Deep multimodal fusion for semantic image segmentation: A survey

Yifei Zhang, Désiré Sidibé, Olivier Morel, Fabrice Mériaudeau

Deep Multimodal Fusion for Semantic Image Segmentation: A Survey

Yifei Zhang[a,*], Désiré Sidibé[b], Olivier Morel[a], Fabrice Mériaudeau[a]

[a] *VIBOT ERL CNRS 6000, ImViA, Université de Bourgogne Franche-Comté, 71200, Le creusot, France*
[b] *Université Paris-Saclay, Univ Evry, IBISC, 91020, Evry, France*

Abstract

Recent advances in deep learning have shown excellent performance in various scene understanding tasks. 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 significant performance improvement. These fusion approaches take the benefits of multiple information sources and generate an optimal joint prediction automatically. This paper describes the essential background concepts of deep multimodal fusion and the relevant applications in computer vision. In particular, we provide a systematic survey of multimodal fusion methodologies, multimodal segmentation datasets, and quantitative evaluations on the benchmark datasets. Existing fusion methods are summarized according to a common taxonomy: early fusion, late fusion, and hybrid fusion. 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.

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

*Corresponding author
Email address: Yifei.Zhang@u-bourgogne.fr (Yifei Zhang)

1. Introduction

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 [1, 2, 3]. This terminology was further distinguished from instance-level segmentation [4] that devotes to produce per-instance mask and class label. Recently, panoptic segmentation [5, 6] 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 [7, 8] gains unprecedented success and tops other approaches by a large margin. For example, Convolutional Neural Networks (CNNs) [9] has become one of the most impressive algorithms for image-driven pattern recognition tasks. Besides, Recurrent Neural Networks (RNNs) [10, 11] are commonly used for retrieving contextual features, which remember every information through time. The various milestones in the evolution of deep learning significantly promote the advancement of semantic segmentation research.

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 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 [12]. 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 [13, 14, 15, 16, 17].

Especially in recent years, deep multimodal fusion methods benefit from the

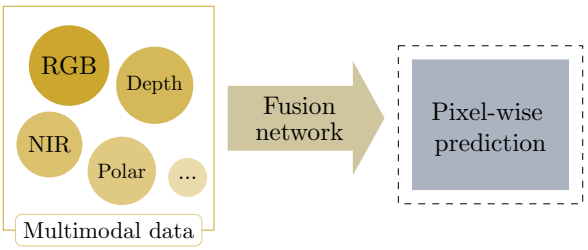


Figure 1: An illustration of deep multimodal segmentation pipeline.

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 is 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 segmentation is to learn the optimal joint representation from rich and complementary features of the same scene. Improving the accuracy and robustness of deep multimodal models is one of the significant challenges in this area. Also, scalability and real-time issues should be taken into consideration for real-world applications. As an illustration, Figure 1 shows the pipeline of deep multimodal segmentation.

Several relevant surveys already exist, such as deep learning-based semantic segmentation [2, 3, 18, 19], indoor scene understanding [20, 21], multimodal perception for autonomous driving [22], multimodal human motion recognition [23], multimodal medical image segmentation [24], and multimodal learning study [25, 26]. However, these review works are mostly focused on unimodal image segmentation, multimodal fusion for specific domains, or multimedia analysis across video, audio, and text. Especially with the advent of low-cost sensors, an increasing number of visible/invisible light and depth cameras are employed in scene understanding. There is a lack of systematic review that focuses explicitly on deep multimodal fusion for 2D/2.5D semantic image segmentation. In summary, the main contributions of this paper are as follows:

- We provide necessary background knowledge on multimodal image segmentation and a global perspective of deep multimodal learning.
- We conduct an extensive literature review on existing deep multimodal fusion methods, with the highlight of their contributions to model design.
- We conduct a comprehensive survey of current semantic segmentation datasets as well as the potential multimodal datasets.
- We gather quantitative experimental results of multimodal fusion methods on different benchmark datasets, including their accuracy, runtime, and memory footprint.

The remainder of this paper is organized as follows. The background concepts of deep multimodal fusion for semantic image segmentation are firstly described in Section 2, including the development, recent advancements as well as related applications. Section 3 reviews the existing deep multimodal segmentation methods according to our taxonomy of fusion strategy, followed by a brief discussion on architectural design. Section 4 provides a broad survey of current unimodal and multimodal image segmentation datasets. Several typical modalities (e.g., RGB-D, Near-InfraRed, thermal and polarization cameras) are highlighted. Quantitative performance evaluations of the fusion methods mentioned earlier are summarized and analyzed in Section 5. Finally, Section 6 concludes this paper.

2. Background

Multimodal fusion systems work like the human brain, which synthesizes multiple sources of information for semantic perception and further decision making. First of all, we adopt the definition of "*modality*" from [27], which refers to each detector acquiring information about the same scene. 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. The primary motivation for multimodal fusion is to obtain rich characteristics of the scenes by integrating multiple sensory modalities.

As a multi-disciplinary research area, the adopted definition 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) [28], to name a few. Benefiting from the complementary and functional information about a target (e.g. an organ), multimodal fusion models can achieve a precise diagnosis and treatment [29, 30, 24]. In multimedia analysis, multimodal data collected from audio, video as well as text modalities [31, 32, 26] are used to tackle semantic concept detection, including human-vehicle interaction [33], biometric identification [34, 35, 36]. In remote sensing applications, multimodal fusion leverages the high-resolution optical data, synthetic aperture radar, and 3D point cloud [37, 38].

In this survey, we clarify the definition of "*modality*" for semantic segmentation tasks as a single image sensor. Relevant modalities reviewed in this survey include RGB-D cameras, Near-InfraRed cameras, thermal cameras, and polarization cameras. Next, we introduce the development of semantic image segmentation from uni-modality to multi-modality and their applications for indoor and outdoor scene understanding.

2.1. Semantic Image Segmentation

There have been many studies addressing the task of semantic image segmentation with deep learning techniques [2, 39]. Fully Convolutional Network (FCN) [40] was first proposed for effective pixel-level classification. In FCN, the last fully connected layer is substituted by convolutional layers. DeconvNet [41], which is composed of deconvolution and unpooling layers, was proposed in the same year. Badrinarayanan et al. [42] introduced a typical encoder-decoder architecture with forwarding pooling indices, mentioned as SegNet. Another typical segmentation network with multi-scale features concatenation, U-Net [43], was initially proposed for biomedical image segmentation. In particular,

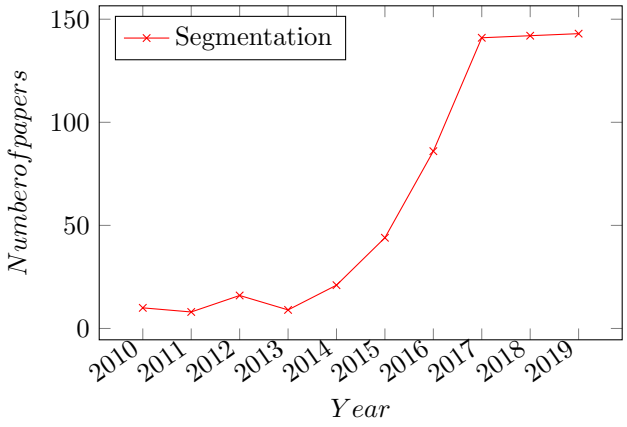


Figure 2: Number of papers published per year. Statistical analysis is based on the work by Caesar [51]. Segmentation includes image/instance/panoptic segmentation and joint depth estimation.

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 [44] was presented for real-time segmentation. In the work of PixelNet, Bansal et al. [45] 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 [46] and DilatedNet [47], which helps to keep the resolution of output feature maps with large receptive fields. Besides, a series of Deeplab models also achieves excellent success on semantic image segmentation [48, 49, 50].

Furthermore, Peng et al. [52] dedicated to employing larger kernels to address both the classification and localization issues for semantic segmentation. RefineNet [53] explicitly exploits multi-level features for high-resolution prediction using long-range residual connections. Zhao et al. [54] 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 [55, 56] and nighttime [57, 58] has also been addressed to perform the generalization capacity and robustness of deep learning models. Figure 2 shows the number of papers about segmentation published in

the past decade.

