Deep Multispectral Semantic Scene Understanding of Forested Environments using Multimodal Fusion

Abhinav Valada, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard

Department of Computer Science, University of Freiburg, Germany

Abstract. Semantic scene understanding of unstructured environments is a highly challenging task for robots operating in the real world. Deep Convolutional Neural Network architectures define the state of the art in various segmentation tasks. So far, researchers have focused on segmentation with RGB data. In this paper, we study the use of multispectral and multimodal images for semantic segmentation and develop fusion architectures that learn from RGB, Near-InfraRed channels, and depth data. We introduce a first-of-its-kind multispectral segmentation benchmark that contains 15,000 images and 366 pixel-wise ground truth annotations of unstructured forest environments. We identify new data augmentation strategies that enable training of very deep models using relatively small datasets. We show that our UpNet architecture exceeds the state of the art both qualitatively and quantitatively on our benchmark. In addition, we present experimental results for segmentation under challenging real-world conditions. Benchmark and demo are publicly available at http://deepscene.cs.uni-freiburg.de.

Keywords: Semantic Segmentation, Convolutional Neural Networks, Scene Understanding, Multimodal Perception

1 Introduction

Semantic scene understanding is a cornerstone for autonomous robot navigation in real-world environments. Thus far, most research on semantic scene understanding has been focused on structured environments, such as urban road scenes and indoor environments, where the objects in the scene are rigid and have distinct geometric properties. During the DARPA grand challenge, several techniques were developed for offroad perception using both cameras and lasers [20]. However, for navigation in forested environments, robots must make more complex decisions. In particular, there are obstacles that the robot can drive over, such as tall grass or bushes, but these must be distinguished safely from obstacles that the robot must avoid, such as boulders or tree trunks.

In forested environments, one can exploit the presence of chlorophyll in certain obstacles as a way to discern which obstacles can be driven over [2]. However,

This work has partly been supported by the European Commission under FP7-267686-LIFENAV and FP7-610603-EUROPA2.

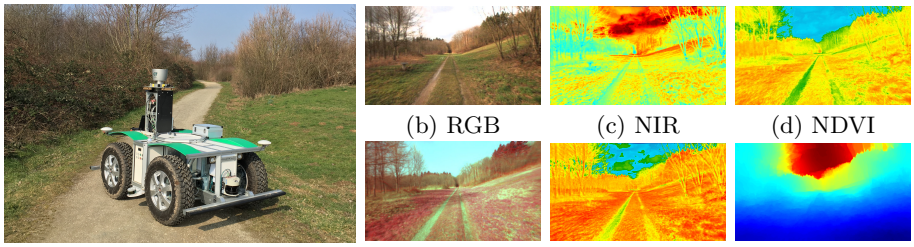
the caveat is the reliable detection of chlorophyll using monocular cameras. This detection can be enhanced by additionally using the Near-InfraRed (NIR) wavelength, which provides a high fidelity description on the presence of vegetation. Potentially, NIR images can also enhance border accuracy and visual quality. We aim to explore the correlation and de-correlation of visible and NIR images frequencies to extract more accurate information about the scene.

In this paper, we address the segmentation problem in forested environments by leveraging deep up-convolutional neural networks and techniques developed in the field of photogrammetry using multispectral cameras to obtain a robust pixel-accurate segmentation of the scene. We present an inexpensive system to capture RGB, NIR, and depth data using two monocular cameras, and introduce a first-of-a-kind multispectral and multimodal segmentation benchmark. We first evaluate the segmentation using our UpNet architecture, individually trained on various spectra and modalities contained in our dataset, then identify the best performing modalities and fuse them using various Deep Convolutional Neural Network (DCNN) fusion architecture configurations. We show that the fusion approach outperforms segmentation using either one of the modalities. Furthermore, we show that the fusion models trained on an extended version of our dataset containing extreme outdoor conditions such as snow, low-lighting and glare, demonstrate higher robustness than their unimodal counterparts.

2 Related Work

In recent years, deep learning approaches have successfully been applied to various robotic vision tasks including object recognition [19, 5], detection [17, 15] and semantic segmentation [14, 10, 11, 1]. For segmentation tasks, Long *et al.* [11] proposed fully convolutional networks (FCNs) that use pooling layers from a classification network to refine the segmentation produced by deconvolution layers. Oliveira *et al.* [14] proposed an improved architecture that further increases the efficiency using parameter reduction and additional refinements. Liu *et al.* [10] introduced a FCN called ParseNet that models global context directly. Kendall *et al.* [1] proposed another extension to FCNs that improves the efficiency by using pooling indices computed in max-pooling for the upsampling step.

