60.3	47.8
CFN [107]	RefineNet-152	-		Hybrid	-	48.1
ACNet [76]	ResNet-50	640x480		Early	-	48.1

high-quality and large-scale multimodal image datasets are expected to cover various scenarios. Meanwhile, inaccuracy and noise should be considered in image processing.

- **Data alignment:** data collected by image sensors should be well aligned before training. Such alignment is often referred to as multi-modality calibration, and is an essential prerequisite for effective multimodal fusion.
- **Dataset construction:** in the construction of multimodal datasets, we should think about 1) what kind of multimodal data do we need for the target scenarios? 2) what kind of multimodal data can provide more efficient information for specific tasks? 3) what kind of multimodal data is easier to collect in practice?

### 3.2.3/ COMPARATIVE ANALYSIS

In this part, we report the evaluations of existing deep multimodal fusion methods on four benchmark datasets: SUN RGB-D [168], NYU Dv2 [163], Cityscapes [29], and Tokyo Multi-Spectral dataset [64]. We also conduct a direct comparison of different unimodal and multimodal methods, aiming to demonstrate the necessity and importance of multimodal fusion approaches. All the results reported in this survey are collected from the original publications to ensure fairness.

#### 3.2.3.1/ ACCURACY

We gathered quantitative results of the aforementioned fusion approaches from the corresponding papers and grouped them according to the benchmark datasets. The mean accuracy (%) and mean IoU (%) are the most reported metrics for a fair comparison. In

Table 3.8: Performance results of deep multimodal fusion methods on NYU Depth v2 dataset.

# of classes	Method	Backbone	Input size	Modality	Fusion strategy	Mean Acc	Mean IoU
13	FuseNet [69]	VGG-16	320x240	RGB-D	Early	67.46	56.01
	Wang[187]	VGG-16	-		Late	52.7	-
	MVCNet [118]	VGG-16	320x240		Early	70.59	59.07
40	Gupta' [63]	-	-	RGB-D	Late	35.1	-
	FuseNet [69]	VGG-16	320x240		Early	44.92	35.36
	Wang[187]	VGG-16	-		Late	47.3	-
	MVCNet [118]	VGG-16	320x240		Early	51.78	40.07
	LSD-GF [25]	VGG-16	417x417		Late	60.7	45.9
	CFN [107]	RefineNet-152	-		Hybrid	-	47.7
	ACNet [76]	ResNet-50	640x480		Early	-	48.3

Table 3.9: Experimental results of deep multimodal fusion methods on Cityscapes dataset. Input images are uniformly resized to  $768 \times 384$ .

Method	Backbone	Modality	Fusion strategy	Mean IoU
ERFnet [9]	-	-	-	62.71
AdapNet [177]	ResNet-50	RGB	-	69.39
AdapNet++ [178]	ResNet-50	-	-	80.80
AdapNet [177]	ResNet-50	Depth	-	59.25
AdapNet++ [178]	ResNet-50	-	-	66.36
AdapNet++ [178]	ResNet-50	HHA	-	67.66
LFC [176]	VGG-16	RGB-D	Late	69.25
CMoDE [177]	AdapNet	-	Late	71.72
SSMA [178]	AdapNet++	-	Hybrid	83.44
SSMA [178]	AdapNet++	RGB-HHA	Hybrid	83.94

the comparison tables, deep multimodal fusion methods are differentiated based on the used backbone network, the type of multimodal input, and the fusion strategy.

- SUN RGB-D dataset

Firstly, we report the experimental results on the indoor scene dataset, SUN RGB-D (see Table 3.7). Ten fusion methods and three unimodal methods are compared on this benchmark dataset. We observe that ACNet and CFN are the two top scorers with a mean IoU score of 48.1%. RedNet and RDFNet are not far behind with a score of 47.8% and 47.7%, respectively. In general, multimodal fusion methods are superior to unimodal methods, which have a similar backbone network.

- NYU Depth v2 dataset

Regarding the NYU Depth v2 dataset, which is also a typical indoor scene dataset with high-quality depth information, we select six methods to make a detailed comparison. Table 3.8 demonstrates the experimental results with 13 and 40 classes. ACNet is again the best performing method with a mean IoU score of 48.3% for 40 classes. Note that when the methods are evaluated on 13 classes only, the performances are higher because most challenging classes are not taken into account.

Table 3.10: Experimental results of deep multimodal fusion methods on Tokyo Multi-Spectral dataset. The image resolution in the dataset is  $640 \times 480$ .

Method	Backbone	Modality	Mean Acc	Mean IoU
SegNet [5]	VGG-16		35.4	31.7
PSPNet [213]	ResNet-50	RGB	44.9	39.0
DUC-HDC [190]	ResNet-101		58.9	47.7
MFNet [64]	VGG-16		45.1	39.7
SegNet-4c [5]	VGG-16		49.1	42.3
FuseNet [69]	VGG-16		52.4	45.6
PSPNet-4c [213]	ResNet-50		51.3	46.1
DUC-HDC-4c [190]	ResNet-101	RGB-Thermal	59.3	50.1
RTFNet [172]	ResNet-152		63.1	53.2

- Cityscapes dataset

Apart from the indoor scene datasets, we also show the segmentation results on a more challenging urban scene dataset, Cityscapes in Table 3.9. For this outdoor dataset, SSMA, as a typical hybrid fusion architecture, achieves the best performance with a mean IoU score of 83.94%. Moreover, we have observed that HHA representation provides more valuable properties than the original depth map. The multimodal fusion methods generally outperform the performance of the unimodal methods.

- Tokyo Multi-Spectral dataset

As shown in Table 3.10, we report the evaluation results on Tokyo Multi-Spectral dataset. Both visible spectral images and thermal images were used in the fusion experiments. We also collect 4-channel early fusion methods for comparative study. The winner, RTFNet, achieves a maximum accuracy of 53.2% mean IoU. Notably, the segmentation accuracy is significantly increased by adding thermal infrared information. These results clearly show the effectiveness of multimodal data and the advancement of deep multimodal methods.

Based on the analysis of these results, we can draw some conclusions. First, depth information is the most commonly used supplementary information for multimodal image fusion. Most deep fusion methods report their results on the large-scale RGB-D datasets for both outdoor and indoor scene understanding. However, other types of multimodal datasets are of varying quality and lack further evaluation. The establishment of standard benchmark datasets is the premise of multimodal fusion study. Also, reported fusion methods employed various backbone networks, input size, and setups for the experiment, making fair performance comparisons difficult. Although many deep learning frameworks and libraries already exist, more multimodal toolkits are expected to facilitate multimodal fusion study.

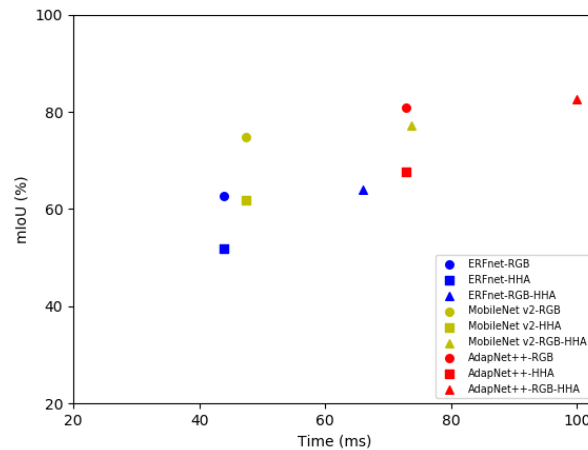


Figure 3.7: Real-time and accuracy performance. Performance of SSMA fusion method using different real-time backbones on the Cityscapes validation set (input image size:  $768 \times 384$ , GPU: NVIDIA TITAN X).

In light of the reported results, we have observed that ACNet and SSMA achieved remarkable results on the RGB-D datasets. A major reason is that these methods adopt many advanced deep learning techniques, such as attention mechanism, multiscale feature aggregation, and skip connection. It can be seen that the development of deep learning technology is of great benefit to multimodal fusion. Moreover, it is worth noting that most methods focus on accuracy, which does not allow for a comprehensive evaluation of fusion models. Multiple metrics can also reflect the effectiveness of multimodal data, which is instructive to the construction of the multimodal data collection platform. In general, deep multimodal fusion methods require higher memory footprint and execution time. We report more detailed results in the following subsections.

### 3.2.3.2/ EXECUTION TIME

In order to evaluate the real-time performance of deep multimodal fusion networks, we summarized and provided the researchers with two sets of execution time comparisons, as shown in Figure 3.73.8. Execution time or runtime, as an essential metric, obviously shows the learning model's execution efficiency. Although this metric is easily ignored in the accuracy-centric algorithm optimization, it should be carefully considered in industrial-level applications, such as self-driving cars. The inference time is usually dependant on the hardware and backend implementation.

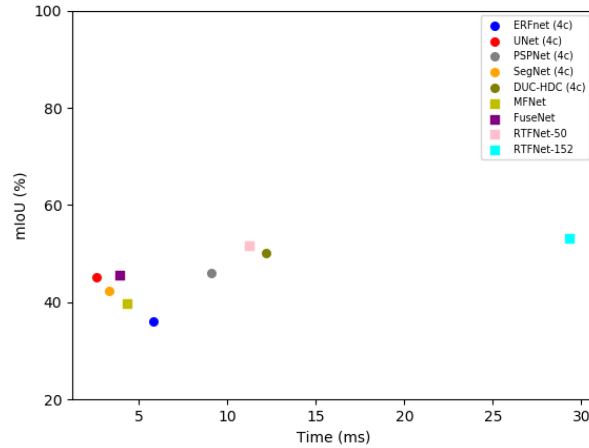


Figure 3.8: Real-time and accuracy performance. Performance of different fusion methods on the Tokyo Multi-Spectral dataset.(input image size:  $640 \times 480$ , GPU: NVIDIA 1080 Ti graphics card).

Table 3.11: Parameters and inference time performance. The reported results on the Cityscapes dataset are collected from [178].

Network	Backbone	mIoU (%)	Parms. (M)	Time (ms)
PSPNet [213]	ResNet-101	81.19	56.27	172.42
DeepLab v3 [19]	ResNet-101	81.34	58.16	79.90
DeepLab v3+ [21]	Modified Xception	82.14	43.48	127.97
AdapNet++ [178]	ResNet-50	81.34	30.20	72.92
SSMA [178]	ResNet-50	82.31	56.44	101.95

### 3.2.3.3/ MEMORY USAGE

Another performance indicator in the implementation aspect is memory usage. Large memory usage may increase computation time during training and testing. In this regard, proper use of deep learning frameworks, GPU acceleration, appropriate batch size, and compressed input may be beneficial for the model training. Table 3.11 demonstrates the comparisons on the number of parameters and inference time for various network architectures. It is worth noting that the elaborated models have higher accuracy but may require higher memory and inference time, which leads to a greater challenge to the real-time performance of the autonomous navigation system.

## 3.3/ SUMMARY

In this chapter, we reviewed deep multimodal image segmentation from two aspects: fusion methodology and dataset. Various existing multimodal fusion methods are cate-

gorized into early fusion, late fusion, and hybrid fusion. We also summarized existing semantic segmentation datasets, covering 12 current multimodal image datasets. Multimodal image data, such as RGB-D image, Near-InfraRed image, thermal image, polarization image, are the primary concerns in this work. We made a comparative analysis of existing fusion approaches in terms of accuracy, execution time, and memory footprint, which evaluate the model performance on different benchmark datasets ranging from indoor scenes to urban street scenes. Based on the reported evaluations, we further discussed architectural design to explore the essentials of deep multimodal fusion.

In conclusion, deep multimodal fusion has gained much attention in recent years. Multimodal images captured from various sensory modalities provide complementary information of the scenes. The collected experimental results show the effectiveness of deep multimodal image fusion. The state-of-the-art methods make efficient use of multimodal data, yielding an improved performance on semantic scene understanding. However, the optimal fusion strategy remains an open question in need of further exploration.





# DEEP MULTIMODAL FUSION FOR SEMANTIC IMAGE SEGMENTATION

**R**obust multimodal fusion is one of the challenging research problems in semantic scene understanding. In real-world applications, the fusion system can overcome the drawbacks of individual sensors by taking different feature representations and statistical properties of multiple modalities. This chapter is dedicated to fully-supervised semantic segmentation with multimodal image input. We seek robust solutions that can effectively learn feature representations from multi-modalities and optimally produce the pixel-wise classification. As detailed in Chapter 3, we conclude the architectural design and input data for multimodal image segmentation from the comprehensive review of the literature. Based on this background knowledge, we explore different fusion architectures with various imaging modalities, including late fusion and central fusion. The former can obtain richer semantic information by using different feature extractors. The latter sequentially maps both low-level and high-level multimodal features into a central branch. Statistical post-processing is employed to reduce model uncertainty, which leads to significant performance improvement.

In the following, we describe the proposed deep multimodal fusion methods for the outdoor scene semantic segmentation task in detail. Multiple fusion architectures are also compared and analyzed in this chapter. Moreover, we introduce a novel multimodal benchmark dataset dedicated to the polarimetric imaging of outdoor road scenes. The performances of the proposed algorithms are evaluated using extensive experiments on the Freiburg Forest dataset, POLABOT dataset, and Cityscapes dataset.

## 4.1/ CMNET: DEEP MULTIMODAL FUSION FOR ROAD SCENE SEGMENTATION

### 4.1.1/ INTRODUCTION

Semantic segmentation is one of the main challenges in computer vision. Along with the appearance and development of Deep Convolutional Neural Network (DCNN) [93], the trained model can predict which class each pixel in the input images belongs to. By learning from massive data sets of diverse samples, this method achieves a good performance on end-to-end image recognition. Robust and accurate scene parsing of outdoor environments paves the way towards autonomous navigation and relationship inference. Compared with indoor scenes, off-road perception is more challenging due to dynamic and complex situations. The outdoor environment may easily change in different time slots with light or color variations. Even in structured environments, for instance on urban roads, there are still several challenges such as the detection of glass and muddy puddles.

Most existing datasets and methods for outdoor scene semantic segmentation are mainly based on RGB camera. They are only well acceptable in general conditions excluding complex environment and small amount of samples. To develop additional practical solutions, one of the main challenges is data fusion from multi-modalities. Therefore, considering the RGB modality as a kind of imperfect sensor, we attempt to fuse the complementary feature information of the same scene from other modalities. Actually, several modalities are ubiquitous in robotic systems, such as RGB-D, LIDAR, near infrared sensor, etc.

In this work, we use a polarimetric camera, as a complementary modality, to provide a richer description of a scene. Polarization of light radiation has more general physical characteristic than intensity and color [191]. We can figure out that windows of a building, the asphalt road, and the puddle of water have reflected polarizations [186]. Plenty of research have demonstrated that the use of polarization camera can significantly enhance the capabilities of scene understanding, especially for reflective areas [65].

Over the past few years, a variety of deep learning-based end-to-end approaches have been proposed. One factor that increased the popularity of deep learning is the availability of massive data. In the case without large amount of samples, we attempt to acquire more features of the same scene using several modalities. To some degree, an effective encoding of complementary information enables learning without the need for massive data, therefore the use of small-scale dataset can also lead to good performances. Recent works have shown promising results in extracting and fusing features from complementary modalities at pixel-level. The idea is to separately or jointly train the model using

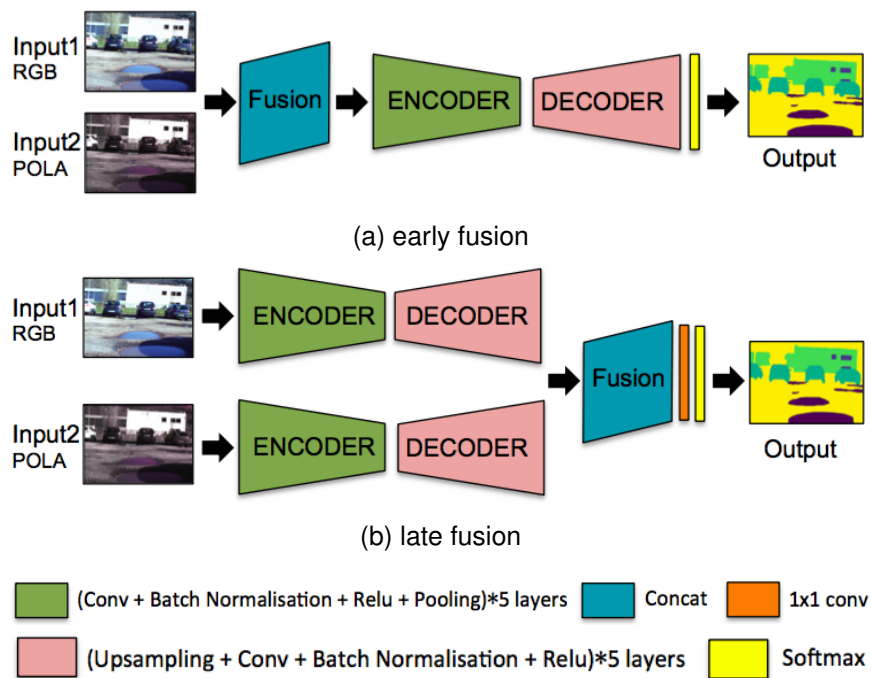


Figure 4.1: Typical early fusion and Late fusion architectures comparison.

data from different sensors and integrate them into a composite feature at early or late stage.

#### 4.1.2/ BASELINE ARCHITECTURES

In this subsection, we describe two baseline architectures in detail, namely early fusion and late fusion. The two simple structures, as well as their extensions, are widely used for deep learning-based fusion. Generally, the backbone network can be an encoder-decoder structure segmentation network, such as SegNet and Deeplab. The encoder is a regular convolutional neural network which contains several layers. Each layer extracts local features, normalizes the data distribution, obtains sparse representations by means of convolution, batch normalization and ReLU accordingly. Afterwards, pooling is used for downsampling the feature map and propagate spacial invariant features. Correspondingly, the decoder unsamples the shrunk feature map and recover the lost spatial information to full-sized segmentation.

##### 4.1.2.1/ EARLY FUSION

As shown in Figure 4.1a, the early fusion architecture has a unitary neural network, fusion takes place before passing into the encoder. Assume that both inputs (for example one RGB image and one polarimetric image) have size  $3 \times H \times W$ , then fused frame will be

$6 \times H \times W$ . So we also call this sort of fusion architecture as channel fusion.

This fusion architecture, combining features before training, seems simple and light. However, it is also more likely to overfit. To see why, let consider the model's complexity. Let  $H$  be a family of functions taking values in  $\{-1, +1\}$  with VC-dimensions  $d_{vc}$  [180]. Then, for any  $\delta > 0$  and all  $h \in H$ , the VC-dimension bound [123] can be derived with a high probability:

$$E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{8}{N} \ln\left(\frac{4(2N)^{d_{vc}}}{\delta}\right)}, \quad (4.1)$$

where  $E_{out}$  denotes out-of-sample error,  $E_{in}$  denotes in-sample error, and  $N$  denotes the data points that the hypothesis space can shatter the set. As the amount of input's dimensions increases, so does the VC-dimensions. Then the model complexity  $\Omega(N, H, \delta)$  rises along with the increase of VC-dimensions. As a result, larger data samples should be fed to fit the deep neural model for less in-sample error. In other words, in the case that samples are not huge enough, the model may be easier to overfit.

#### 4.1.2.2/ LATE FUSION

Figure 4.1b shows a typical late fusion architecture. It has two separated branches of network, with each branch trained to extract features from a special modality. Fusion takes place after a series of downsampling. Assuming that the two feature maps have size  $1 \times H \times W$ , after concatenation, the resulting feature will be  $2 \times H \times W$ . Then a  $1 \times 1$  convolution is applied to reduce the number of channels.

This approach has the advantages that each network computes weights separately while encoding. Compared with early fusion, to some extent, it may reduce the difficulty of model fitting and yield a better outcomes. Furthermore, thanks to the scalability and flexibility of this architecture, the model can be designed in accordance with requirements and easily extend to multi-inputs without a large dimension increase.

#### 4.1.3/ PROPOSED METHOD

We propose a new approach for multimodal data fusion, Complex Modality Neural Network (CMnet), based on late fusion architecture since it has aforementioned merits.