In addition to the aforementioned networks, many practical deep learning techniques (e.g., Spatial Pyramid Pooling [59], CRF-RNN [60], Batch Normalization [61], Dropout [62]) were proposed for improving the effectiveness of learning models. Notably, multi-scale feature aggregation was frequently used in semantic segmentation [63, 64, 65, 66, 67]. These learning models experimentally achieve significant performance improvement. Lin et al. [68] 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 [69, 70, 71]. Also, the 1×1 convolution layer is commonly used to allow complex and learnable interaction across modalities and channels [72, 70]. Besides, attention mechanism has become a powerful tool for image recognition [73, 74, 75]. The attention distribution enables the model to selectively pick valuable information [76], achieving more robust feature representation and more accurate prediction.

2.2. Deep Multimodal 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* [77]. Machine learning techniques used for such fusion include Principal Component Analysis (PCA), Independent Components Analysis (ICA), and Canonical Correlation Analysis (CCA) [25]. As the discriminative classifiers [78] become increasingly popular (e.g. SVM [79] and Random Forest [80]), 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 [81] and ResNet [82], 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 [83]. We will present a detailed review of recent achievements in the following sections, covering various existing fusion methodologies and multimodal image datasets.

We conclude this part by pointing 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. During data acquisition, image sensors have different sensitivity to scene information. The poor performance of individual modality and the absence of modalities should be seriously considered.

Effectiveness. In practical applications, multimodal fusion models need to satisfy certain requirements, including simplicity of implementation, scalability, real-time, etc. Moreover, ensuring network convergence can be a significant challenge with the use of redundant multimodal data.

2.3. Applications for Scene Understanding

As one of the major challenges in scene understanding, deep multimodal fusion for semantic segmentation cover a wider variety of scenarios. For instance,

Hazirbas et al. [84] address the problem of pixel-level prediction of indoor scenes using color and depth data. Schneider et al. [85] present a mid-level fusion architecture for urban scene segmentation. Similar works in both indoor and outdoor scene segmentation can be found in [70]. Furthermore, the work by Valada et al. [86] led to a new research topic in scene understanding of unstructured forested environments. Considering non-optimal weather conditions, Pfeuffer and Dietmayer [56] investigated a robust fusion approach for foggy scene segmentation.

Besides the image segmentation task mentioned above, there are many other scene understanding tasks that benefit from multimodal fusion, such as object detection [13, 87, 88], human detection [89, 14, 90, 91], salient object detection [92, 93], trip hazard detection [94] and object tracking [69]. Especially for autonomous systems, LiDAR is always employed to provide highly accurate three-dimensional point cloud information [95, 96]. Patel et al. [97] 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., [98, 99, 100]). It is reasonably foreseeable that deep multimodal fusion of homogeneous and heterogeneous information sources can be a strong emphasis for intelligent mobility [22, 101] in the near future.

3. Fusion Methodologies

In this section, we provide a comprehensive review of deep multimodal fusion methods for semantic image segmentation. We highlight their benefits and drawbacks, providing interested readers with a complete overview of deep fusion strategies.

3.1. Taxonomy

In the early works [31, 95, 102], 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

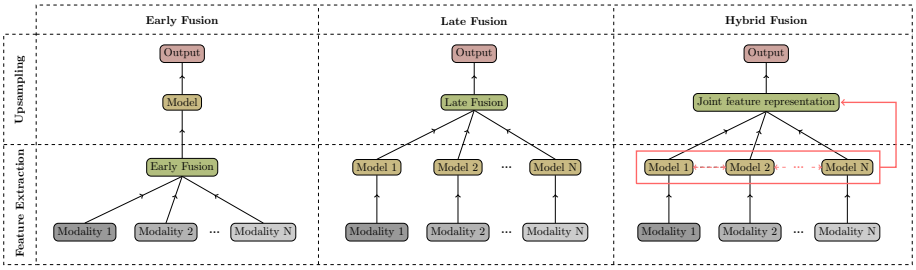


Figure 3: An illustration of different fusion strategies for deep multimodal learning.

divide the deep multimodal fusion methods into early fusion, late fusion, and hybrid fusion, according to the fusion stage and motivation (see Figure 3).

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, $(x_1, x_2, ..., x_n)$ is a set of n modalities as input, then the final prediction y can be defined as:

$$y = I(x_1, x_2, ..., x_n). \tag{1}$$

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.) [22], followed by a series of convolutional layers. 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., [103]) are therefore categorized into late fusion strategy in this review. Suppose that the segmentation networks $(I_1, \ldots, I_n)$ are used to process the multimodal

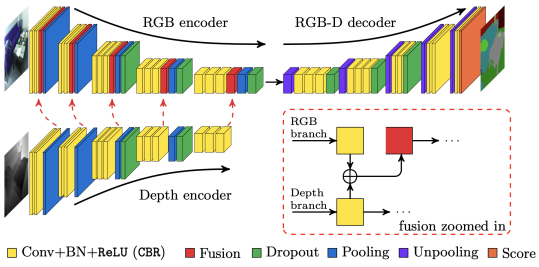


Figure 4: FuseNet architecture with RGB-D input. Figure reproduced from [84].

data $(x_1, x_2, ..., x_n)$ from different modalities, and P is the fusion operation as well as the following convolutional layers, the final output y can be formulated as:

$$y = P(I_1(x_1), I_2(x_2), ..., I_n(x_n)). \tag{2}$$

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.

3.2. Fusion strategies

Based on the common taxonomy of fusion strategy in Section 3.1, we systematically review the existing deep multimodal fusion networks for semantic image segmentation.

3.2.1. Early fusion

The first attempt at deep multimodal fusion was made by Couprie et al. [83] 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

Table 1: Typical early fusion methods reviewed in this paper.

Ref.	Method	Backbone	Contribution(s)	Year	Source Code
[83]	Couprie'	-	Initial attempt	2013	Available
[84]	FuseNet	VGG-16	Dense fusion/Sparse fusion	2016	Available
[104]	MVCNet	VGG-16	Multi-view consistency	2017	-
[72]	LDFNet	VGG-16	D&Y Encoder	2018	Available
[105]	RFBNet	AdapNet++	Bottom-up interactive fusion structure	2019	-
[71]	ACNet	ResNet-50	Multi-branch attention based network	2019	Available
[106]	RTFNet	ResNet-152	RGB-Thermal fusion with Upcepton blocks	2019	Available

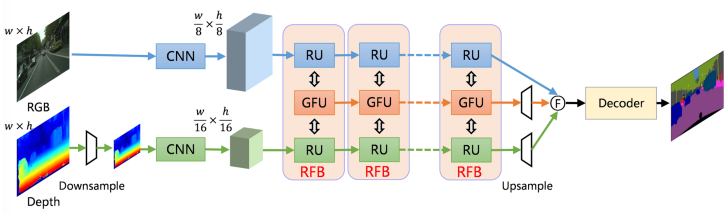


Figure 5: RFBNet architecture with three bottom-up streams (RGB stream, depth stream, and interaction stream). Figure extracted from [105].

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 [84] was proposed in 2016 (see Figure 4). 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. [104] proposed MVCNet to predict multi-view consistent semantics. Then Hung et al. [72] 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 [105] was proposed with an efficient fusion mechanism that explores the interdependence between the encoders (see Figure 5). The Residual Fusion Block (RFB), which consists of two modality-specific

Table 2: Typical late fusion methods reviewed in this paper.

Ref.	Method	Backbone	Contribution(s)	Year	Source Code
[107]	Gupta'	-	CNN+SVM	2014	Available
[108]	LSTM-CF	Deeplab	LSTM-based context fusion	2016	Available
[86]	LFC	VGG-16	Late-fused convolution	2016	Available
[103]	Wang'	VGG-16	Feature transformation network	2016	-
[109]	CMoDE	AdapNet	Class-wise adaptive gating network	2017	Available
[110]	LSD-GF	VGG-16	Locality-sensitive DeconvNet with gated fusion	2017	-
[111]	CMnet	VGG-16/ ResNet-10	RGB-Polarization fusion/Different encoders	2019	-

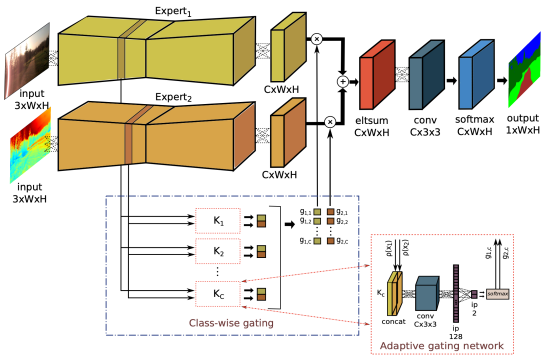


Figure 6: Convoluted Mixture of Deep Experts framework. Figure extracted from [109].

