TAS-NIR: A VIS+NIR Dataset for Fine-grained Semantic Segmentation in Unstructured Outdoor Environments

Peter Mortimer
Universität der Bundeswehr München
Institute for Autonomous Systems Technology
peter.mortimer@unibw.de

Hans-Joachim Wünsche
Universität der Bundeswehr München
Institute for Autonomous Systems Technology
joe.wuensche@unibw.de

***Abstract*— Vegetation Indices based on paired images of the visible color spectrum (VIS) and near infrared spectrum (NIR) have been widely used in remote sensing applications. These vegetation indices are extended for their application in autonomous driving in unstructured outdoor environments. In this domain we can combine traditional vegetation indices like the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) with Convolutional Neural Networks (CNNs) pre-trained on available VIS datasets. By laying a focus on learning calibrated CNN outputs, we can provide an approach to fuse known hand-crafted image features with CNN predictions for different domains as well. The method is evaluated on a VIS+NIR dataset of semantically annotated images in unstructured outdoor environments. The dataset is available at mucar3.de/iros2022-ppniv-tas-nir.**

I. Introduction

Most computer vision approaches to autonomous driving consider the visible spectrum (VIS) during image acquisition. This led to many large-scale semantic segmentation datasets for urban driving scenarios. Only few annotated datasets have been released for unstructured outdoor driving scenarios [1], [2], [3], [4] and even fewer consider image data beyond the visible spectrum [5].

Foliage has a high reflectivity in the near infrared (NIR) spectrum, which is also known as the Wood effect [6]. Partly based on this observation, a few vegetation indices using visible light and near infrared light (VIS+NIR) were developed for Remote Sensing. This motivates the use of the VIS+NIR spectrum for vegetation and ground surface segmentation in autonomous driving in unstructured outdoor environments. The lack of large training datasets of VIS+NIR image pairs prevent end-to-end deep learning approaches, especially for a fine-grained semantic segmentation.

To alleviate the imbalance of semantically segmented VIS images and NIR images, we suggest a late fusion of predictions made by neural networks pre-trained on larger VIS datasets with predictions coming from hand-crafted vegetation indices based on smaller VIS+NIR datasets. Additionally, the VIS image prediction should produce calibrated outputs to more accurately resemble the confidence of a given class prediction. The performance improvement of our approach will be quantitatively compared with previous approaches on a fine-grained semantic segmentation VIS+NIR dataset which we release as part of this publication. We make the following contributions in this paper:

- A method for the late fusion of calibrated neural network predictions on VIS images and hand-crafted VIS+NIR features like NDVI and EVI. We apply this for the fine-grained semantic segmentation of ground surfaces and vegetation types in unstructured outdoor environments.
- A novel dataset consisting of 209 semantically segmented and aligned VIS+NIR images in different driv-

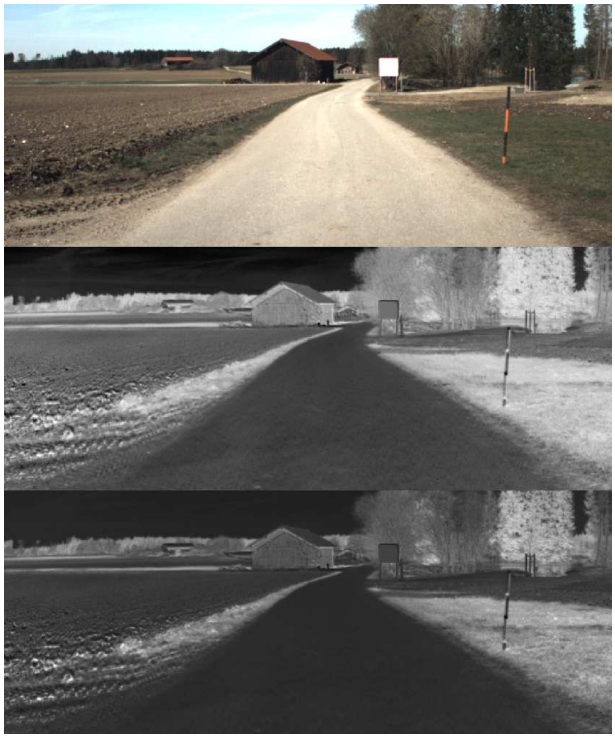

Fig. 1: For the fine-grained semantic segmentation of ground surfaces and vegetation we leverage both calibrated neural networks using images from the visible spectrum (top) and hand-crafted vegetation indices like the Normalized Difference Vegetation Index (middle) and the Enhanced Vegetation Index (bottom).

