Exploration of deep learning-based multimodal fusion for semantic road scene segmentation

Yifei Zhang, Olivier Morel, Marc Blanchon, Ralph Seulin, Mojdeh Rastgoo and Désiré Sidibé
ImViA Laboratory EA 7535, ERL VIBOT CNRS 6000, Université de Bourgogne Franche-Comté, France
{*Yifei.Zhang*}*@u-bourgogne.fr*

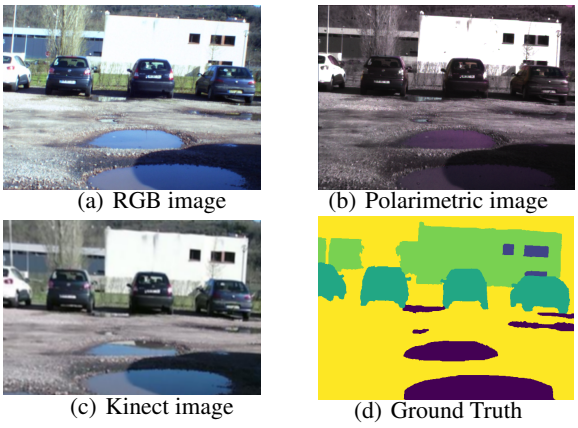
Keywords: Semantic Segmentation, Multimodal Fusion, Deep Learning, Road Scenes

Abstract: Deep neural networks have been frequently used for semantic scene understanding in recent years. Effective and robust segmentation in outdoor scene is prerequisite for safe autonomous navigation of autonomous vehicles. In this paper, our aim is to find the best exploitation of different imaging modalities for road scene segmentation, as opposed to using a single RGB modality. We explore deep learning-based early and later fusion pattern for semantic segmentation, and propose a new multi-level feature fusion network. Given a pair of aligned multimodal images, the network can achieve faster convergence and incorporate more contextual information. In particular, we introduce the first-of-its-kind dataset, which contains aligned raw RGB images and polarimetric images, followed by manually labeled ground truth. The use of polarization cameras is a sensory augmentation that can significantly enhance the capabilities of image understanding, for the detection of highly reflective areas such as glasses and water. Experimental results suggest that our proposed multimodal fusion network outperforms unimodal networks and two typical fusion architectures.

1 INTRODUCTION

Semantic segmentation is one of the main challenges in computer vision. Along with the appearance and development of Deep Convolutional Neural Network (DCNN) (Krizhevsky et al., 2012), the trained model can predict which class each pixel in the input images belongs to. By learning from massive data sets of diverse samples, this method achieves a good performance on end-to-end image recognition. Robust and accurate scene parsing of outdoor environments paves the way towards autonomous navigation and relationship inference. Compared with indoor scenes, off-road perception is more challenging due to dynamic and complex situations. The outdoor environment may easily change in different time slots with light or color variations. Even in structured environments, for instance on urban roads, there are still several challenges such as the detection of glass and muddy puddles.

Most existing datasets and methods for outdoor scene semantic segmentation are mainly based on RGB camera. They are only well acceptable in general conditions excluding complex environment and small amount of samples. To develop additional practical solutions, one of the main challenges is data fusion from multi-modalities. Therefore, considering the RGB modality as a kind of imperfect sensor, we attempt to fuse the complementary feature information of the same scene from other modalities. Actually, several modalities are ubiquitous in robotic systems, such as RGB-D, LIDAR, near infrared sensor, etc. Figure 1 shows the multimodal images of our POLABOT dataset.



(a) RGB image (b) Polarimetric image (c) Kinect image (d) Ground Truth

Figure 1: Multimodal images in POLABOT dataset.

In this work, we use a polarimetric camera, as a complementary modality, to provide a richer description of a scene. Polarization of light radiation has more general physical characteristic than inten-

sity and color (Wolff, 1997). We can figure out that windows of a building, the asphalt road, and the puddle of water have reflected polarizations (Walraven, 1977). Plenty of research have demonstrated that the use of polarization camera can significantly enhance the capabilities of scene understanding, especially for reflective areas (Harchanko and Chenault, 2005).