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. [71] 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 [106] 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.

3.2.2. Late fusion

As early as 2014, Gupta et al. [107] 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 an SVM classifier

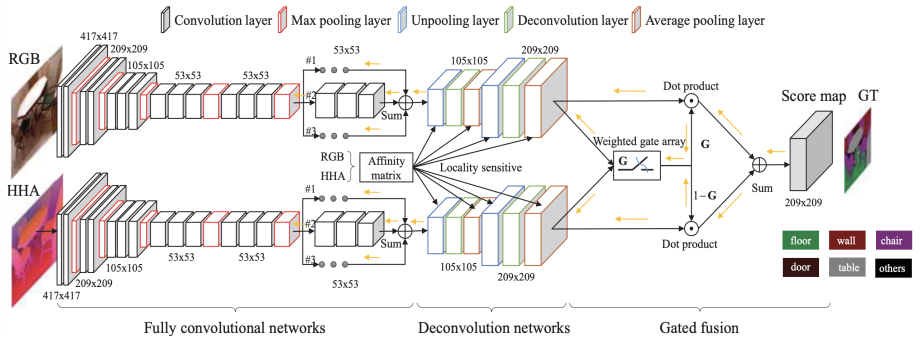


Figure 7: LSD-GF architecture. Figure extracted from [110].

at the late stage. Then the work by Li et al. [108] 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 RGB stream are fused by simply concatenating.

Besides, Wang et al. [103] proposed a feature transformation network for learning the 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. [86]. 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 convoluted mixture of deep expert units, referred to as CMoDE [109]. This deep fusion framework was inspired by the work [112, 113], 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 6). In the work of LSD-GF, Cheng et al. [110] proposed a gated fusion module to adaptively merge RGB and depth score maps according to their weighted contributions (see Figure 7). More recently, CMnet [111] made a new

Table 3: Typical hybrid fusion methods reviewed in this paper.

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

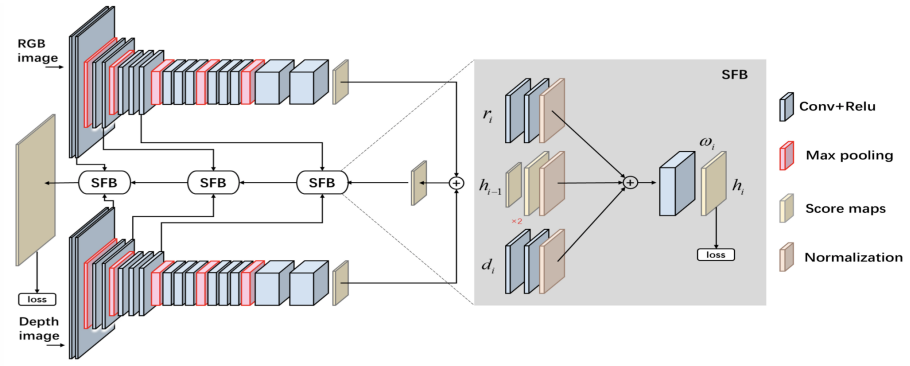


Figure 8: Semantics-guided multi-level fusion. Figure extracted from [116]

attempt to explore the complementary characteristics of polarimetric data. Different backbones are used for multimodal feature extraction.

3.2.3. 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. [114] extended the core idea of residual 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. [115] introduced a fusion structure combining

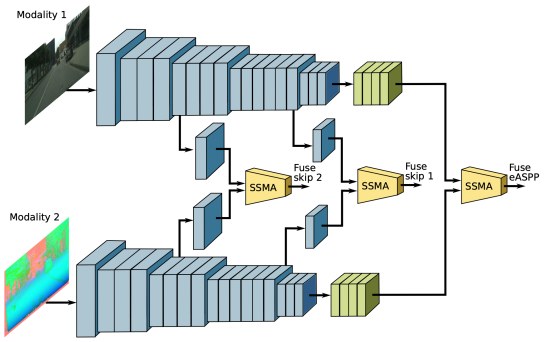


Figure 9: Fusion architecture with self-supervised model sdaptation modules. Figure extracted from [109].

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. [117]. 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 [116], referred to as S-M Fusion, was proposed to learn the feature representation in a bottom-up manner (see Figure 8). This fusion strategy employed the cascaded Semantics-guided Fusion Block (SFB) to fuse lower-level features across modalities sequentially.

Moreover, Jiang et al. [118] 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) [70]. This fusion framework dynamically adapts the fusion of semantically mature multiscale representations. The latent

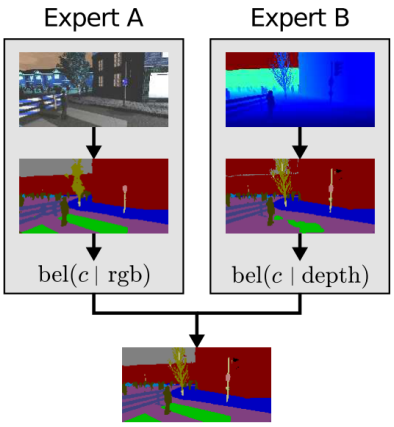


Figure 10: Individual semantic segmentation experts are combined modularly using different statistical methods. Figure extracted from [119]

joint representation generated from the SSMA block is integrated into decoder by two skip connections (see Figure 9). 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.

3.2.4. Statistical fusion

As an alternative post-processing approach, statistical fusion is proposed to reduce the uncertainty of model and modalities at decision-level. Blum et al. [119] introduced statistical fusion methods to integrate deep learning-based segmentation prediction, including Bayes categorical fusion and Dirichlet fusion. The presented methods allow different training sets per expert (modality). Without extra training on aligned data, only a small subset is needed for calibration of the statistical models (see Figure 10). Combining multiple classifiers in a statistical way is not a new concept [120], but this work leads to an interesting research direction in the combination of deep learning and statistics.

3.2.5. Discussion

Deep multimodal fusion for scene understanding is a complex issue that involves several factors, including the spatial location of objects, the semantic

context of scenes, the effectiveness of fusion models, the physical properties of modalities, etc. The fusion strategies mentioned above follow different design concepts to tackle this challenge. Early fusion methods make an effort to optimally integrate information from multimodal sources during feature extraction. Namely, the representative features from complementary modality 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.

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. [121] offers some ideas based on learning with the privileged information paradigm to tackle this challenge. Otherwise, the trade-offs between accuracy/speed [122, 123, 22] 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 Section 5.

4. Datasets

Over the last decade, a large number of datasets have been proposed to meet the needs of deep learning-based methods. Since the quantity and quality of training data significantly affect the performance of learning models, 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.

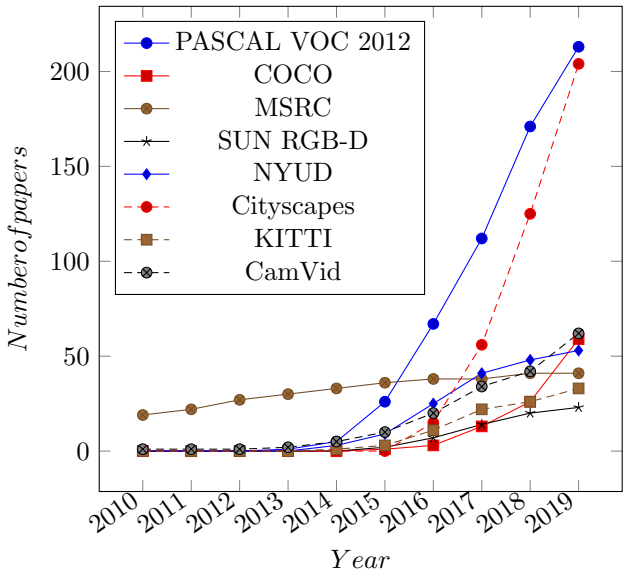


Figure 11: Accumulated dataset importance. Statistical analysis is based on the work by [51].

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 11 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 [124] and Cityscapes [125] are two of the most popular datasets for semantic segmentation. As the representative RGB-D dataset, NYU-D [126] and SUN RGB-D [127] 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.

Table 4: Summary of popular datasets for image segmentation task.