Although, DCNNs have achieved state-of-the-art performance in various perception tasks, so far they have only been applied to and demonstrated on standard datasets that primarily contain RGB or at most depth images, collected in ideal conditions without aggressive changes in weather and illumination. There are limited number of DCNN architectures that explore the fusion of multiple modalities or spectra [3, 16, 18]. Eitel *et al.* [3] proposed a late-fusion approach for object detection using RGB-D data. Their approach utilizes a two-stream convolutional neural network (RGB and colorized depth image), first trained individually on each modality, followed by the fusion of their predictions using a set of inner-product layers. Schwarz *et al.* [16] proposes a similar approach that uses a two-stream network for RGB-D fusion, where the DCNN is only used for feature extraction followed by an SVM to determine the category, instance, and pose. Socher *et al.* [18] proposes technique that uses RGB and depth features



(a) Viona Autonomous Platform (b) RGB (c) NIR (d) NDVI (e) NRG (f) EVI (g) DEPTH

Fig. 1. Robot platform used for experimentation and sample images from our benchmark showing various spectra and modalities contained in our benchmark.

extracted from a single layer CNN and feeds both representations to a set of recurrent neural networks (RNNs). The concatenation of all the vectors from the RNNs forms the final representation which is then given to a softmax classifier.

In contrast to these approaches, our techniques learn highly discriminative features for semantic segmentation. We perform comprehensive evaluations on these two fusion approaches using combinations of multiple modalities and spectra contained in our benchmark. To the best of our knowledge, this is the first work to explore the use of multimodal and multispectral data for end-to-end semantic segmentation.

3 Multispectral Segmentation Benchmark

We collected the dataset using our Viona autonomous mobile robot platform equipped with a Bumblebee2 stereo vision camera and a modified dashcam with the NIR-cut filter removed for acquiring RGB and NIR data respectively. We use a Wratten 25A filter in the dashcam to capture the NIR wavelength in the blue and green channels. Both cameras are time synchronized and frames were captured at 20Hz. In order to match the images captured by both cameras, we first compute SIFT [12] correspondences between the images using the Difference-of-Gaussian detector to provide similarity-invariance. We then filter the detected keypoints with the nearest neighbours test, followed by requiring consistency between the matches with respect to an affine transformation. The matches are further filtered using Random Sample Consensus (RANSAC) [4] and the transformation is estimated using the Moving Least Squares method by rendering through a mesh of triangles. We then transform the RGB image with respect to the NIR image and crop to the intersecting regions of interest. Although our implementation uses two cameras, it is the most cost-effective solution compared to commercial single multispectral cameras. Fig. 1 shows our autonomous robot platform that we used and some examples from our benchmark from each spectrum and modality.

We collected data on three different days to have enough variability in lighting conditions as shadows and sun angles play a crucial role in the quality of acquired images. Our raw dataset contains over $15,000$ images sub-sampled at 1Hz, which

corresponds to traversing about 4.7km each day. Our benchmark contains 366 images with pixel level groundtruth annotations which were manually annotated. As there is an abundant presence of vegetation in our environment, we compute global-based vegetation indices such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) to extract consistent spatial and global information. NDVI is resistant to noise caused due to changing sun angles, topography and shadows but is susceptible to error due to variable atmospheric and canopy background conditions [7]. EVI was proposed to compensate for these defects with improved sensitivity to high biomass regions and improved detection though decoupling of canopy background signal and reduction in atmospheric influences. For all the images in our dataset, we calculate NDVI and EVI as shown by Huete *et al.* [7].

Although our dataset contains images from the Bumblebee stereo pair, the processed disparity images were substantially noisy due to several factors such as rectification artifacts, motion blur, etc. We compared the results from semi-global matching [6] to a DCNN approach that predicts depth from single images and found that for an unstructured environment such as ours, the DCNN approach gave better results. In our work, we use the approach from Liu *et. al,* [9] that employs a deep convolutional neural field model for depth estimation by constructing unary and pairwise potentials of conditional random fields. Our dataset is publicly available at `http://deepscene.cs.uni-freiburg.de/\#datasets`.