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

In this paper, we firstly review the existing fusion methods and datasets in section 2. Next, in section 3, we explore the two typical early and late fusion architectures, and propose our multi-stage Complex Modality network (CMnet), which has an encoder-decoder pattern and takes advantage of the state-of-art segmentation network. We evaluate the performances of the different fusion schemes using two different datasets in section 4. In particular, we introduce a new dataset, which to the best of our knowledge is the first multimodal dataset containing polarimetric images. Finally, the paper ends with concluding remarks in section 5.

2 RELATED WORK

In this section, we go through some of semantic segmentation methods, more details can be found in the review of (Garcia-Garcia et al., 2017). Then we summarize existing deep learning-based fusion schemes and various outdoor scene multimodal datasets.

Deep Neural Network Before deep learning achieved its current tremendous success, traditional computer vision methods were widely used, these methods are base on classifiers which operates on fixed-size feature inputs and a sliding-window. From the beginning with FCN (Long et al., 2015), the end-to-end fully convolutional network has become one of the most popular models for image segmentation. Recent years have witnessed a series of new encoder-decoder architectures along this line, including SegNet (Badrinarayanan et al., 2017), and U-Net (Ronneberger et al., 2015). Followed by the dilated convolutions proposed in (Yu and Koltun, 2015). Based on this technology, the series of DeepLab (Chen et al., 2014; Chen et al., 2018a; Chen et al., 2018b) achieves the state of the art performance in semantic segmentation.

Multimodal Fusion Architecture Benefiting from the improvement of unimodal neural network, excellent progress has been made on multimodal fusion architecture. Several common spectral sensors, such as RGB-D and near-infrared sensor, were applied to pixel-level data fusion of the same scene. For example, FuseNet (Hazirbas et al., 2016) and multi-view neural network (Ma et al., 2017) were proposed to incorporate complementary depth information into RGB segmentation framework. These fusion networks are based on an early fusion architecture, the feature maps from depth are constantly fused into the RGB branch in the encoder part.

Besides, a late fusion based model, Long Short-Term Memorized Context Fusion, also called LSTM-CF, was proposed by (Li et al., 2016). This network extracts multimodal features from depth and photometric data sources separately, then concatenates the feature map at three different scales. Another simple late fusion network (Eitel et al., 2015) was proposed for robust RGB-D object recognition. Furthermore, a convoluted mixture of deep experts technique (Valada et al., 2016a) was used in the late fusion architecture. These early and late fusion architectures were studied and applied to various scenarios and fields, for instance, forested environments navigation (Valada et al., 2016b), urban driving assistance (Jaritz et al., 2018).

Datasets Along with the development of computer vision techniques, a series of high-quality outdoor scene datasets have appeared, such as CamVid (Brostow et al., 2008b; Brostow et al., 2008a), Cityscapes (Cordts et al., 2016), etc. They are widely used in outdoor semantic scene understanding. In addition, some research institutes publish their scenario-based multimodal dataset. For instance, KAIST dataset (Hwang et al., 2015) is a multi-spectral pedestrian dataset of real traffic scenes, which was collected by a co-aligned RGB/Thermal camera, RGB stereo, 3D LiDAR and inertial sensors. Especially for semantic segmentation, there is KITTI dataset (Geiger et al., 2013) which contains high-resolution RGB data, grayscale stereo cameras data, and 3D point cloud; Freiburg Multi-spectral Forest dataset (Valada et al., 2016b) is also a multi-spectral dataset for forested environment semantic