Ref.	Dataset	Classes	Resolution	Images	Scene	Data	Year
[128]	MSRC	23	320x213	591	Outdoor	2D	2006
[133]	Stanford background	8	320x240	715	Outdoor	2D	2009
[134]	CamVid	32	960x720	701	Outdoor	2D	2009
[124]	PASCAL VOC	20	Variable	11K	Variable	2D	2012
[126]	NYU Depth v2	40	480x640	1449	Indoor	2.5D	2012
[132]	Microsoft COCO	80	Variable	330K	Variable	2D	2014
[135]	KITTI	11	Variable	400	Outdoor	2D/3D	2015
[125]	Cityscapes	30	2048x1024	5K	Outdoor	2.5D	2015
[136]	SYNTHIA	13	960x720	13K	Outdoor(synthetic)	2.5D	2016
[137]	GTA5	19	1914x1052	13K	Outdoor(synthetic)	2D	2016
[127]	SUN RGB-D	37	Variable	10K	Indoor	2.5D	2015
[138]	ADE20K	150	Variable	22K	Variable	2D	2017
[139]	Mapillary Vistas	66	1920x1080	25K	Outdoor	2D	2017
[140]	WildDash	28	Variable	1.8K	Outdoor	2D	2018

4.1. Popular datasets for image segmentation

As one of the earliest pixel-wise labeled image databases, MSRC dataset [128] 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 [124] 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 [129], PASCAL-Part [130], SBDB [131]. Another similar large-scale dataset is Microsoft COCO dataset [132], 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.

Furthermore, several outdoor road scene datasets are constantly emerging

during the last decade, e.g. CamVid [134], KITTI [135], Cityscapes [125], Mapillary Vistas [139], toward promoting the commercialization and advancement of autonomous driving technology [141]. 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 [139] 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 [140] 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 [142] 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 4. Further information are provided, including numbers of classes, size of the database, and the type of scenes.

Table 5: Summary of popular 2D/2.5D multimodal datasets for scene understanding.

Ref.	Dataset	Images	Scene	Multi-modal data	Year
[126]	NYUDv2	1449	Indoor	RGB/Depth	2012
[127]	SUN RGB-D	10K	Indoor	RGB/Depth	2015
[125]	Cityscapes	5K	Urban street	RGB/Depth	2015
[136]	SYNTHIA	13K	Urban street	RGB/Depth	2016
[86]	Freiburg Forest	5K	Forest	RGB/Depth/NIR	2016
[143]	ScanNet	19K	Indoor	RGB/Depth	2017
[144]	Tokyo Multi-Spectral	1569	Urban street	RGB/Thermal	2017
[145]	CATS 2	686	Variable	RGB/Depth/Thermal	2018
[146]	RANUS	40k	Urban street	RGB/NIR	2018
[111]	POLABOT	175	Outdoor	RGB/NIR/Polarization	2019
[147]	PST900	894	Subterranean	RGB/Thermal	2019
[148]	DISCOMAN	600K	Indoor	RGB/Depth	2019

4.2. 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 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 6.

4.2.1. 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. [107] 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 [107, 13] and pose estimation [149]. The first row in Table 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 advantage 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 [126] 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 [127]. 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 [136]) are generated for driving scenes semantic segmentation.

4.2.2. Near-InfraRed datasets

Infrared imaging captured from multi-spectral cameras shows high contrast of natural and artificial objects [150, 151]. 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 [86] 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 [146] 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 [152, 153], face recognition [154], image dehazing [155, 156], video surveillance [157], to name a few.

4.2.3. 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 [158]. The wavelength is generally detected up to $14\mu m$. In the early years, thermal imaging cameras are invented for military uses. With the cost of sensors decreasing, many scene understanding tasks can now benefit from thermal information [106].

Tokyo Multi-Spectral [144] 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

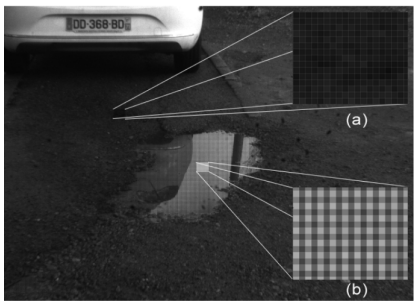


Figure 12: Reflection influence on polarimetry. (a) and (b) represent a zoom on the non-polarized and polarized area, respectively. Figure extracted from [164].

labeled to eight classes: car, person, bike, curve, car stop, guardrail, color cone, and bump. Then Shivakumar et al. [147] 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.

4.2.4. Polarization datasets

As a universal phenomenon existing in natural scenes, polarimetric imaging is highly sensitive to the vibration pattern of the light [159]. In the natural environment, the polarisation of light is generally obtained by reflection or scattering. The polarization images carry crucial information of reflection surface, including object shape and surface material. As shown in 12, the micro-grid appears on the polarized surface and reveals an intensity change according to the polarizer affected. To tackle practical problems in computer vision, polarization images have been widely applied to object detection [160], image dehazing [161], depth estimation [162, 163].

Zhang et al. [111] 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. [165] 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 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 [166, 167] in scene understanding. For future perspectives, polarization cameras may be extremely valuable in autonomous driving [164, 165] and robotics [168, 169].

4.2.5. 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, 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.

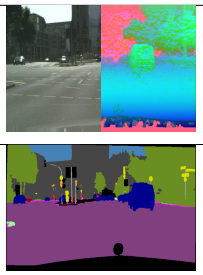
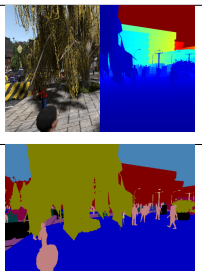
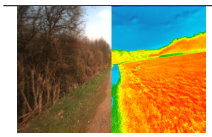
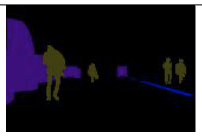
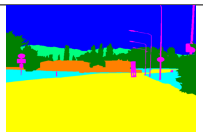
NYUDv2 (RGB+Depth)	SUN RGB-D (RGB+Depth)	Cityscapes (RGB+Depth)	SYNTHIA (RGB+Depth)
			
Freiburg Forest (RGB+NIR)	Tokyo Multi-Spectral (RGB+Thermal)	RANUS (RGB+NIR)	POLABOT (RGB+Polarization)
			
			

Table 6: Examples of multimodal image datasets mentioned in Section 4.2. For each dataset, the top image shows two modal representations of the same scene. The bottom image is the corresponding groundtruth.

- **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?

5. Evaluation

In this section, we report the evaluations of deep multimodal fusion methods that are mentioned in Section 3 on four benchmark datasets: SUN RGB-D [127], NYU Dv2 [126], Cityscapes [125], and Tokyo Multi-Spetral dataset [144]. 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.

5.1. Evaluation metrics and backbone networks

It is well known that how to evaluate the performance of a segmentation algorithm is a critical issue. A series of benchmark datasets promote the standardization of comparison metrics, providing a fair comparison of the state-of-the-art methods. More precisely, the performance of deep learning-based approaches can be reflected in all aspects [2], such as accuracy, memory usage, and runtime. Among these factors, accuracy may be the most common evaluation criteria to measure the performance of pixel-level prediction [170]. As a reference, a general analysis of accuracy metrics for classification tasks can be found in [171]. For multimodal image segmentation, the most popular metrics have no difference with 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 [40].

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. Then we can define these 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

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

Table 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 [175]	VGG-16	-		-	45.9	30.7
Context [176]	VGG-16	-	RGB	-	53.4	42.3
RefineNet [53]	ResNet-152	-		-	58.5	45.9
LSTM-CF [108]	VGG-16	426x426		Late	48.1	-
FuseNet [84]	VGG-16	224x224		Early	48.30	37.3
DFCN-DCRF [115]	VGG-16	480x480		Early	50.6	39.3
S-M Fusion [116]	VGG-16	449x449		Hybrid	53.93	40.98
LSD-GF [110]	VGG-16	417x417	RGB-D	Late	58.0	-
SSMA [70]	ResNet-50	768x384		Hybrid	-	44.52
RDFNet [114]	ResNet-152	-		Early	60.1	47.7
RedNet [118]	ResNet-50	640x480		Hybrid	60.3	47.8
CFN [117]	RefineNet-152	-		Hybrid	-	48.1
ACNet [71]	ResNet-50	640x480		Early	-	48.1

Besides, many elaborated backbone networks, such as VGGNet [81], ResNet [82], and Xception [172], are widely used in a variety of segmentation network design [173]. These backbone networks not only extract effective semantic information and spatial details but also simplify the training.