4 Technical Approach

In this section, we first describe our base network architecture for segmenting unimodal images and then elaborate our approaches for learning from multimodal and multispectral images. We represent the training set as $S = \{(X_n, Y_n), n = 1, \ldots, N\}$, where $X_n = \{x_j, j = 1, \ldots, |X_n|\}$ denotes the raw image, $Y_n = \{y_i, j = 1, \ldots, |X_n|\}, y_j \in \{0, C\}$ denotes the corresponding ground truth mask with C classes, θ are the parameters of the network and $f(x_j; \theta)$ is the activation function. The goal of our network is to learn features by minimizing the cross-entropy (*softmax*) loss that can be computed as $\mathcal{L}(u, y) = -\sum_k y_k log u_k$. Using stochastic gradient decent, we then solve

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}\left(\left(f(x^i; \theta)\right), y^i\right). \tag{1}$$

Recently, approaches that employ DCNNs for semantic segmentation have achieved state-of-the-art performance on segmentation benchmarks including PASCAL VOC, PASCAL Parts, PASCAL-Context, Sift-Flow and KITTI [11, 14]. These networks are trained end-to-end and do not require multi-stage techniques. Due to their unique architecture they take the full context of the image into account while providing pixel-accurate segmentations. We build upon our UpNet architecture, following this general principle with two main components: contraction and expansion. Given an input image, the contraction is responsible

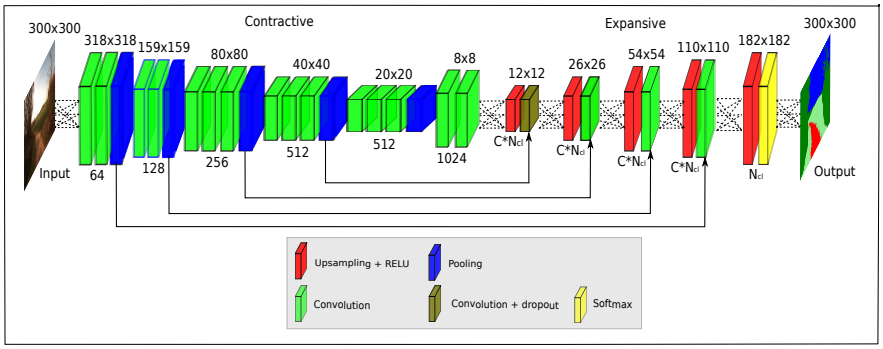
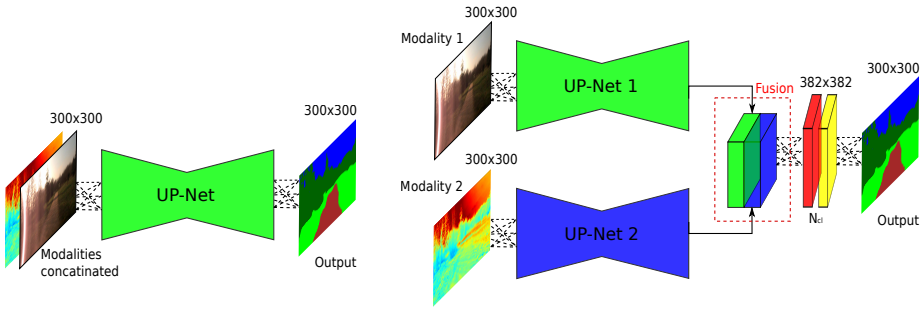


Fig. 2. Our UpNet architecture with up-convolutional layers of size $C \times N_{cl}$, where N_{cl} is the number of classes and C is a scalar factor of filter augmentations. The contractive segment of the network contains convolution and pooling layers, while the expansive segment of the network contains upsampling and convolution layers.

for generating a low resolution segmentation mask. We use the 13-layer VGG [19] architecture as basis on the contraction side. The expansion side consists of five up-convolutional refinement segments that refine the coarse segmentation masks generated by the contraction segment. Each up-convolutional refinement is composed of one up-sampling layer followed by a convolution layer. We add a rectified linear unit (ReLU) after each refinement and to avoid overfitting and we use spatial dropout after the first and last refinement layers.