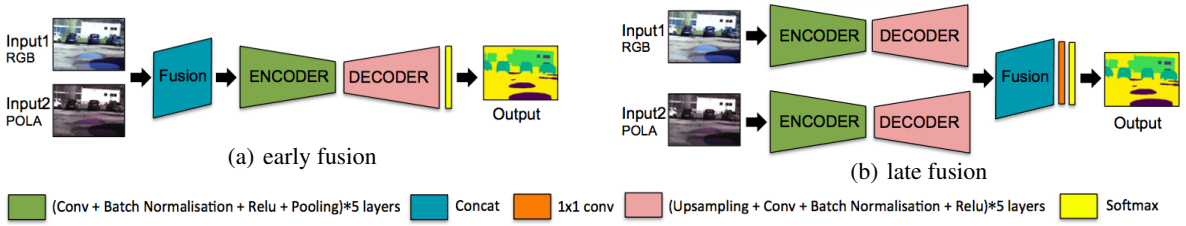


Figure 2: Early fusion and Late fusion architectures comparison.

segmentation, it contains RGB, Depth, NIR, Near-Infrared, Red, Green (NRG), Enhanced Vegetation Index (EVI), and Normalized Difference Vegetation Index (NDVI) images. However, none of these datasets contains polarimetric data.

3 MULTIMODAL FUSION

In this section, we describe the fusion architectures for multi-modalities and the training procedure in details. In essence, the process of training is to minimize the error while regularizing the parameters. Let $S = \{(X_n, y_n)|n = 1,2,...,N\}$ be a set of N training examples, where X_n is the feature vector of n-th example extracted from different modalities, and $y_n \in \{1,2,...,c\}$ is the corresponding segmentation class. Then the training problem can be framed as an optimization one, which can be formulated as:

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

where the loss is computed as $L(u,y) = -\sum_k y_k log u_k$. Then we can use, for example, gradient descent algorithm to find local minimum.

3.1 Fusion architectures

In this part, we describe two typical fusion strategies, namely early fusion and late fusion. The two simple structures, as well as their extensions, are widely used for deep learning-based fusion. Here we use SegNet as baseline network to construct such architectures. SegNet has a classical Encoder-Decoder structure followed by a Softmax classifier. The encoder is a regular convolutional neural network which contains five layers. Each layer extracts local features, normalizes the data distribution, obtains sparse representations by means of convolution, batch normalization and ReLU accordingly. Afterwards, pooling is used for downsampling the feature map and propagate spacial invariant features. Correspondingly, the decoder unsamples the shrunk feature map and recover the lost spatial information to full-sized segmentation.

3.1.1 Early fusion

As shown in Figure 2(a), the early fusion architecture has a unitary neural network, fusion takes place before passing into the encoder. Assume that both inputs (for example one RGB image and one polarimetric image) have size 3×H×W, then fused frame will be 6×H×W. So we also call this sort of fusion architecture as channel fusion.

This fusion architecture, combining features before training, seems simple and light. However, it is also more likely to over-fit. To see why, let consider the model's complexity. Let H be a family of functions taking values in $\{-1,+1\}$ with VC-dimensions d_{vc} (Vapnik, 1998). Then, for any $\delta > 0$ and all $h \in H$, the VC-dimension bound (Mohri et al., 2012) can be derived with a high probability:

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

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

3.1.2 Late fusion

Figure 2(b) shows the late fusion architecture which was used in this paper. It has two separated branches of network, with each branch trained to extract features from a special modality. Fusion takes place after a series of downsampling. Assuming that the two feature maps have size 1×H×W, after concatenation, the resulting feature will be 2×H×W. Then a 1×1 convolution is applied to reduce the number of channels.

This approach has the advantages that each network computes weights separately while encoding. Compared with early fusion, to some extent, it may

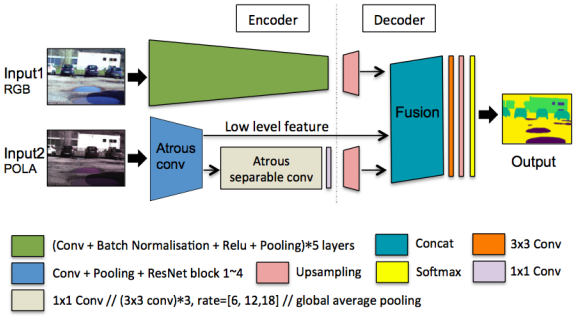


Figure 3: Our proposed fusion architecture: CMnet for multimodal fusion based on late fusion architecture.

reduce the difficulty of model fitting and yield a better outcomes. Furthermore, thanks to the scalability and flexibility of this architecture, the model can be designed in accordance with requirements and easily extend to multi-inputs without a large dimension increase.