Let  $S = \{(X_n, y_n) | n = 1, 2, \dots, N\}$  denotes the training set, and  $X_n = \{x_a, x_b\}$  is the training example, where  $x_a$  and  $x_b$  are the vector of input images from modality  $a$  and  $b$ , respectively. Also let  $M_1$ , and  $M_2$ , denote the map between the input and output of the first, and second branch of the encoder-decoder network, respectively. Then the output of the

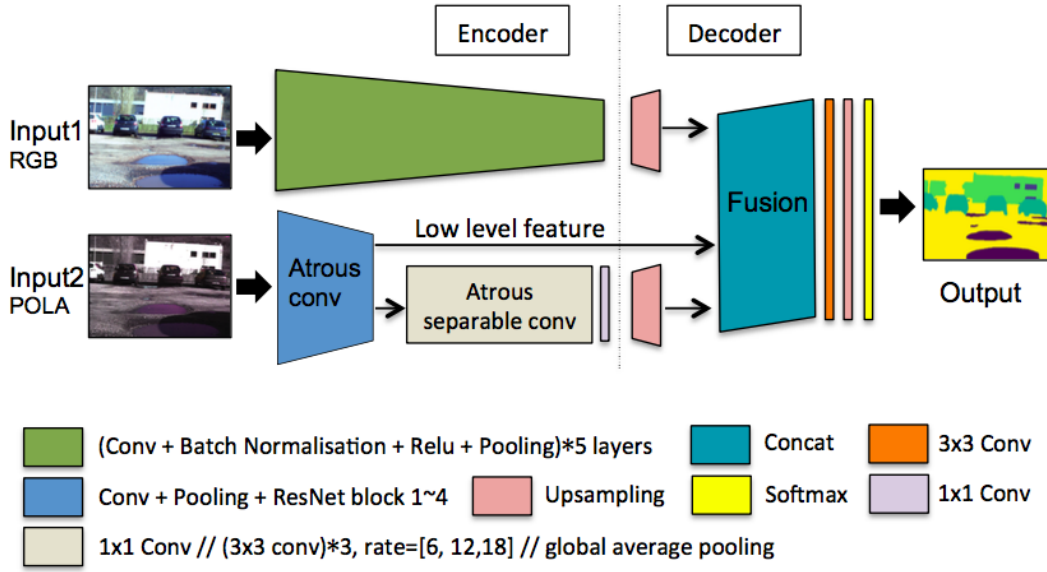


Figure 4.2: Our proposed fusion architecture: CMnet for multimodal fusion based on late fusion architecture.

fusion module can be written as:

$$\hat{y}_n = f(X_n) = \text{softmax}[W * (M_1(x_a) + M_2(x_b))], \quad (4.2)$$

where,  $W$  is a series of convolution kernels for upsampling. The **softmax** function is introduced to represent the categorical distribution, and is defined as:

$$\text{softmax}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}, \quad (4.3)$$

where  $z = [z_1, \dots, z_K]^T$ .

Moreover, the process of model training is to minimize the error while regularizing the parameters. It can be framed as an optimization one, which can be formulated as:

$$\theta^* = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y_n, f(x_n; \theta)), \quad (4.4)$$

where the loss is computed as:

$$L(u, y) = - \sum_k y_k \log \hat{y}_k. \quad (4.5)$$

Then we can use gradient descent algorithm to find local minimum.

Figure 4.2 presents the whole architecture of CMnet. It has an Encoder-Decoder structure and two separated branches. The encoder is used for mapping raw inputs to feature

representations. The decoder integrates three feature maps, then recovers the feature representation to final segmentation results. That is a reliable method to extract different modality features and recover sharp object boundaries for end-to-end segmentation.

On the one hand, the branch for RGB modality incorporates a SegNet-like encoder. By copying the indices from max-pooling, it can capture and store boundary information in the encoder feature maps before sub-sampling. We keep this strength to make the network more memory efficient and improve boundary delineation. On the other hand, we focus on the feature quality of the extra modality. Other modalities can provide rich complementary information on low level appearance features.

However, how to captures rich contextual information from extra modality is a challenging task. We refer to the state-of-the-art segmentation network Deeplab v3+ , which uses a new pooling method named ASPP (Atrous Spatial Pyramid Pooling) to incorporate the multi-scale contextual information. We experimentally apply this network structure as the other branch's encoder for the complementary modality and achieved improved results. The first upsampling stage is subsequently applied to each branch to recover the feature representation to the same fusion size, then we fuse these three feature maps, which contains high-level and low-level multimodal features information simultaneously. The second upsampling stage and softmax are applied to the fused feature map, which produces the final results.

#### 4.1.4/ DATASET

##### 4.1.4.1/ POLARIZATION FORMALISM

Polarization is a common property of light waves that specifies the geometrical orientation of the oscillations. Figure 4.3 indicates the electric and magnetic field component of a light which oscillates in phase perpendicular to each other and to the direction in which the radiation propagates. In nature scenes, many light sources such as sunlight, LED spotlights, and incandescent bulbs produce unpolarized light because their electric field's direction fluctuates randomly in time. When the direction of the electric field of light is restricted to a single plane by filtration, then it is called polarized light. If an unpolarized light is reflected by an object, it becomes partially linearly polarized. In general, we can classify the states of polarization of the light into unpolarized, partially polarized, and fully polarized. Moreover, polarimetric imaging carries not only the color and shape information of objects, but also characterizes the special physical information for reflecting surface. We can find the nature of the object's roughness, orientation and the reflection in each pixel of polarimetric images [192].

The electrical field of a progressive transverse wave in its propagation plan [12] can be

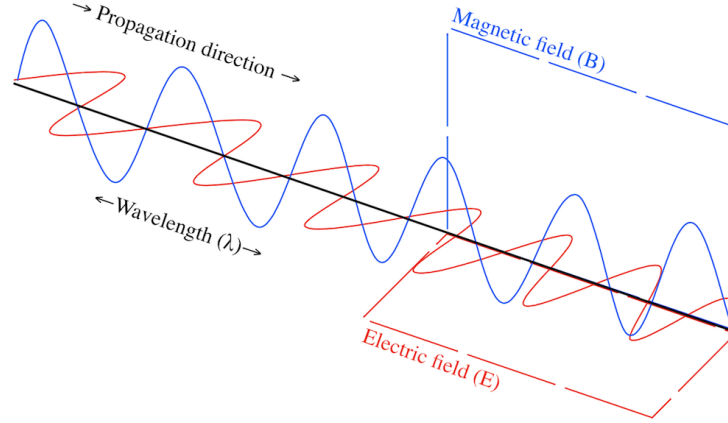


Figure 4.3: The electric and magnetic field of light as well as their continuous self-propagating.

defined as:

$$\vec{E}(t) = E_x(t)\cos(-k\vec{z} + \omega t)\vec{u}_x + E_y(t)\sin(-k\vec{z} + \omega t + \phi(t))\vec{u}_y, \quad (4.6)$$

where  $k$  is the wave number,  $\vec{z}$  denotes the direction of propagation.  $\omega$  and  $\phi$  are a pulsation and a phase in the orthonormal basis  $B = \{u_x, u_y\}$ .  $E_x$  and  $E_y$  are the amplitudes of  $\vec{E}(t)$  according to  $u_x$  and  $u_y$ , respectively.

In nature, light determined by scattering or reflection is generally unpolarized or partially linearly polarized and can be described by the linear Stokes vector,  $S = [S_0 \ S_1 \ S_2]^T$ . Using the parameters in Equation 4.6, it can be rewritten as:

$$S = \begin{bmatrix} S_0 \\ S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} \langle E_x^2 \rangle + \langle E_y^2 \rangle \\ \langle E_x^2 \rangle - \langle E_y^2 \rangle \\ 2\langle E_x E_y \cos(\phi) \rangle \end{bmatrix}. \quad (4.7)$$

Depending on the Stokes parameters in Equation 4.7, we can determine the Angle Of Polarization (AOP) and the Degree Of Polarization (DOP) as:

$$AOP = \frac{1}{2} \text{atan2}(S_2, S_1), \quad (4.8)$$

$$DOP = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}. \quad (4.9)$$

The AOP belongs to  $[-\frac{\pi}{2}; \frac{\pi}{2}]$ , which identifies the orientation of the polarized light with regards to the incident plan. The DOP indicates the quantity of polarized light in a wave. Namely, DOP equals to one means a fully polarized light, while it is between 0 and 1 means the partially polarized light and up to zero for the unpolarized light. Using polari-



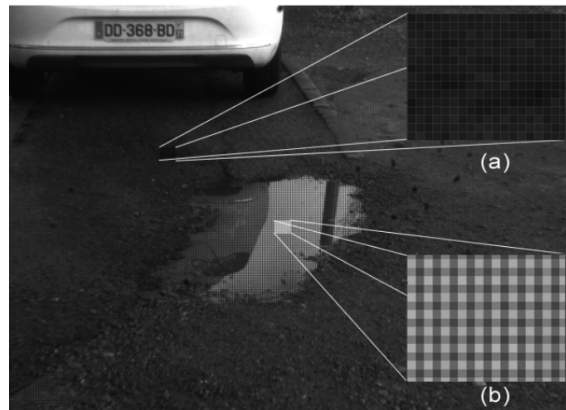


Figure 4.4: Reflection influence on polarimetry. (a) and (b) represent a zoom on the non-polarized and polarized area, respectively. Figure extracted from [11].

metric cameras, the micro-grid appears on the polarized surface and reveals an intensity change according to the polarizer affected. Figure 4.4 shows the detected polarization performance with a polarimetric camera.

#### 4.1.4.2/ POLABOT DATASET

To fully explore the multimodal image fusion for semantic segmentation task, we built a novel outdoor road scene dataset called POLABOT. As shown in Figure 4.5, we collected multimodal images using a mobile robot platform equipped with four cameras: the RGB camera (IDS Ucam), a polarimetric camera (PolarCam), a depth camera (Kinect 2.0), and a near-infrared camera. The raw dataset contains over 700 multimodal images of the same scene. All the images were acquired, synchronized, and calibrated using the Robot Operating System (ROS) framework. In particular, the three gray-scale description images of the raw polarimetric data can be obtained by calculating the AOP, DOP, and the intensity, which enables further generate the HSL (Hue Saturation Luminance) images [11]. The benchmark dataset contains 175 images with pixel-wise ground truth annotation.

Moreover, the images have been semantically dispatched into 8 classes: unlabeled, sky, water, windows, road, car, buildings and others. Benefiting from the use of a polarimetric camera, our mobile robot platform is more capable of discerning on windows, water and other reflective areas. That facilitates exploratory research on the use of polarimetric cameras in semantic scene understanding domain. In this thesis, we mainly employ aligned RGB and polarimetric images as inputs to train and evaluate the deep fusion models. For integrating the acquired images, we apply an automatic homographic method to image alignment [124]. This method allows to transform the RGB images with respect to the polarimetric images, and crop to the intersecting regions of interest. Moreover,

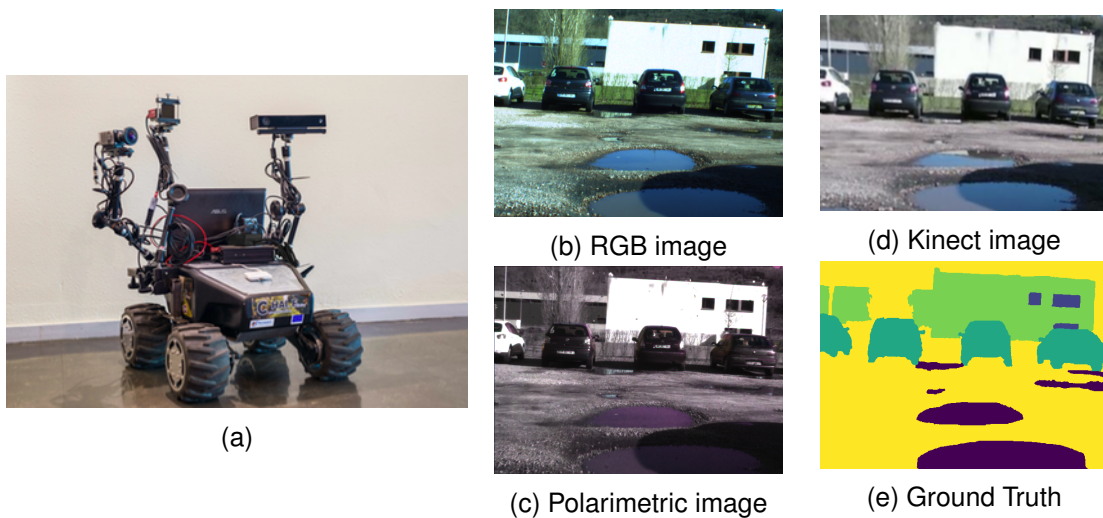


Figure 4.5: (a) Mobile robot platform used for the acquisition of the POLABOT dataset. It is equipped with the IDS Ucam, PolarCam, Kinect 2 and a NIR camera. (b)(c)(d) Multi-modal images in the POLABOT dataset.

as deep learning models need large data sets of diverse examples, a certain amount of data should be guaranteed. For this reason, we employ geometric data augmentations to increase the effective number of training samples, including rotation and flipping. Data augmentation and multimodal data fusion help to train deep neural networks on small scale datasets.

#### 4.1.5/ EXPERIMENTS

In this subsection, we evaluate the different fusion models, and report a series of experimental results on two benchmark datasets. One is the publicly available Freiburg Multi-spectral Forest dataset [175] and the second one is a new multimodal dataset containing polarimetric and RGB data, called POLABOT dataset. In this work, all the networks are implemented based on Pytorch framework with a Nvidia Titan Xp graphics processing unit (GPU) acceleration. The input data was randomly shuffled after each epoch. We initialize the learning rate as 0.0001 and use the contraction segments of pre-trained VGG-16 model and ResNet-101 as encoders. Then we fine-tuned the weights of the decoders until convergence.

##### 4.1.5.1/ EVALUATION ON FREIBURG FOREST DATASET

We train the segmentation architectures on the public Freiburg Forest dataset first. This dataset was collected by a modified RGB dashcam with NIR-cut filter in outdoor forested environment. It consists of over 15,000 raw images, and 325 images with pixel level

Table 4.1: Performance of segmentation models on Freiburg Multispectral Forest dataset. EF, LF refer to early fusion and late fusion respectively. We report pixel accuracy (PA), mean accuracy (MA), mean intersection over union (MIoU), frequency weighted IoU (FWIoU) as metric to evaluate the performance.

	PA	MA	MIoU	FWIoU
RGB	92.07	89.56	79.87	86.19
EVI	92.05	88.76	79.66	85.82
EF	91.80	88.02	78.95	85.67
LF	92.26	89.52	80.36	86.34
CMnet	<b>93.02</b>	<b>90.06</b>	<b>81.64</b>	<b>87.68</b>

Table 4.2: Comparison of deep unimodal and multimodal fusion approaches by class. We report MIoU as metric to evaluate the performance.

	Road	Grass	Veg/Tree	Sky
RGB	77.18	73.47	89.78	80.66
EVI	81.55	73.50	88.08	76.39
EF	80.78	74.07	86.90	78.68
LF	<b>82.27</b>	75.66	88.54	77.68
CMnet	81.01	<b>76.55</b>	<b>90.64</b>	<b>83.25</b>

ground truth annotations for 6 classes, which are the **sky**, **trail**, **grass**, **vegetation**, **obstacle** and others. In this unstructured forest environment, Enhanced Vegetation Index (EVI) was proposed to improve sensitivity to high biomass regions and vegetation monitoring. It shows stronger capacities on feature representation than NIR in the previous work. To extract more accurate information, here in our case, we select EVI images as the second modality input besides the visible input. We crop the RGB and EVI images as size  $3 \times 256 \times 256$ , and use them as inputs correspondingly. We report several metrics to assess segmentation models: pixel accuracy (PA), mean accuracy (MA), mean intersection over union (MIoU), frequency weighted IoU (FWIoU). These metrics are defined in Section 2.5.

The results shown in Table 4.1 show that segmentation using RGB images yields better results than EVI images on the whole. This shows that RGB images provide better high-level features while training. For fusion architectures, late fusion methods outperform channel fusion method as we analyzed in the previous section. Our network yields around 1% ~ 2% comprehensive improvements comparing with other methods.

Furthermore, the results in Table 4.2 demonstrate the evaluations by class. We report the main four classes as Road, Grass, Veg/Tree and Sky. For uni-modality network, we can find that EVI shows good performance on Road and Grass classes, and RGB modality has a significant advantage on Sky class, which is susceptible to lighting changes. Moreover, the fusion architecture outperforms uni-modality scheme by integrating complemen-

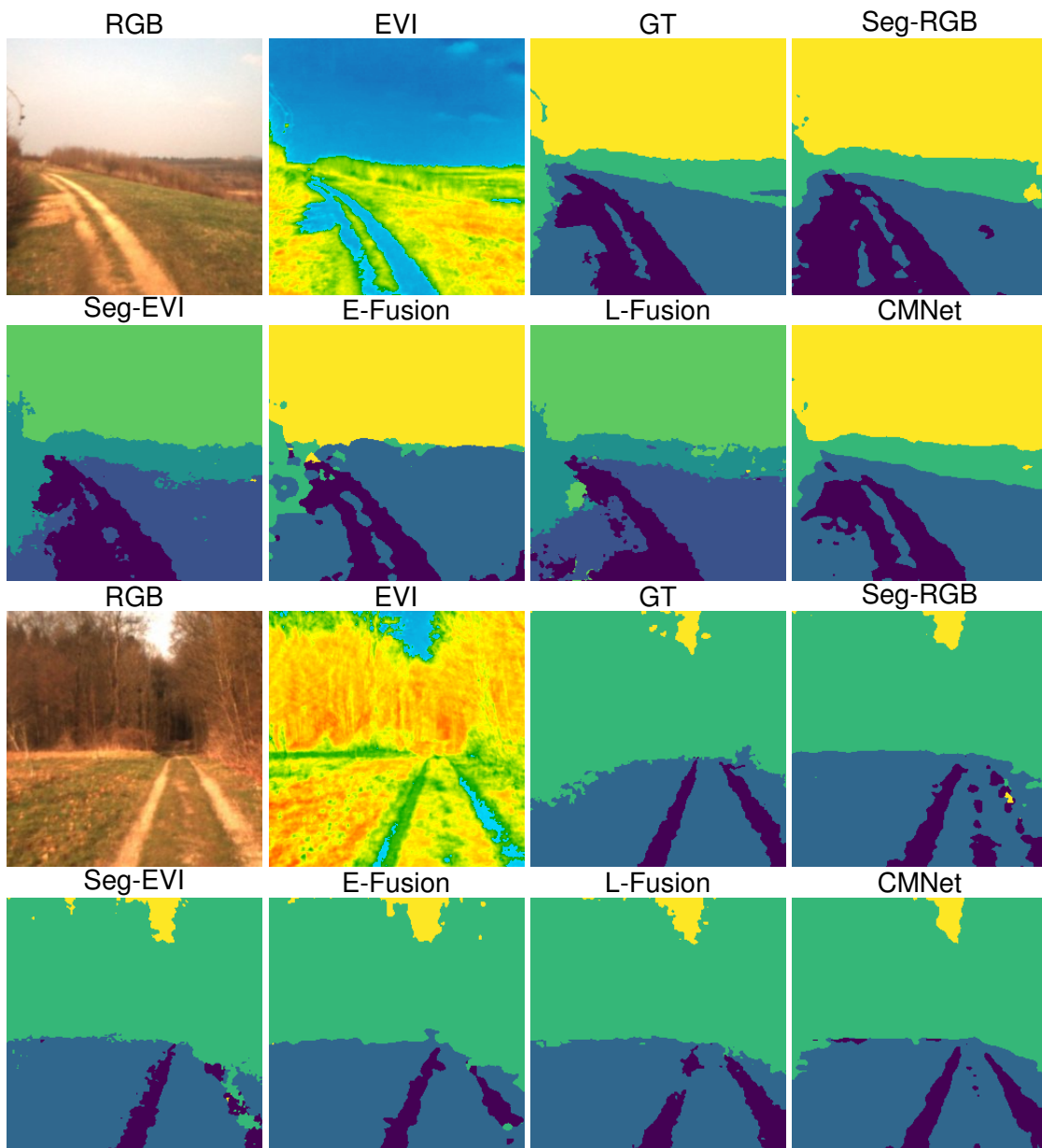


Figure 4.6: Two segmented examples from Freiburg Forest dataset. RGB and/or EVI images were given as inputs.

tary multimodal information. In particular, our CMnet model achieved a remarkable results on segmentation comparing with other fusion architectures, especially for Veg/Tree and Sky class.