The segmentation performance is generally affected by many factors, such as the preprocessing of data, fusion strategy, the choice of the backbone network, the practice of state-of-the-art deep learning technologies, etc. In the following, we summarize the evaluation results of multimodal fusion models and also analyze the influence of the algorithms and the multimodal data on the performances. Almost all of the models are pre-trained on a large-scale image database such as ImageNet [174].

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

Table 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 [84]	VGG-16	320x240	RGB-D	Early	67.46	56.01
	Wang'[103]	VGG-16	-		Late	52.7	-
	MVCNet [104]	VGG-16	320x240		Early	70.59	59.07
40	Gupta' [107]	-	-	RGB-D	Late	35.1	-
	FuseNet [84]	VGG-16	320x240		Early	44.92	35.36
	Wang'[103]	VGG-16	-		Late	47.3	-
	MVCNet [104]	VGG-16	320x240		Early	51.78	40.07
	LSD-GF [110]	VGG-16	417x417		Late	60.7	45.9
	CFN [117]	RefineNet-152	-		Hybrid	-	47.7
	ACNet [71]	ResNet-50	640x480		Early	-	48.3

Table 9: Experimental results of deep multimodal fusion methods on Cityscapes dataset. Input images are uniformly resized to 768×384.

Method	Backbone	Modality	Fusion strategy	Mean IoU
ERFnet [177]	-	RGB	-	62.71
AdapNet [109]	ResNet-50		-	69.39
AdapNet++ [70]	ResNet-50		-	80.80
AdapNet [109]	ResNet-50	Depth	-	59.25
AdapNet++ [70]	ResNet-50		-	66.36
AdapNet++ [70]	ResNet-50	HHA	-	67.66
LFC [86]	VGG-16	RGB-D	Late	69.25
CMoDE [109]	AdapNet		Late	71.72
SSMA [70]	AdapNet++		Hybrid	83.44
SSMA [70]	AdapNet++	RGB-HHA	Hybrid	83.94

- SUN RGB-D dataset

  Firstly, we report the experimental results on the indoor scene dataset, SUN RGB-D (see Table 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

Table 10: Experimental results of deep multimodal fusion methods on Tokyo Multi-Spectral dataset. The image resolution in the dataset is 640×480.

Method	Backbone	Modality	Mean Acc	Mean IoU
SegNet [42]	VGG-16		35.4	31.7
PSPNet [178]	ResNet-50	RGB	44.9	39.0
DUC-HDC [179]	ResNet-101		58.9	47.7
MFNet [144]	VGG-16		45.1	39.7
SegNet-4c [42]	VGG-16		49.1	42.3
FuseNet [84]	VGG-16		52.4	45.6
PSPNet-4c [178]	ResNet-50		51.3	46.1
DUC-HDC-4c [179]	ResNet-101	RGB-Thermal	59.3	50.1
RTFNet [106]	ResNet-152		63.1	53.2

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 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.

- Cityscapes dataset

  Apart from the indoor scene datasets, we also show the segmentation results on a more challenging urban scene dataset, Cityscapes in Table 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 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.

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.

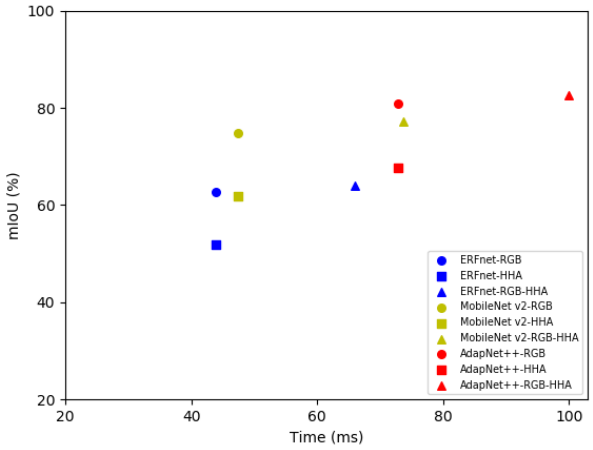


Figure 13: Real-time and accuracy performance. Performance of SSMA fusion method using different real-time backbones on the Cityscapes validation set (input image size: 768×384, GPU: NVIDIA TITAN X).

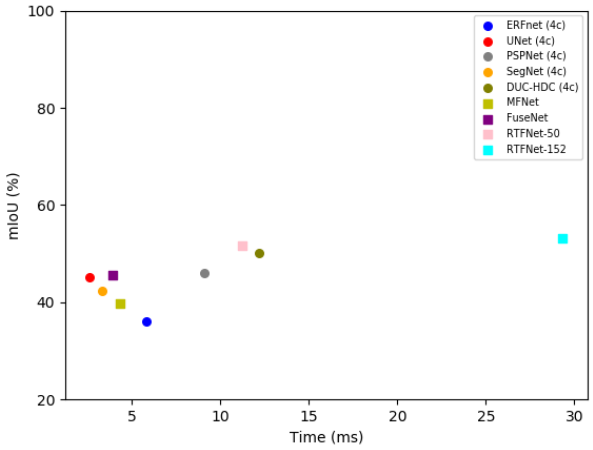


Figure 14: Real-time and accuracy performance. Performance of different fusion methods on Tokyo Multi-Spectral dataset.(input image size: 640×480, GPU: NVIDIA 1080 Ti graphics card).

Table 11: Parameters and inference time performance. The reported results on the Cityscapes are collected from [70].

Network	Backbone	mIoU (%)	Parms. (M)	Time (ms)
PSPNet [178]	ResNet-101	81.19	56.27	172.42
DeepLab v3 [49]	ResNet-101	81.34	58.16	79.90
DeepLab v3+ [50]	Modified Xception	82.14	43.48	127.97
AdapNet++ [70]	ResNet-50	81.34	30.20	72.92
SSMA [70]	ResNet-50	82.31	56.44	101.95

5.3. Real-time consideration

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 1314. 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.

5.4. Memory footprint

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 11 demonstrates the comparisons on the number of parameters and inference time for various network architectures.

6. Conclusion

We have reviewed deep multimodal image segmentation from two aspects: fusion methodology and dataset. Multimodal image data, such as RGB-D image, Near-InfraRed image, thermal image, polarization image, are the primary concerns in this paper. To the best of our knowledge, this is the first review

paper that focuses on deep learning-based multimodal fusion for semantic image segmentation. In this work, we first described the development of multimodal fusion and provided the reader with relevant background knowledge to support text comprehension. Then we categorized 20 deep multimodal fusion methods into *early fusion*, *late fusion*, and *hybrid fusion*. We further discussed architectural design to explore the essentials of deep multimodal fusion. Besides, the existing image segmentation datasets are summarized, covering 12 current multimodal datasets. 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.

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 experimental results collected in this survey show the effectiveness of the deep multimodal fusion method. 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. 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.

References

[1] Y.-i. Ohta, T. Kanade, T. Sakai, An analysis system for scenes containing objects with substructures, in: Proceedings of the Fourth International Joint Conference on Pattern Recognition, 1978, pp. 752–754.

[2] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J. Garcia-Rodriguez, A Review on Deep Learning Techniques Applied to Semantic Segmentation, arXiv preprint arXiv:1704.06857 (2017).

[3] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun, Y. Tang, Methods and datasets on semantic segmentation: A review, Neurocomputing 304 (2018) 82–103.

[4] S. Edelman, T. Poggio, Integrating visual cues for object segmentation and recognition, Optics News 15 (1989) 8.

[5] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic Segmentation, 2018. arXiv:1801.00868.

[6] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, L.-C. Chen, Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation, 2019. arXiv:1911.10194.

[7] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436.

[8] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (1998) 2278–2324.

[10] Z. C. Lipton, J. Berkowitz, C. Elkan, A critical review of recurrent neural networks for sequence learning, 2015. arXiv:1506.00019.

[11] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, A. Courville, Reseg: A recurrent neural network-based model for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 41–48.

[12] R. Gade, T. B. Moeslund, Thermal cameras and applications: a survey, Machine vision and applications 25 (2014) 245–262.

[13] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard, Multimodal deep learning for robust RGB-D object recognition, in: IEEE International Conference on Intelligent Robots and Systems, volume

2015-Decem, IEEE, 2015, pp. 681–687. doi:10.1109/IROS.2015.7353446. arXiv:1507.06821.

