Towards Robust Semantic Segmentation using Deep Fusion

Abhinav Valada, Gabriel L. Oliveira, Thomas Brox, and Wolfram Burgard
Department of Computer Science, University of Freiburg, Germany
{valada, oliveira, brox, burgard}@cs.uni-freiburg.de

***Abstract*—Robust semantic scene understanding of unstructured environments is critical for robots operating in the real world. Several inherent natural factors such as shadows, glare and snow make this problem highly challenging, especially using RGB images. In this paper, we propose the use of multispectral and multimodal images to increase robustness of segmentation in real-world outdoor environments. Deep Convolutional Neural Network (DCNN) architectures define the state of the art in various segmentation tasks. However, architectures that incorporate fusion have not been sufficiently explored. We introduce early and late fusion architectures for dense pixel-wise segmentation from RGB, Near-InfraRed (NIR) channels, and depth data. We identify data augmentation strategies that enable training of very deep fusion models using small datasets. We qualitatively and quantitatively evaluate our approach and show it exceeds several other state-of-the-art architectures. In addition, we present experimental results for segmentation under challenging real-world conditions. The dataset and demos are publicly available at http://deepscene.cs.uni-freiburg.de.**

I. 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 unstructured outdoor environments such as forests, 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 of them can be driven over [3]. However, the caveat is the reliable detection of chlorophyll using monocular cameras. This detection can be enhanced by additionally using the NIR wavelength $(0.7 - 1.1\mu m)$, 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.

Fusion of multiple modalities and spectra has not been sufficiently explored in the context of semantic segmentation. We can classify fusion strategies into early and late fusion approaches. Each of these approaches have their own benefits and drawbacks. Early fusion involves providing the network with multiple modalities from the beginning or from the initial layers so that the network learns the combined features all through. The major benefit of early fusion is minimal computational burden. Late fusion approaches have multiple streams of networks, one for each modality, followed by a series of fusion layers. The individual streams are first trained separately, followed by the training of the fusion layers to yield a combined prediction. This approach can potentially learn better complementary features along with fusion specific features but the early fusion technique can only learn the combined features from the beginning.

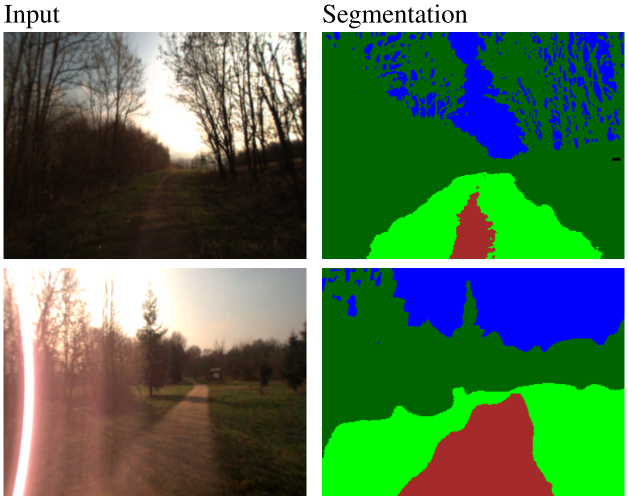


Fig. 1. Examples of segmentation in extreme lightning conditions using our deep-fused convolution approach. First row shows the segmentation in the presence of poor illumination and the second row shows a scene in presence of high saturation. In both cases, our multispectral fusion approach accurately segments the image.

In this paper, we address the problem of robust segmentation by leveraging deep up-convolutional neural networks and techniques developed in the field of photogrammetry using multispectral cameras to obtain robust pixel-accurate segmentation of the scene. We developed 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 dataset. We first evaluate the segmentation using our UpNet architecture, individually trained on various spectra and modalities contained in our dataset. We then identify the best performing modalities and fuse them using DCNN fusion architecture configurations. We show that the fused result outperforms segmentation using only RGB data in extreme outdoor conditions.

The rest of the paper is organized as follows. We first review the related work in Section II and describe our network architectures in Section III. We detail our data collection methodology in Section IV and the results from our experiments in Section V. Finally, in Section VI we report the conclusions and discuss future work.

II. Related Works

Recently, deep learning approaches have achieved state of the art performance in semantic segmentation. Such techniques perform segmentation on the whole image and are capable of end-to-end learning [13, 15, 12, 1]. Long *et al.* [13] proposed the so-called fully convolutional network (FCN) which is one of the first attempts that uses earlier layers in the hierarchy for refining the segmentation. FCNs do not require any pre or post-processing and allow the network to refine the coarse segmentation mask to the same resolution of the input. Oliveira *et al.* [15] applied a similar approach to human part segmentation and proposed improvements with regard to segmentation of occluded parts and over-fitting. Liu *et al.* [12] proposed a network called ParseNet which models global context directly, such approaches have also demonstrated near state-of-the-art results. Kendall *et al.* [1] proposed a variation of the FCN architecture geared towards increasing the performance. The main contribution resides on the use of pooling indices computed in the max-pooling step to perform upsampling. This approach eliminates the need for learning to upsample and reduces the systems memory requirement. However, approaches such as UpNet [15] and ParseNet [12] achieve superior performance.

