SEMANTIC VEHICLE SEGMENTATION IN VERY HIGH RESOLUTION MULTISPECTRAL AERIAL IMAGES USING DEEP NEURAL NETWORKS

Nina Merkle[*1], *Seyed Majid Azimi*[*1], *Sebastian Pless*[2], *Franz Kurz*[1]

[1]Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen
[2] Institute of Optical Sensor Systems, German Aerospace Center (DLR), Berlin-Adlershof

ABSTRACT

The fusion of complementary information from co-registered multi-modal image data enables a more detailed and more robust understanding of an image scene or specific objects, and is important for several applications in the field of remote sensing. In this paper, the benefits of combining RGB, near infrared (NIR) and thermal infrared (TIR) aerial images for the task of semantic vehicle segmentation through deep neural networks are investigated. Therefore, RGB, NIR and TIR image triplets acquired by the Modular Aerial Camera System (MACS) are precisely co-registered through the application of a virtual camera system and subsequently used for the training of different neural network architectures. Various experiments were conducted to investigate the influence of the different sensor characteristics and an early or late fusion within the network on the quality of the segmentation results.

Index Terms— Aerial Imagery, Data Fusion, Deep Learning, Multispectral Imagery, Vehicle Segmentation

1. INTRODUCTION

By providing a more detailed and more robust understanding of an image scene or specific objects, multi-modal image fusion has proven to be beneficial for a variety of remote sensing applications and is therefore often used for tasks such as land cover classification, change detection or urban surface modeling. For the specific task of vehicle segmentation or detection from aerial or satellite images, previous research studies such as [1, 2] mainly utilizes RGB images. The advantage of RGB images is a high spatial resolution, which enables a reliable and accurate segmentation and detection of vehicles. On the other hand, these images capture only a small region of the spectrum and are affected by varying illumination conditions and the appearance of shadowed areas.

In contrast, our deep learning based vehicle segmentation approach is based on a combined usage of RGB, near infrared (NIR) and thermal infrared (TIR) aerial images in order to overcome individual shortcomings of each sensor. More precisely, the usage of the NIR region or the spectrum enables a more efficient separation between vegetation and non-vegetation, e.g. for a better identification of vehicle partially parked under bushes or trees or on grass. Contrary to RGB and NIR sensors, TIR sensors measure the emitted radiation from the surface of a target and therefore provide special spectral characteristics of an object, usually with a lower spatial resolution. In the context of vehicle segmentation, these characteristics could help for a better identification of vehicle parked in shadow or partly covered areas. Figure 1 exemplifies the difference in visibility of vehicles in RGB, NIR and TIR images.

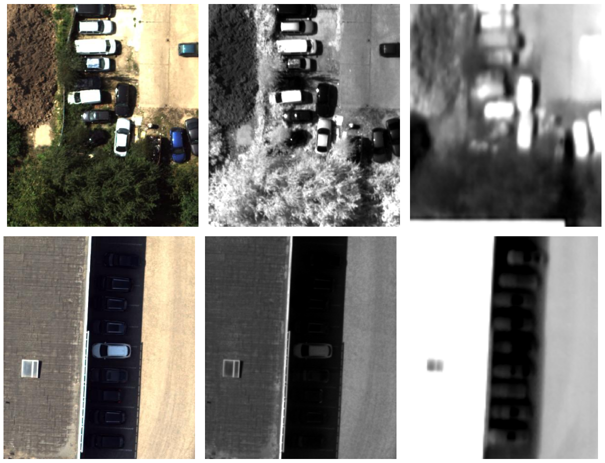


Fig. 1: Comparison of vehicle appearance and visibility in RGB, NIR and TIR images.

In the following a conceptional overview of the proposed method is provided in Section 2, which is followed by a detailed description of the utilized dataset in Section 3. The results achieved are presented and discussed in Section 4 and an outlook of future experimenters is provided in Section 5.