A note about the results is that Freiburg Forest dataset was collected from a series of frames, the scene of these frames are homogenized, the structure of each class in these images does not fluctuate a lot. The specialization of certain scenes may also reduce the demand on the number of samples.

Some segmentation results on the Freiburg dataset are shown in Figure 4.6.

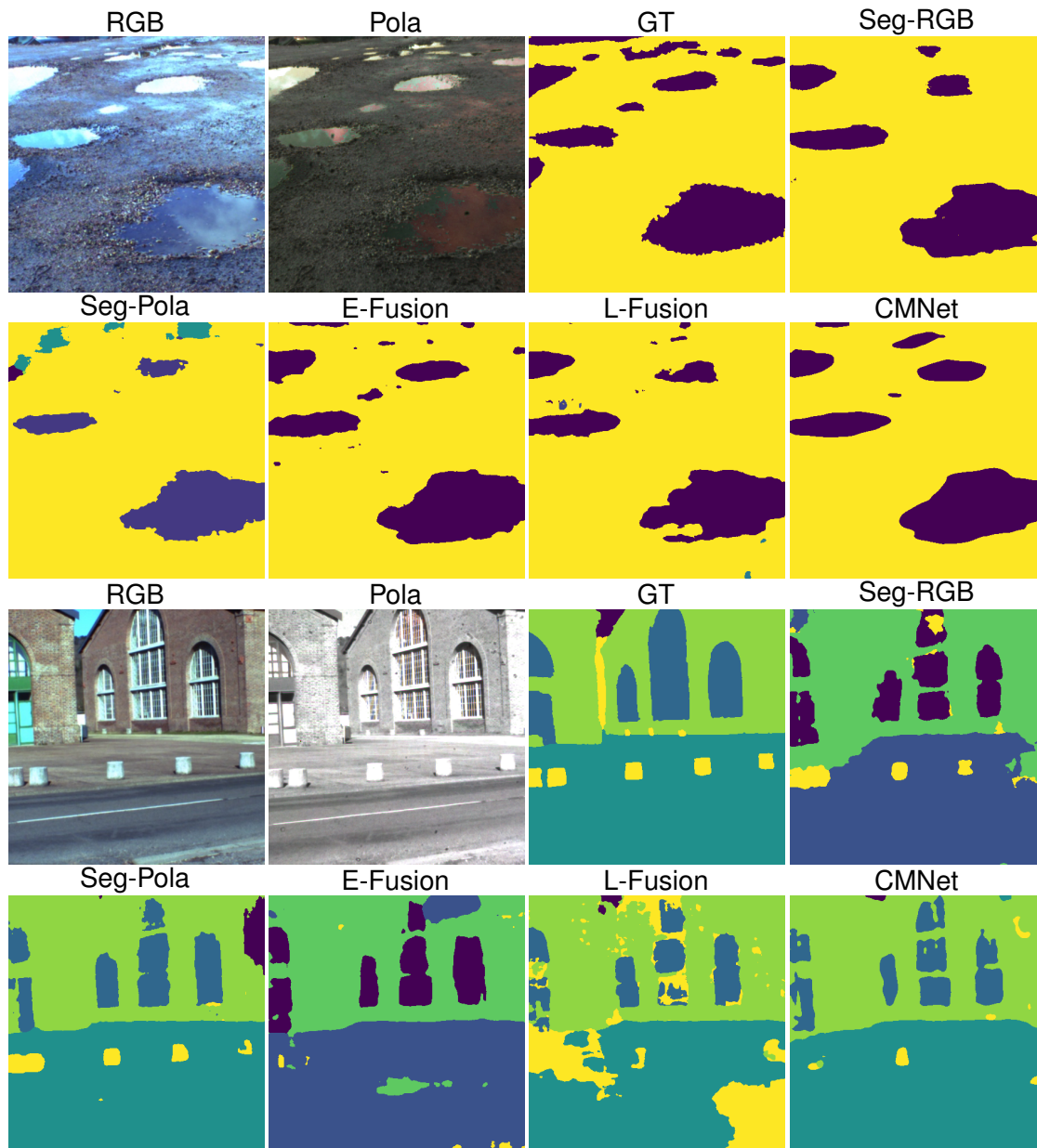


Figure 4.7: Two segmented examples from POLABOT dataset. RGB and/or POLA images were given as inputs.

#### 4.1.5.2/ EVALUATION ON POLABOT DATASET

In the following, we report several experimental results on our POLABOT dataset. The metrics shown in Table 4.3 correspond to pixel accuracy (PA), mean accuracy (MA), F1 score (F1) and mean intersection over Union (MIoU).

We process the RGB and polarimetric images with size  $3 \times 448 \times 448$ . While training the networks, we experimentally found that stochastic gradient descent (batch size=1) does not work well. It is reasonable that online learning adds too much instability to the learning process as the weights widely vary with each batch, especially for small scale dataset with



Table 4.3: Segmentation performance on POLABOT dataset

Input	Methods	PA	MA	F1	MIoU
RGB	SegNet	87.76	81.44	87.67	64.79
POLA	SegNet	90.51	84.15	90.77	68.58
RGB	E-Fusion	90.25	85.06	90.64	69.48
+	L-Fusion	90.02	84.28	90.11	68.81
POLA	CMnet	<b>90.70</b>	<b>85.90</b>	<b>90.92</b>	<b>72.59</b>

multi-classes. As a complement of previous analysis of training on small scale dataset, a novel data augmentation technology [11] applied to POLABOT dataset gives the additional guarantee for weights learning. As a result, we can find that polarimetric images in our dataset provide high quality feature information, it is a beneficial premise for further data fusion. The overall best performance in this dataset was obtained with CMnet integrating RGB and polarimetric inputs, achieving a mean IoU of 72.59%. It yields around 3% comprehensive improvements comparing with the second best methods.

Some segmentation results on the POLABOT dataset are shown in Figure 4.7.

## 4.2/ A CENTRAL MULTIMODAL FUSION FRAMEWORK

### 4.2.1/ INTRODUCTION

Robust and reliable scene understanding of robotic systems in the real world has been among the most challenging tasks in computer vision. A wealth of research on deep learning focuses on image segmentation of outdoor road scenes, especially the perception of autonomous vehicles. However, robotic systems still exhibit a limited ability to the semantic understanding of complex environments because they may be negatively affected in challenging situations such as varying illumination conditions or seasons changes. Specific sensors display much better performance than standard ones in some scenarios. For instance, numerous studies have shown that the use of a polarization camera delivers outstanding performance on semantic scene understanding in reflective areas [65, 11]. Over the past few years, academia and industry have expressed a significant interest in multimodal data acquisition systems and analysis methods. Previous works [63, 68, 175, 210] experimentally demonstrate that qualified multimodal fusion models yield better results than the unimodal ones, due to richer information representation of scenes provided by complementary modalities.

State-of-the-art segmentation models have demonstrated remarkable pixel-level classification results. However, most of the existing methods focus on segmentation with RGB or 3D data. With the advent of low-cost RGB-depth and multi-spectral cameras, there is

an increasing study on 2D/2.5D multimodal fusion. Existing deep fusion methods generally incorporate the features at an early or late stage, which may cause a loss of semantic information. As shown in Figure 4.8, we employ a centralized fusion strategy to continuously correlate multimodal information. Besides, traditional neural networks lack probabilistic considerations, while the representation of uncertainty should be taken into account to tackle this problem.

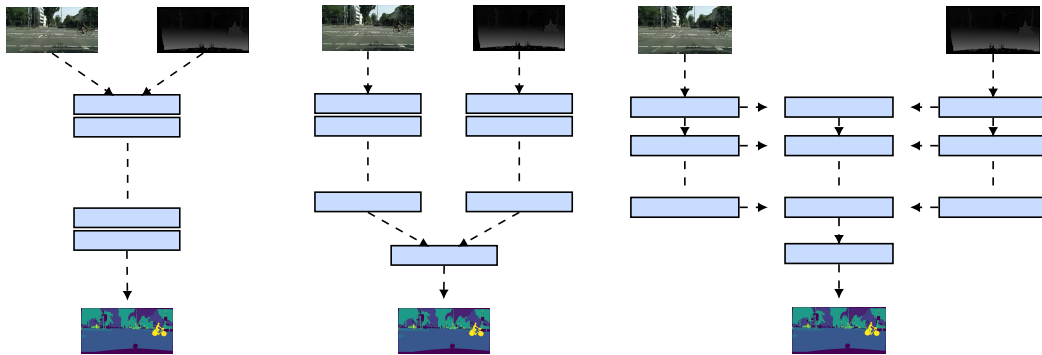


Figure 4.8: Typical fusion strategies with RGB and depth input. (a) Early fusion. (b) Late fusion. (c) Central fusion. As a comparison, the proposed fusion structure integrates the feature maps in a succession of layers into a central branch.

## 4.2.2/ METHOD

### 4.2.2.1/ CENTRAL FUSION

In general, our multimodal fusion framework consists of two main modules, a central fusion network and statistical fusion as post-processing. The central multimodal fusion network, referred to as CMFnet, continuously incorporates multimodal features into a central branch from the modality-specific branches. A lightweight gating unit is employed to compute the weights of feature maps in each layer. In order to provide more spatial information and enhance the model learning ability, multi-scale feature maps of the central branch are concatenated to the last layer in the decoder. Then in the statistical fusion module, the output of CMFnet is fused with the segmentation prediction of the optimal modality-specific network based on Bayesian prior probability distribution. Namely, we statistically merge the probability maps from CMFnet and qualified unimodal baseline network. Next, we introduce the proposed framework in detail.

### 4.2.2.2/ ADAPTIVE CENTRAL FUSION NETWORK

The main idea of CMFnet is to project the multimodal information into a common feature space. By integrating the feature maps in a succession of layers into the central branch,

the fusion network can sequentially learn the joint feature representation. Contrary to the typical structure of early and late fusion, such central fusion can automatically identify which are the best levels for feature integration and how these feature maps should be combined [183]. The overall network architecture is illustrated in Figure 4.9.

Suppose that we have two modalities  $x_i$  and  $x_j$ , each modality goes through a series of convolutional layers to produce different feature maps. Let  $l_i^k$  and  $l_j^k$  be the feature maps obtaining at layer  $k$  from each modality. To evaluate the value of feature maps and obtain a joint representation, we employ an adaptive gating unit (AGU) to assign the weights. This unit is the stack of a concatenating layer, a convolutional layer with the ReLu activation function, and a global average pooling layer followed by a softmax layer. The global average pooling layer (GAP) [110] is applied to compute the spatial average of feature maps from the previous convolutional layer. It is used to prevent overfitting and enforce the correspondence between feature maps and weights [110]. The output of AGU is a set of two weights  $w_i^k$  and  $w_j^k$ , therefore the first central layer can be defined as

$$F^1 = [\tilde{l}_i^1 \vee \tilde{l}_j^1] \quad (4.10)$$

where  $\tilde{l}_i^1 = w_i^1 l_i^1$  and  $\vee$  denotes the tensor concatenation. This operation is repeated for the different layers of the central fusion network, each time concatenating the weighted feature maps of layer  $k$  with the output of the previous layer:

$$F^k = [F^{k-1} \vee \tilde{l}_i^k \vee \tilde{l}_j^k] \quad (4.11)$$

In particular, considering the size of the last central layer is  $N \times W \times H$ , we compute a  $1 \times W \times H$  feature map for each central layer in the encoder part. The feature maps containing multi-scale spatial details are integrated into the last central layer for further refinement. Finally, we add a softmax layer for loss computation.

#### 4.2.2.3/ STATISTICAL PRIOR FUSION

To further exploit the multimodal prior probability distribution, we aim to statistically combine the output of the central fusion network with unimodal baselines. Similar work can be found in [13]. In a series of experiments, we have seen significant improvements in the accuracy of certain categories. The statistical methods are proved to be effective in dealing with model uncertainty. Especially for deep learning-based multimodal fusion, multiple modalities bring more model uncertainty in decision making. Different from the previous work, we employ Bayesian fusion as post-processing to statistically integrate the outputs of the optimal unimodal baseline and CMFnet. This fine-tuning process, to some extent, can optimize the pixel-wise prediction and achieve robust performance.



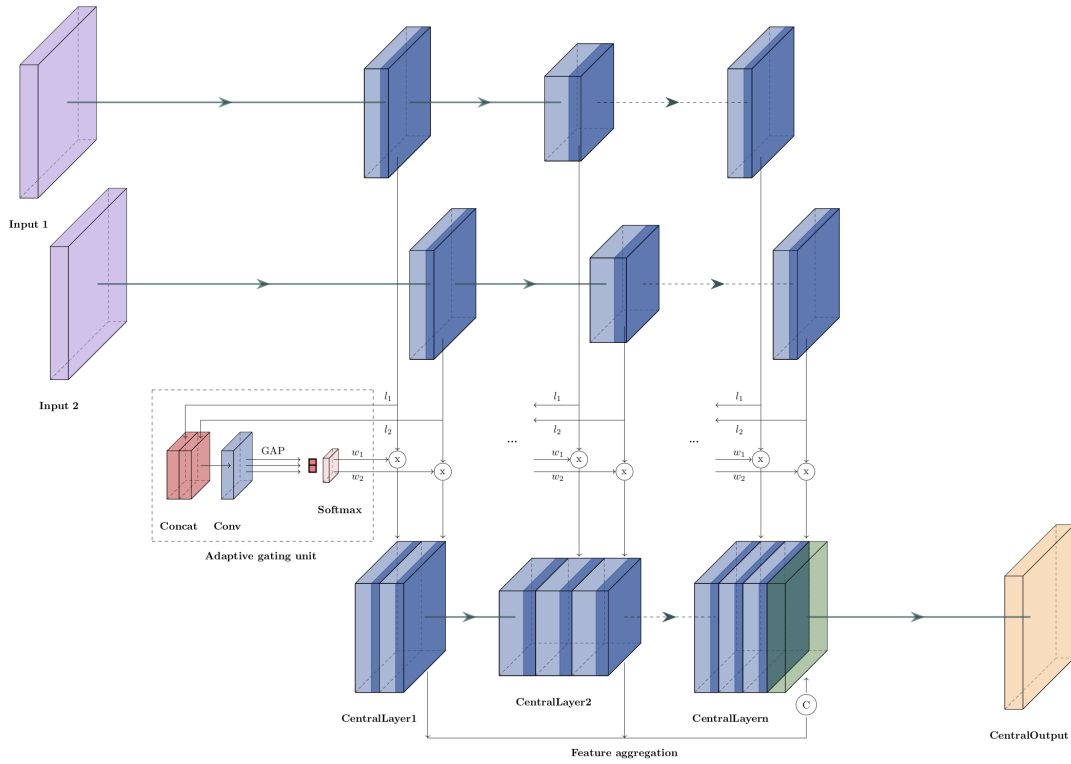


Figure 4.9: Overview of the central fusion network (CMFnet). The adaptive gating unit automatically produces the weights of modality-specific branch in each layer. GAP denotes to Global Average Pooling, "x" means multiplication, "C" means concatenation.

Next, we formulate the statistical post-processing. Assuming that the performance of unimodal baseline  $O_s$  is the best in comparison of baselines, we further combine  $O_s$  with the output of CMFnet, which referred to as  $O_c$ . The prediction probability for each possible class  $k \in 1, \dots, K$  can be defined as  $p(k|\{O_c, O_s\})$ . For every feature, the class with the highest probability is chosen as the final prediction. By maximizing the probability  $p(k|\{O_c, O_s\})$ , we can obtain optimized segmentation results. The statistical fusion process can be described as:

$$\begin{aligned}
 k^* &= \arg \max_k p(k|\{O_c, O_s\}) \\
 &= \arg \max_k p(k)p(O_c|k)p(O_s|k) \\
 &= \arg \max_k [\log p(k) + \log p(O_c|k) + \log p(O_s|k)]
 \end{aligned} \tag{4.12}$$

where  $p(O_c|k)$  denotes the categorical distribution over the output of CMFnet,  $p(O_s|k)$  is the categorical distribution over the optimal baseline output, and  $p(k)$  is the prior probability of class  $k$ .

Note that given the class  $k$ , we consider  $O_c$  and  $O_s$  are conditional independent because they were trained separately. Every term in the Equation (4.12) can be obtained from

the confusion matrix of  $O_c$  and  $O_s$ , which referred to as  $M_c$  and  $M_s$ . In more detail, if the first dimension of the confusion matrix denotes the actual output (i.e.,  $O_c$  or  $O_s$ ) and the second dimension is the ground truth class  $k$ , the conditional probabilities can be computed as:

$$p(O_c|k) = \frac{M_c[O_c][k]}{\sum_{j=1}^K M_c[O_j][k]} \quad (4.13)$$

$$p(O_s|k) = \frac{M_s[O_s][k]}{\sum_{j=1}^K M_s[O_j][k]} \quad (4.14)$$

Based on the Equation (4.13)(4.14), we can obtain the final segmentation prediction.

### 4.2.3/ EXPERIMENTS

#### 4.2.3.1/ IMPLEMENTATION DETAILS

We conduct experiments on two public outdoor road scene datasets: POLABOT [11, 210] and Cityscapes [29]. In order to fully evaluate the effectiveness and robustness of multimodal fusion models, we apply different types of image inputs for training and testing, including RGB, depth, and polarimetric images. We use the VGG-16 model [164] pre-trained on ImageNet [83] as backbone. All the methods are implemented using Pytorch [132] deep learning library on dual Nvidia TITAN Xp GPU (12GB memory). We choose the batch size according to the computing power of hardware and the learning rate is initialized to 1e-4.

In this work, we employ SegNet [5] and ENet [131] as unimodal baseline network. Additionally, we compare the proposed framework with several fusion approaches based on the corresponding baseline network, including average, LFC [175], and CMoDE [177]. The simple averaging fuses the RGB and depth baseline by taking the mean over the outputs of the last convolutional layer, followed by a shared softmax layer, while LFC and CMoDE are two competitive late fusion approaches. Specifically, we perform the ablation studies on the proposed central fusion method:

- CMFnet: the central multimodal fusion without post-processing, i.e., the output of the central fusion network is the final segmentation result.
- CMFnet+BF2: the central multimodal fusion with Bayesian fusion, i.e., the output of CMFnet is statistically fused with the output of the optimal unimodal baseline.
- CMFnet+BF3: an extension of CMFnet+BF2, where the CMFnet is fused with all available unimodal baselines.

#### 4.2.3.2/ EVALUATION ON POLABOT DATASET

We present a detailed comparison of different fusion approaches on the POLABOT [210] dataset. This dataset dedicates to polarimetric imaging, containing synchronized RGB and polarimetric images. It provides 175 aligned image pairs with ground truth annotation and focuses on the segmentation of reflective areas such as windows and water. In order to enhance the generalization ability of models, we randomly apply a series of data augmentation while training [11], including rotation and flipping. In this experiment, we adopt SegNet as the unimodal baseline and the backbone of fusion methods. The segmentation models were trained with batch size 4 at a resolution of  $448 \times 448$ .

The experimental results are shown in Table 4.4. We report the mean IoU as the primary metric to evaluate the overall performance of different segmentation methods. Firstly, we observe that the segmentation results of the polarimetric baseline are better than the RGB ones. This is particularly the case for the categories *Cars*, *Windows*, and *Water*. Therefore in post-processing, we adopt the polarimetric baseline to provide complementary statistical information for CMFnet+BF2. Regarding these fusion methods, we can observe that the central fusion method without statistical fusion, CMFnet, is slightly better than other fusion approaches, such as simple average, LFC and CMoDE, but the improvement is marginal. With statistical fusion, the proposed framework CMFnet+BF2 achieves excellent performance with a mean IoU of 86.62%, a relative improvement of +3.93% and +2.97% over the RGB and polarimetric baseline, respectively. Compared with CMFnet, we find that the statistical fusion module leads to a relative improvement of +2.47% overall. Furthermore, CMFnet+BF2 performs reliably for most of the categories, especially for the category *Water*, *Windows*, *Cars* and *Buildings*. The results demonstrate that our fusion framework effectively exploits RGB and polarimetric information, yielding better segmentation accuracy than the previous fusion methods.

Furthermore, our central fusion strategy speeds up the convergence while training (see Figure 4.10), which also reduces the time cost of model training to some extent. Figure 4.12 shows the experimental results for two pairs of RGB and polarimetric images from the POLABOT dataset.