The inner-product layers of the VGG-16 architecture has 4096 filters of 7×7 size, which is primarily responsible for relatively slow classification times. We reduce the number of filters to 1024 and the filter size to 3×3 to accelerate the network. There was no noticeable performance drop due to this change. Recent work have demonstrated improved performance by having variable number of filters as in the contraction segment [13, 14]. We experimented with this relationship and now use a $C \times N_{cl}$ mapping scheme, where C is a scalar constant and N_{cl} is the number of classes in the dataset. This makes the network learn more feature maps per class and hence increases the efficiency in the expansion segment. In the last layer we use the number of filters as N_{cl} in order to calculate the loss only over the useful classes. The structure of our base UpNet architecture is shown in Fig. 2. We train our segmentation network individually on RGB, NIR and depth data, as well as on various combinations of these spectra and modalities, as shown in section 5. To provide a more informative and sharper segmentation, we introduce two approaches;

- *Channel Stacking*: The most intuitive paradigm of fusing data using DCNNs is by stacking them into multiple channels and learning combined features end-to-end. However, previous efforts have been unsuccessful due to the difficulty in propagating gradients through the entire length of the model [11].
- *Late-Fused-Convolution:* In the late-fused-convolution approach, each model is first learned to segment using a specific spectrum/modality. Afterwards, the feature maps are summed up element-wise before a series of convolu-



(a) Channel Stacking (b) Late-fused Convolution

Fig. 3. Deep fusion architecture configurations proposed. Channel Stacking involves concatenating multiple modalities into channels and learning combined features from the beginning, while Late-fused Convolution involves individually learning to segment using separate streams, followed by further learning fused representations.

tion, pooling and up-convolution layers. This approach has the advantage as features in each model may be good at classifying a specific class and combining them may yield a better throughput, even though it necessitates heavy parameter tuning.

Fig. 3 shows a depiction of both these approaches. Our experiments provide an in-depth analysis of the advantages and disadvantages of each of these approaches in the context of semantic segmentation.

5 Experimental Results

In this section, we report results using the various spectra and modalities contained in our benchmark. We use the Caffe [8] deep learning framework for the implementation. Training on an NVIDIA Titan X GPU took about 4 days with cuDNN acceleration.

5.1 Comparison to the state of the art

To compare with the state-of-the-art, we train models using the *RGB RSC* (Rotation, Scale, Color) set from our benchmark which contains 60,900 RGB images with rotation, scale and color augmentations applied. We selected the baseline networks by choosing the top three end-to-end deep learning approaches from the PASCAL VOC 2012 leaderboard. We explored the parameter space to achieve the best baseline performance. We trained our network with both fixed and poly learning rate policies with a initial learning rate $\lambda_0 = 10^{-9}$, which can be given as $\lambda_n = \lambda_0 \times \left(\frac{1-N}{N_{max}}\right)^c$, where λ_n is the current learning rate, N is the iteration number, N_{max} is the maximum number of iterations and c is the power. We train the network using stochastic gradient descent with a momentum of 0.9 for

Table 1. Performance of our proposed model in comparison to the state-of-the-art

Baseline	IoU	PA	PRE	REC	FPR	FNR	Time
FCN-8 [11]	77.46	90.95	87.38	85.97	10.32	12.12	$\sim$ 255ms
SegNet [1]	74.81	88.47	84.63	86.39	13.53	11.65	$\sim$ 156ms
ParseNet [10]	83.65	93.43	90.07	91.57	8.94	7.41	$\sim$ 90ms
Ours Fixed lr	84.90	94.47	91.16	91.86	7.80	7.40	$\sim$ 52ms
Ours Poly lr	**85.31**	**94.47**	**91.54**	**91.91**	**7.40**	**7.30**	$\sim$ **52**ms

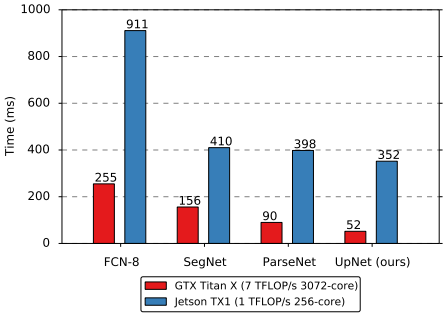


Fig. 4. Comparison of forward pass time with baseline networks.

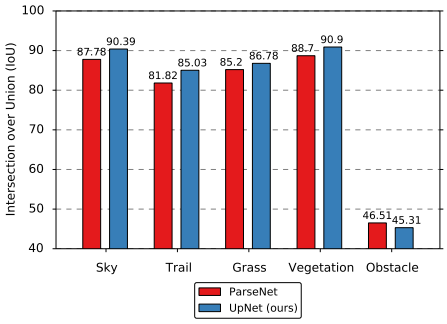


Fig. 5. Comparison of per-class IoU of best baseline (Parsenet) with ours.