2. METHODOLOGY

2.1. Image Co-registration

In order to fuse the information from individual sensors exactly co-registered multi-channel images are required. The images used in this paper are captured by the Mod-

[*] denotes equal contribution

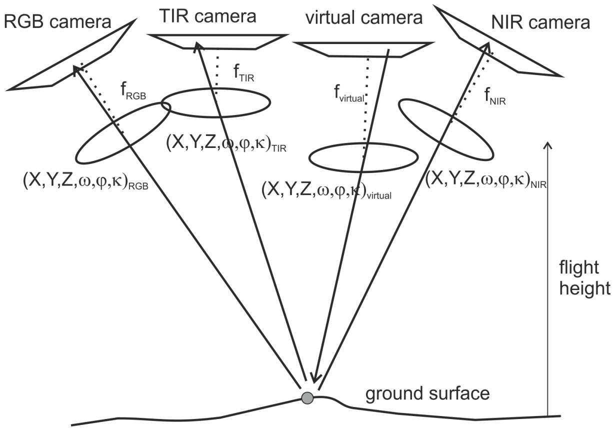


Fig. 2: Setup of the MACS virtual camera

ular Aerial Camera Systems (MACS), which acquires the visual (400-680nm), near (700-950nm) and thermal infrared (7.5-14.0μm) spectrum with three camera heads mounted on an airplane [3]. As the RGB, NIR and TIR images are taken from different camera heads with slightly different exterior orientation and strongly different interior camera properties the co-registration is a crucial step.

Commonly, the co-registration of images from multi-head camera systems is solved by generating one virtual image out of the single images, as demonstrated e.g. for the DMC camera [4] or for oblique UAV images [5]. Similar to the described approaches, one virtual camera for each acquisition time was created for the MACS camera system, into which the RGB, NIR and TIR camera images were projected. Before, a self-calibrating bundle adjustment was performed using ground control information from a reference data set and using the directly measured GNSS/IMU data, but without using any laboratory parameters or relative orientations. The bundle adjustment uses tie points matched between and within the RGB and NIR images as well as tie points only within the TIR images.

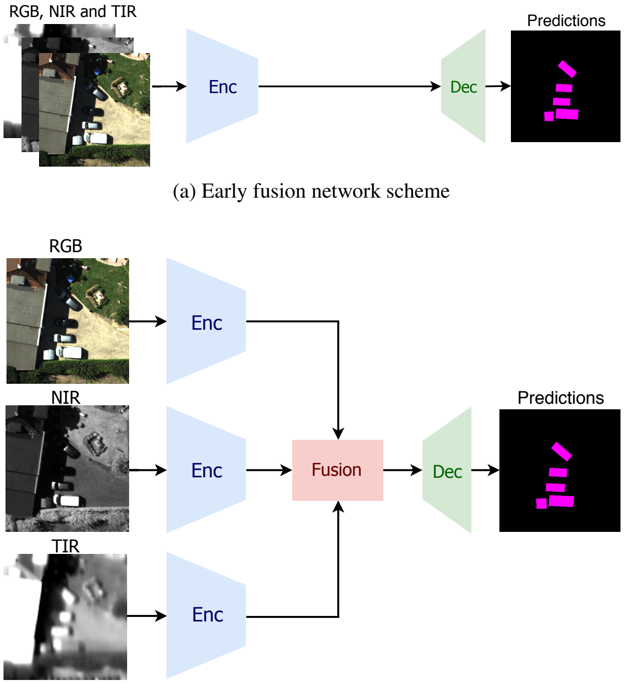
The setup for the virtual camera is illustrated in Fig. 2. Starting from a pixel in the virtual camera image, the corresponding pixel in the RGB, NIR and TIR image are assigned by intersection of the ray with a reference DSM and reprojecting the point into the RGB, NIR and TIR camera using the interior $(f, x_p, y_p, ...)$ and exterior $(X, Y, Z, \omega, \phi, \kappa)$ orientation parameter. Finally, the virtual camera image has five multispectral bands, three from RGB, one from NIR and one from TIR. The parameters of the virtual camera have been selected so that the field of view (FOV) is smaller than the FOV of every real camera and the GSD of the virtual camera reaches the GSD of the NIR camera. The exterior orientation of the virtual camera were taken from the TIR camera.