Although there are numerous traditional learning approaches relating to recognition from RGB-D data, there are limited amount of DCNN approaches that have explored the use of multiple modalities or spectra. Eitel *et al.* [4] proposed a late fusion approach that employ two streams of DCNNs which are first individually trained to classify using a certain modality, and a stack of inner product layers are used in the end to fuse features from both networks. They fuse an RGB image and a colorized depth image for object recognition applications. In a similar approach [18], the authors use a pretrained two stream ImageNet network for object recognition from RGB-D images. Bo *et al.* [2] proposed an approach called hierarchical matching pursuit, which uses hierarchical spare coding to learn features from multimodal data. In [16], the authors use recurrent neural networks to combine convolutional filters for object classification from RGB-D data. A popular HHA encoding scheme was introduced in [7] where a CNN trained on RGB images is first used to extract features from depth data and the information is encoded into three channels. For each pixel they encode the height above the ground, the horizontal disparity and the pixelwise angle between a surface normal and the gravity.

In contrast to these multimodal object recognition approaches, we employ a late-fused convolution technique to learn highly discriminative features even after the fusion, for semantic segmentation. To the best of our knowledge this is the first work to explore both multimodal and multispectral images for end-to-end semantic segmentation.

III. Technical Approach

In this section, we first describe our base network architecture for segmenting unimodal images and then explore fusion architectures that learn 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, $Y_n = \{y_j, j = 1, \ldots, |X_n|\}, y_j \in \{0, C\}$ denotes the corresponding groundtruth 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 descent, we then solve

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

Our UpNet architecture has a similar form as that of the recently proposed fully convolutional neural networks [13, 15]. The architecture follows this general principle of being composed of two main components, a contraction segment and an expansion segment. Given an input image, the contraction segment generates a low resolution segmentation mask. We use the 13-layer VGG [19] architecture as basis for this contraction segment and initialize the layer parameters from the pretrained VGG network. The expansion segment consists of five upconvolutional refinement layers that refine the coarse segmentation masks generated by the contraction segment. Each upconvolutional refinement is composed of one up-sampling layer followed by a convolution layer. We add a rectified linear unit (ReLU) after each upconvolutional refinement. To avoid overfitting, we use dropout after the first and last refinement layers. The base UpNet architecture is shown in figure 2.

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. The architecture in [15] has a one-to-one mapping between the number of filters and classes in expansion segment. However, the recently proposed U-nets [17] architecture has demonstrated improved performance by having variable number of filters as in the contraction segment. 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.

We use a multi-stage training techinique to train our model. We use the Xavier [6] weight initialization for the convolution layers and a bilinear weight initialization for the upconvolution layers. We train our network with a initial learning rate $\lambda_0 = 10^{-9}$ and with the poly learning rate policy as

$$\lambda_n = \lambda_0 \times \left(\frac{1-N}{N_{max}}\right)^c, \tag{2}$$

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 (SGD) with a momentum of 0.9 for $300,000$ iterations for each refinement stage. 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 V. To provide a more informative and sharper segmentation, we introduce two strategies to make the network learn the integration of multiple spectra and modalities:

- *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 [13].
- *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 convolution, 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.

Our experiments provide an in-depth analysis of the advantages and disadvantages of each of these approaches in the context of semantic segmentation.

IV. Data Collection

Since existing datasets with RGB, NIR and depth data are not available, we extensively gathered data using our Viona autonomous mobile robot platform. The platform shown in figure 3 is equipped with a Bumblebee2 stereo vision camera and a modified dashcam with the NIR-cut filter removed for acquiring RGB and NIR images respectively. We use a Wratten 25A filter in the dashcam to capture the NIR wavelength in the blue and green channels. Both cameras are software time synchronized and frames were captured at 20Hz. In order to match the images captured by both cameras, we first compute SIFT [14] correspondences between the images using the Difference-of-Gaussian detector to provide similarity-invariance and 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) [5] 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.