#### 4.2.3.3/ EVALUATION ON CITYSCAPES DATASET

In this subsection, we adopt ENet as the baseline network and further evaluate the segmentation performance of the proposed fusion framework on the Cityscapes dataset. Cityscapes is a standard urban street scenes benchmark dataset that contains RGB, disparity images, and pixel-wise semantic segmentation annotation. Due to the limitation of computing memory, we resize the full input images to the resolution of  $768 \times 384$  and use the stochastic gradient descent method with one batch size for training. We group

Table 4.4: Performance comparison of our method with baseline models on the PO-LABOT dataset. Note that SegNet is used as the unimodal baseline and the backbone of multimodal fusion methods.

Class	RGB	Pola	Average	LFC	CMoDE	CMFnet	CMFnet+BF2	CMFnet+BF3
Mean IoU (%)	82.69	83.65	84.07	83.92	84.15	84.49	<b>86.62</b>	85.89
Sky	94.08	94.91	94.67	93.92	95.30	94.22	94.61	<b>95.43</b>
Water	83.42	86.61	84.21	82.83	83.21	86.11	<b>88.03</b>	85.31
Windows	73.28	77.71	73.21	73.14	72.72	77.49	<b>79.80</b>	79.54
Road	80.61	75.32	83.71	<b>85.16</b>	84.94	83.76	84.50	83.80
Cars	83.89	87.29	86.41	85.64	86.78	85.65	<b>90.42</b>	86.83
Building	80.53	81.82	82.26	81.83	81.69	82.39	<b>84.05</b>	83.92
None	83.01	81.91	84.02	84.94	84.41	85.53	84.96	<b>86.41</b>

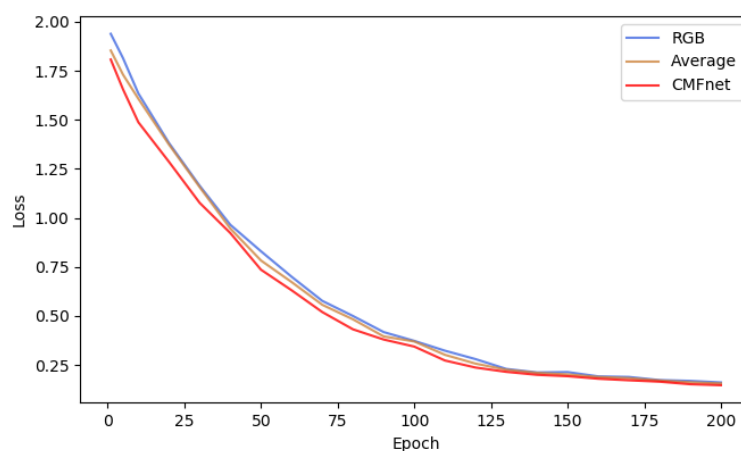


Figure 4.10: Training loss of models on the POLABOT dataset.

the general 33 classes into a set of 12 common categories following the experiments in [177, 13]: **background**, **sky**, **building**, **road**, **sidewalk**, **fence**, **vegetation**, **pole**, **vehicle**, **traffic sign**, **person**, **bicycle**. Also, we take a random 10% sample out of the given training set as the development set to validate the central fusion network.

Table 4.5 shows the overall performance of the unimodal and multimodal fusion approaches. Our proposed fusion framework, CMFnet+BF2, yields a mean IoU of 58.97%, which constitutes an improvement of +4.80% over the best unimodal baselines, i.e., RGB. It also outperforms other fusion approaches by around +2%. We can also observe that fusing all the unimodal networks, i.e., CMFnet+BF3, does not exhibit superior performance due to the poor performance of the depth baseline. The ablation studies indicate that the advanced capabilities of central fusion framework are based on the premise that modalities can extract high-quality semantic features. As shown in Table 4.6, we further provide the qualitative comparisons of segmentation per class, including the unimodal methods and central multimodal fusion methods. The proposed framework achieves significant improvements in most of the classes. Notably, CMFnet+BF2 earns a large gain of +11.53% in **fence** class, +9.00% in **bicycle** class, +9.82% in **sidewalk** class, compared

Table 4.5: Ablation study of our method. Per class performance of our proposed framework in comparison to individual modalities with ENet baseline on Cityscapes dataset.

Input	Mean IoU	Background	Bicycle	Person	Traffic Sign	Vehicle	Pole	Vegetation	Fence	Sidewalk	Road	Building	Sky
RGB	54.17	<b>67.17</b>	33.89	25.71	20.97	80.09	22.03	83.33	23.46	54.90	95.00	62.39	83.09
DEPTH	34.64	54.42	2.10	18.63	5.80	42.19	16.63	54.39	0.00	30.00	88.74	34.95	67.73
CMFnet	56.58	67.16	38.48	26.57	<b>26.01</b>	70.71	<b>35.62</b>	84.56	24.54	61.37	95.35	65.32	83.20
CMFnet+BF2	<b>58.97</b>	66.19	<b>42.89</b>	<b>28.81</b>	21.70	<b>83.49</b>	26.50	<b>85.15</b>	<b>34.99</b>	<b>64.72</b>	<b>96.38</b>	<b>68.78</b>	87.99
CMFnet+BF3	50.18	64.89	25.92	3.87	15.45	83.01	19.78	83.76	14.21	41.98	93.78	67.34	<b>88.24</b>

Table 4.6: Performance of fusion models with ENet backbone on Cityscapes dataset.

Input	Method	MIoU
RGB	Unimodal	54.17
DEPTH	Unimodal	34.64
RGB-D	Average	55.72
	LFC	55.19
	CMoDE	56.42
	CMFnet	56.58
	CMFnet+BF2	<b>58.97</b>
	CMFnet+BF3	50.18

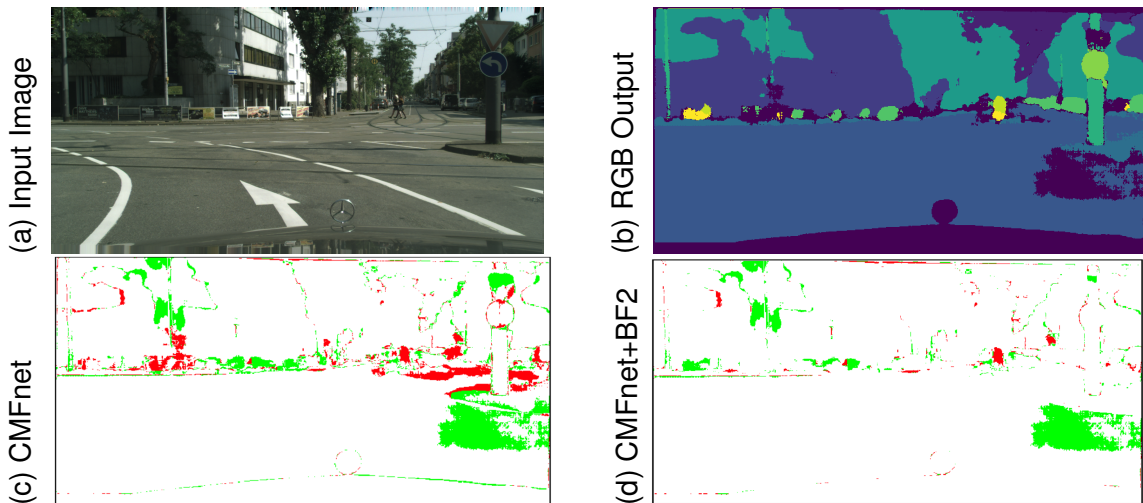


Figure 4.11: Improvement/error maps of the proposed CMFnet and CMFnet+BF2 in comparison to the RGB baseline. Note that the improved pixels and the misclassified pixels are recorded in green and red, respectively.

to the RGB baseline.

Figure 4.11 illustrates the improvement/error maps of CMFnet and CMFnet+BF2, which denote the improvement over the output of the RGB baseline in green and the misclassifications in red. From the visualized improvement/error maps, we can clearly see our fusion network gains in certain classes, such as **vegetation**, **traffic sign**, and **road**. After statistical post-processing, the segmentation prediction achieves a significant improvement due to the reduction in model uncertainty. We show more segmentation results on

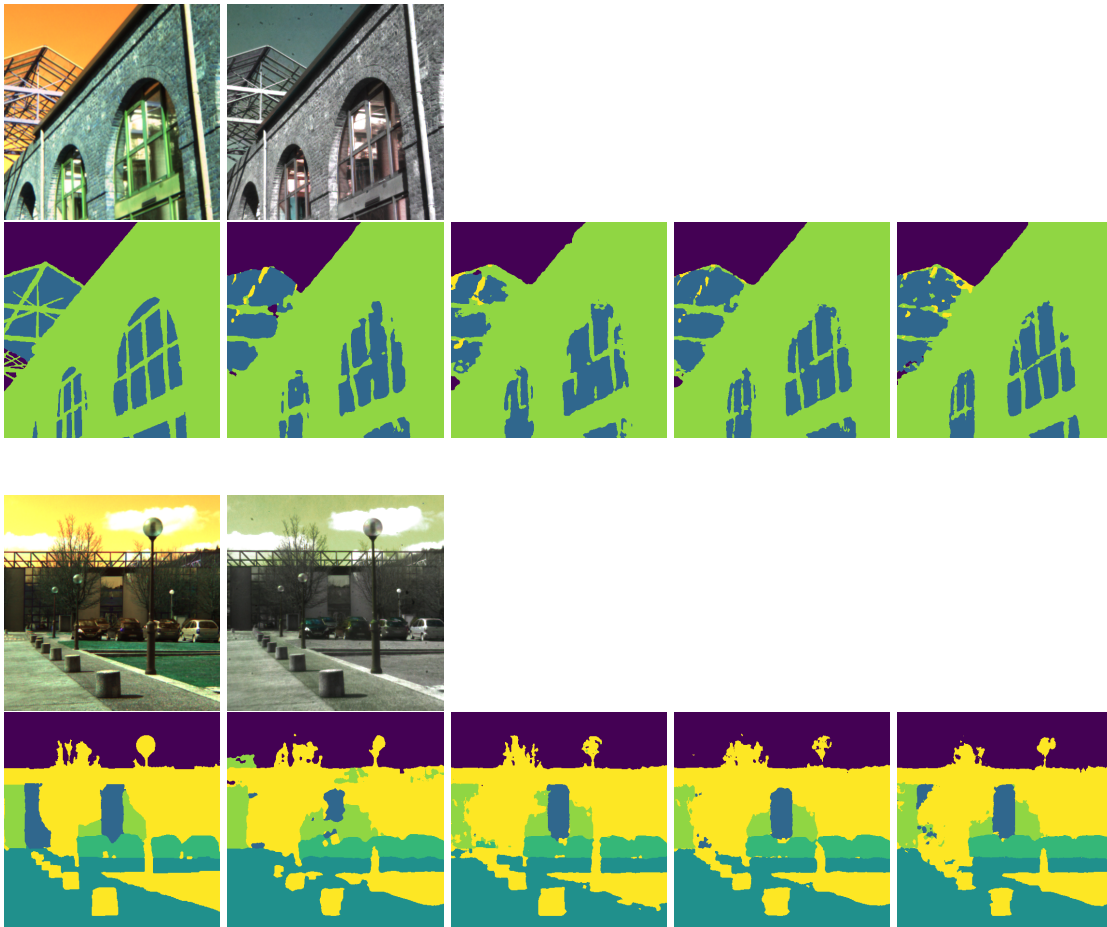


Figure 4.12: Segmentation results on the POLABOT dataset. The first row of examples contains the RGB image and the corresponding polarimetric image. The second row of examples shows, from left to right, the ground truth image, the segmentation outputs of the individual experts (RGB and Polar), and the fusion results of CMFnet-BF2 and CMFnet-BF3.

the Cityscape dataset in Figure 4.13.

### 4.3/ SUMMARY

In this chapter, we first explored the typical early fusion and late fusion architectures that extract features from multi-modalities, and extensively evaluated their merits and deficiencies. We also proposed an extensible late fusion scheme for outdoor road scene semantic segmentation. It provides design choices for future research directions. We presented comprehensive quantitative evaluations of multimodal fusion on the two small-scale benchmark datasets. In addition, we introduced a first-of-a-kind outdoor scene segmentation dataset for road scene navigation, which contains high-quality aligned polarimetric images. We empirically demonstrate that the use of polarization camera enhance

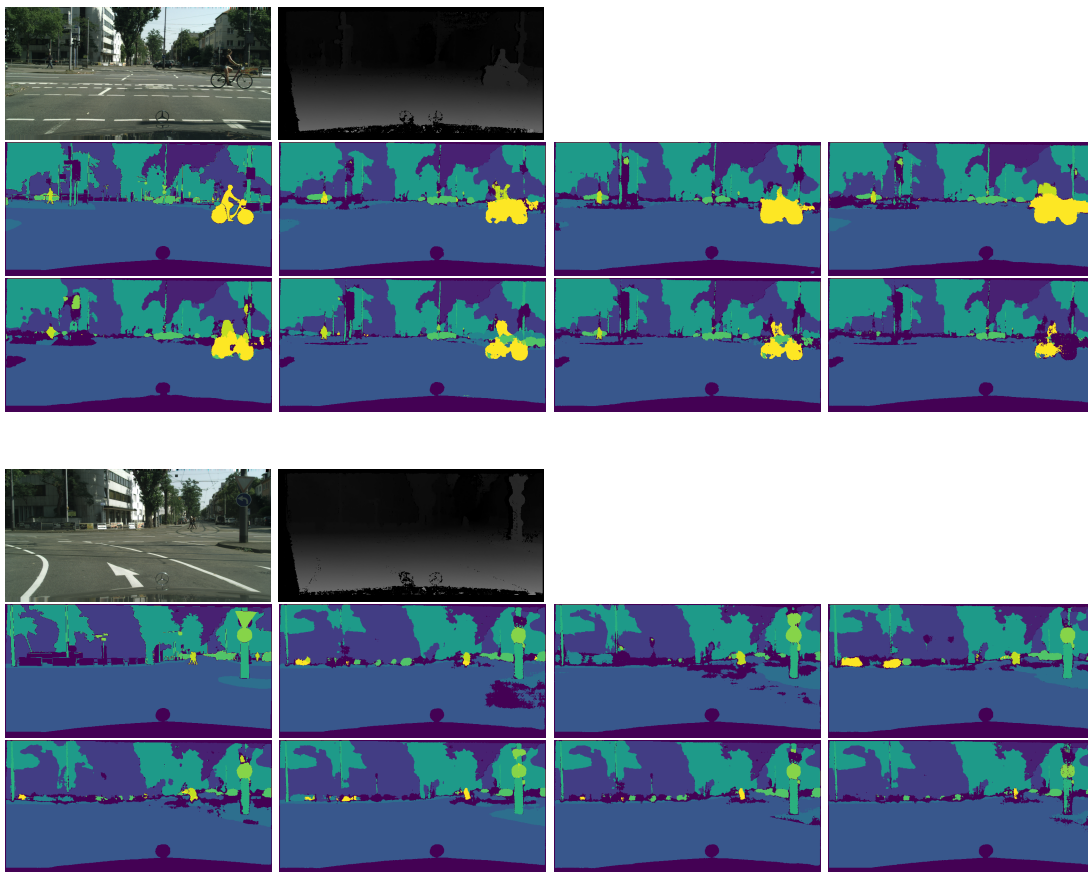


Figure 4.13: Two sets of segmentation results on Cityscapes dataset. The first row of examples contains the RGB image and the corresponding depth image. From left to right of the second row of examples: ground truth, the prediction of RGB input, average and LFC. From left to right of the second row of examples: the fusion results of CMoDE, CMFnet, CMFnet-BF2 and CMFnet-BF3, respectively.

the capabilities of scene understanding.

Furthermore, we presented a novel central fusion framework for multimodal image segmentation. The network module sequentially learns joint feature representations from the modality-specific branch. A lightweight gating unit is employed to assign the weights of feature maps in each layer. To fully take advantage of the prior probability distribution of multiple modalities, statistical fusion is applied to integrate the output of the central fusion network and qualified uni-modality. We perform extensive experiments and ablation studies on two public datasets, including POLABOT and Cityscapes dataset. The experimental results demonstrate that the proposed framework outperforms existing methods, leading to an improved segmentation performance.



# FEW-SHOT SEMANTIC IMAGE SEGMENTATION

**T**he previous chapter addresses the problems of fully-supervised semantic segmentation for multimodal image input. Such image understanding techniques require a large number of labeled images for training and is difficult to generalize to new categories. A step forward is the exploration of semi-supervised semantic segmentation. Few-shot segmentation presents a significant challenge for semantic scene understanding under limited supervision. Namely, this task targets at generalizing the segmentation ability of the model to new categories given a few samples.

This chapter first introduces the background knowledge on few-shot segmentation then presents a novel few-shot semantic segmentation method based on the prototypical network and a few-shot RGB-D image segmentation algorithm, which consists of two mirrored streams. The former algorithm employs a multiscale feature enhancement module to extract rich contextual information of labeled support images. The learned representative features of target classes provide further semantic guidance on the query image. Multiple similarity-guided probability maps are adaptively integrated by the attention mechanism. The latter algorithm aims to extend the RGB-centric methods to take advantage of complementary depth data for complete scene information. The proposed method learns class-specific prototype representations within RGB and depth embedding spaces, respectively. Furthermore, we report extensive experimental results on the PASCAL-5<sup>i</sup> and Cityscapes-3<sup>i</sup> dataset to show the effectiveness of the proposed methods. Ablation studies also demonstrate that the supplementary geometric cues lead to more accurate segmentation performance.



## 5.1/ INTRODUCTION ON FEW-SHOT SEGMENTATION

This section summarizes the existing methods for few-shot semantic segmentation. Many approaches for few-shot learning are proposed to generalize prior knowledge to new tasks using only a few examples. Some research [184, 101] introduced the metric learning-based matching network for the few-shot classification task. The non-parametric structure facilitates the generalization of models to new training sets. Snell et al. [166] presented a method to represent the prototypes per class in a representation space, known as Prototypical Networks. Moreover, several studies such as [53] have focused on the graph-based methods for few-shot learning.

For segmentation tasks, few-shot semantic segmentation refers to the pixel-level prediction of new categories on the query set, given only a few labeled support images. For example, Shaban et al. [158] first presents a dual branch parallel network for one-shot segmentation, known as OSLSM, including a conditioning branch and a segmentation branch. The conditioning branch extracts representative high-level features from the supporting image-label pair, whilst the segmentation branch integrates the parameters learned from the conditioning branch and performs a segmentation mask on the query image. Other variants of OSLSM include Co-FCN [144], PL+SEG [33] and MDL [34]. All of which extend such dual branch structure to achieve a substantial performance improvement. In the AMP model, Siam et al. [162] replaces the guidance branch with a multi-resolution weight. Moreover, SG-One [207] proposed a Masked Average Pooling block (MAP) to extract the representative vectors of support objects. Then the segmentation mask was predicted via a similarity guidance network. More recently, Wang et al. [188] presents a novel prototype alignment network, called PANet, based on non-parametric metric learning.

## 5.2/ MAPNET: A MULTISCALE ATTENTION-BASED PROTOTYPICAL NETWORK

### 5.2.1/ INTRODUCTION

Despite the undeniable success of deep learning-based methods in various application domains, much research dedicates to exploring advanced technologies in limited-data and challenging scenarios, such as robotics [35], natural language processing [184, 220], and drug discovery in medical applications [1]. Recently, semi-supervised learning [105, 48, 147] has emerged as a hot topic in the computer vision community. Contrary to leveraging a large amount of data, few-shot learning aims to recognize new categories under limited supervision. Especially for few-shot segmentation task, the trained model

predicts pixel-level mask of new categories on the query image, given only a few labeled support images. The semantic guidance ability of support images and the generalizability to unseen class may significantly affect the segmentation performance.

Existing methods generally address this problem by learning a set of parameters or prototypes from **support** images and guiding the pixel-wise segmentation on the **query** image. However, most of the previous studies do not explore the support information sufficiently, which is not taking enough advantage of potential semantic information of support images. Usually, they only consider a simple connection between the support set and query set (e.g., cosine similarity), which is adverse to the generalizability.