3.2 Proposed fusion model

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

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

$$\hat{y}_n = f(X_n) = \text{softmax}[W * (M_1(x_a) + M_2(x_b))], \quad (3)$$

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

$$\text{softmax}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \quad (4)$$

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

Figure 3 presents the whole architecture of CMnet. It has an Encoder-Decoder structure and two separated branches. The encoder is used for mapping raw inputs to feature representations. The decoder integrates three feature maps, then recovers the feature representation to final segmentation results. That is a reliable method to extract different modality features and recover sharp object boundaries for end-to-end segmentation.

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

However, how to captures rich contextual information from extra modality is a challenging task. We refer to the state-of-the-art segmentation network Deeplab v3+ (Chen et al., 2018b), which uses a new pooling method named ASPP (Atrous Spatial Pyramid Pooling) to incorporate the multi-scale contextual information. We apply this network structure as the other branch's encoder for the complementary modality. The first upsampling stage is subsequently applied to each branch to recover the feature representation to the same fusion size, then we fuse these three feature maps, which contains high-level and low-level multimodal features information simultaneously. The second upsampling stage and softmax are applied to the fused feature map, which produces the final results.

4 EXPERIMENTAL RESULTS

In this section, we evaluate the different fusion models, and report a series of results on two datasets. One is the publicly available Freiburg multispectral forest dataset (Valada et al., 2016b), and the second one is a new multimodal dataset containing polarimetric and RGB data, called POLABOT dataset. In this work, all the networks are implemented based on Pytorch framework with a Nvidia Titan Xp graphics processing unit (GPU) acceleration. The input data was randomly shuffled after each epoch. We initialize the learning rate as 0.0001 and use the contraction segments of pre-trained VGG-16 model and ResNet-101 as encoders. Then we fine-tuned the weights of the decoders until convergence.

4.1 POLABOT dataset

As shown in Figure 4, we collected multimodal images using a mobile robot platform equipped with four cameras: the RGB camera (IDS Ucam), a polarimetric camera (PolarCam), a depth camera (Kinect 2.0), and a near-infrared camera. Our raw dataset contains over 700 multi-modalities images. All the images were acquired, synchronized and calibrated using the Robot Operating System (ROS) framework. Our benchmark also contains 175 images with pixel level ground truth annotations which were generated

Figure 4: Mobile robot platform used for the acquisition of the POLABOT dataset. It is equipped with the IDS Ucam, PolarCam, Kinect 2 and a NIR camera.

manually. These images have been dispatched into 8 classes: unlabeled, sky, water, windows, road, car, buildings and others. Benefiting from the use of a polarimetric camera, our mobile robot platform is more capable of discerning on windows, water and other reflective areas. That allows us to do much more exploratory research on polarimetric images in semantic scene understanding domain. In this paper, we use aligned RGB and polarimetric images as inputs to train the fusion models.

For integrating the acquired images, we apply an automatic homographic method to image alignment (Moisan et al., 2012). This method allows to transform the RGB images with respect to the polarimetric images, and crop to the intersecting regions of interest. Moreover, as deep learning models need large data sets of diverse examples, a certain amount of data should be guaranteed. For this reason, we employ geometric data augmentations to increase the effective number of training samples, including rotation and flipping. Data augmentation and multimodal data fusion help to train deep neural networks on small scale datasets.

4.2 Experimental evaluation

4.2.1 Freiburg Multispectral Forest dataset

We train the segmentation architectures on the public Freiburg Forest dataset first. This dataset was collected by a modified RGB dashcam with NIR-cut filter in outdoor forested environment. It consists of over 15,000 raw images, and 325 images with pixel level ground truth annotations for 6 classes, which are the sky, trail, grass, vegetation, obstacle and others. In this unstructured forest environment, Enhanced Vegetation Index(EVI) was proposed to improve sensitivity to high biomass regions and vegetation monitoring. It shows stronger capacities on feature representation than NIR in the previous work. To extract more accurate information, here in our case, we select EVI images as the second modality input besides the visible input.