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 dataset contains 325 images with pixel level groundtruth masks which were manually annotated. As there is an abundant presence of vegetation in our environment, we can 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 [9]. 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 as

$$NDVI = \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + \rho_{red}}, \tag{3}$$

where ρ_{nir} is the reflectance at the NIR wavelength $(0.7 - 1.1\mu m)$ and ρ_{red} is the reflectance at the red wavelength $(0.6 - 0.7\mu m)$. EVI can be computed as

$$EVI = G \times \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + (C_1 \times \rho_{red} - C_2 \times \rho_{blue}) + L}, \tag{4}$$

where ρ_{blue} is the reflectance at the blue wavelength $(0.45 - 0.52\mu m)$, G is the gain factor, L is a soil adjustment factor, C_1 and C_2 are coefficients used to correct for aerosol scattering in the red band by the use of the blue band.

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 [8] 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,* [11] that employs a deep convolutional neural field model for depth estimation by constructing unary and pairwise potentials of conditional random fields. Let an image x model the conditional probability of n superpixels of depth $y = [y_1, \ldots, y_n] \in \mathbb{R}^n$ by the

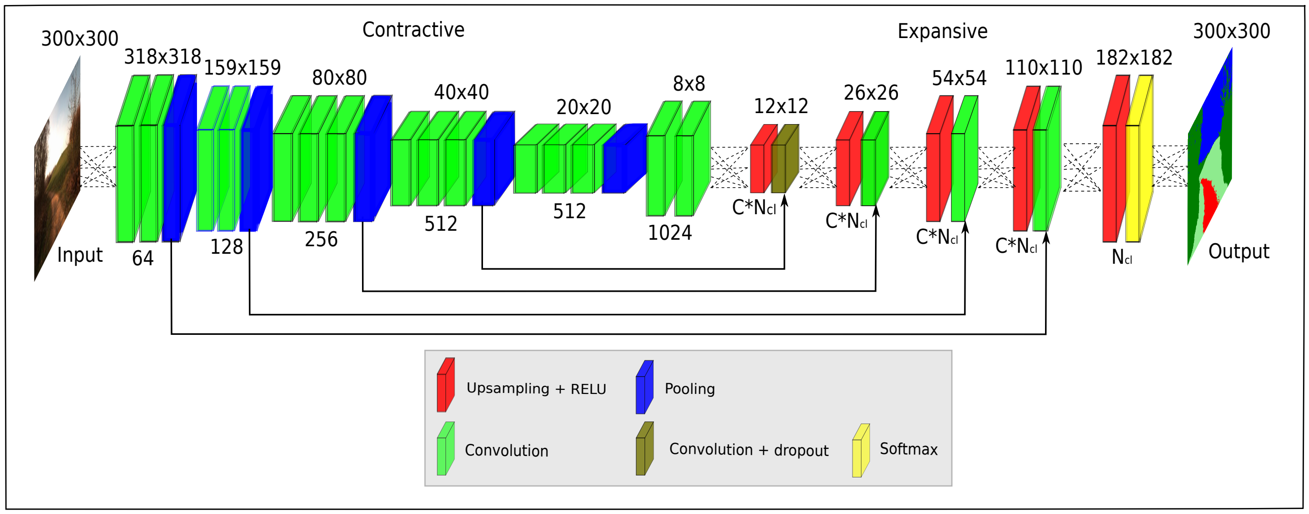


Fig. 2. Depiction of our UpNet architecture. Up-convolutional layers have size $C \times N_{cl}$, where N_{cl} is the number of classes and C is a scalar factor of filter augmentations. The contractive segment of the network contain convolution and pooling layers, while the expansive segment of the network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

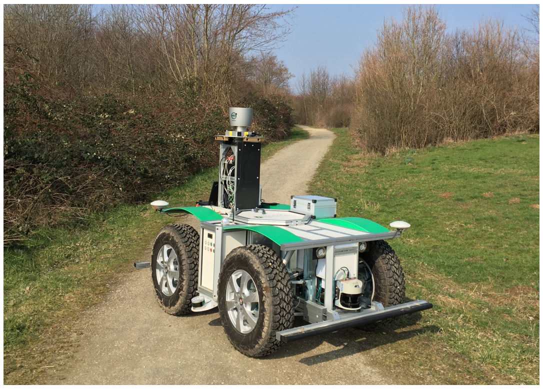


Fig. 3. The Viona autonomous mobile robot platform equipped with bumblebee stereo cameras and a modified dashcam with the NIR-cut filter removed for capturing RGB and NIR images respectively.

density function

$$Pr(y|x) = \frac{1}{Z(x)} exp(-E(y,x)), \tag{5}$$

where E is the energy function and Z is the partition function. The energy function is given as a combination of unary potentials V over the superpixels in $\mathcal{N}$ and edges $\mathcal{S}$ in x, i.e.,

$$E(x,y) = \sum_{p \in N} U(y_p, x) + \sum_{(p,q) \in S} V(y_p, y_q, x) \tag{6}$$

$$Z(s) = \int_y exp(-E(y,x))dy \tag{7}$$

A unified DCNN framework learns the value of U and V. The network is composed of a unary component, a pairwise component and CRF loss layer. The unary component is consists of a CNN that regresses depth values of superpixels, while the pairwise component outputs a vector containing the similarities for each of the neighbouring superpixels. The CRF loss layer minimizes the negative log-likelihood by taking the outputs of the unary and pairwise components. The depth of the new image is predicted by solving the maximum-a-posteriori inference problem, i.e.,

$$y^* = \underset{y}{\text{argmax}} Pr(y|x) \tag{8}$$

For our prediction we use the network pretrained on the Make3D dataset. Figure 4 shows some examples from our dataset from each spectrum and modality.