2.2. Semantic Vehicle Segmentation

The principle aim of our semantic segmentation network is to identify all vehicles in an image scenes by allocating each pixel to one of three classes: ”car”, ”truck” or ”non-vehicle”. To reach this goal our proposed network is composed of three consecutive parts: 1) one or more encoders to down-sample the input images for the extraction of more and more high-level features, 2) a fusion sub-network to combine the features extracted from the single encoder streams, and 3) a decoder to gradually up-sample and thereby restores the spatial properties of the predictions.

In order to determine the optimal moment for the data fusion, the influence of two fusion schemes is investigated. The first type is based on an early fusion, where the input images are stacked and directly passed to one common encoder. Note that in this case no fusion sub-network is used. The second type is based on a late fusion, where several encoder networks extract the features from the different input modalities independently. Afterwards, the extracted features are fused in a fusion sub-network. For both fusion schemes the same decoders were used. An overview of the conceptual fusion approaches is provided in Figure 3.

The building blocks of our en- and decoders are based on the fully convolutional neural network architecture (FCNs) proposed in [6]. In the case of the late fusion stream we investigated the influence of three fusion sub-networks: 1) simple concatenation of the encoders output, 2) concatenation followed by three convolutional layers and 3) concatenation followed by five convolutional layers.



(a) Early fusion network scheme

(b) Late fusion network scheme

Fig. 3: Illustration of the early and late fusion networks. Abbr.: Encoder (Enc), Decoder (Dec)

3. DATASET

For data acquisition, a 22km long section of the German motorway $A2$ between the cities of Hanover and Brunswick was imaged on $07/08/2017$ by the MACS camera system at a flight height of $350m$ above ground. 411 visual and near infrared images were captured synchronously with $1Hz$, whereas 3096 thermal infrared images were captured asynchronously with $7.5Hz$, which led to small time offsets and according to this position differences of moving vehicles in the co-registered multispectral images. Table 1 lists the most important parameters of the dataset including the original GSD of the MACS cameras and the GSD of the virtual camera at a flight height of $350\,m$. The TIR images finally are up-sampled to the GSD of the NIR images, whereas the original GSD of the TIR images is $22.5\,cm$.

	f $[mm]$	image size	FOV	pixel size $[\mu m]$	GSD $[cm]$
RGB	52.2	3232/4864	38°	7.4	5.0
NIR	59.2	2472/3296	34°	11.0	6.5
TIR	32.4	768/1024	35°	20.0	22.5
virt.	57.2	2304/3072	33°	11.0	6.7

Table 1: Parameters of the MACS camera setup. The GSD refers to a flight height of $350\,m$ above ground level.

For the training and the evaluation of our network, the acquired images where divided into two subsets. The first 251 images serve as training data from which 5% was considered as validation set, while 116 as testing set. The NIR and TIR images were normalized based on their minimum and maximum values. The images of both datasets show different landscapes, e.g. highways, parking lots, industrial and rural areas, and were acquired under varying illumination conditions.

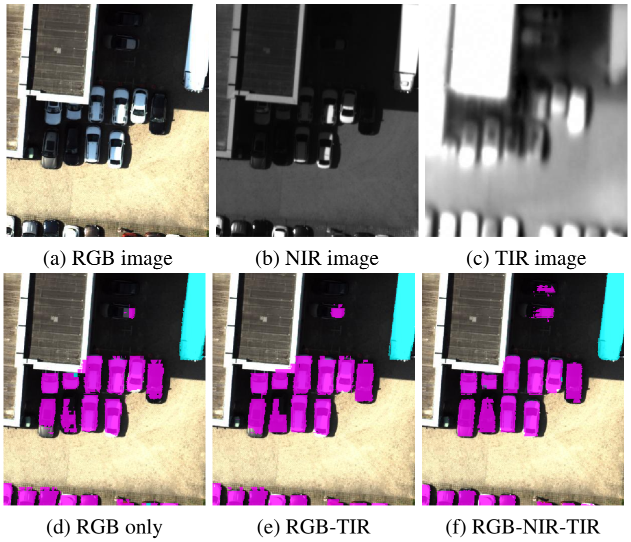
In order to investigate the influence of the different sensors characteristics on the segmentation results seven training sets are created out of the RGB, NIR and TIR training data: Training sets 1-3) contain images of one of the sensors only (RGB, NIR or TIR), 4)-6) contain pairs of images from different sensors (RGB&NIR, RGB&TIR or NIR&TIR), and 7) contains RGB, NIR and TIR image triplets.