For the above reasons, we propose a novel few-shot segmentation network called multiscale attention-based prototypical network (MAPnet). To fully exploit the representative features from labeled support images, our method extracts rich contextual information via a multiscale feature enhancement module. This module consists of three elaborated branches that aggregate multiscale features of target classes. Multiple learned prototypes provide further similarity-based guidance on the query feature, containing multiscale feature attention. Then we employ the attention mechanism to adaptively weight the probability maps for the final mask prediction. We find that this method effectively strengthens the segmentation model's generalizability, especially for the 5-shot setting. Moreover, the use of attention-based gating accelerates the convergence to a lower loss. The network was trained in an end-to-end manner without any post-processing steps. Figure 5.1 illustrates the overall workflow of our MAPnet.

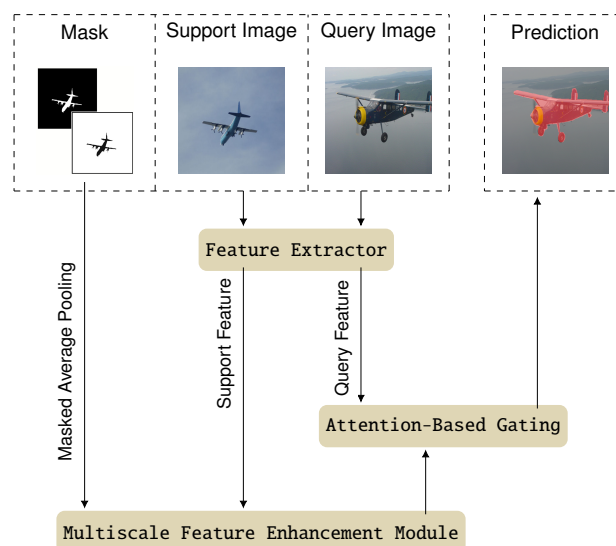


Figure 5.1: An overview of the proposed method (MAPnet). Given a query image of a new category, e.g., aeroplane, the goal of few-shot segmentation is to predict a mask of this category regarding only a few labeled samples.

### 5.2.2/ PROBLEM SETTING

The primary motivation of the few-shot segmentation is to develop a segmentation model with high generalizability. Given only one or a few examples, the model can produce pixel-level prediction with sufficient accuracy on a new category. Usually, few-shot learning is considered as a  $N$ -way- $K$ -shot classification task that discriminates between  $N$  classes with  $K$  examples per category.

In this work, we adopt the problem definition proposed in [158, 162, 188]. Suppose there are two semantic class sets  $L_{train}$  and  $L_{test}$ , the few-shot segmentation model deals with a dataset  $D = \{D^{train}, D^{test}\}$  where  $D^{train}$  and  $D^{test}$  are composed of image samples including at least one pixel belonging to  $L_{train}$  and  $L_{test}$ , respectively. The training set, which contains  $N_{train}$  image samples, can be defined as  $D^{train} = (X^i, Y(l)^i)_i^{N_{train}}$  where  $X^i$  denotes the  $i^{th}$  training image and  $Y(l)^i$  is the corresponding segmentation mask of class  $l$ . The test set is given as  $D^{test} = (X^i, Y(l)^i)_i^{N_{test}}$ . It is important to note that the model is tested on new semantic classes that do not belong to the training set, i.e.  $L_{train} \cap L_{test} = \emptyset$ .

Both the training and test sets contain several episodes that consist of a set of labeled support images  $S = (x_s^i, y(l)_s^i)_{i=1}^K$  and a set of query images  $Q = (x_q, y_q(l))$  where  $l$  is the semantic class. The support set comprises  $K$  labeled examples for each of the  $N$  classes, which defines a  $N$ -way- $K$ -shot segmentation. During training,  $episodes = (S, Q)$  randomly sampled from  $D^{train}$  are used to perform segmentation on the query set, namely  $\hat{y}_q = (S, x_q)$ . The performance is measured by a loss function  $loss(\hat{y}_q, y_q)$  where  $y_q$  is the corresponding segmentation mask. Therefore, the optimal parameters of few-shot segmentation model are  $\theta^* = \arg \min_{\theta} loss(\hat{y}_q, y_q)$ . While testing, the model is given a set of labeled support images sampled from  $D^{test}$ . Then the few-shot segmentation model is expected to predict the segmentation mask on the query image for the relative class. By taking advantage of prior knowledge, the model can rapidly generalize for a new class of limited supervised information.

### 5.2.3/ METHOD

The proposed MAPnet, as shown in Figure 5.2, is based on the prototypical network. The prototype learning-based methods enable the network to learn a set of feature vectors with adequate discriminative information. In the early work of [166], researchers proposed a prototypical network that learns a common metric space. Few-shot classification can be achieved by computing distances to prototype representations of each class. However, such methods do not explore potential semantic information of support images in sufficient depth. The learned prototypes provide limited semantic guidance on the query feature, constraining the segmentation model's generalizability. Therefore we adopt the idea of prototype learning and introduce a novel few-shot segmentation method

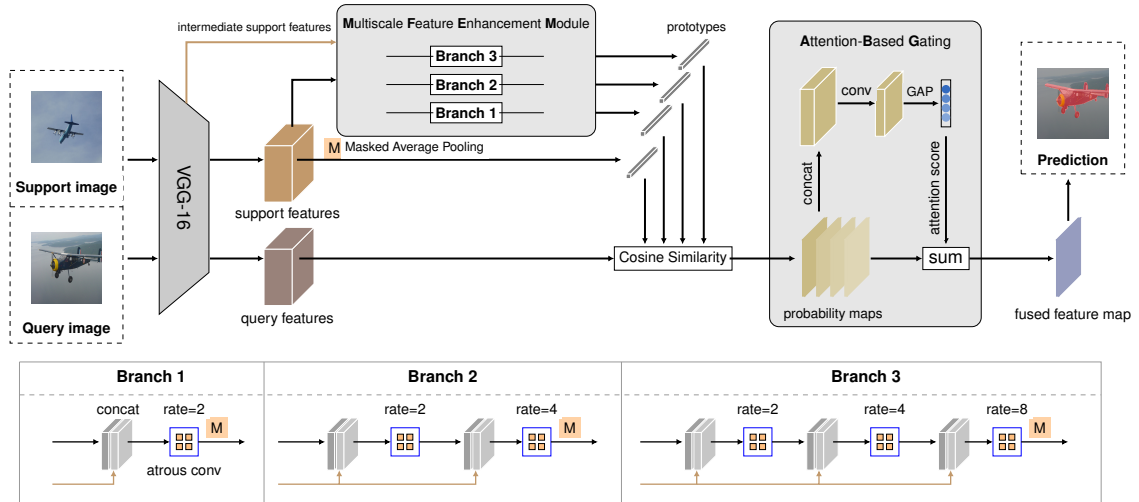


Figure 5.2: Illustration of the proposed method (MAPnet) for few-shot semantic segmentation.

with multiscale feature attention.

In general, our method contains a feature extractor, a multiscale feature enhancement module, and an attention-based gating (see Figure 5.2). The support images and query images are embedded into a high-level feature space via a shared feature extractor. Concerning the practical implementation, we employ the first five convolutional blocks of VGG-16 [164] as the backbone network. The convolutions in the fifth convolutional blocks are replaced by atrous convolutions with a rate of 2. Besides, we retain the third convolutional block’s output as intermediate features for further multiscale feature enhancement.

#### \* Multiscale Feature Enhancement Module

Generally, each category that appeared in the input images differs in shape and size. The uniscale filters learned by the neural network may lead to many restrictions on the similarity-guided semantic guidance. Thus, we define a multiscale feature enhancement module (MFE module) to provide multiscale feature supervision. In consideration of the trade-off between high performance and computational cost, we elaborate three branches in MFE module. Each branch takes the intermediate support features and high-level support features as input. The support features are resized and concatenated for providing effective feature aggregation. This module enables the few-shot segmentation network more expressive as the model becomes deeper and wider. Empirically, we employ multiscale atrous convolution with rates  $r = 2, 4, 8$  to preserve more spatial and contextual information.

In order to enhance the discriminative power of the model, we leverage both foreground and background information of support images, known as  $y(l)(+, -)$ , to extract the representative prototypes of target classes  $l$ . The background information provides complementary clues for semantic understanding. Masked Average Pool-

ing [207] is used in the process to compute these feature vectors. The set of feature vectors can be defined as  $V = \{v_0(+, -), v_1(+, -), \dots, v_n(+, -)\}$  where  $v_n(+, -)$  is the  $n_{th}$  pair of support feature vectors. Each pair of prototypes is used to generate the corresponding probability map with multiscale feature attention.

#### \* Attention-Based Gating

Different from existing few-shot segmentation methods, our model produces a set of similarity-guided probability maps by estimating the distance between a series of representative prototypes and the high-level query features. Following the work in [207, 188], we employ the cosine similarity as the non-parametric nearest neighbor classifier. Thus we employ the attention-based gating block as a combination strategy to generate an optimal mask prediction. Suppose that the concatenation of probability maps is  $P$ , the attention score  $g$  can be defined as  $g = softmax(w * P)$ , where  $w$  denotes a convolutional layer and global average pooling layer (GAP) [110]. Then our network learns the convolutional kernels  $\rho$  over the fused probability map. More formally,

$$\hat{y}_q = softmax[\rho * \sum_i^I (g_i \cdot p_i)] \quad (5.1)$$

where  $g_i$  and  $p_i$  denote the  $i_{th}$  attention score and probability map, respectively.

## 5.2.4/ EXPERIMENTS

### 5.2.4.1/ SETUP

**Dataset** We evaluate the proposed method on the PASCAL-5<sup>i</sup> dataset, which derives from PASCAL VOC 2012 [42] with SBD [67] augmentation. This dataset was firstly created by Shaban et al. [158], then widely used in the few-shot segmentation task. Similar to the setup of OSLSM [158], we sample 5 classes out of all 20 categories as test label-set  $L_{test} = \{5i + 1, \dots, 5i + 5\}$  with  $i$  being the folder number. The remaining 15 classes form the train label-set  $L_{train}$ . As shown in Table 5.1, our model is trained on three splits, then tested on the rest one in a cross-validation manner. In this work, we evaluate the performance of our model on 1,000 randomly sampled episodes for each folder.

**Implementation details** We conduct the experiments with implementations in PyTorch [133]. The backbone network (i.e., VGG-16) was initialized with pre-trained weights on ImageNet [153]. We resized the input images to  $320 \times 320$  with random horizontal flipping. All the few-shot segmentation models were trained on a single Nvidia TITAN Xp GPU with 12GB memory, using stochastic gradient descent (SGD) with a batch size of 1, a momentum of 0.9, and weight decay of 0.0005 for a maximum of 40,000 iterations. The

Table 5.1: Training and evaluation on PASCAL-5<sup>i</sup> dataset using 4-fold cross-validation, where  $i$  denotes the number of subsets.

Dataset	Test classes
Pascal-5 <sup>0</sup>	aeroplane, bicycle, bird, boat, bottle
Pascal-5 <sup>1</sup>	bus, car, cat, chair, cow
Pascal-5 <sup>2</sup>	diningtable, dog, horse, motorbike, person
Pascal-5 <sup>3</sup>	potted plant, sheep, sofa, train, tv/monitor

initial learning rate was set to 1e-3 and reduced by 0.1 every 10,000 iterations.

**Evaluation metrics** Following the previous works on the few-shot segmentation [158, 207, 188], we apply two standard metrics to evaluate the performance of learning models: mean-IoU and binary-IoU. Generally, the mean Intersection-over-Union (mean-IoU) is used to measure each foreground class’s accuracy and average over all the categories. Binary-IoU deals uniformly with all object categories as one foreground class and averages the IoU of both foreground and background. Based on these two metrics, we can fairly compare the accuracy in terms of 1-way N-shot semantic segmentation.

#### 5.2.4.2/ EVALUATION

In this part, we report the experimental evaluations on the benchmark dataset. Table 5.2 shows the comparison result of our method MAPnet and other previous methods in terms of 1-way 1-shot and 1-way 5-shot segmentation. We observe that our model achieves 48.2% on the whole for the 1-way 1-shot setting, which substantially outperforms the baseline network OSLSM by +7.4%. Also, the performance of MAPnet is competitive to the state-of-the-art method PANet. Our model earns the largest gain of +1.4% on PASCAL-5<sup>3</sup> compared to PANet, where the test classes are **potted plant**, **sheep**, **sofa**, **train** and **tv/monitor**. Our method yields a mean IoU of 56.0% overall with five support images, which achieves significant improvement over other baseline networks. Compared to SG-One, which has a simple but effective prototypical structure, we can see that the proposed method leads to a relative improvement of +9.7% on PASCAL-5<sup>0</sup>, +6.5% on PASCAL-5<sup>1</sup>, +9.8% on PASCAL-5<sup>2</sup>, and +9.4% on PASCAL-5<sup>3</sup>.

Besides, we report the averaged binary-IoU on the four-fold cross-validation in Table 5.3. Our method shows remarkable improvement in 1-way 5-shot, which gains an increment of 5.1% comparing to 1-way 1-shot. The main reason behind the increase of accuracy is that our multiscale feature enhancement module provides richer contextual information for the semantic guidance of target classes. Namely, the attention-based multiscale feature aggregation becomes more prominent as the number of support images increases. Moreover, we replace the attention-based gating with element-wise addition, aiming to

Table 5.2: Results of 1-way 1-shot and 1-way 5-shot semantic segmentation on PASCAL-5<sup>i</sup> using mean-IOU(%) metric. The results of 1-NN and LogReg are reported by [158].

Methods	1-shot					5-shot				
	Pascal-5 <sup>0</sup>	Pascal-5 <sup>1</sup>	Pascal-5 <sup>2</sup>	Pascal-5 <sup>3</sup>	Mean	Pascal-5 <sup>0</sup>	Pascal-5 <sup>1</sup>	Pascal-5 <sup>2</sup>	Pascal-5 <sup>3</sup>	Mean
1-NN	25.3	44.9	41.7	18.4	32.6	34.5	53.0	46.9	25.6	40.0
LogReg	26.9	42.9	37.1	18.4	31.4	35.9	51.6	44.5	25.6	39.3
OSLSM [158]	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
co-FCN [144]	36.7	50.6	44.9	32.4	41.1	37.5	50.0	44.1	33.9	41.4
SG-One [207]	40.2	<b>58.4</b>	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
PANet [188]	42.3	58.0	<b>51.1</b>	41.2	48.1	<b>51.8</b>	64.6	<b>59.8</b>	46.5	55.7
MAPnet	<b>42.9</b>	58.3	48.8	<b>42.6</b>	<b>48.2</b>	51.6	<b>65.1</b>	58.4	<b>48.8</b>	<b>56.0</b>

Table 5.3: Results of 1-way 1-shot and 1-way 5-shot segmentation on PASCAL-5<sup>i</sup> using binary-IOU(%) metric.  $\Delta$  denotes the difference between 1-shot and 5-shot.

Mehtods	1-shot	5-shot	$\Delta$
co-FCN [144]	60.1	60.2	0.1
OSLSM [158]	61.3	61.5	0.2
MDL [34]	63.2	63.7	0.5
PL+SEG [33]	61.2	62.3	1.1
AMP-2+FT [162]	62.2	63.8	1.6
SG-One [207]	63.1	65.9	2.8
PANet [188]	66.5	70.7	4.2
MAPnet	<b>66.7</b>	<b>71.8</b>	<b>5.1</b>

compare the convergence speed of our method trained with and without multiscale attention. As shown in Figure 5.4, we observe that the attention-based gating speeds up the convergence and reaches a lower loss, which also earns a 3.5% gain in accuracy.

Furthermore, we demonstrate qualitative results on PASCAL-5<sup>i</sup> in Figure 5.3. We present different cases involving outdoor scenes and indoor scenes. These examples show the high discriminatory power and generalizability of our method. It is capable of extracting sufficient contextual information of the target class, and a challenging case is shown in Figure 5.8 row 4. We also present two typical failure cases in our experiments. Based on the observation of the first failure case, we find that it is not easy to distinguish the objects with similar characteristics, especially when these objects are placed in an overlapping manner. Another failure case shows that the model has a limited capacity to recognize irregular objects and their boundary delineation. These failure cases may be challenging issues in future work.

#### 5.2.4.3/ TEST WITH WEAK ANNOTATIONS

To further validate the generalization ability of our model, we report the experimental results on different weak annotations in Table 5.4, including scribbles and bounding box annotations. Different from tedious and inefficient per-pixel annotating, these weak annotations are frequently used in interactive image segmentation [106]. In our experiments,













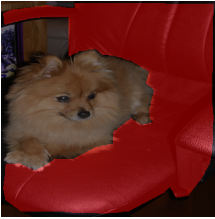

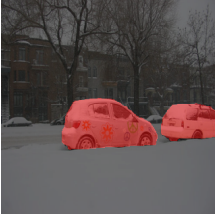
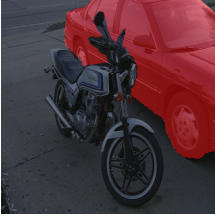

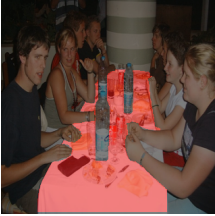


Class	Support Image	Query GT	Prediction
Sheep			
			
Sofa			
			
Failure cases			
Car			
			

Figure 5.3: Qualitative results of our method for 1-way 1-shot segmentation on the PASCAL-5<sup>i</sup> dataset.

the pixel-wise masks of support images are replaced by the corresponding weak annotations at the test time.

In general, our model's performance using weak annotations is comparable to the result with pixel-wise annotations, indicating the robustness of MAPnet. We also observe that



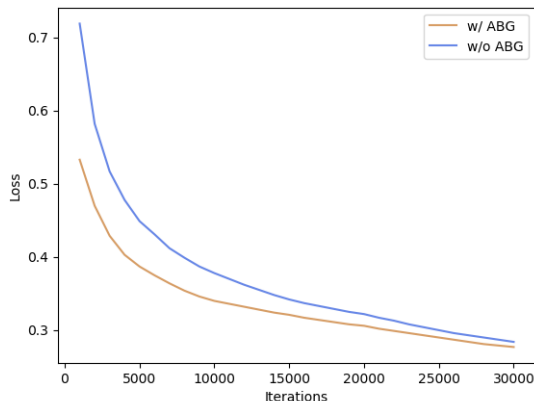


Figure 5.4: Training loss of models with and without attention-based gating (ABG) for 1-way 1-shot segmentation on PASCAL-5<sup>0</sup>.

Table 5.4: Evaluation results of using different types of annotations in mean-IoU(%) metric.

Methods	1-shot			5-shot		
	Dense	Scribble	Bbox	Dense	Scribble	Bbox
PANet [188]	48.1	44.8	45.1	55.7	54.6	52.8
MAPnet	48.2	44.1	45.7	56.0	53.5	53.7

using bounding box annotations achieves higher accuracy than using scribble annotations. A potential reason could be that our method learns more representative prototypes within the valid region of the bounding box. Figure 5.5 shows some qualitative examples of the segmentation results.

## 5.3/ RDNET: INCORPORATING DEPTH INFORMATION INTO FEW-SHOT SEGMENTATION

### 5.3.1/ INTRODUCTION

With the advent of multiple sensory modalities, multimodal data has attracted much attention in the computer vision domain. As one of the most commonly-used modalities, depth-sensing cameras provide rich geometric information of the scenes. Several deep neural networks exploit these depth maps as an addition image channel [68, 118] or point cloud in 3D space [140, 141]. Arguably, the integration of additional depth features in semantic image segmentation leads to significant performance improvement. Different from fully supervised semantic segmentation, few-shot segmentation concentrates on the generalization of segmentation ability to unseen categories given only a few samples. To be specific, some existing few-shot segmentation methods learn the representative features













Class	Support Image	Query GT	Prediction
Train			
			
Potted plant			
			

Figure 5.5: Qualitative results of our model using scribble and bounding box annotations for 1-way 5-shot setting. The chosen example in support images shows the annotation types.

for each target class in the support images, then guide the pixel-level prediction on the query image. However, the generalization and discrimination abilities of these methods still remain to be improved, especially for complex scenes.

For the above reasons, we take inspiration from existing RGB-centric methods for few-shot semantic segmentation and propose a two-stream deep neural network based on metric learning, called RDNet. The original intention of our work is to incorporate supplementary depth information into a few-shot segmentation model. As shown in Figure 5.6, the proposed RDNet employs both RGB and depth images of the same scene in the support and query set. The abstract foreground and background features of target classes are embedded into the corresponding embedding space. These prototype representations learned from RGB and depth inputs provide further similarity guidance on the query feature. Then our RDNet fuses multiple probability maps generated by the two streams into a joint prediction. In this way, our method outperforms the baseline networks with higher accuracy.