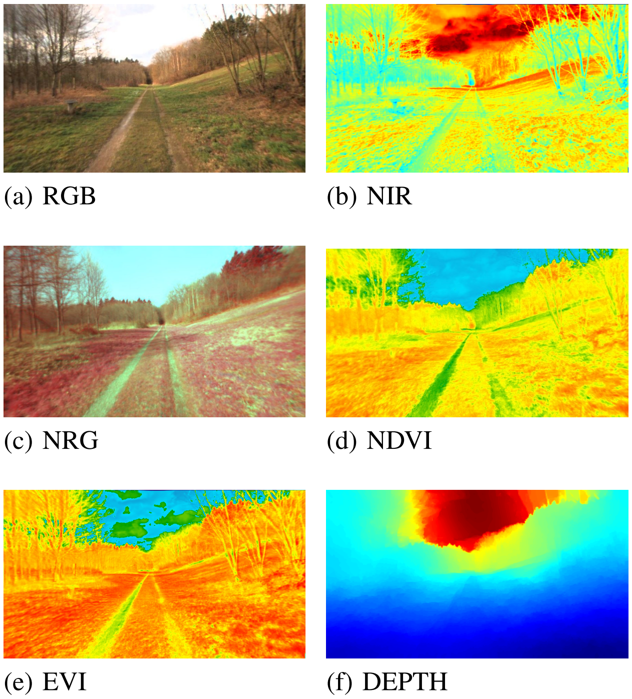
V. Experimental Results

We use the Caffe [10] deep learning framework for the DCNN implementation and ROS for capturing and synchronizing the images. Training our network on a NVIDIA Titan X GPU took about 7 days.

A. Baseline Comparison

To compare with the state-of-the-art, we train models using the *RGB RSC* set from our dataset 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 found the poly learning rate policy to converge much faster than fixed or step policy and yield a slight improvement in performance. The metrics shown in Table I correspond to Mean Intersection over Union (IoU), Mean Pixel Accuracy (PA), Precision (PRE), Recall (REC), False Positive Rate (FPR), False Negative Rate (FNR).

Figure 5 shows the forward pass time comparisons of our network to other state-of-the-art models. Our network



(a) RGB

(b) NIR

(c) NRG

(d) NDVI

(e) EVI

(f) DEPTH

Fig. 4. Sample images from our dataset showing various spectra and modalities. NIR refers to near-infrared, NDVI refers to normalized difference vegetation index, EVI refers to enhanced vegetation index, and NRG refers to a combination of near-infrared, red and green channels

TABLE I
PERFORMANCE OF OUR PROPOSED UNIMODAL MODEL IN COMPARISON TO THE STATE-OF-THE-ART

Baseline	IoU	PA	PRE	REC	FPR	FNR
FCN-8 [13]	77.46	90.95	87.38	85.97	10.32	12.12
SegNet [1]	74.81	88.47	84.63	86.39	13.53	11.65
ParseNet [12]	83.65	93.43	90.07	91.57	8.94	7.41
Ours Fixed lr	84.90	94.47	91.16	91.86	7.80	7.40
Ours Poly lr	**85.31**	**94.47**	**91.54**	**91.91**	**7.40**	**7.30**

has a run-time of almost twice as fast as the second best approach. Fast run-times are critical for outdoor navigation in unstructured environments as a slow perception system slows down the entire autonomy.

B. Parameter Estimation

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. Table II shows the results from these experiments. Data augmentation helps train very large networks on small datasets. However, on the present dataset it has a smaller impact on performance than on PASCAL VOC or human body part segmentation [15]. In our network, we replace the dropout in the VGG architecture with spatial dropout [21] 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.

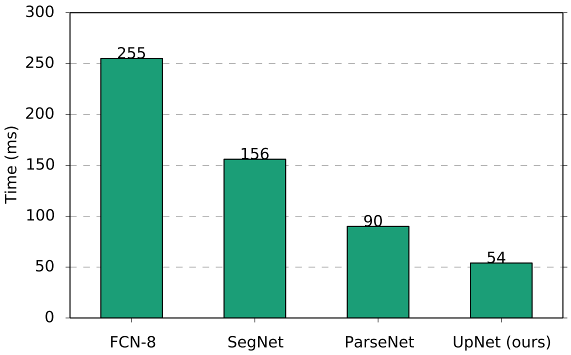


Fig. 5. Comparison of forward pass times with various state-of-the-art networks. Our network is significantly faster than other approaches.