[14] J. Liu, S. Zhang, S. Wang, D. N. Metaxas, Multispectral deep neural networks for pedestrian detection, British Machine Vision Conference 2016, BMVC 2016 2016-Septe (2016) 73.1–73.13.

[15] X. Xu, Y. Li, G. Wu, J. Luo, Multi-modal deep feature learning for rgb-d object detection, Pattern Recognition 72 (2017) 300–313.

[16] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, U. J. Nunes, Multi-modal vehicle detection: fusing 3d-lidar and color camera data, Pattern Recognition Letters 115 (2018) 20–29.

[17] M. Y. Yang, B. Rosenhahn, V. Murino, Multimodal Scene Understanding: Algorithms, Applications and Deep Learning, Elsevier Science, 2019. URL: https://books.google.fr/books?id=lPKiDwAAQBAJ.

[18] X. Liu, Z. Deng, Y. Yang, Recent progress in semantic image segmentation, Artificial Intelligence Review 52 (2018) 1089–1106.

[19] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: A review, arXiv preprint arXiv:1910.07655 (2019).

[20] M. Naseer, S. Khan, F. Porikli, Indoor Scene Understanding in 2.5/3D for Autonomous Agents: A Survey, IEEE Access 7 (2019) 1859–1887.

[21] F. Fooladgar, S. Kasaei, A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks, Multimedia Tools and Applications (2019) 1–26.

[22] D. Feng, C. Haase-Schuetz, L. Rosenbaum, H. Hertlein, F. Duffhauss, C. Glaeser, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, arXiv preprint arXiv:1902.07830 (2019).

[23] P. Wang, W. Li, P. Ogunbona, J. Wan, S. Escalera, Rgb-d-based human motion recognition with deep learning: A survey, Computer Vision and Image Understanding 171 (2018) 118 – 139.

[24] T. Zhou, S. Ruan, S. Canu, A review: Deep learning for medical image segmentation using multi-modality fusion, Array (2019) 100004.

[25] D. Ramachandram, G. W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Processing Magazine 34 (2017) 96–108.

[26] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2018) 423–443.

[27] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proceedings of the IEEE 103 (2015) 1449–1477.

[28] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.

[29] P. Moeskops, J. M. Wolterink, B. H. M. van der Velden, K. G. A. Gilhuijs, T. Leiner, M. A. Viergever, I. Išgum, Deep learning for multi-task medical image segmentation in multiple modalities, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 478–486.

[30] Y. Liu, X. Chen, J. Cheng, H. Peng, A medical image fusion method based on convolutional neural networks, in: 2017 20th International Conference on Information Fusion (Fusion), IEEE, 2017, pp. 1–7.

[31] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: A survey, Multimedia Systems 16 (2010) 345–379.

[32] Y. Mroueh, E. Marcheret, V. Goel, Deep multimodal learning for Audio-Visual Speech Recognition, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, volume 2015-Augus, IEEE, 2015, pp. 2130–2134. doi:10.1109/ICASSP.2015.7178347. arXiv:1501.05396.

[33] E. Erzin, Y. Yemez, A. M. Tekalp, A. Ercil, H. Erdogan, H. Abut, Multimodal person recognition for human-vehicle interaction, IEEE MultiMedia 13 (2006) 18–31.

[34] R. W. Frischholz, U. Dieckmann, Biold: a multimodal biometric identification system, Computer 33 (2000) 64–68.

[35] S. Soleymani, A. Dabouei, H. Kazemi, J. Dawson, N. M. Nasrabadi, Multi-level feature abstraction from convolutional neural networks for multi-modal biometric identification, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 3469–3476.

[36] P. Connor, A. Ross, Biometric recognition by gait: A survey of modalities and features, Computer Vision and Image Understanding 167 (2018) 1 – 27.

[37] N. Audebert, B. Le Saux, S. Lefèvre, Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks, ISPRS Journal of Photogrammetry and Remote Sensing 140 (2018) 20–32.

[38] W. Zhang, H. Huang, M. Schmitz, X. Sun, H. Wang, H. Mayer, Effective fusion of multi-modal remote sensing data in a Fully convolutional network for semantic labeling, Remote Sensing 10 (2018) 52.

[39] S. Ghosh, N. Das, I. Das, U. Maulik, Understanding Deep Learning Techniques for Image Segmentation, 2019. arXiv:1907.06119.

[40] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[41] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, volume 2015 Inter, 2015, pp. 1520–1528. doi:10.1109/ICCV.2015.178. arXiv:1505.04366.

[42] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 2481–2495.

[43] O. Ronneberger, P. Fischer, T. 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, Springer, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-4_28. arXiv:1505.04597.

[44] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation, arXiv preprint arXiv:1606.02147 (2016).

[45] A. Bansal, X. Chen, B. Russell, A. Gupta, D. Ramanan, Pixelnet: Towards a general pixel-level architecture, arXiv preprint arXiv:1609.06694 (2016).

[46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, arXiv preprint arXiv:1412.7062 (2014).

[47] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings (2016).

[48] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. 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 (2018) 834–848.

[49] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking Atrous Convolution for Semantic Image Segmentation, arXiv preprint arXiv:1706.05587 (2017).

[50] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. 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 (2018) 833–851.

[51] H. Caesar, really-awesome-semantic-segmentation, `https://github.com/nightrome/really-awesome-semantic-segmentation`, 2018.

[52] C. Peng, X. Zhang, G. Yu, G. Luo, J. 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, 2017, pp. 1743–1751. doi:`10.1109/CVPR.2017.189`. `arXiv:1703.02719`.

[53] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1925–1934.

[54] H. Zhao, X. Qi, X. Shen, J. Shi, J. Jia, ICNet 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, 2018, pp. 418–434. doi:`10.1007/978-3-030-01219-9_25`. `arXiv:1704.08545`.

[55] C. Sakaridis, D. Dai, S. Hecker, L. 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), 2018, pp. 687–704.

[56] A. Pfeuffer, K. Dietmayer, Robust Semantic Segmentation in Adverse Weather Conditions by means of Sensor Data Fusion, 2019. arXiv:1905.10117.

[57] D. Guan, Y. Cao, J. Yang, Y. Cao, M. Y. Yang, Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, Information Fusion 50 (2019) 148–157.

[58] L. Sun, K. Wang, K. Yang, K. 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, International Society for Optics and Photonics, 2019, p. 111690A.

[59] K. He, X. Zhang, S. Ren, J. Sun, Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (2015) 1904–1916.

[60] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. S. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1529–1537.

[61] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).

[62] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, 2016, pp. 1050–1059.

[63] Z. Zhang, X. Zhang, C. Peng, X. Xue, J. Sun, Exfuse: Enhancing feature fusion for semantic segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 269–284.

[64] H. Li, P. Xiong, H. Fan, J. Sun, DFANet: Deep Feature Aggregation for Real-Time Semantic Segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9522–9531. URL: http://arxiv.org/abs/1904.02216. arXiv:1904.02216.

[65] X. Li, H. Zhao, L. Han, Y. Tong, K. Yang, GFF: Gated Fully Fusion for Semantic Segmentation, arXiv preprint arXiv:1904.01803 (2019).

[66] Z. Tian, T. He, C. Shen, Y. Yan, Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation, 2019. arXiv:1903.02120.

[67] X. Li, L. Zhang, A. You, M. Yang, K. Yang, Y. Tong, Global Aggregation then Local Distribution in Fully Convolutional Networks, 2019. arXiv:1909.07229.

[68] M. Lin, Q. Chen, S. Yan, Network in network, 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (2014).

[69] Y. Zhu, C. Li, Y. Lu, L. Lin, B. Luo, J. Tang, FANet: Quality-Aware Feature Aggregation Network for RGB-T Tracking, arXiv preprint arXiv:1811.09855 (2018).

[70] A. Valada, R. Mohan, W. Burgard, Self-supervised model adaptation for multimodal semantic segmentation, International Journal of Computer Vision (2019).

[71] X. Hu, K. Yang, L. Fei, K. Wang, ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation, arXiv preprint arXiv:1905.10089 (2019).

[72] S.-W. Hung, S.-Y. Lo, H.-M. Hang, Incorporating Luminance, Depth and Color Information by a Fusion-based Network for Semantic Segmentation, 2018. arXiv:1809.09077.

[73] L.-C. Chen, Y. Yang, J. Wang, W. Xu, A. L. Yuille, Attention to scale: Scale-aware semantic image segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3640–3649.