300,000 iterations for each refinement stage. We found the poly learning rate policy to converge faster and yield a slight improvement in performance.

The metrics shown in Tab. 1 correspond to Mean Intersection over Union (IoU), Mean Pixel Accuracy (PA), Precision (PRE), Recall (REC), False Positive Rate (FPR), False Negative Rate (FNR). The time reported is for a forward pass through the network. The results demonstrate that our network outperforms all the state-of-the-art approaches and with a runtime of almost twice as fast as the second best technique.

5.2 Parameter Estimation and Augmentation

To increase the effective number of training samples, we employ data augmentations including scaling, rotation, color, mirroring, cropping, vignetting, skewing, and horizontal flipping. We evaluated the effect of augmentation using three different subsets in our benchmark: RSC (Rotation, Scale, Color), Geometric augmentation (Rotation, Scale, Mirroring, Cropping, Skewing, Flipping) and all aforementioned augmentations together. Tab. 2 shows the results from these experiments. Data augmentation helps train very large networks on small datasets. In our network, we replace the dropout in the VGG architecture with spatial dropout which gives us an improvement of 5.7%. Furthermore, we initialize the convolution layers in the expansion part of the network with Xavier initialization, which makes the convergence faster and also enables us to use a higher learning rate. This yields a 1% improvement.

Table 2. Comparison on the effects of augmentation on our benchmark.

	Sky	Trail	Grass	Veg	Obst	IoU	PA
Ours Aug.RSC	90.46	84.51	86.72	90.66	44.39	84.90	94.47
Ours Aug.Geo	89.60	84.47	86.03	90.40	42.23	84.39	94.15
Ours Aug.All	90.39	85.03	86.78	90.90	45.31	85.30	94.51

Table 3. Comparison of deep fusion approaches. D, N, E refer to depth, NIR and EVI respectively. CF and LFC refer channel fusion and late-fused-convolution.

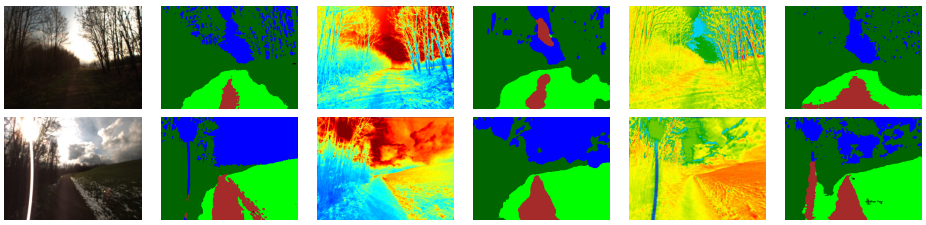
	Sky	Trail	Grass	Veg	Obst	IoU	FPR	FNR
RGB	90.46	84.51	86.72	90.66	44.39	84.90	7.80	7.40
NIR	86.08	75.57	81.44	87.05	42.61	80.22	10.22	9.60
DEPTH	88.24	66.47	73.35	83.13	46.13	76.10	12.76	11.14
NRG	89.88	85.08	86.27	90.55	47.56	85.23	7.70	7.10
EVI	88.00	83.40	84.59	87.68	44.9	83.25	8.70	8.10
NDVI	87.79	83.86	83.57	87.45	48.19	83.39	8.62	8.00
3CF RGB-N-D	89.23	**85.86**	86.08	90.32	**61.68**	86.35	7.50	6.20
4CF RGB-N	89.64	83.37	85.83	**90.67**	59.85	85.79	**7.00**	7.20
5CF RGB-N-D	89.40	84.30	85.84	89.40	60.62	86.00	7.20	6.80
LFC RGB-D	90.21	79.14	83.46	88.67	57.73	84.04	9.40	6.55
LFC RGB-N	90.67	83.31	86.19	90.30	58.82	85.94	7.50	6.56
LFC RGB-E	**90.92**	85.75	**87.03**	90.50	59.44	**86.90**	**7.00**	**5.76**
LFC NRG-D	90.34	80.64	84.81	89.08	56.60	84.77	7.58	7.65