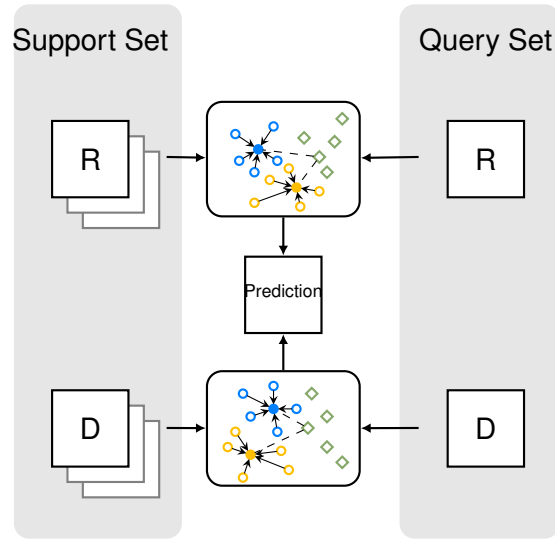


Figure 5.6: Overview of the proposed RDNet approach. R and D indicate the RGB and depth image input, respectively. The abstract features of labeled support images are mapped into the corresponding embedding space (circles). Multiple prototypes (blue and yellow solid circles) are generated to perform semantic guidance (dashed lines) on the corresponding query features (rhombus). RDNet further produces the final prediction by combining the probability maps from RGB and depth stream.

Furthermore, we report the experimental results on a new benchmark dataset, Cityscapes-3<sup>i</sup>. Different from the frequently-used PASCAL-5<sup>i</sup> dataset for object segmentation, Cityscapes-3<sup>i</sup> is derived from the large-scale Cityscapes dataset, which consists of diverse urban street scenes at varying times. Complex category information greatly increases the difficulty of scene understanding, especially with limited supervisory samples. To tackle this challenge, we conduct various comparative experiments to exploit the potential of depth information and effective fusion pattern. To the best of our knowledge, we are the first to facilitate the few-shot segmentation problem with additional depth cues. This work also promotes the use of multimodal data in the few-shot learning field.

### 5.3.2/ PROBLEM SETTING

In this work, few-shot semantic segmentation involves three datasets: a training set  $D_{train}$ , a support set  $D_s$ , and a query set  $D_q$ . The segmentation model is trained on  $D_{train}$ , and evaluated on  $D_s$  and  $D_q$ . Moreover, we adopt the training and testing protocols in Suppose the set of semantic classes in  $D_{train}$  is  $C_{seen}$ . We assume that the set of classes at test time,  $C_{unseen}$ , does not overlap with  $C_{seen}$ , i.e.  $C_{seen} \cap C_{unseen} = \emptyset$ . We formally define these datasets in the following lines:

- $D_{train} = (x_i^R, x_i^D, y(l)_i)_{i=1}^N$ , where  $x_i^R$  is a color image,  $x_i^D$  is a depth image of the same scene,  $y(l)_i$  denotes the corresponding segmentation mask of class  $l$  ( $l \in C_{seen}$ ), and

$N$  indicates the number of training examples.

- $D_s = (x_j^R, x_j^D, y(l)_j)_{j=1}^M$ , where  $x_j^R$  and  $x_j^D$  denote the corresponding RGB and depth image,  $y(l)_j$  is the mask for the semantic class  $l$  ( $l \in C_{unseen}$ ), and  $M$  indicates the number of labeled samples given in the test phase.
- $D_q = (x_j^R, x_j^D)_{j=1}^n$  is the query set of  $n$  pairs of RGB and depth images. Evaluations on  $D_q$  show the relative performance of the models.

The goal of few-shot segmentation is to train a model  $f$  with high discriminative power and generalizability from  $D_{train}$ , then produces a segmentation prediction  $\hat{y}_q$  on  $D_q$  given a support set  $D_s$ . The performance is measured by a loss function  $l(\hat{y}_q, y_q)$ , where  $y_q$  is the corresponding annotation. The optimal parameters of few-shot segmentation model are  $\theta^* = \arg \min_{\theta} l(\hat{y}_q, y_q)$ . Usually, if the support set consists of  $K$  labeled samples for each of  $C$  semantic classes, we consider such few-shot learning problem as  $C$ -way  $K$ -shot segmentation task.

### 5.3.3/ METHOD

The main motivation of our work is to facilitate the few-shot segmentation task by incorporating complementary depth information. Existing supervised semantic segmentation approaches for RGB-D data do not offer a satisfactory solution to learn new categories rapidly from limited data. For this reason, we employ ideas from previous work of non-parametric metric learning and propose a two-stream deep neural network (RDNet). The main novelty of this study is to separately learn the RGB and depth prototype representations in different embedding spaces. The learned prototypes are applied to the corresponding query features as semantic guidance. Then we integrate the results from these two streams for an improved segmentation performance.

- **RGB-D input:** As shown in Figure 5.7, the proposed RDNet consists of two mirrored prototypical networks, which process RGB and depth input separately. Note that the support and query set through the depth stream provide the same scene information as the RGB Stream. Then the support images are embedded into high-level abstract features via a base network. For efficient implementation, we adopt a VGG-16 as the backbone network following the setup in [188]. In this way, we can map RGB and depth data into different embedding spaces.
- **Prototype learning for RGB-D data:** Snell et al. [166] proposed a prototypical network that learns a common metric space. Few-shot classification can be achieved by computing distances to prototype representations of each class. We employ

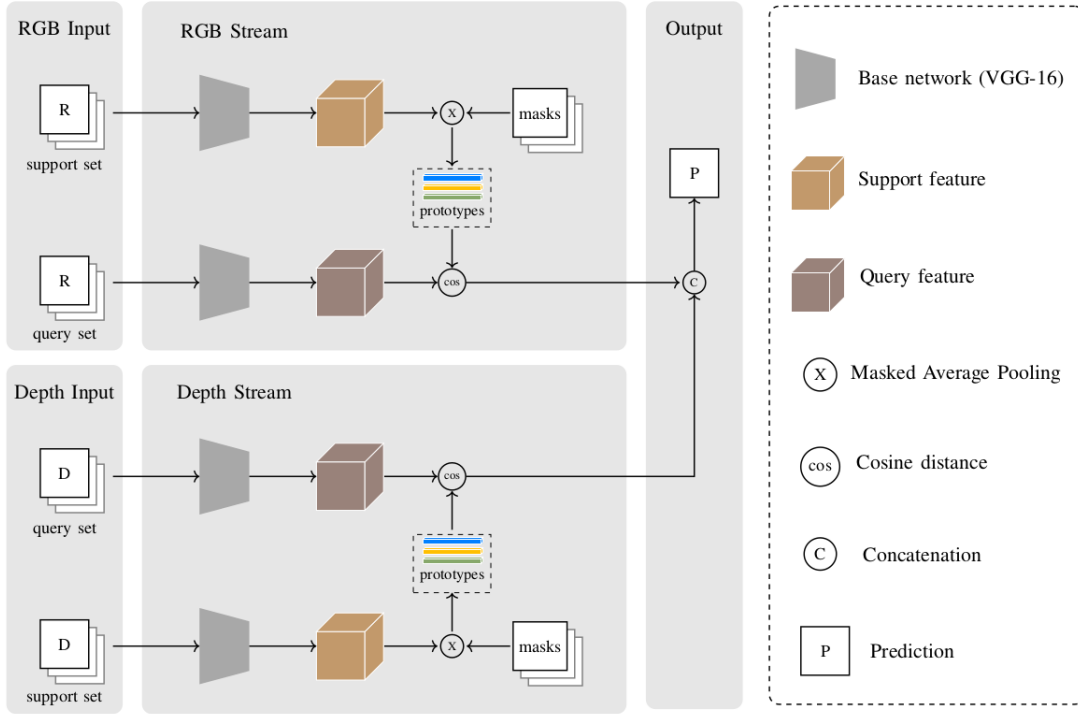


Figure 5.7: Details of the proposed RDNet architecture. It includes two mirrored streams: an RGB stream and a depth stream. Each stream processes the corresponding input data, including a support set and a query set. The prototypes of support images are obtained by masked average pooling. Then the semantic guidance is performed on the query feature by computing the relative cosine distance. The results from these two streams are combined at the late stage.

the Masked Average Pooling [207] to build pre-class prototypes from both foreground and background information of the support images. Given a support set  $D_s = (x_j^R, x_j^D, y(l)_j)_{j=1}^M$ , let  $F(l)_j$  be the output feature maps of the base network with support RGB or depth input. Then  $F(l)_j$  denotes the resized feature maps, which have the same width  $w$  and height  $h$  as the semantic mask  $y(l)_j \in \{0, 1\}^{W \times H}$ . The prototype of target class  $l$  can then be defined via Masked Average Pooling by the following equation:

$$p_c = \frac{1}{M} \sum_j \frac{\sum_{w=0, h=0}^{(w, h)} F(l)_j^{(w, h)} \mathbb{1}[y(l)_j^{(w, h)} = l]}{\sum_{w=0, h=0}^{w, h} \mathbb{1}[y(l)_j^{(w, h)} = l]} \quad (5.2)$$

where  $\mathbb{1}(\cdot)$  is the indicator function that equals to 1 if the argument is true or 0 otherwise. Similarly, the prototypes for the background can be computed with  $\mathbb{1}[y(l)_j^{(w, h)} \neq l]$ . It is notable that both foreground and background information of RGB and depth images should be considered in this work. These representative prototypes are the premise of reliable semantic guidance. To take an example, Figure shows the visualization of RGB and depth prototype representations in our experi-

Table 5.5: Training and evaluation on Cityscapes-3<sup>i</sup> dataset using 3-fold cross-validation, where  $i$  denotes the number of subsets.

Dataset	Test classes
Cityscapes-3 <sup>0</sup>	road, sidewalk, bus
Cityscapes-3 <sup>1</sup>	vegetation, terrain, sky
Cityscapes-3 <sup>2</sup>	human, car, building

ments.

- **Similarity guidance and feature fusion:** We compare the abstract query feature with expressive prototypes using distance metric learning method. To be specific, we map the query feature vector into the corresponding embedding space. The computed cosine distance indicates the similarity of target class. Besides, according to the previous work in fully-supervised semantic segmentation with RGB-D data [68, 32], there are two main fusion strategies, i.e., early fusion and late fusion [6, 145]. In our work, we employ the late fusion strategy, and concatenate all the probability maps generated from RGB and depth streams for a joint prediction. More formally,

$$\hat{y}_q = \text{softmax}[w * (p^R \vee p^D)] \quad (5.3)$$

where  $w$  denotes the convolution kernels for upsampling.

#### 5.3.4/ CITYSCAPES-3I DATASET

To fully exploit few-shot semantic segmentation with additional depth information, we create a new dataset, named Cityscapes-3<sup>i</sup>. We adopt the annotated RGB images and the depth images of the same scene from the Cityscapes dataset [29]. Cityscapes is a popular benchmark dataset for semantic understanding of outdoor scenes, which consists of thousands of precise depth images and pixel-wise semantic segmentation. Compared with object segmentation datasets such as PASCAL VOC [42] and COCO [111], it is more challenging to predict a pixel-wise mask for semantic classes in the image of Cityscapes. First, Cityscapes contains more complex urban street scenes. Images provide a broader perspective from the ground to the sky, involving a variety of categories. Then, most of the categories in the image have irregular shapes and lack distinct boundaries. Objects may overlap and be arranged randomly. Therefore it is a difficult task for segmentation models to learn characteristic features from only a few labeled samples and generalize to unseen classes.

We adopt all the RGB-D image pairs as well as the corresponding segmentation masks from Cityscapes training set for training, referred to as  $D_{train}$ . The test set  $D_{test}$  is formed by including all the samples in Cityscapes validation set. Then we choose 9 typical cate-

gories out of 30 as our target classes, containing **road**, **sidewalk**, **bus**, **vegetation**, **terrain**, **sky**, **human**, **car**, **building**. Following the setup of few-shot segmentation dataset PASCAL-5<sup>i</sup>, we sample 3 classes out of all 9 categories as test label-set  $L_{test} = \{3i + 1, 3i + 2, 3i + 3\}$  where  $i \in [1, 3]$  denotes the number of subsets, and the remaining 6 classes form the train label-set  $L_{train}$  (see Table 5.5). Namely,  $L_{train} \cap L_{test} = \emptyset$ . The images in  $D_{train}$  and  $D_{test}$  contain at least one pixel in the semantic mask from the label-set  $L_{train}$  and  $L_{test}$ , respectively. Moreover, we reset the pixels in segmentation masks that not belong to the corresponding label-sets as the background. In our experiments, we train and evaluate the proposed model on 3 folders in a cross-validation manner. For each folder, we take a random 500 samples and average the results from 5 runs to evaluate the performance of the models.

### 5.3.5/ EXPERIMENTS

#### 5.3.5.1/ SETUP

**Implementation details** We conduct the experiments with implementations in PyTorch [133]. The backbone network (i.e., VGG-16) was initialized with pre-trained weights on ImageNet [153]. We resized the input images to 768×384 and trained on a single Nvidia TITAN Xp GPU with 12GB memory. All the few-shot segmentation models were trained using stochastic gradient descent (SGD) with a batch size of 1, a momentum of 0.9, and weight decay of 0.0005 for a maximum of 30,000 iterations. The initial learning rate was set to 0.0001 and reduced by 0.1 every 10,000 iterations.

**Evaluation metrics** Following the previous works on few-shot segmentation [158, 207, 188], we apply two standard metrics to evaluate the performance of learning models: mean-IoU and binary-IoU. Generally, the mean Intersection-over-Union (mean-IoU) is used to measure the accuracy of each foreground class and average over all the classes. Binary-IoU deals uniformly with all object categories as one foreground class and averages the IoU of both foreground and background. Based on these two metrics, we can fairly compare the accuracy and efficiency of baselines in terms of 1-way N-shot semantic segmentation.

#### 5.3.5.2/ EXPERIMENTAL RESULTS

In Table 5.6, we illustrate the performance of our proposed RDNet and other baseline methods on Cityscapes-3<sup>i</sup>, including 1-way 1-shot and 1-way 2-shot semantic segmentation. First, we observe that using RGB data provides better segmentation results than using depth data as input. Moreover, one can also notice that a simple concatenation of



Table 5.6: Results of 1-way 1-shot and 1-way 2-shot semantic segmentation on Cityscapes-5<sup>i</sup> using mean-loU(%) metric.

Methods	Modality	1-way 1-shot				1-way 2-shot			
		Cityscapes-3 <sup>0</sup>	Cityscapes-3 <sup>1</sup>	Cityscapes-3 <sup>2</sup>	Mean	Cityscapes-3 <sup>0</sup>	Cityscapes-3 <sup>1</sup>	Cityscapes-3 <sup>2</sup>	Mean
PANet	RGB	35.2	19.7	32.1	29.0	37.2	23.2	36.7	32.4
RDNet-R	RGB	35.7	22.3	32.6	30.2	36.7	24.1	37.5	32.8
PANet	Depth	32.6	14.5	19.3	22.1	34.2	15.8	22.5	24.2
RDNet-D	Depth	35.1	15.8	21.0	24.0	33.7	17.3	25.3	25.4
RDNet-concat	RGB-D	33.8	15.7	20.7	23.4	34.3	17.9	26.9	26.4
RDNet (ours)	RGB-D	<b>36.8</b>	<b>23.5</b>	<b>33.3</b>	<b>31.2</b>	<b>37.3</b>	<b>26.1</b>	<b>37.6</b>	<b>33.7</b>

Table 5.7: Per-class mean-loU(%) comparison of ablation studies for 1-way 1-shot semantic segmentation

Class	RDNet	RDNet-R	RDNet-D
Mean	<b>31.2</b>	30.2	24.0
Road	83.0	80.9	<b>84.4</b>
Sidewalk	<b>17.8</b>	15.7	15.7
Bus	9.5	<b>10.6</b>	5.3
Vegetation	<b>43.1</b>	40.2	26.9
Terrain	8.3	<b>10.1</b>	6.8
Sky	<b>19.1</b>	16.7	13.7
Human	<b>47.8</b>	46.6	36.9
Car	<b>12.1</b>	12.1	5.0
Building	<b>39.9</b>	39.2	21.1

RGB and depth features, RDNet-concat, does not provide satisfactory results. Indeed, RDNet-concat achieves a mIoU score of 26.4%, which is higher than the score obtained with RDNet-D (25.4%) but much lower than the score obtained by RDNet-R (32.8%) for 1-way 2-shot semantic segmentation. Our method, RDNet, outperforms other unimodal networks and concatenated approach overall. RDNet achieves a mIoU score of 31.2% for 1-way 1-shot and 33.7% for 1-way 2-shot, which represents an increase of +7% compared to RDNet-concat.

We further conduct ablation studies to investigate the validity of RDNet. The results are shown in Table 5.7. We can observe a satisfactory performance enhancement of our method for most of the classes. In particular, the **vegetation**, **sidewalk** and **sky** classes. These experimental results illustrate the effectiveness of our method and the potential of depth information in scene understanding with limited supervision.

Compared with RDNet-concat, our proposed method provides an improvement of +7.8% and +7.3% in terms of mIoU and binary IoU for 1-way 1-shot segmentation (see Table 5.8). The results also show that simple concatenation has no significant improvement in the segmentation prediction. Besides, Figure 5.8 shows the qualitative results of our method, including multimodal input and the segmentation prediction. Our model yields promising segmentation results in 1-shot settings. However, it is still challenging to distinguish the irregular objects and categories with similar characteristics in the complex



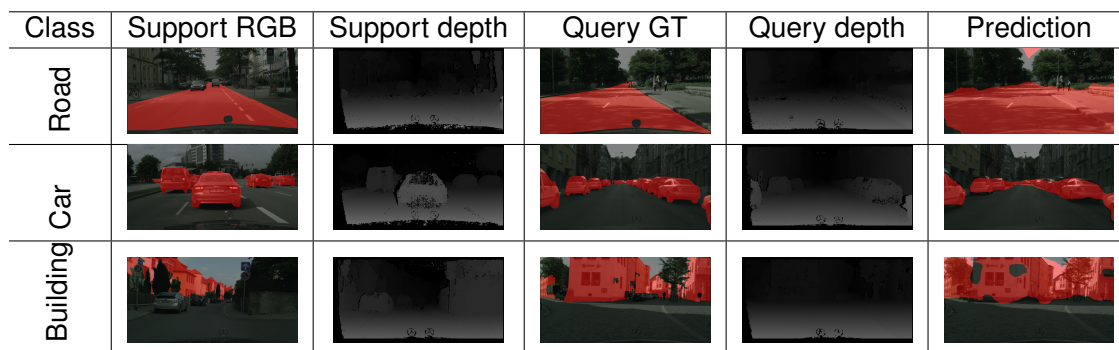


Figure 5.8: Qualitative results of our method for 1-way 1-shot semantic segmentation on Cityscapes-3<sup>i</sup>.

Table 5.8: Results of 1-way 1-shot semantic segmentation using binary IoU and the runtime.

Methods	Modality	binary IoU	Runtime
PANet	RGB	55.0	71ms
RDNet-R		56.5	65ms
RDNet-concat	RGB-D	51.9	67ms
RDNet (ours)		57.9	135ms

scenes, such as **car** and **bus**.

### 5.3.5.3/ FEATURE VISUALIZATION

To clearly demonstrate the generalization and discrimination of the proposed model, we visualize the prototype representations of target classes in the RGB and depth embedding space using t-SNE (see Figure 5.9). In our work, each figure was generated using 500 samples of test classes in Cityscapes-3<sup>i</sup>. On the whole, the prototypes generated from support RGB input can be well separated, especially for **vegetation**, **terrain**, **sky** in Figure 5.9c. Although it is challenging to produce distinctive prototypes in the depth embedding space, these prototype representations provide complementary cues regarding depth information. For example, the depth embeddings in Cityscapes-3<sup>0</sup> clearly show the discrimination on the classes **vegetation**, **terrain** and **sky** (see Figure 5.9b). Consequently, the generalizability of our few-shot segmentation network gets improved by incorporating supplementary depth information, leading to more promising prediction results.