4. EXPERIMENTS AND RESULTS

As the FCN-based en- and decoder, we choose the 8s variation with VGG-16 [7] base-network. The weights of the FCN encoders are initialized by pre-trained models (trained on the ImageNet ILSVRC 2012). Due to GPU memory limits, we crop each input image into $512 \times 512\,$px and ignore the patches that contain pixels belonging to vehicles less than threshold. We add horizontally flipped patches to the dataset. All experiments are carried out using NVIDIA Titan XP by training for 50 epochs using the Adam optimizer with 0.0001 learning rate. We use the Intersection over Union (IoU), recall and precision as the criteria to evaluate the different methods. Table 2 provides on overview of the performance of different configurations and Figure 5 a corresponding qualitative evaluation.

The results in Table 2 show that a late fusion of RGB, NIR and TIR achieves the best performance (81.05% mIoU) compared to the other setups with almost a 3% higher mIoU than the early fusion of the three image modalities. Using RGB data alone achieves even better performance compared to the three-image-modality early fusion. These observations indicate although different image modalities contain complementary data, by early fusion the similar features are extracted in higher layers. Using TIR without fusion achieves the worst results, which intuitively is because of the lower-resolution compared to RGB and to the time shift in image acquisition. As the ground truth was generated on the basis of the RGB images, the ground truth for the TIR images is less inaccurate. In the future, this could be addressed by secondary loss function and using an independent ground truth for TIR images. Interestingly, in both early and late fusion scenarios, the fusion of RGB and TIR achieves better performance than RGB and NIR fusion. This could be due to the fact that RGB and NIR contain more similar information than TIR. The TIR images on the other hand, add extra information which are not present in the RGB or NIR images.

The qualitative evaluation in Figure 4 and 5 supports the theoretical assumptions made about the benefits of fusing the three image modalities (see Section 1). More precisely, the provided samples show that fusing NIR and TIR with RGB images helps to improves the segmentation results especially in image areas with difficult illumination conditions such as shadowed or tree covered areas (see sample in 4).


(a) RGB image (b) NIR image (c) TIR image

(d) RGB only (e) RGB-TIR (f) RGB-NIR-TIR

Fig. 4: Qualitative influence of the input modularities (first row) on the segementation results (second row).

fusion scheme	input modularities	fusion sub-network	IoU [%] mean	car	truck	non-vehicle	average [%] recall	precision
early	RGB	-	78.51	70.63	56.64	99.54	87.00	86.18
early	NIR	-	75.87	62.39	57.96	99.45	84.66	85.06
early	TIR	-	54.75	22.93	36.19	99.03	62.20	68.58
early	RGB-NIR	-	77.97	69.09	57.12	99.51	88.26	84.47
early	RGB-TIR	-	78.67	70.59	57.52	99.54	88.07	85.30
early	NIR-TIR	-	77.31	67.41	56.34	99.50	85.60	85.99
early	RGB-NIR-TIR	-	78.18	68.79	58.07	99.49	87.88	84.98
late	RGB-NIR	concat + 3 conv layers	79.60	72.24	58.63	99.52	88.76	86.18
late	RGB-TIR	concat + 5 conv layers	80.13	74.96	57.23	99.55	**89.47**	86.16
late	NIR-TIR	concat + 3 conv layers	77.47	67.00	57.64	99.48	85.97	85.33
late	RGB-NIR-TIR	concat	**81.05**	**75.40**	**59.37**	**99.59**	89.46	**87.22**

Table 2: Quantitative comparison of the early and late fusion scheme, the different image modalities and the different fusion sub-networks. Abbr.: concatenation (concat), convolutional (conv).