5.3 Evaluations on Multi-Spectra/Modality Benchmark

Segmentation using RGB yields best results among all the individual spectra and modalities that we experimented with. The low representational power of depth images causes poor performance in the grass, vegetation and trail classes, bringing down the mean IoU. The results of the unimodal images shown in Tab. 3 demonstrate the need for fusion. Multispectrum channel fusion such as NRG (Near-Infrared, Red, Green) shows greater performance when compared to their individual counterparts and better recognition of obstacles. The best channel fusion we obtained was using a three channel input, composed of grayscaled RGB, NIR and depth data. It achieved an IoU of 86.35% and most importantly a considerable gain (over 13%) on the obstacle class, which is the hardest to segment in our benchmark. The overall best performance was from the late-fused-convolution of RGB and EVI, achieving a mean IoU of 86.9% and comparably top results in individual class IoUs as well. This approach also had the lowest false positive and false negative rates.

5.4 Robustness Evaluation

We performed extensive evaluations on an extended version of our dataset containing adverse conditions including snow, glare, motion blur and low lighting.



(a) RGB Output (b) NIR Output (c) EVI Output

Fig. 6. Segmented examples from our benchmark. Each spectrum provides valuable information. First row shows the image and the corresponding segmentation in highly shadowed areas. Second row shows the performance in the presence of glare and snow.

Fig. 6 shows some qualitative results from this subset. It can be seen that each of the spectra performs well in different conditions. Segmentation using RGB images shows remarkable detail, although being easily susceptible to lighting changes. NIR images on the other hand show robustness to lighting changes but often show false positives between the sky and trail classes. EVI images are good at detecting vegetation but show a large amount of false positives for the sky. We retrained the models presented in Tab. 3 on our adverse conditions dataset. All the models demonstrate improved performance as they learn probable distributions of corruption patterns that occur due to a change in conditions that take place throughout the day and across seasons.

Fig.7 shows the improvement in the mean IoU for both the fusion approaches after the addition of the adverse conditions subset. For unimodal data, segmentation with NIR images have the largest improvement of 3.91% mean IoU, followed by a 2.49% improvement for segmentation using RGB images. The model trained using the NIR images also showed a 5.27% decrease in the false-positive rate. For the Channel-stacking approach, segmentation using NRG images yielded the highest mean IoU of 87.27%, which is an improvement of 2.04% compared to the model trained on the dataset without the adverse conditions. Finally, for our Late-fused convolution approach, similar to the results reported in Tab. 3, segmentation using RGB and EVI yields the overall best results, achieving a mean IoU of 88.16%.

Fig. 8 shows qualitative comparisons between the segmentation obtained using the RGB model and the two deep fusion approaches that have demonstrated the best results in the quantitative experiments (Channel-stacking: NRG, Late-fused convolution: RGB and EVI). Fig. 8 (a) shows results in low lighting conditions and in the presence of shadows. In this scenario, models trained on RGB data or using Channel-stacking often have difficulty in identifying the pixels that belong to the trail class. This is especially evident in the images that have very narrow trail paths (Fig. 8 (a), (c) and (d)). It can also be seen that the results using the RGB model and the Late-fused Convolution have the highest segmentation granularity, which is noticeable in the segmentation of trees.

Fig. 8 (b) exemplifies segmentation with high saturation. RGB and Late-fused Convolution models demonstrate close similarity. Although it can be observed

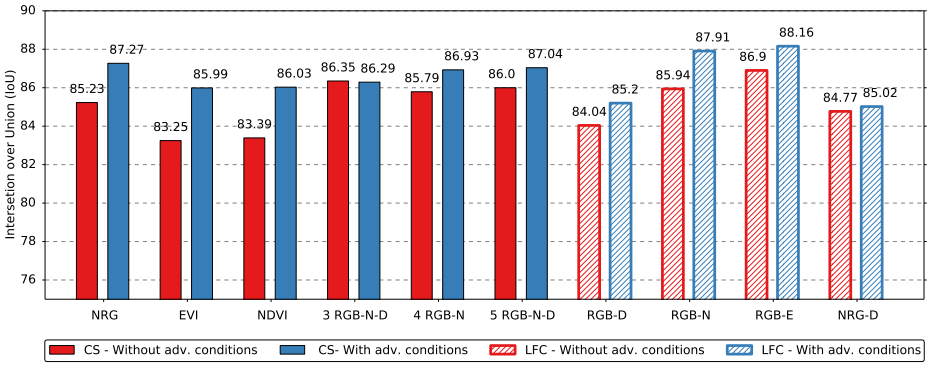


Fig. 7. Comparison of deep fusion models trained with and without the adverse conditions subset. Late-fused convolution of RGB and EVI yields the overall best fusion result, followed by of RGB and NIR. Acronyms D, N, E used in the model names refer to depth, NIR and EVI respectively and the digits indicate the number of channels.