Table 1: Performance of segmentation models on Freiburg Multispectral Forest dataset. EF, LF refer to early fusion and late fusion respectively. We report pixel accuracy (PA), mean accuracy (MA), mean intersection over union (MIoU), frequency weighted IoU (FWIoU) as metric to evaluate the performance.

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

Table 2: Comparison of deep unimodal and multimodal fusion approaches by class. We report MIoU as metric to evaluate the performance.

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

We crop the RGB and EVI images as size $3\times256\times256$, and use them as inputs correspondingly. We report several metrics to assess segmentation models: pixel accuracy (PA), mean accuracy (MA), mean intersection over union (MIoU), frequency weighted IoU (FWIoU). They are frequently used in semantic segmentation domain.

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

The results in Table 2 demonstrate the evaluations by class. We report the main four classes as Road, Grass, Veg/Tree and Sky. For uni-modality network, we can find that EVI shows good performance on Road and Grass classes, and RGB modality has a significant advantage on Sky class, which is susceptible to lighting changes. Moreover, the fusion architecture outperforms uni-modality scheme by integrating complementary multimodal information. In particular, our CMnet model achieved a remarkable results

Table 3: Segmentation performance on POLABOT dataset

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

on segmentation comparing with other fusion architectures, espe.

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

Some segmentation results on the Freiburg dataset are shown in Figure 5.

4.2.2 POLABOT dataset

In the following part, we report several experimental results on our POLABOT dataset. The metrics shown in Table 3 correspond to pixel accuracy (PA), mean accuracy (MA), F1 score (F1) and mean intersection over Union (MIoU).

We process the RGB and polarimetric images with size $3\times448\times448$. While training the networks, we experimentally found that stochastic gradient descent (batch size=1) doesn't work well. It is reasonable that online learning adds too much instability to the learning process as the weights widely vary with each batch, especially for small scale dataset with multi-classes. As a complement of previous analysis of training on small scale dataset, the data augmentation technology applied to POLABOT dataset gives the additional guarantee for weights learning. As a result, we can find that polarimetric images in our dataset provide high quality feature information, it is a beneficial premise for further data fusion. The overall best performance in this dataset was obtained with CMnet integrating RGB and polarimetric inputs, achieving a mean IoU of 72.59%. It yields around 3% comprehensive improvements comparing with the second best methods.

Some segmentation results on the POLABOT dataset are shown in Figure 6.

5 CONCLUSIONS

In this paper, we explored the typical early fusion and late fusion architectures that extract features from multi-modalities, and extensively evaluated theirs merits and deficiencies. We also proposed an extensible multi-level fusion scheme for semantic segmentation, which adopts advanced deep neural network techniques. It provides design choices for future research directions. We presented comprehensive quantitative evaluations of multimodal fusion on two datasets. The results show the benefits of fusing multimodal features to achieve state-of-the-art segmentation performance on small scale datasets. In addition, we introduced a first-of-a-kind outdoor scene segmentation dataset for road scene navigation, which contains high-quality aligned polarimetric images. We empirically demonstrate that the use of polarization camera enhance the capabilities of scene understanding.

Future work concerns deeper analysis of multimodal fusion network, since there is still plenty room for greater precision. One direction is to add the weights for each input while integrating. Moreover, it is possible to optimize the fusion pattern based on the physical properties of modalities and real-world scenarios.

ACKNOWLEDGEMENTS

This work was supported by the French Agence Nationale de la Recherche(ANR), under grant ANR-15-CE22-0009 (project VIPeR), as well as a hardware grant from NVIDIA.