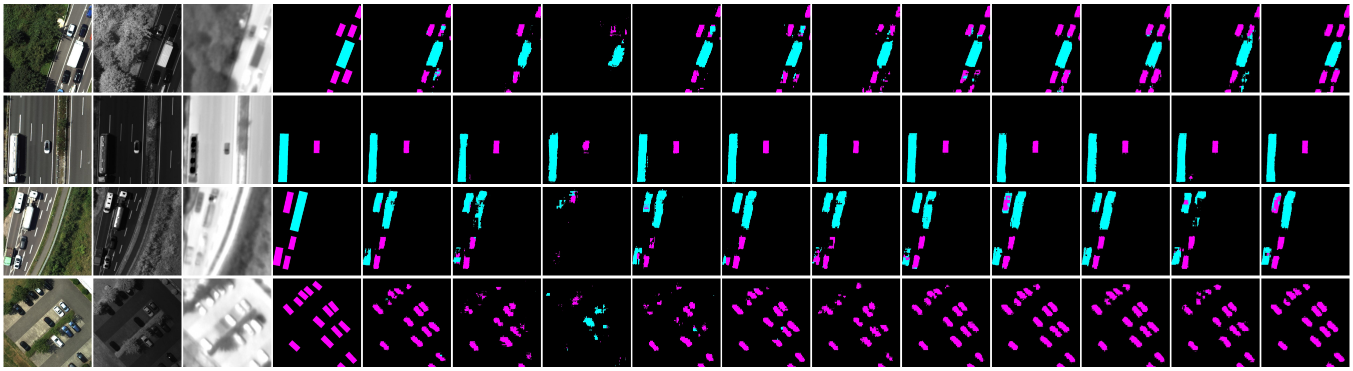


Fig. 5: Qualitative performance of different configuration trained on RGB, NIR and TIR data or a subset of them. The predicted classes ”car”, ”truck” and ”non-vehicle” are marked in cyan, pink and black, respectively. The first four columns display RGB, NIR and TIR patches as well as the ground truth. After that, the order of the predictions is the same as in Table 2 (from top to bottom).

5. CONCLUSION

In this paper, we investigated the benefits of fusing RGB, NIR, and TIR images through different fusion strategies for the segmentation of multi-class vehicles in aerial imagery. As a first step, we proposed the usage of a virtual camera to alleviate the positing shift of the individual camera heads. One the basis of the resulting co-registered image set, we created a pixel-wise labeled dataset, which we used to investigated the performances of each sensor individually as well as the different fusion approaches in a fully convolutional neural network. The results show that fusing the images in a later stage leads the a better performance than an early fusion. Moreover, the combination of RGB-TIR achieves higher accuracy than RGB-NIR to the complementary information in TIR. In the future, we are planing to create a more diverse and larger dataset and to investigate more complex and specified fusion strategies as well as different en- and decoder approaches.

6. REFERENCES

[1] N. Ammour, H. Alhichri, Y. Bazi, B. Benjdira, N. Alajlan, and M. Zuair, “Deep Learning Approach for Car Detection in UAV Imagery,” *Remote Sensing*, vol. 9, no. 4, 2017.

[2] S. Azimi, E. Vig, F. Kurz, and P. Reinartz, “Segment-and-Count: Vehicle Counting in Aerial Imagery Using Atrous Convolutional Neural Networks,” *ISPRS - Inter. Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 19–23, 2018.

[3] J. Brauchle, S. Bayer, D. Hein, R. Berger, and S. Pless, “MACS-Mar: a real-time remote sensing system for maritime security applications,” *CEAS Space Journal*, vol. 11, no. 1, pp. 35–44, 2019.

[4] C. Doerstel, W. Zeitler, and K. Jacobsen, “Geometric Calibration of the DMC: Method and Results,” *Inter. Archives of Photogrammetry and Remote Sensing*, vol. 34, pp. 324–333, 2002.

[5] A. Tommaselli, M. Galo, M.and de Moraes, J. Marcato, and Rodrigo F. Caldeira, C.and Lopes, “Generating Virtual Images from Oblique Frames,” *Remote Sensing*, vol. 5, no. 4, pp. 1875–1893, 2013.

[6] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[7] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks For Large-Scale Image Recognition,” *ICRL*, 2015.