that only the Late-Fused Convolution model is able to accurately segment the entire obstacle (building in the right background). It can often be seen that when an obstacle is in the far background (Fig. 8 (b), (e) and (f), the RGB model is only able to segment a small part of the obstacle and the Channel-stacking model completely fails to detect it. Fig. 8 (c) is an example of where both RGB and Channel-stacking models fail to segment a challenging transition from grass to vegetation. In this example, the RGB model also shows difficulty in accurately segmenting the trail class, especially in the areas that have tall grass. The Channel-stacking model is able to detect the entire trail, however it is unable to accurately detect the grass-vegetation transition and it shows false positives in the obstacle class. The Late-fused Convolution model is able to accurately detect such challenging transitions with a low false positive rate.

The examples shown in Figs. 8 (d), (e) and (f) illustrate adverse conditions such as glare, motion blur and snow. Fig. 8 (d) and (e) show an example of a scene with glare directly on the optics, which is common scenario for robots operating in real-world outdoor environments. Both the RGB and the Channel-stacking models are unable to accurately segment the classes in the presence of these disturbances. Fig. 8 (e) shows a similar scenario with motion blur and in the obstacles. Fig. 8 (f) is characterized by the presence of snow on the ground. RGB and Channel-stacking models misclassify snow as a part of the trail class and fail to detect the obstacles. The Late-fused Convolution model on the other hand, demonstrates invariance to glare and snow. This highlights the advantage of this approach, as it fuses feature maps further down the network, it is likely to make less mistakes and learn complementary features. In Channel-stacking, if there is a discrepancy in the features learned it cannot be corrected as the multimodal and multispectral features are learned together from the beginning.

In addition to these experiments, a live demo can be accessed at `http://deepscene.cs.uni-freiburg.de/#demo`, where a user can upload any image of an unstructured forest environment for segmentation or choose a random example from the benchmark.