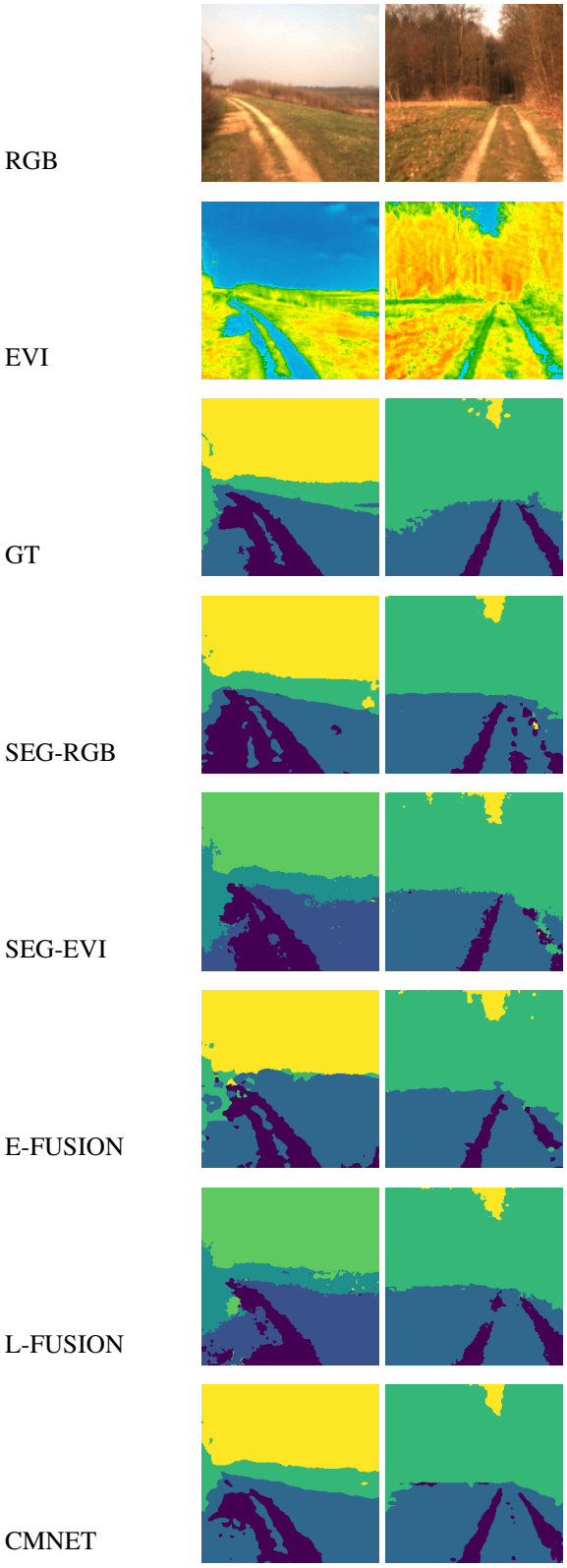


Figure 5: Two segmented examples from Freiburg Forest dataset. RGB and/or EVI images were given as inputs.