## 5.4/ SUMMARY

Our work has presented MAPnet, a novel few-shot segmentation method based on the prototypical network. The proposed method provides effective semantic guidance on the

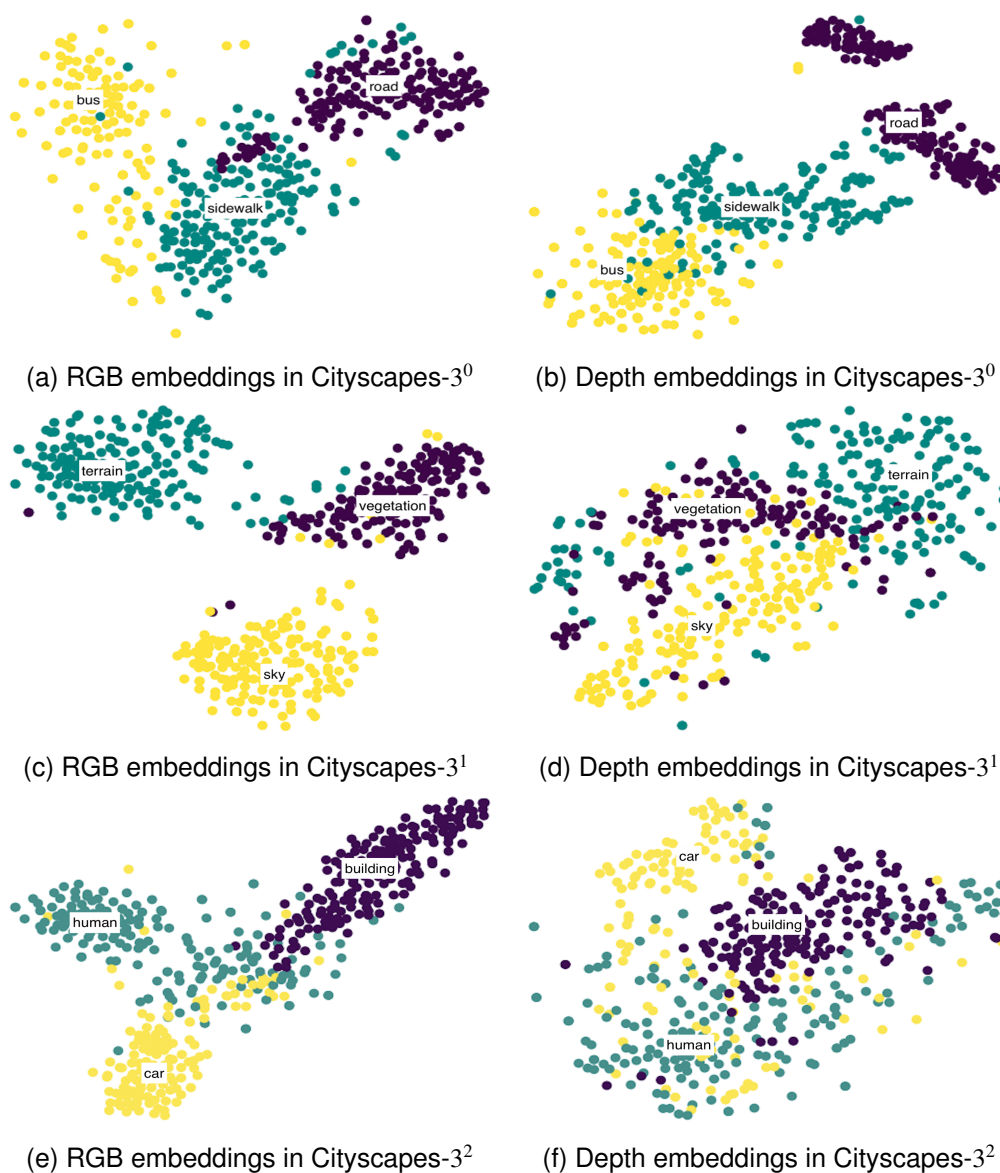


Figure 5.9: Visualization using t-SNE [179] for RGB and depth prototype representations in our RDNet.

query feature by a multiscale feature enhancement module. We elaborate the branches in this module to fully exploit the support information. Moreover, we employ the attention mechanism on the similarity-guided probability maps to produce an optimal pixel-wise prediction, which also speeds up the convergence. Extensive experiments demonstrate the improved generalizability and discriminating ability of the proposed method. Our model achieves a comparable accuracy with the state-of-the-art, outperforming most of the previous methods.

Moreover, in order to explore the multimodal image fusion in few-shot segmentation, we proposed a novel segmentation network to incorporate complementary depth information, which consists of two mirrored streams based on metric learning. To fully take advantage

of color and geometric information of the scenes, we mapped the representative features of target classes into different embedding spaces. The learned prototype representations provide effective semantic guidance on the corresponding query feature. Then we integrated the generated probability maps at a late stage. Comprehensive experiments and ablation studies on Cityscapes-3<sup>i</sup> dataset demonstrate the improved generalizability and discriminating ability of our method. The proposed RDNet is simple yet effective, and explore the use of depth information in few-shot segmentation task.

# CONCLUSION AND FUTURE WORK

*“As a technologist, I see how AI and the fourth industrial revolution will impact every aspect of people’s lives.”*

– Fei-Fei Li, *Professor of Computer Science at Stanford University*

## 6.1/ GENERAL CONCLUSION

The general aim of this thesis was to improve the semantic scene understanding for outdoor road scene using complementary multimodal data. In Chapter 2, we first introduced the fundamental background knowledge of deep neural networks. Chapter 3 explored existing multimodal image fusion methods using state-of-the-art deep learning techniques. Compared with traditional machine learning methods, deep multimodal fusion performs better in terms of accuracy when trained with huge amounts of data. Then a comprehensive review and analysis of the related work for multimodal images was made to provide the reader with a broad overview of input data [211]. Moreover, we have studied the problem of multimodal semantic segmentation from two aspects, including data and algorithms. Chapter 4 presented different semantic segmentation methods and related ablation studies with multiple data input. Chapter 5 further explored the unimodal and multimodal semantic segmentation under limited supervision.

In more detail, we have presented an extensible multi-level fusion network for fully-supervised multimodal semantic segmentation, known as CMnet, to integrate RGB and polarimetric images [210]. Moreover, a central multimodal fusion framework was introduced to adaptively learn the joint feature representations of low-level and high-level modality-specific features. We also employed statistical methods as post-processing. With regard to few-shot semantic segmentation, we first developed a novel segmentation method based on the prototypical network. The proposed MAPnet contains a multiscale feature enhancement module to fully exploit the support features, and attention mechanism to fuse multiple probability maps for an optimal pixel-wise prediction [209]. Then

we extended the RGB-centric methods to take advantage of complementary depth information [208]. Our experiments have demonstrated the effectiveness of the proposed RDNet to integrate complementary depth cues in few-shot semantic segmentation. The rich color and geometric information of scenes can provide valuable semantic features.

All the proposed deep multimodal fusion methods show excellent performance in segmentation tasks with the support of a large number of labeled image data and computing power. We employed a variety of multimodal image inputs in the experiments, such as depth maps, near-infrared images, polarization images, etc. These multimodal image data provide valuable supplementary information of the same scene, making the segmentation model more robust and efficient. Our work illustrates the effectiveness and necessity of multi-modality in outdoor scene understanding.

## 6.2/ FUTURE PERSPECTIVES

Based on the work presented in this thesis, we give some recommendations for future research.

In Chapter 4, we have introduced two multimodal image fusion methods for semantic segmentation tasks. Although the proposed CMnet and CMFnet+BF2 achieve satisfactory results, there still remain inherent drawbacks of such methods: (i) The trade-off between prediction accuracy and cost-effectiveness should be considered in practical applications. The simplicity of implementation, real-time, and scalability are very challenging problems for autonomous UGV navigation. (ii) The deep multimodal fusion requires a large number of high-quality labeled images. The training data greatly influences the stability of the segmentation model, and it is especially necessary to be wary of noisy data and missing data. Thus, it would be interesting to investigate more robust multimodal image segmentation algorithms. Furthermore, we have shown the critical role of multimodal image data in the outdoor scene understanding and indicated that different sensory modalities could affect the segmentation performance. Hence, we expect to employ multimodal data in other high-level image segmentation tasks such as panoptic segmentation. Such segmentation unifies the typically distinct tasks of semantic segmentation and instance segmentation, which represents an important step toward real-world vision systems.

Besides, we have presented advanced few-shot semantic segmentation methods in Chapter 5. Although extensive experiments and ablation studies demonstrate the improved generalizability and discriminating ability of few-shot segmentation algorithms by integrating depth information, how to fuse multimodal data in high-level few-shot learning tasks for optimal performance is still in the preliminary stage. As a future perspective, we will focus on the impact of multimodal data in semi-supervised semantic understand-

ing as well as the optimal network architecture for image fusion. As we know that deep learning-based artificial intelligence is gradually evolving from perception to cognitive intelligence, we expect deep multimodal fusion to facilitate this evolution and offer a host of innovations in the following years.



# BIBLIOGRAPHY

- [1] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.
- [2] Alireza Asvadi, Luis Garrote, Cristiano Premebida, Paulo Peixoto, and Urbano J Nunes. Multimodal vehicle detection: fusing 3d-lidar and color camera data. *Pattern Recognition Letters*, 115:20–29, 2018.
- [3] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, 2010. ISSN 09424962. doi: 10.1007/s00530-010-0182-0.
- [4] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140:20–32, 2018. ISSN 09242716. doi: 10.1016/j.isprsjprs.2017.11.011.
- [5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2644615.
- [6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.
- [7] Aayush Bansal, Xinlei Chen, Bryan Russell, Abhinav Gupta, and Deva Ramanan. Pixelnet: Towards a general pixel-level architecture. *arXiv preprint arXiv:1609.06694*, 2016.
- [8] Yannick Benezeth, Désiré Sidibé, and Jean-Baptiste Thomas. Background subtraction with multispectral video sequences. 2014.
- [9] Luis M Bergasa, Roberto Arroyo, Eduardo Romera, and M Alvarez. Efficient ConvNet for Real-time Semantic Segmentation. In *IEEE Intelligent Vehicles Symposium, Proceedings*, number Iv, pages 1789–1794. IEEE, 2017. ISBN 9781509048038. URL <http://www.robosafe.uah.es/personal/eduardo.romera/pdfs/Romera17iv.pdf>.



- [10] Mario Bijelic, Fahim Mannan, Tobias Gruber, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing Through Fog Without Seeing Fog: Deep Sensor Fusion in the Absence of Labeled Training Data. [arXiv preprint arXiv:1902.08913](#), 2019.
- [11] Marc Blanchon, Olivier Morel, Yifei Zhang, Ralph Seulin, Nathan Crombez, and Désiré Sidibé. Outdoor Scenes Pixel-wise Semantic Segmentation using Polarimetry and Fully Convolutional Network. In [VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications](#), volume 5, pages 328–335, 2019. ISBN 9789897583544. doi: 10.5220/0007360203280335.
- [12] Rachel Blin, Samia Ainouz, Stephane Canu, and Fabrice Meriaudeau. A new multi-modal rgb and polarimetric image dataset for road scenes analysis. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops](#), pages 216–217, 2020.
- [13] Hermann Blum, Abel Gawel, Roland Siegwart, and Cesar Cadena. Modular sensor fusion for semantic segmentation. In [2018 IEEE/RSJ International Conference on Intelligent Robots and Systems \(IROS\)](#), pages 3670–3677. IEEE, 2018.
- [14] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In [European conference on computer vision](#), pages 44–57. Springer, 2008.
- [15] Holger Caesar. really-awesome-semantic-segmentation. <https://github.com/nightrome/really-awesome-semantic-segmentation>, 2018.
- [16] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In [Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition](#), pages 3051–3060, 2018.
- [17] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. [arXiv preprint arXiv:1412.7062](#), 2014.
- [18] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In [Proceedings of the IEEE conference on computer vision and pattern recognition](#), pages 3640–3649, 2016.
- [19] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. [arXiv preprint arXiv:1706.05587](#), 2017. URL <http://arxiv.org/abs/1706.05587>.
- [20] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional

- Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. ISSN 01628828. doi: 10.1109/TPAMI.2017.2699184.
- [21] Liang Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11211 LNCS:833–851, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01234-2\_49.
- [22] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1978, 2014.
- [23] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 6526–6534, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.691.
- [24] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation, 2019.
- [25] Yanhua Cheng, Rui Cai, Zhiwei Li, Xin Zhao, and Kaiqi Huang. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3029–3037, 2017.
- [26] Gyeongmin Choe, Seong Heum Kim, Sunghoon Im, Joon Young Lee, Srinivasa G. Narasimhan, and In So Kweon. RANUS: RGB and NIR urban scene dataset for deep scene parsing. *IEEE Robotics and Automation Letters*, 3(3):1808–1815, 2018. ISSN 23773766. doi: 10.1109/LRA.2018.2801390.
- [27] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018. ISSN 15249050. doi: 10.1109/TITS.2018.2791533.
- [28] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *Computer Vision and Image Understanding*, 167:1 – 27, 2018. ISSN 1077-3142. doi: <https://doi.org/10.1016/j.cviu.2018.01.007>. URL <http://www.sciencedirect.com/science/article/pii/S1077314218300079>.

- [29] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 3213–3223, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.350.
- [30] Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor Semantic Segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. URL <http://arxiv.org/abs/1301.3572>.
- [31] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.
- [32] Liuyuan Deng, Ming Yang, Tianyi Li, Yuesheng He, and Chunxiang Wang. RFB-Net: Deep Multimodal Networks with Residual Fusion Blocks for RGB-D Semantic Segmentation. *arXiv preprint arXiv:1907.00135*, 2019.
- [33] Nanqing Dong and Eric Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018.
- [34] Zihao Dong, Ruixun Zhang, Xiuli Shao, and Hongyu Zhou. Multi-scale discriminative location-aware network for few-shot semantic segmentation. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, volume 2, pages 42–47. IEEE, 2019.
- [35] Yan Duan, Marcin Andrychowicz, Bradly C. Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, Pieter Abbeel, and Wojciech Zaremba. One-shot imitation learning, 2017.
- [36] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
- [37] Frederike Dübgen, Majed El Helou, Natalija Gucevska, and Sabine Süsstrunk. Near-infrared fusion for photorealistic image dehazing. *Electronic Imaging*, 2018 (16):321–1, 2018.
- [38] S Edelman and T Poggio. Integrating visual cues for object segmentation and recognition. *Optics News*, 15(5):8, 1989.
- [39] David Eigen, Marc’Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

- [40] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust RGB-D object recognition. In *IEEE International Conference on Intelligent Robots and Systems*, volume 2015-Decem, pages 681–687. IEEE, 2015. ISBN 9781479999941. doi: 10.1109/IROS.2015.7353446.
- [41] Engin Erzin, Yucel Yemez, A Murat Tekalp, Aytul Ercil, Hakan Erdogan, and Huseyin Abut. Multimodal person recognition for human-vehicle interaction. *IEEE MultiMedia*, 13(2):18–31, 2006.
- [42] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- [43] Mark Everingham, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [44] Sajad Farokhi, Jan Flusser, and Usman Ullah Sheikh. Near infrared face recognition: A literature survey. *Computer Science Review*, 21:1–17, 2016.
- [45] Di Feng, Christian Haase-Schuetz, Lars Rosenbaum, Heinz Hertlein, Fabian Duffhauss, Claudius Glaeser, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *arXiv preprint arXiv:1902.07830*, 2019.
- [46] Eduardo Fernandez-Moral, Renato Martins, Denis Wolf, and Patrick Rives. A new metric for evaluating semantic segmentation: leveraging global and contour accuracy. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1051–1056. IEEE, 2018.
- [47] Julian Fierrez-Aguilar, Javier Ortega-Garcia, and Joaquin Gonzalez-Rodriguez. Fusion strategies in multimodal biometric verification. In *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, volume 3, pages III–5. IEEE, 2003.
- [48] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.
- [49] Robert W Frischholz and Ulrich Dieckmann. Biold: a multimodal biometric identification system. *Computer*, 33(2):64–68, 2000.
- [50] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. *Machine Vision & Applications*, 25(1):245–262.

- [51] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: A survey. *Machine Vision & Applications*, 25(1):245–262, 2014. ISSN 0932-8092.
- [52] Yarín Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.
- [53] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.
- [54] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv preprint arXiv:1704.06857*, 2017. URL <http://arxiv.org/abs/1704.06857>.
- [55] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013. ISSN 02783649. doi: 10.1177/0278364913491297.
- [56] Swarnendu Ghosh, Nibaran Das, Ishita Das, and Ujjwal Maulik. Understanding Deep Learning Techniques for Image Segmentation, 2019.
- [57] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.
- [58] Alejandro González, David Vázquez, Antonio M López, and Jaume Amores. On-board object detection: Multicue, multimodal, and multiview random forest of local experts. *IEEE transactions on cybernetics*, 47(11):3980–3990, 2016.
- [59] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [60] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *2009 IEEE 12th international conference on computer vision*, pages 1–8. IEEE, 2009.
- [61] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. ISSN 15662535. doi: 10.1016/j.inffus.2018.11.017.
- [62] Joris Guerry, Bertrand Le Saux, and David Filliat. "Look at this one" detection sharing between modality-independent classifiers for robotic discovery of people. In *2017 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2017.

- [63] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European conference on computer vision*, pages 345–360. Springer, 2014.
- [64] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *IEEE International Conference on Intelligent Robots and Systems*, volume 2017-Septe, pages 5108–5115. IEEE, 2017. ISBN 9781538626825. doi: 10.1109/IROS.2017.8206396.
- [65] John S. Harchanko and David B. Chenault. Water-surface object detection and classification using imaging polarimetry. In *Polarization Science and Remote Sensing II*, volume 5888, page 588815. International Society for Optics and Photonics, 2005. doi: 10.1117/12.623542.
- [66] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 991–998, 2011. ISBN 9781457711015. doi: 10.1109/ICCV.2011.6126343.
- [67] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011.
- [68] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusetnet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016.
- [69] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10111 LNCS, 2017. ISBN 9783319541808. doi: 10.1007/978-3-319-54181-5\_14.
- [70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
- [71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015. URL <http://arxiv.org/abs/1512.03385>. cite arxiv:1512.03385Comment: Tech report.



- [73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. ISSN 01628828. doi: 10.1109/TPAMI.2015.2389824.
- [74] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
- [75] Andrew Holliday, Mohammadamin Barekatin, Johannes Laurmaa, Chetak Kandaswamy, and Helmut Prendinger. Speedup of deep learning ensembles for semantic segmentation using a model compression technique. *Comput. Vis. Image Underst.*, 164:16–26, 2017.
- [76] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation. *arXiv preprint arXiv:1905.10089*, 2019.
- [77] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.
- [78] Shang-Wei Hung, Shao-Yuan Lo, and Hsueh-Ming Hang. Incorporating Luminance, Depth and Color Information by a Fusion-based Network for Semantic Segmentation, 2018.
- [79] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.
- [80] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, Geoffrey E Hinton, and Others. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.
- [81] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art, 2017. URL <http://arxiv.org/abs/1704.05519>.
- [82] Dong-Won Jang and Rae-Hong Park. Colour image dehazing using near-infrared fusion. *IET Image Processing*, 11(8):587–594, 2017.
- [83] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer*

- vision and pattern recognition, pages 248–255. Ieee, 2009. doi: 10.1109/cvprw.2009.5206848.
- [84] Jindong Jiang, Zhijun Zhang, Yongqian Huang, and Lunan Zheng. Incorporating depth into both cnn and crf for indoor semantic segmentation. In *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 525–530. IEEE, 2017.
- [85] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation. *arXiv preprint arXiv:1806.01054*, 2018. URL <http://arxiv.org/abs/1806.01054>.
- [86] Xin Jin, Qian Jiang, Shaowen Yao, Dongming Zhou, Rencan Nie, Jinjin Hai, and Kangjian He. A survey of infrared and visual image fusion methods. *Infrared Physics & Technology*, 85:478–501, 2017.
- [87] Takumi Karasawa, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Thematic Workshops 2017 - Proceedings of the Thematic Workshops of ACM Multimedia 2017, co-located with MM 2017*, pages 35–43. ACM, 2017. ISBN 9781450354165. doi: 10.1145/3126686.3126727.
- [88] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *British Machine Vision Conference 2017, BMVC 2017*, 2017.
- [89] Bahador Khaleghi, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1): 28–44, 2013. ISSN 15662535. doi: 10.1016/j.inffus.2011.08.001.
- [90] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [91] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic Segmentation, 2018.
- [92] Pavel Kirsanov, Airat Gaskarov, Filipp Konokhov, Konstantin Sofiiuk, Anna Vorontsova, Igor Slinko, Dmitry Zhukov, Sergey Bykov, Olga Barinova, and Anton Konushin. DISCOMAN: Dataset of Indoor SCenes for Odometry, Mapping And Navigation, 2019.
- [93] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.