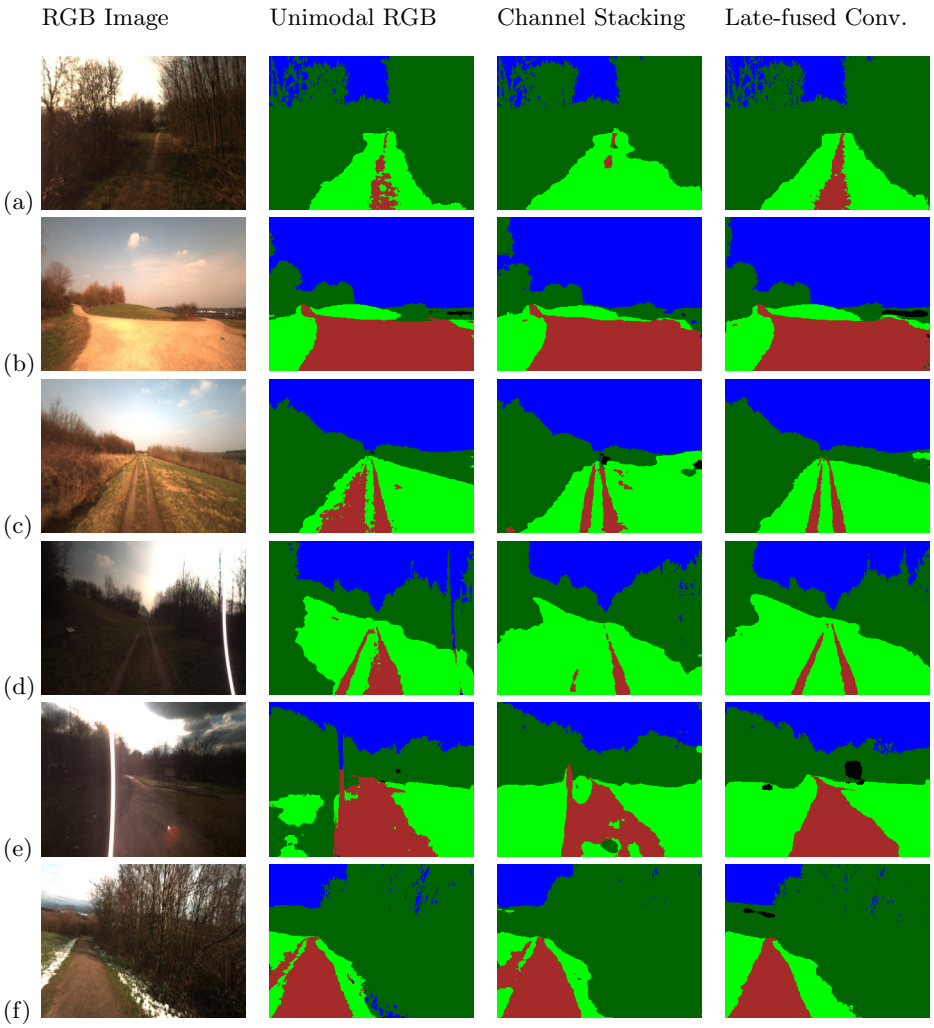


Fig. 8. Qualitative comparison of segmentation from the unimodal RGB model and our two fusion strategies. Our late-fused convolution model consistently yields unparalleled performance even in conditions such as snow, low-lighting, glare and motion blur.

6 Conclusions

In this paper, we presented a DCNN architecture for semantic segmentation of outdoor environments. Our network outperforms several state-of-the-art architectures with near real-time performance which is critical for robotic applications. We extensively evaluated the benefits and drawbacks of deep early and late-fusion architectures for dense pixel-wise segmentation using multiple modalities and spectra. Our late-fused convolution technique exceeds channel stacking by achieving the lowest false detection rate. We additionally trained models on an

extended version of our dataset containing images captured in adverse weather conditions such as snow, low-lighting, glare and motion blur. We showed that our networks learn to leverage features from complementary modalities and spectra to yield robust segmentation in the presence of these disturbances. Furthermore, we qualitatively demonstrated the benefits of multispectral fusion in several adverse conditions. The results demonstrate that fusing the NIR wavelength with RGB to obtain yields a more robust segmentation in unstructured outdoor environments. We publicly released a first-of-a-kind multispectral and multimodal semantic segmentation benchmark to accelerate further research on deep fusion.

References

1. V. Badrinarayanan et. al., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", arXiv preprint arXiv:1511.00561, 2015.
2. D.M. Bradley et al., "Vegetation Detection for Mobile Robot Navigation", Tech Report CMU-RI-TR-05-12, Carnegie Mellon University, 2004.
3. A. Eitel et al., "Multimodal Deep Learning for Robust RGB-D Object Recognition", Intl. Conf. on Intelligent Robots and Systems, 2015.
4. M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis" Comm. of the ACM, 1981.
5. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", arXiv preprint arXiv:1512.03385, 2015.
6. H. Hirschmüller, "Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information", CVPR, 2005.
7. A. Huete, C.O. Justice, and W.J D. van Leeuwen, MODIS Vegetation Index (MOD 13), Algorithm Theoretical Basis Document (ATBD), Version 3.0, pp. 129, 1999.
8. Y. Jia et. al, "Caffe: Convolutional Architecture for Fast Feature Embedding", arXiv preprint arXiv:1408.5093, 2014.
9. F. Liu, C. Shen and G. Lin, "Deep Convolutional Neural Fields for Depth Estimation from a Single Image", arXiv:1411.6387, 2014.
10. W. Liu et. al, "ParseNet: Looking Wider to See Better", preprint arXiv:1506.04579, 2015.
11. J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR, Nov 2015.
12. D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, Vol 60, Issue 2, pp 91-110, Nov 2004.
13. O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", MICCAI, 2015.
14. G.L. Oliveira, W. Burgard, and T. Brox, "Efficient Deep Methods for Monocular Road Segmentation", Intl. Conf. on Intelligent Robots and Systems, 2016.
15. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS, 2015.
16. M. Schwarz, H.Schulz, and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features", ICRA, 2015.
17. P. Sermanet et al.,"Overfeat: Integrated recognition, localization and detection using convolutional networks", arXiv preprint arXiv:1312.6229, 2013.
18. R. Socher et al., "Convolutional-Recursive Deep Learning for 3D Object Classification', NIPS 25, 2012.
19. K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv:1409.1556, 2014.
20. S. Thrun, M. Montemerlo1, H. Dahlkamp, et. al, "Stanley: The robot that won the DARPA Grand Challenge", JFR, Vol 23, Issue 9, pp 661-692, 2006.