TABLE II
COMPARISON ON THE EFFECTS OF AUGMENTATION ON OUR DATASET.

	Sky	Trail	Grass	Veg	Obst	IoU	PA
ParseNet	87.78	81.82	85.20	88.70	46.51	83.65	93.43
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

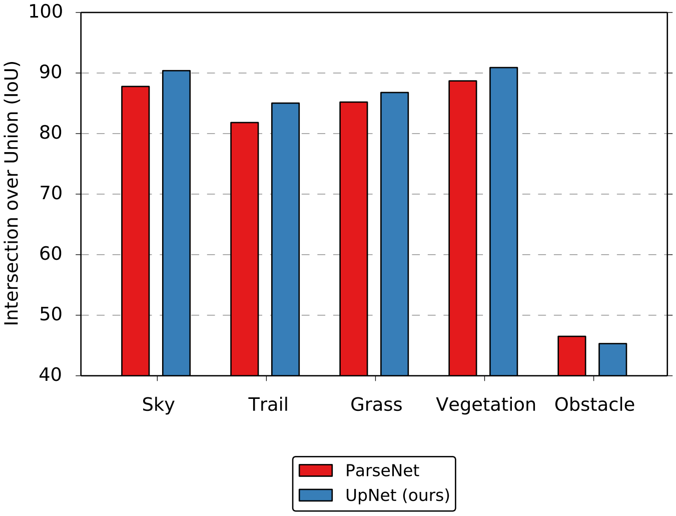


Fig. 6. Comparison of per-class intersection over union with the second best approach. Our network outperforms all the classes with the exception of Obstacle.

Scaling scales the image by a factor between 0.7 and 1.4. Rotation is applied with a up to 30 degrees range clockwise

TABLE III
COMPARISON OF DEEP MULTISPECTRUM AND MULTIMODAL FUSION APPROACHES. D, N, E REFER TO DEPTH, NIR AND EVI RESPECTIVELY. CS AND LFC REFER CHANNEL STACKING 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
3CS RGB-N-D	89.23	**85.86**	86.08	90.32	**61.68**	86.35	7.50	6.20
4CS RGB-N	89.64	83.37	85.83	**90.67**	59.85	85.79	**7.00**	7.20
5CS RGB-N-D	89.40	84.30	85.84	89.40	60.62	86.00	7.20	6.80
LFC RGB-N	90.67	83.31	86.19	90.30	58.82	85.94	7.50	6.56
LFC RGB-D	90.21	79.14	83.46	88.67	57.73	84.04	9.40	6.55
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

and anti-clockwise. Color augmentation is performed adding a value between -0.1 and 0.1 to the hue value channel of the HSV representation. Cropping provides C different crops, $C/2$ crops at the original image and $C/2$ crops with images horizontally flipped. The Skewing augmentation is calculated with a value ranging from 0 to 0.1. The final augmentation performs vignetting with a scale ranging from 0.1 and 0.6. Figure 6 shows the comparison of per-class intersection over union between ParseNet and our network. Our network outperforms ParseNet in all the classes other than the obstacle class.

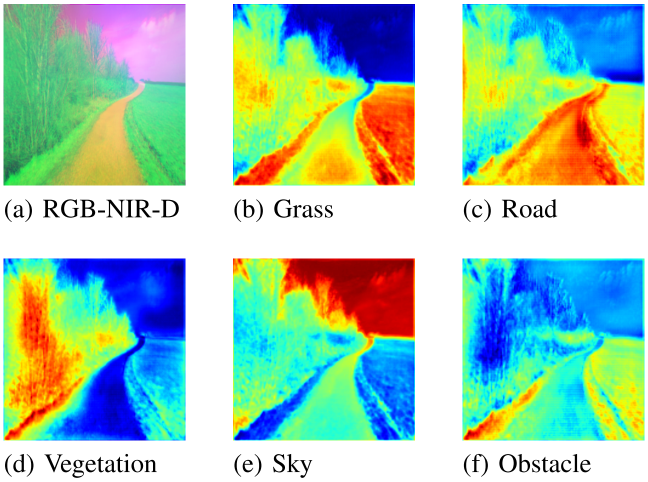

(a) RGB-NIR-D (b) Grass (c) Road (d) Vegetation (e) Sky (f) Obstacle

Fig. 7. Activation maps for various classes from the last layer of our channel stacked fusion network. Figure (a) shows our channel stacked input consisting of grayscaled RGB, NIR and DEPTH data. Figures (b), (c), (d), (e) and (f) show the activation maps for each of the classes in our dataset. High activations are shown in red and low activations as shown in blue.