- [94] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014.
- [95] Dana Lahat, Tülay Adalı, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.
- [96] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
- [97] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.
- [98] Hanchao Li, Pengfei Xiong, Haoqiang Fan, and Jian Sun. DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2019. URL <http://arxiv.org/abs/1904.02216>.
- [99] Xiangtai Li, Li Zhang, Ansheng You, Maoke Yang, Kuiyuan Yang, and Yunhai Tong. Global Aggregation then Local Distribution in Fully Convolutional Networks, 2019.
- [100] Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, and Kuiyuan Yang. GFF: Gated Fully Fusion for Semantic Segmentation. *arXiv preprint arXiv:1904.01803*, 2019.
- [101] Xiaomeng Li, Lequan Yu, Chi-Wing Fu, Meng Fang, and Pheng-Ann Heng. Re-visiting metric learning for few-shot image classification. *ArXiv*, abs/1907.03123, 2019.
- [102] Yabei Li, Junge Zhang, Yanhua Cheng, Kaiqi Huang, and Tieniu Tan. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. In *Proceedings - International Conference on Image Processing, ICIP*, volume 2017-Septe, pages 1262–1266. IEEE, 2018. ISBN 9781509021758. doi: 10.1109/ICIP.2017.8296484.
- [103] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. *CoRR*, abs/1812.03904, 2018. URL <http://arxiv.org/abs/1812.03904>.
- [104] Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9906 LNCS, pages 541–557. Springer, 2016. ISBN 9783319464749. doi: 10.1007/978-3-319-46475-6\_34.

- [105] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.
- [106] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *CoRR*, abs/1604.05144, 2016. URL <http://arxiv.org/abs/1604.05144>.
- [107] Di Lin, Guangyong Chen, Daniel Cohen-Or, Pheng Ann Heng, and Hui Huang. Cascaded Feature Network for Semantic Segmentation of RGB-D Images. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 1320–1328, 2017. ISBN 9781538610329. doi: 10.1109/ICCV.2017.147.
- [108] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.
- [109] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Exploring context with deep structured models for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1352–1366, 2017.
- [110] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.
- [111] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [112] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A W M van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. ISSN 1361-8415. doi: 10.1016/j.media.2017.07.005. URL <http://dx.doi.org/10.1016/j.media.2017.07.005>.
- [113] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. *British Machine Vision Conference 2016, BMVC 2016*, 2016-September:73.1–73.13, 2016. doi: 10.5244/c.30.73.
- [114] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018.

- [115] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. ParseNet: Looking Wider to See Better. [arXiv preprint arXiv:1506.04579](https://arxiv.org/abs/1506.04579), 2015. URL <http://arxiv.org/abs/1506.04579>.
- [116] Yu Liu, Xun Chen, Juan Cheng, and Hu Peng. A medical image fusion method based on convolutional neural networks. In [2017 20th International Conference on Information Fusion \(Fusion\)](#), pages 1–7. IEEE, 2017.
- [117] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In [Proceedings of the IEEE conference on computer vision and pattern recognition](#), pages 3431–3440, 2015.
- [118] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In [2017 IEEE/RSJ International Conference on Intelligent Robots and Systems \(IROS\)](#), pages 598–605. IEEE, 2017.
- [119] Sean McMahon, Niko Sunderhauf, Ben Upcroft, and Michael Milford. Multimodal Trip Hazard Affordance Detection on Construction Sites. [IEEE Robotics and Automation Letters](#), 3(1):1–8, 2018. ISSN 23773766. doi: 10.1109/LRA.2017.2719763.
- [120] Oier Mees, Andreas Eitel, and Wolfram Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In [IEEE International Conference on Intelligent Robots and Systems](#), volume 2016-Novem, pages 151–156. IEEE, 2016. ISBN 9781509037629. doi: 10.1109/IROS.2016.7759048.
- [121] Dmytro Mishkin and Jiri Matas. All you need is a good init. [arXiv preprint arXiv:1511.06422](#), 2015.
- [122] Pim Moeskops, Jelmer M Wolterink, Bas H M van der Velden, Kenneth G A Gilhuijs, Tim Leiner, Max A Viergever, and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In [International Conference on Medical Image Computing and Computer-Assisted Intervention](#), pages 478–486. Springer, 2016.
- [123] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. [Foundations of machine learning](#). 2018.
- [124] Lionel Moisan, Pierre Moulon, and Pascal Monasse. Automatic homographic registration of a pair of images, with a contrario elimination of outliers. [Image Processing On Line](#), 2:56–73, 2012.
- [125] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detec-

- tion and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.
- [126] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for Audio-Visual Speech Recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2015-Augus, pages 2130–2134. IEEE, 2015. ISBN 9781467369978. doi: 10.1109/ICASSP.2015.7178347.
- [127] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kotschieder. The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes. In *International Conference on Computer Vision (ICCV)*, 2017. URL <https://www.mapillary.com/dataset/vistas>.
- [128] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 1520–1528, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.178.
- [129] Yu-ichi Ohta, Takeo Kanade, and Toshiyuki Sakai. An analysis system for scenes containing objects with substructures. In *Proceedings of the Fourth International Joint Conference on Pattern Recognitions*, pages 752–754, 1978.
- [130] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4980–4989, 2017.
- [131] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv preprint arXiv:1606.02147*, 2016. URL <http://arxiv.org/abs/1606.02147>.
- [132] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017.
- [133] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.
- [134] Naman Patel, Anna Choromanska, Prashanth Krishnamurthy, and Farshad Khorrami. Sensor modality fusion with CNNs for UGV autonomous driving in indoor environments. In *IEEE International Conference on Intelligent Robots and Systems*, volume 2017-Septe, pages 1531–1536. IEEE, 2017. ISBN 9781538626825. doi: 10.1109/IROS.2017.8205958.

- [135] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters - Improve semantic segmentation by global convolutional network. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 1743–1751, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.189.
- [136] Andreas Pfeuffer and Klaus Dietmayer. Optimal sensor data fusion architecture for object detection in adverse weather conditions. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2018.
- [137] Andreas Pfeuffer and Klaus Dietmayer. Robust Semantic Segmentation in Adverse Weather Conditions by means of Sensor Data Fusion, 2019.
- [138] Nathan Piasco, Désiré Sidibé, Valérie Gouet-Brunet, and Cédric Demonceaux. Learning scene geometry for visual localization in challenging conditions. In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 9094–9100. IEEE, 2019. doi: 10.1109/ICRA.2019.8794221. URL <https://doi.org/10.1109/ICRA.2019.8794221>.
- [139] Pedro Pinheiro and Ronan Collobert. Recurrent convolutional neural networks for scene labeling. In *International conference on machine learning*, pages 82–90, 2014.
- [140] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.
- [141] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.
- [142] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.
- [143] Simeng Qiu, Qiang Fu, Congli Wang, and Wolfgang Heidrich. Polarization demosaicking for monochrome and color polarization focal plane arrays. 2019.
- [144] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. 2018.
- [145] Dhanesh Ramachandram and Graham W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6): 96–108, 2017. ISSN 10535888. doi: 10.1109/MSP.2017.2738401.

- [146] Mojdeh Rastgoo, Cedric Demonceaux, Ralph Seulin, and Olivier Morel. Attitude estimation from polarimetric cameras. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8397–8403. IEEE, 2018.
- [147] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Learning Representations (ICLR)*, 2017.
- [148] Mengye Ren and Richard S. Zemel. End-to-end instance segmentation and counting with recurrent attention. *CoRR*, abs/1605.09410, 2016. URL <http://arxiv.org/abs/1605.09410>.
- [149] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9906 LNCS, pages 102–118. Springer, 2016. ISBN 9783319464749. doi: 10.1007/978-3-319-46475-6\_7.
- [150] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9351, pages 234–241. Springer, 2015. ISBN 9783319245737. doi: 10.1007/978-3-319-24574-4\_28.
- [151] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 3234–3243, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.352.
- [152] Noah A Rubin, Gabriele D’Aversa, Paul Chevalier, Zhujun Shi, Wei Ting Chen, and Federico Capasso. Matrix fourier optics enables a compact full-stokes polarization camera. *Science*, 365(6448):eaax1839, 2019.
- [153] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and Others. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [154] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 687–704, 2018.

- [155] Yoav Y. Schechner, Srinivasa G. Narasimhan, and Shree K. Nayar. Instant dehazing of images using polarization. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, 1:I–I, 2001.
- [156] Lukas Schneider, Manuel Jasch, Björn Fröhlich, Thomas Weber, Uwe Franke, Marc Pollefeys, and Matthias Räscht. Multimodal neural networks: RGB-D for semantic segmentation and object detection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10269 LNCS, pages 98–109. Springer, 2017. ISBN 9783319591254. doi: 10.1007/978-3-319-59126-1\_9.
- [157] Max Schwarz, Hannes Schulz, and Sven Behnke. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 1329–1335. IEEE, 2015.
- [158] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.
- [159] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. PST900: RGB-Thermal Calibration, Dataset and Segmentation Network. *arXiv preprint arXiv:1909.10980*, 2019.
- [160] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.
- [161] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *European conference on computer vision*, pages 1–15. Springer, 2006.
- [162] Mennatullah Siam, Boris Oreshkin, and Martin Jagersand. Adaptive masked proxies for few-shot segmentation. *arXiv preprint arXiv:1902.11123*, 2019.
- [163] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.
- [164] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [165] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Linear depth estimation from an uncalibrated, monocular polarisation image. In *European Conference on Computer Vision*, pages 109–125. Springer, 2016.



- [166] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017.
- [167] Sobhan Soleymani, Ali Dabouei, Hadi Kazemi, Jeremy Dawson, and Nasser M Nasrabadi. Multi-level feature abstraction from convolutional neural networks for multimodal biometric identification. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3469–3476. IEEE, 2018.
- [168] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 567–576, 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298655.
- [169] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [170] Dongming Sun, Xiao Huang, and Kailun Yang. A multimodal vision sensor for autonomous driving. In *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies III*, volume 11166, page 111660L. International Society for Optics and Photonics, 2019.
- [171] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion. In *Artificial Intelligence and Machine Learning in Defense Applications*, volume 11169, page 111690A. International Society for Optics and Photonics, 2019.
- [172] Yuxiang Sun, Weixun Zuo, and Ming Liu. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. ISSN 23773766. doi: 10.1109/LRA.2019.2904733.
- [173] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation, 2019.
- [174] Wayne Treible, Philip Saponaro, Yi Liu, Agnijit Das Gupta, Vinit Veerendraveer, Scott Sorensen, and Chandra Kambhamettu. Cats 2: Color and thermal stereo scenes with semantic labels. In *CVPR Workshops*, 2019.
- [175] Abhinav Valada, Gabriel Oliveira, Thomas Brox, and Wolfram Burgard. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *The 2016 International Symposium on Experimental Robotics (ISER 2016)*, Tokyo, Japan, October 2016. URL <http://ais.informatik.uni-freiburg.de/publications/papers/valada16iser.pdf>.



- [176] Abhinav Valada, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard. Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion. In *International Symposium on Experimental Robotics*, pages 465–477. Springer, 2017. doi: 10.1007/978-3-319-50115-4\_41.
- [177] Abhinav Valada, Johan Vertens, Ankit Dhall, and Wolfram Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651. IEEE, 2017.
- [178] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision*, 07 2019. doi: 10.1007/s11263-019-01188-y.
- [179] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL <http://www.jmlr.org/papers/v9/vandermaaten08a.html>.
- [180] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [181] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL <http://arxiv.org/abs/1706.03762>.
- [182] Maurice Velte. *Semantic image segmentation combining visible and near-infrared channels with depth information*. PhD thesis, Ph. D. Dissertation. bibinfoschoolBonn-Rhein-Sieg University of Applied Sciences, 2015.
- [183] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Central-Net: a Multilayer Approach for Multimodal Fusion, 2018.
- [184] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016.
- [185] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN 2016 - 24th European Symposium on Artificial Neural Networks*, pages 509–514, 2016. ISBN 9782875870278.
- [186] Robert Walraven. Polarization Imagery. *Optical Polarimetry \$- \$ Instrum and Appl*, 112(1):164–167, 1977. ISSN 0091-3286. doi: 10.1117/12.7972655.

- [187] Jinghua Wang, Zhenhua Wang, Dacheng Tao, Simon See, and Gang Wang. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In *European Conference on Computer Vision*, pages 664–679. Springer, 2016.
- [188] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9197–9206, 2019.
- [189] Ningning Wang and Xiaojin Gong. Adaptive Fusion for RGB-D Salient Object Detection, 2019.
- [190] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1451–1460. IEEE, 2018.
- [191] Lawrence B. Wolff. Polarization vision: A new sensory approach to image understanding. *Image and Vision Computing*, 15(2):81–93, 1997. ISSN 02628856. doi: 10.1016/s0262-8856(96)01123-7.
- [192] Lawrence B Wolff and Andreas G Andreou. Polarization camera sensors. *Image and Vision Computing*, 13(6):497–510, 1995.
- [193] Lawrence B. Wolff and Terrance E. Boult. Constraining object features using a polarization reflectance model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(7):635–657, July 1991. ISSN 0162-8828. doi: 10.1109/34.85655. URL <https://doi.org/10.1109/34.85655>.
- [194] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. Multimodal End-to-End Autonomous Driving. *arXiv preprint arXiv:1906.03199*, 2019.
- [195] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00033.
- [196] Lei Xu, Adam Krzyzak, and Ching Y Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics*, 22(3):418–435, 1992.
- [197] Xiangyang Xu, Yuncheng Li, Gangshan Wu, and Jiebo Luo. Multi-modal deep feature learning for rgb-d object detection. *Pattern Recognition*, 72:300–313, 2017.

- [198] Kailun Yang, Luis M. Bergasa, Eduardo Romera, Xiao Huang, and Kaiwei Wang. Predicting Polarization beyond Semantics for Wearable Robotics. In *IEEE-RAS International Conference on Humanoid Robots*, volume 2018-Novem, pages 96–103. IEEE, 2019. ISBN 9781538672839. doi: 10.1109/HUMANOIDS.2018.8625005.
- [199] M Y Yang, B Rosenhahn, and V Murino. *Multimodal Scene Understanding: Algorithms, Applications and Deep Learning*. Elsevier Science, 2019. ISBN 9780128173596. URL <https://books.google.fr/books?id=IPKiDwAAQBAJ>.
- [200] Hang Yin and Christian Berger. When to use what data set for your self-driving car algorithm: An overview of publicly available driving datasets. In *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, volume 2018-March, pages 1–8. IEEE, 2018. ISBN 9781538615256. doi: 10.1109/ITSC.2017.8317828.
- [201] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.
- [202] Hongshan Yu, Zhengeng Yang, Lei Tan, Yaonan Wang, Wei Sun, Mingui Sun, and Yandong Tang. Methods and datasets on semantic segmentation: A review. *Neurocomputing*, 304:82–103, 2018.
- [203] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. WildDash-creating hazard-aware benchmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–416, 2018.
- [204] Oliver Zendel, Markus Murschitz, Marcel Zeilinger, Daniel Steininger, Sara Abbasi, and Csaba Beleznai. Railsem19: A dataset for semantic rail scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [205] Richard Zhang, Stefan A Candra, Kai Vetter, and Avidesh Zakhori. Sensor fusion for semantic segmentation of urban scenes. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1850–1857. IEEE, 2015.
- [206] Wenkai Zhang, Hai Huang, Matthias Schmitz, Xian Sun, Hongqi Wang, and Helmut Mayer. Effective fusion of multi-modal remote sensing data in a Fully convolutional network for semantic labeling. *Remote Sensing*, 10(1):52, 2018. ISSN 20724292. doi: 10.3390/rs10010052.
- [207] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *arXiv preprint arXiv:1810.09091*, 2018.

- [208] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Meriaudeau. Incorporating depth information into few-shot semantic segmentation. In *25th International Conference on Pattern Recognition (ICPR 2020)*, .
- [209] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Meriaudeau. Multiscale attention-based prototypical network for few-shot semantic segmentation. In *25th International Conference on Pattern Recognition (ICPR 2020)*, .
- [210] Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo, and Désiré Sidibé. Exploration of Deep Learning-based Multimodal Fusion for Semantic Road Scene Segmentation. In *VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 336–343, 2019. ISBN 9789897583544. doi: 10.5220/0007360403360343.
- [211] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multi-modal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, page 104042, 2020.
- [212] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Ex-fuse: Enhancing feature fusion for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–284, 2018.
- [213] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 6230–6239, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.660.
- [214] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. IC-Net for Real-Time Semantic Segmentation on High-Resolution Images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11207 LNCS, pages 418–434, 2018. ISBN 9783030012182. doi: 10.1007/978-3-030-01219-9\_25.
- [215] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H S Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.
- [216] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 5122–5130, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.544.

- [217] Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, page 100004, 2019.
- [218] Dizhong Zhu and William AP Smith. Depth from a polarisation + rgb stereo pair. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [219] Yabin Zhu, Chenglong Li, Yijuan Lu, Liang Lin, Bin Luo, and Jin Tang. FANet: Quality-Aware Feature Aggregation Network for RGB-T Tracking. *arXiv preprint arXiv:1811.09855*, 2018.
- [220] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events, 2017.



**Title:** Real-time multimodal semantic scene understanding for autonomous navigation

**Keywords:** Semantic segmentation, Image fusion, Multi-modal, Deep learning

**Abstract:**

Robust semantic scene understanding is challenging due to complex object types, as well as environmental changes caused by varying illumination and weather conditions. This thesis studies the problem of deep semantic segmentation with multimodal image inputs. Multimodal images captured from various sensory modalities provide complementary information for complete scene understanding. We provided effective solutions for fully-supervised multimodal image segmentation and few-shot semantic segmentation of the outdoor road scene. Regarding the former case, we proposed a multi-level fusion network to integrate RGB and polarimetric images. A central fusion framework was also introduced to adaptively learn the joint representations of modality-specific features

and reduce model uncertainty via statistical post-processing. In the case of semi-supervised semantic scene understanding, we first proposed a novel few-shot segmentation method based on the prototypical network, which employs multiscale feature enhancement and the attention mechanism. Then we extended the RGB-centric algorithms to take advantage of supplementary depth cues. Comprehensive empirical evaluations on different benchmark datasets demonstrate that all the proposed algorithms achieve superior performance in terms of accuracy as well as demonstrating the effectiveness of complementary modalities for outdoor scene understanding for autonomous navigation.

**Titre :** Analyse et fusion d'images multimodales pour la navigation autonome

**Mots-clés :** Segmentation sémantique, Fusion d'images, Multimodalité, Apprentissage profond

**Résumé :**

Une analyse sémantique robuste des scènes extérieures est difficile en raison des changements environnementaux causés par l'éclairage et les conditions météorologiques variables, ainsi que par la variation des types d'objets rencontrés. Cette thèse étudie le problème de la segmentation sémantique à l'aide de l'apprentissage profond et avec des d'images de différentes modalités. Les images capturées à partir de diverses modalités d'acquisition fournissent des informations complémentaires pour une compréhension complète de la scène. Nous proposons des solutions efficaces pour la segmentation supervisée d'images multimodales, de même que pour la segmentation semi-supervisée de scènes routières en extérieur. Concernant le premier cas, nous avons proposé un réseau de fusion multi-niveaux pour intégrer des images couleur et polarimétriques. Une méthode de fusion centrale a également été introduite pour apprendre de manière adaptative

les représentations conjointes des caractéristiques spécifiques aux modalités et réduire l'incertitude du modèle via un post-traitement statistique. Dans le cas de la segmentation semi-supervisée, nous avons d'abord proposé une nouvelle méthode de segmentation basée sur un réseau prototypique, qui utilise l'amélioration des fonctionnalités multi-échelles et un mécanisme d'attention. Ensuite, nous avons étendu les algorithmes centrés sur les images RGB, pour tirer parti des informations de profondeur supplémentaires fournies par les caméras RGBD. Des évaluations empiriques complètes sur différentes bases de données de référence montrent que les algorithmes proposés atteignent des performances supérieures en termes de précision et démontrent le bénéfice de l'emploi de modalités complémentaires pour l'analyse de scènes extérieures dans le cadre de la navigation autonome.