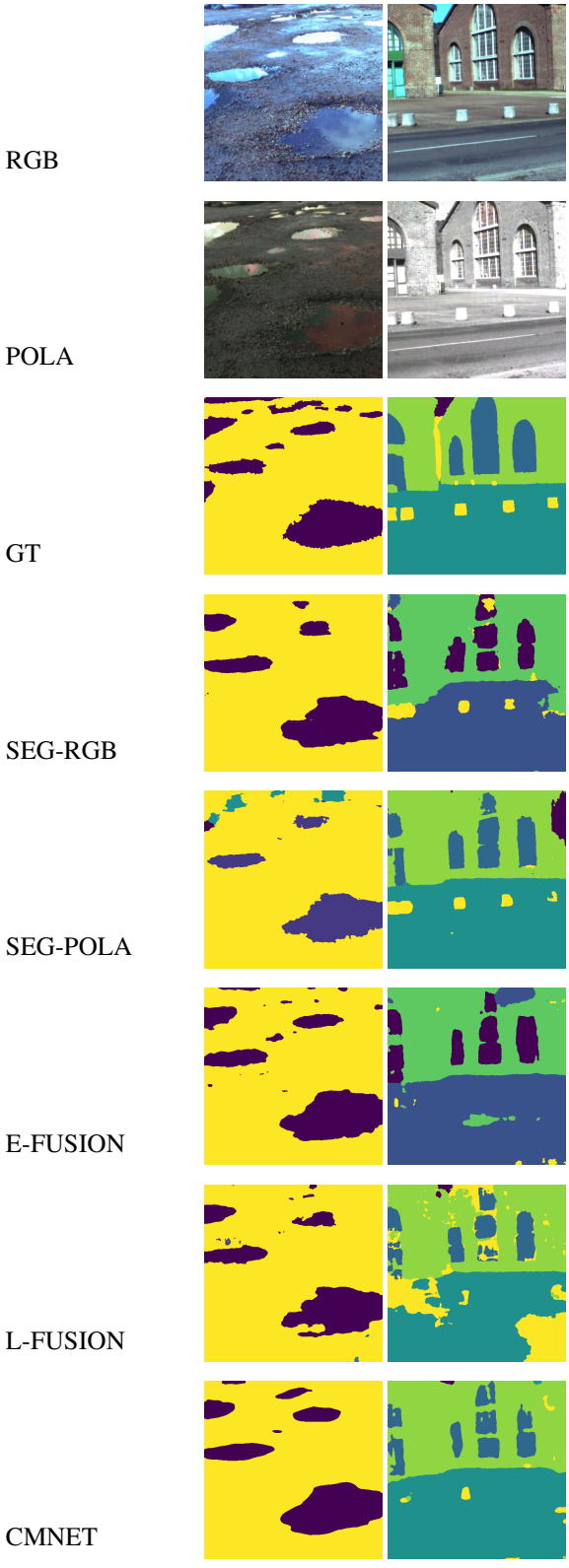


Figure 6: Two segmented examples from POLABOT dataset. RGB and/or POLA images were given as inputs.

REFERENCES

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2481–2495.

Brostow, G. J., Fauqueur, J., and Cipolla, R. (2008a). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx.

Brostow, G. J., Shotton, J., Fauqueur, J., and Cipolla, R. (2008b). Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pages 44–57.

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

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. (2015). Multimodal deep learning for robust rgb-d object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 681–687. IEEE.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*.

Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*.

Harchanko, J. S. and Chenault, D. B. (2005). Water-surface object detection and classification using imaging polarimetry. In *Polarization Science and Remote Sensing II*, volume 5888, page 588815. International Society for Optics and Photonics.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2016). Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, pages 213–228. Springer.

Hwang, S., Park, J., Kim, N., Choi, Y., and So Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1037–1045.

Jaritz, M., De Charette, R., Wirbel, E., Perrotton, X., and Nashashibi, F. (2018). Sparse and dense data with cnns: Depth completion and semantic segmentation. *arXiv preprint arXiv:1808.00769*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Li, Z., Gan, Y., Liang, X., Yu, Y., Cheng, H., and Lin, L. (2016). Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *European Conference on Computer Vision*, pages 541–557. Springer.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Ma, L., Stückler, J., Kerl, C., and Cremers, D. (2017). Multi-view deep learning for consistent semantic mapping with rgb-d cameras. In *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*, pages 598–605. IEEE.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

Moisan, L., Moulon, P., and Monasse, P. (2012). Automatic homographic registration of a pair of images, with a contrario elimination of outliers. *Image Processing On Line*, 2:56–73.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Valada, A., Dhall, A., and Burgard, W. (2016a). Convoluted mixture of deep experts for robust semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) Workshop, State Estimation and Terrain Perception for All Terrain Mobile Robots*.

Valada, A., Oliveira, G., Brox, T., and Burgard, W. (2016b). Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In *The 2016 International Symposium on Experimental Robotics (ISER 2016)*, Tokyo, Japan.

Vapnik, V. (1998). *Statistical learning theory. 1998*, volume 3. Wiley, New York.

Walraven, R. (1977). Polarization imagery. In *Optical Polarimetry: Instrumentation and Applications*, volume 112, pages 164–168. International Society for Optics and Photonics.

Wolff, L. B. (1997). Polarization vision: a new sensory approach to image understanding. *Image and Vision computing*, 15(2):81–93.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.