[74] M. Ren, R. S. Zemel, End-to-end instance segmentation and counting with recurrent attention, CoRR abs/1605.09410 (2016).

[75] Y. Li, X. Chen, Z. Zhu, L. Xie, G. Huang, D. Du, X. Wang, Attention-guided unified network for panoptic segmentation, CoRR abs/1812.03904 (2018).

[76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017).

[77] B. Khaleghi, A. Khamis, F. O. Karray, S. N. Razavi, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (2013) 28–44.

[78] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms, John Wiley & Sons, 2014.

[79] J. Fiérrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Fusion strategies in multimodal biometric verification, in: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698), volume 3, IEEE, 2003, pp. III–5.

[80] A. González, D. Vázquez, A. M. López, J. Amores, On-board object detection: Multicue, multimodal, and multiview random forest of local experts, IEEE transactions on cybernetics 47 (2016) 3980–3990.

[81] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (2015).

[82] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.

[83] C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor Semantic Segmentation using depth information, arXiv preprint arXiv:1301.3572 (2013).

[84] C. Hazirbas, L. Ma, C. Domokos, D. 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. doi:10.1007/978-3-319-54181-5_14.

[85] L. Schneider, M. Jasch, B. Fröhlich, T. Weber, U. Franke, M. Pollefeys, M. Rätsch, 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, Springer, 2017, pp. 98–109. doi:10.1007/978-3-319-59126-1_9.

[86] A. Valada, G. L. Oliveira, T. Brox, W. Burgard, Deep Multispectral Semantic Scene Understanding of Forested Environments Using Multimodal Fusion, in: International Symposium on Experimental Robotics, Springer, 2017, pp. 465–477. doi:10.1007/978-3-319-50115-4_41.

[87] T. Karasawa, K. Watanabe, Q. Ha, A. Tejero-De-Pablos, Y. Ushiku, T. 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, ACM, 2017, pp. 35–43. doi:10.1145/3126686.3126727.

[88] M. Bijelic, F. Mannan, T. Gruber, W. Ritter, K. Dietmayer, F. Heide, Seeing Through Fog Without Seeing Fog: Deep Sensor Fusion in the Absence of Labeled Training Data, arXiv preprint arXiv:1902.08913 (2019).

[89] O. Mees, A. Eitel, W. Burgard, Choosing smartly: Adaptive multimodal fusion for object detection in changing environments, in: IEEE International Conference on Intelligent Robots and Systems, volume 2016-Novem, IEEE, 2016, pp. 151–156. doi:10.1109/IROS.2016.7759048. arXiv:1707.05733.

[90] J. Wagner, V. Fischer, M. Herman, S. Behnke, Multispectral pedestrian detection using deep fusion convolutional neural networks, in: ESANN 2016 - 24th European Symposium on Artificial Neural Networks, 2016, pp. 509–514.

[91] J. Guerry, B. Le Saux, D. Filliat, " Look at this one" detection sharing between modality-independent classifiers for robotic discovery of people, in: 2017 European Conference on Mobile Robots (ECMR), IEEE, 2017, pp. 1–6.

[92] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3051–3060.

[93] N. Wang, X. Gong, Adaptive Fusion for RGB-D Salient Object Detection, 2019. arXiv:1901.01369.

[94] S. McMahon, N. Sunderhauf, B. Upcroft, M. Milford, Multimodal Trip Hazard Affordance Detection on Construction Sites, IEEE Robotics and Automation Letters 3 (2018) 1–8.

[95] R. Zhang, S. A. Candra, K. Vetter, A. Zakhor, Sensor fusion for semantic segmentation of urban scenes, in: 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2015, pp. 1850–1857.

[96] J. Janai, F. Güney, A. Behl, A. Geiger, Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art, 2017. URL: http://arxiv.org/abs/1704.05519. arXiv:1704.05519.

[97] N. Patel, A. Choromanska, P. Krishnamurthy, F. 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, IEEE, 2017, pp. 1531–1536. doi:10.1109/IROS.2017.8205958.

[98] X. Chen, H. Ma, J. Wan, B. Li, T. 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, 2017, pp. 6526–6534. doi:10.1109/CVPR.2017.691. arXiv:1611.07759.

[99] A. Pfeuffer, K. Dietmayer, Optimal sensor data fusion architecture for object detection in adverse weather conditions, in: 2018 21st International Conference on Information Fusion (FUSION), IEEE, 2018, pp. 1–8.

[100] D. Xu, D. Anguelov, A. Jain, PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 244–253. doi:10.1109/CVPR.2018.00033. arXiv:1711.10871.

[101] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, A. M. López, Multimodal End-to-End Autonomous Driving, arXiv preprint arXiv:1906.03199 (2019).

[102] K. Liu, Y. Li, N. Xu, P. Natarajan, Learn to combine modalities in multimodal deep learning, arXiv preprint arXiv:1805.11730 (2018).

[103] J. Wang, Z. Wang, D. Tao, S. See, G. Wang, Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks,

in: European Conference on Computer Vision, Springer, 2016, pp. 664–679.

[104] L. Ma, J. Stückler, C. Kerl, D. 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), IEEE, 2017, pp. 598–605.

[105] L. Deng, M. Yang, T. Li, Y. He, C. Wang, RFBNet: Deep Multimodal Networks with Residual Fusion Blocks for RGB-D Semantic Segmentation, arXiv preprint arXiv:1907.00135 (2019).

[106] Y. Sun, W. Zuo, M. Liu, RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes, IEEE Robotics and Automation Letters 4 (2019) 2576–2583.

[107] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from rgb-d images for object detection and segmentation, in: European conference on computer vision, Springer, 2014, pp. 345–360.

[108] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, L. 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, Springer, 2016, pp. 541–557. doi:10.1007/978-3-319-46475-6_34. arXiv:1604.05000.

[109] A. Valada, J. Vertens, A. Dhall, W. Burgard, Adapnet: Adaptive semantic segmentation in adverse environmental conditions, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 4644–4651.

[110] Y. Cheng, R. Cai, Z. Li, X. Zhao, K. 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, 2017, pp. 3029–3037.

[111] Y. Zhang, O. Morel, M. Blanchon, R. Seulin, M. Rastgoo, D. 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, 2019, pp. 336–343. doi:10.5220/0007360403360343.

[112] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton, Others, Adaptive mixtures of local experts., Neural computation 3 (1991) 79–87.

[113] D. Eigen, M. Ranzato, I. Sutskever, Learning factored representations in a deep mixture of experts, arXiv preprint arXiv:1312.4314 (2013).

[114] S.-J. Park, K.-S. Hong, S. Lee, RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4980–4989.

[115] J. Jiang, Z. Zhang, Y. Huang, L. 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), IEEE, 2017, pp. 525–530.

[116] Y. Li, J. Zhang, Y. Cheng, K. Huang, T. Tan, Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation, in: Proceedings - International Conference on Image Processing, ICIP, volume 2017-Septe, IEEE, 2018, pp. 1262–1266. doi:10.1109/ICIP.2017.8296484.

[117] D. Lin, G. Chen, D. Cohen-Or, P. A. Heng, H. Huang, Cascaded Feature Network for Semantic Segmentation of RGB-D Images, in: Proceedings of the IEEE International Conference on Computer Vision, volume 2017-Octob, 2017, pp. 1320–1328. doi:10.1109/ICCV.2017.147.

[118] J. Jiang, L. Zheng, F. Luo, Z. Zhang, RedNet: Residual Encoder-Decoder Network for indoor RGB-D Semantic Segmentation, arXiv preprint arXiv:1806.01054 (2018).

[119] H. Blum, A. Gawel, R. Siegwart, C. Cadena, Modular sensor fusion for semantic segmentation, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 3670–3677.

[120] L. Xu, A. Krzyzak, C. Y. Suen, Methods of combining multiple classifiers and their applications to handwriting recognition, IEEE transactions on systems, man, and cybernetics 22 (1992) 418–435.

[121] N. Piasco, D. Sidibé, V. Gouet-Brunet, C. 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, IEEE, 2019, pp. 9094–9100. URL: https://doi.org/10.1109/ICRA.2019.8794221. doi:10.1109/ICRA.2019.8794221.

[122] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al., Speed/accuracy trade-offs for modern convolutional object detectors, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7310–7311.