C. Comparison of Fusion Approaches

In this section, we report results on segmentation using individual spectra and modalities, namely RGB, NIR, depth, and fusion with its combinations. 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 in Table III demonstrate the need for fusion. Multispectrum channel stacked 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 stacked 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.

Figure 7 shows the channel stacked input composed of grayscaled RGB, NIR and depth data and their corresponding activation maps. The maps represent the specific activations of the network for each class in the dataset. High activations are shown in red and the low activations are shown in blue. It can be seen that each of the maps have high activations for the specific class, also depicting the certainty in prediction. The channel stacked fusion architecture consumes 15ms more per forward pass when compared to their unimodal counterparts. The overall best performance was obtained with the late-fused-convolution of RGB and EVI, achieving a mean IoU of 86.9% and comparably high results in individual class IoUs as well. This approach also had the lowest false positive and false negative rates.

D. Qualitative Evaluation

We performed a series of stress testing experiments in a variety of weather conditions to evaluate the robustness of our approach in real-world environments. Specifically, we collected an additional dataset in a previously unseen place in low lighting, glare, shadows and snow. Figure 8 shows the segmented output from our network for various unimodal inputs. It can be seen that each of the spectra perform 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

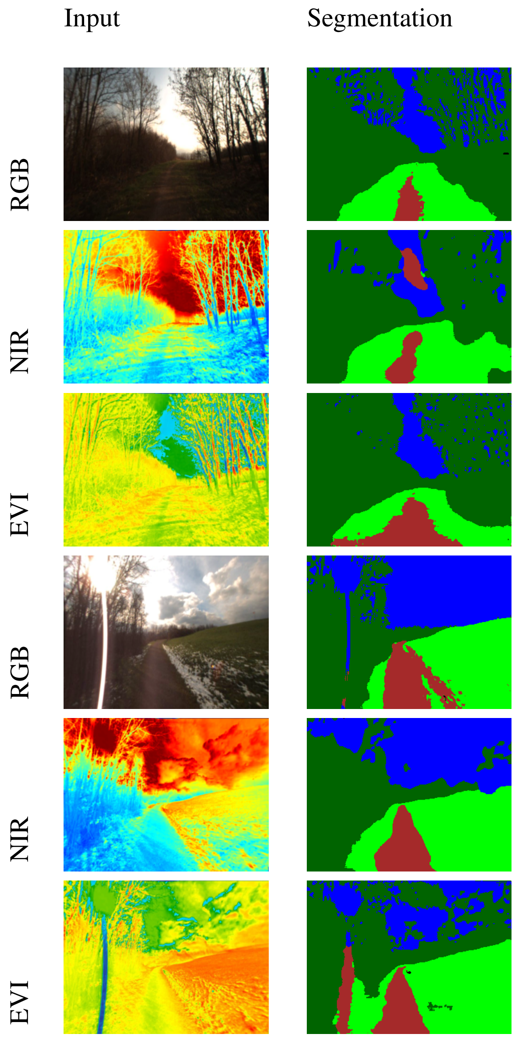


Fig. 8. Segmented examples from our dataset. Each spectrum provides valuable information. The first three rows shows the segmentation in highly shadowed areas and the last three rows show the performance in the presence of glare and snow. EVI refers to enhanced vegetation index and NIR refers to near-infrared.

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, especially when there are abundant clouds. This further demonstrates the need for fusion to increase the robustness in these extreme conditions.

Figure 9 shows some qualitative comparison between unimodal segmentation and the two deep fusion approaches. The figure also shows the channel stacked input of grayscaled RGB, NIR and depth which is fed into the channel stacked fusion network. For the late fused convolution results shown, we use RGB and EVI, which has yielded the best results in the quantitative evaluation experiments. It can be seen that the channel stacked input image already has a clear distinction among sky, trail and vegetation classes. However the segmentation from this network often has false positives between the vegetation and grass classes. It also has difficulty in recognizing the trail class when the path is too narrow. Figures 9 (a), (b) and (c) show this effect. On the other hand, our late fused convolution approach demonstrates the best results for all classes in this dataset as well. The channel stacked fusion shows less amount of detail in the vegetation class while compared to the segmentation from the unimodal RGB network and the late-fused convolution network. This clearly highlights the advantage of the late-fused convolution, as it fuses the feature maps further down the network, it is likely to make less mistakes. But in channel stacked fusion, if there is a discrepancy in the feature learned it cannot be corrected as the features are learned together from the beginning.