[123] A. Holliday, M. Barekatain, J. Laurmaa, C. Kandaswamy, H. Prendinger, Speedup of deep learning ensembles for semantic segmentation using a model compression technique, Comput. Vis. Image Underst. 164 (2017) 16–26.

[124] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, International journal of computer vision 88 (2010) 303–338.

[125] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. 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, 2016, pp. 3213–3223. doi:10.1109/CVPR.2016.350. arXiv:1604.01685.

[126] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.

[127] S. Song, S. P. Lichtenberg, J. 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, 2015, pp. 567–576. doi:10.1109/CVPR.2015.7298655.

[128] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: European conference on computer vision, Springer, 2006, pp. 1–15.

[129] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, A. Yuille, The role of context for object detection and semantic segmentation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 891–898.

[130] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, A. 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, 2014, pp. 1971–1978.

[131] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 991–998. doi:10.1109/ICCV.2011.6126343.

[132] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[133] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: 2009 IEEE 12th international conference on computer vision, IEEE, 2009, pp. 1–8.

[134] G. J. Brostow, J. Shotton, J. Fauqueur, R. Cipolla, Segmentation and recognition using structure from motion point clouds, in: European conference on computer vision, Springer, 2008, pp. 44–57.

[135] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The KITTI dataset, International Journal of Robotics Research 32 (2013) 1231–1237.

[136] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A. 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, 2016, pp. 3234–3243. doi:10.1109/CVPR.2016.352.

[137] S. R. Richter, V. Vineet, S. Roth, V. 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, Springer, 2016, pp. 102–118. doi:10.1007/978-3-319-46475-6_7. arXiv:1608.02192.

[138] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Scene parsing through ADE20K dataset, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, 2017, pp. 5122–5130. doi:10.1109/CVPR.2017.544.

[139] G. Neuhold, T. Ollmann, S. Rota Bulò, P. Kontschieder, The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes, in: Inter-

national Conference on Computer Vision (ICCV), 2017. URL: https://www.mapillary.com/dataset/vistas.

[140] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, G. Fernandez Dominguez, WildDash-creating hazard-aware benchmarks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 402–416.

[141] H. Yin, C. 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, IEEE, 2018, pp. 1–8. doi:10.1109/ITSC.2017.8317828.

[142] O. Zendel, M. Murschitz, M. Zeilinger, D. Steininger, S. Abbasi, C. Beleznai, Railsem19: A dataset for semantic rail scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[143] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5828–5839.

[144] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, T. 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, IEEE, 2017, pp. 5108–5115. doi:10.1109/IROS.2017.8206396.

[145] W. Treible, P. Saponaro, Y. Liu, A. D. Gupta, V. Veerendraveer, S. Sorensen, C. Kambhamettu, Cats 2: Color and thermal stereo scenes with semantic labels, in: CVPR Workshops, 2019.

[146] G. Choe, S. H. Kim, S. Im, J. Y. Lee, S. G. Narasimhan, I. S. Kweon,

RANUS: RGB and NIR urban scene dataset for deep scene parsing, IEEE Robotics and Automation Letters 3 (2018) 1808–1815.

[147] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, C. J. Taylor, PST900: RGB-Thermal Calibration, Dataset and Segmentation Network, arXiv preprint arXiv:1909.10980 (2019).

[148] P. Kirsanov, A. Gaskarov, F. Konokhov, K. Sofiiuk, A. Vorontsova, I. Slinko, D. Zhukov, S. Bykov, O. Barinova, A. Konushin, DISCOMAN: Dataset of Indoor SCenes for Odometry, Mapping And Navigation, 2019. arXiv:1909.12146.

[149] M. Schwarz, H. Schulz, S. 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), IEEE, 2015, pp. 1329–1335.

[150] M. Velte, Semantic image segmentation combining visible and near-infrared channels with depth information, Ph.D. thesis, Ph. D. Dissertation. bibinfoschoolBonn-Rhein-Sieg University of Applied Sciences, 2015.

[151] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, K. He, A survey of infrared and visual image fusion methods, Infrared Physics & Technology 85 (2017) 478–501.

[152] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, I. S. Kweon, KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving, IEEE Transactions on Intelligent Transportation Systems 19 (2018) 934–948.

[153] C. Li, D. Song, R. Tong, M. Tang, Illumination-aware faster r-cnn for robust multispectral pedestrian detection, Pattern Recognition 85 (2019) 161–171.

[154] S. Farokhi, J. Flusser, U. U. Sheikh, Near infrared face recognition: A literature survey, Computer Science Review 21 (2016) 1–17.

[155] D.-W. Jang, R.-H. Park, Colour image dehazing using near-infrared fusion, IET Image Processing 11 (2017) 587–594.

[156] F. Dümbgen, M. E. Helou, N. Gucevska, S. Süsstrunk, Near-infrared fusion for photorealistic image dehazing, Electronic Imaging 2018 (2018) 321–1.

[157] Y. Benezeth, D. Sidibé, J.-B. Thomas, Background subtraction with multispectral video sequences, 2014.

[158] R. Gade, T. Moeslund, Thermal cameras and applications: A survey, Machine Vision & Applications 25 (2014) 245–262.

[159] L. B. Wolff, Polarization vision: A new sensory approach to image understanding, Image and Vision Computing 15 (1997) 81–93.

[160] L. B. Wolff, T. E. Boult, Constraining object features using a polarization reflectance model, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 635–657.

[161] Y. Y. Schechner, S. G. Narasimhan, S. 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 (2001) I–I.

[162] W. A. Smith, R. Ramamoorthi, S. Tozza, Linear depth estimation from an uncalibrated, monocular polarisation image, in: European Conference on Computer Vision, Springer, 2016, pp. 109–125.

[163] D. Zhu, W. A. Smith, Depth from a polarisation + rgb stereo pair, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[164] M. Blanchon, O. Morel, Y. Zhang, R. Seulin, N. Crombez, D. 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, 2019, pp. 328–335. doi:10.5220/0007360203280335.

[165] D. Sun, X. Huang, K. Yang, A multimodal vision sensor for autonomous driving, in: Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies III, volume 11166, International Society for Optics and Photonics, 2019, p. 111660L.

[166] S. Qiu, Q. Fu, C. Wang, W. Heidrich, Polarization demosaicking for monochrome and color polarization focal plane arrays (2019).

[167] N. A. Rubin, G. D'Aversa, P. Chevalier, Z. Shi, W. T. Chen, F. Capasso, Matrix fourier optics enables a compact full-stokes polarization camera, Science 365 (2019) eaax1839.

[168] K. Yang, L. M. Bergasa, E. Romera, X. Huang, K. Wang, Predicting Polarization beyond Semantics for Wearable Robotics, in: IEEE-RAS International Conference on Humanoid Robots, volume 2018-Novem, IEEE, 2019, pp. 96–103. doi:10.1109/HUMANOIDS.2018.8625005.

[169] M. Rastgoo, C. Demonceaux, R. Seulin, O. Morel, Attitude estimation from polarimetric cameras, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 8397–8403.

[170] E. Fernandez-Moral, R. Martins, D. Wolf, P. Rives, A new metric for evaluating semantic segmentation: leveraging global and contour accuracy, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 1051–1056.

[171] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing and Management 45 (2009) 427–437.

[172] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings - 30th IEEE Conference on Computer Vision and

Pattern Recognition, CVPR 2017, volume 2017-Janua, 2017, pp. 1800–1807. doi:10.1109/CVPR.2017.195. arXiv:1610.02357.

[173] L. Fan, W.-C. Wang, F. Zha, J. Yan, Exploring new backbone and attention module for semantic segmentation in street scenes, IEEE Access 6 (2018) 71566–71580.

[174] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

[175] A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, British Machine Vision Conference 2017, BMVC 2017 (2017).

[176] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Exploring context with deep structured models for semantic segmentation, IEEE transactions on pattern analysis and machine intelligence 40 (2017) 1352–1366.

[177] L. M. Bergasa, R. Arroyo, E. Romera, M. Alvarez, Efficient ConvNet for Real-time Semantic Segmentation, in: IEEE Intelligent Vehicles Symposium, Proceedings, Iv, IEEE, 2017, pp. 1789–1794. URL: http://www.robesafe.uah.es/personal/eduardo.romera/pdfs/Romera17iv.pdf.

[178] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, 2017, pp. 6230–6239. doi:10.1109/CVPR.2017.660. arXiv:1612.01105.

[179] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, Understanding convolution for semantic segmentation, in: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, 2018, pp. 1451–1460.