The unimodal RGB network shows a difficulty in estimating the boundaries of the vegetation and grass classes accurately. This can be seen in all the images shown in figure 9. The late-fused convolution has the most accurate estimation of the boundary for the trail class. Figure 9 (e) shows an example where the grass is covered with snow. It can be seen that the channel-stacked fusion still accurately segments grass even if it is covered with snow. The segmentation from the unimodal RGB network shows a very high amount of false positives in this case. The channel-stacked fusion on the other hand has difficulty in identifying grass when it is covered with snow. This demonstrates the advantage of fusing EVI with RGB as EVI is very accurate for segmenting vegetation and grass, while RGB yields the highest amount of detail in the ideal case. Figure 9 (f) shows another extreme case where the image is affected by glare from the sun. This is a very common scenario for robots operating in real-world outdoor environments. The results show that the channel-stacked fusion has no difficulty in segmentation in the presence of glare and saturation as EVI which is one of the modalities used for fusion is unaffected by these factors.

In addition, a live demo can be accessed at http://deepscene.cs.uni-freiburg.de, where a user can upload any image of an unstructured outdoor environment for segmentation or choose a random example.

VI. Conclusions

We presented a deep end-to-end architecture for semantic segmentation of outdoor environments. Our network outperforms several state-of-the-art architectures with near real-time performance. We extensively evaluated the benefits and drawbacks of 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. Furthermore, we qualitatively evaluated the benefits of multispectral fusion in extreme outdoor conditions. The results demonstrate our hypothesis of fusing the NIR wavelength with RGB to obtain robust segmentation in unstructured outdoor environments.

Future work will include extending our late-fused convolution network. Currently the network only has one convolution layer after the fusion, adding a pooling and upconvolution layer would introduce more invariance and discriminability

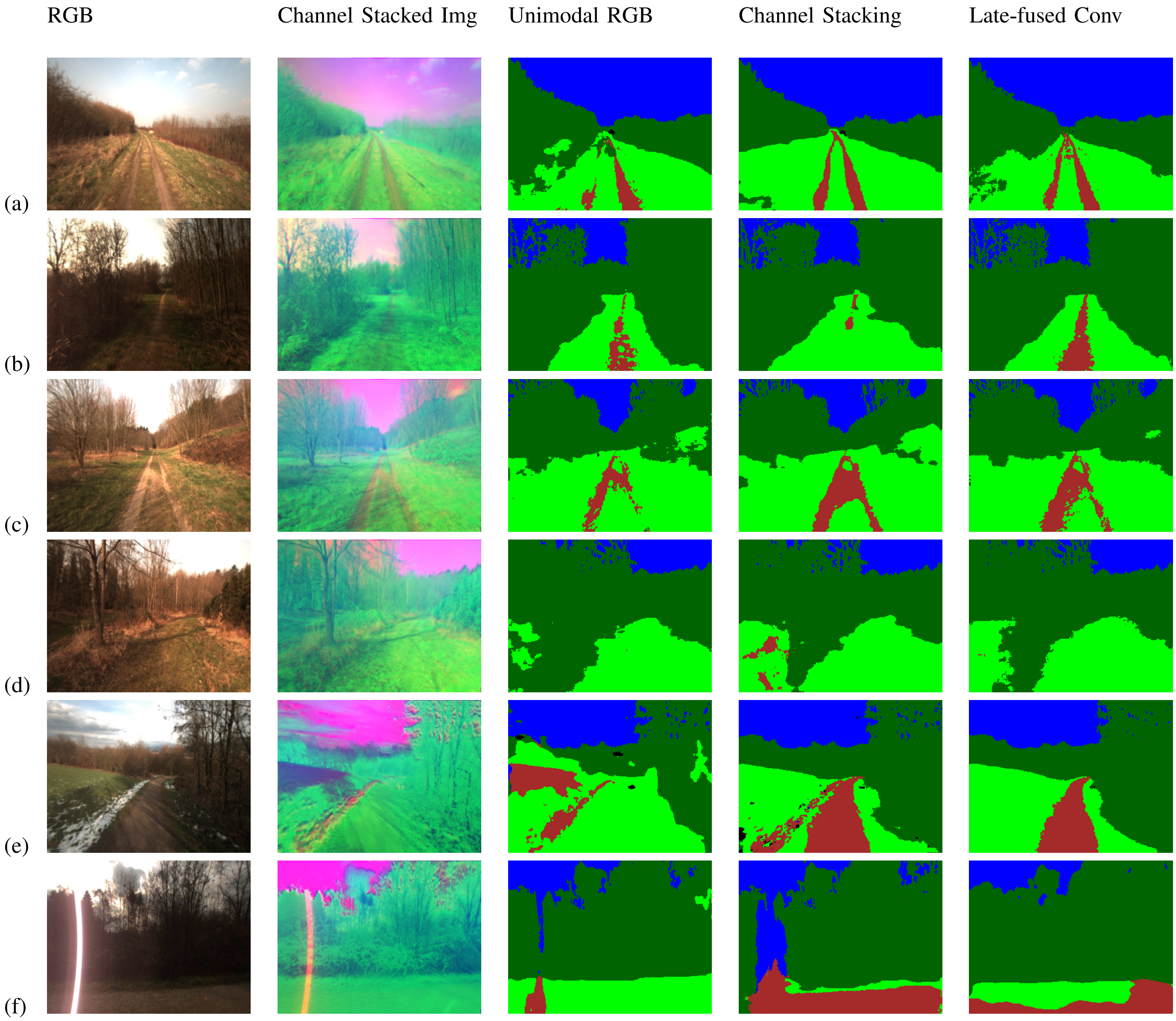


Fig. 9. Qualitative comparison of segmentation results between our unimodal deep network and our two fusion strategies. The late-fused convolution model surpasses the performance of others in almost all the cases. The last two rows show the results in extreme conditions of snow and glare. Our multispectral fusion model is still able to successfully segment under these conditions.

to the filters learned after the fusion. Recently adaptive fusion strategies demonstrated improved performance for fusing multiple modalities for detection tasks, however they have not been explored in the context of semantic segmentation. It would be of interest to evaluate such architectures in comparison to channel stacking and late-fused convolution.

Acknowledgments

This work has partly been supported by the European Commission under ERC-AGPE7-267686-LIFENAV and FP7-610603-EUROPA2.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv: 1511.00561*, 2015. URL http://arxiv.org/abs/1511.00561.

[2] L. Bo, X. Ren, and D. Fox. Unsupervised Feature Learning for RGB-D Based Object Recognition. In *ISER*, June 2012.

[3] David Bradley, Scott Thayer, Anthony (Tony) Stentz, and Peter Rander. Vegetation detection for mobile robot navigation. Technical Report CMU-RI-TR-04-12, Robotics Institute, Pittsburgh, PA, February 2004.

[4] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *Proc. of the IEEE Int. Conf. on Intelligent Robots and*

Systems (IROS), 2015.

[5] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981. ISSN 0001-0782. URL http://doi.acm.org/10.1145/358669.358692.

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International conference on artificial intelligence and statistics*, pages 249–256, 2010.

[7] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*, chapter Learning Rich Features from RGB-D Images for Object Detection and Segmentation, pages 345–360. 2014. URL http://dx.doi.org/10.1007/978-3-319-10584-0_23.

[8] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814, June 2005.

[9] Justice C.O. Huete, A. and W.J.D. van Leeuwen. Modis vegetation index (mod 13). *Algorithm Theoretical Basis Document (ATBD)*, Version 3.0:129, 1999.

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[11] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. *arXiv preprint arXiv: 1411.6387*, 2014. URL http://arxiv.org/abs/1411.6387.

[12] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv: 1506.04579*, 2015. URL http://arxiv.org/abs/1506.04579.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv: 1411.4038*, 2014. URL http://arxiv.org/abs/1411.4038.

[14] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. URL http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

[15] G. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. Deep learning for human part discovery in images. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016. URL http://lmb.informatik.uni-freiburg.de//Publications/2016/OB16a.

[16] Richard Socher and Brody Huval and Bharath Bhat and Christopher D. Manning and Andrew Y. Ng. Convolutional-Recursive Deep Learning for 3D Object Classification. In *Advances in Neural Information Processing Systems 25*. 2012.

[17] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. Technical Report 1505.04597, arXiv, 2015. URL http://lmb.informatik.uni-freiburg.de//Publications/2015/RFB15. arXiv preprint arXiv:1505.04597.

[18] Max Schwarz, Hannes Schulz, and Sven Behnke. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In *ICRA*, 2015.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*, 2014. URL http://arxiv.org/abs/1409.1556.

[20] Sebastian Thrun, Mike Montemerlo, Hendrik Dahlkamp, David Stavens, Andrei Aron, James Diebel, Philip Fong, John Gale, Morgan Halpenny, Gabriel Hoffmann, Kenny Lau, Celia Oakley, Mark Palatucci, Vaughan Pratt, Pascal Stang, Sven Strohband, Cedric Dupont, Lars-Erik Jendrossek, Christian Koelen, Charles Markey, Carlo Rummel, Joe van Niekerk, Eric Jensen, Philippe Alessandrini, Gary Bradski, Bob Davies, Scott Ettinger, Adrian Kaehler, Ara Nefian, and Pamela Mahoney. Stanley: The robot that won the darpa grand challenge. *Journal of Field Robotics*, 23(9):661–692, 2006. URL http://dx.doi.org/10.1002/rob.20147.

[21] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. *arXiv preprint arXiv:1411.4280*, 2014. URL http://arxiv.org/abs/1411.4280.