The authors gratefully acknowledge funding by the Federal Office of Bundeswehr Equipment, Information Technology and In-Service Support (BAAINBw).

(a) The ROI from the VIS image. (b) The ROI from the NIR image. (c) VIS+NIR after homography transform $H_{\text{NIR}\to\text{VIS}}$.

Fig. 2: The overlapping regions of interest (ROI) are matched, and a homography matrix $H_{\text{NIR}\to\text{VIS}}$ is applied onto the NIR image to also match the perspective of the VIS image (see Eq. (1)). The perspective transform based on a ground surface homography can lead to pixel mismatches especially along the border of close obstacles like the pole in Fig. 2c.

ing scenarios in unstructured outdoor environments. The fine-grained semantic segmentation of the different vegetation and ground surface types allows closer analysis of VIS+NIR based features.

II. Related Work

A. NIR in Computer Vision

NIR images are used in security applications, because NIR images can reveal content similar to VIS images while not requiring a light source visible to humans [7]. The difference of radiance between the NIR spectrum and the visible red color spectrum has been used to detect vegetation in remote sensing applications [8]. This observation formed the basis for vegetation indices such as NDVI [9] and EVI [10]. The common use of the VIS+NIR spectrum for vegetation detection in remote sensing motivated the use of VIS+NIR imagery for ground-level robotics in forested environments [11]. The NIR spectrum has also been shown to enhance VIS imagery by dehazing and providing higher contrast along object boundaries [12], [13]. In our work, we present the application of NIR imagery for autonotmous driving in unstructured outdoor environments.

B. VIS+NIR Semantic Segmentation

A semantic segmentation using a conditional random field (CRF) of VIS+NIR scenery images has shown an improvement in prediction accuracy over VIS images for semantic classes like *sky*, *vegetation* and *water*, whose response in the NIR domain is discriminant [14]. For unmanned ground vehicles (UGVs), neural network architectures training on multimodal images of forested environments have been developed. These relied on multiple independent expert networks trained on the different image modalities and included depth images as well [5]. This is extended by self-supervised fusion mechanisms trained to adapt to the spatial location and object class in the image [15]. The more intricate fusion strategies were deemed necessary compared to a channel-wise stacking of the input modalities, where networks did not learn to leverage complementary features and cross-modal interdependencies [16]. These works segment the scene using very broad object classes like *grass* and *tree* for vegetation types and *soil* and *road* for surface types. In our work, we make use of the NIR domain for a more fine-grained segmentation of the vegetation and surface types found in unstructured outdoor environments.

C. Calibrated Neural Network Predictions

Modern neural networks for image classification have shown to be overconfident in their prediction probability estimates, when compared to their true correctness likelihood. To obtain more accurate confidence scores, a so-called temperature scaling optimized over the negative log-likelihood (NLL) on the validation set can be added as a post-processing after the network's prediction [17]. Other approaches use network ensembles [18] or Monte Carlo dropout [19] at test time to estimate predictive uncertainty. The temperature scaling, which represents a single-parameter variant of Platt Scaling [20], was extended for semantic segmentation tasks by generating a temperature map in local temperature scaling (LTS) [21]. LTS incorporates an image and location dependent temperature map to account for different miscalibration effects, such as accurate predictions for object interiors and ambiguities at near-boundary locations in semantic maps. Our approach relies on LTS to produce calibrated prediction outputs for our probabilistic model.

III. Dataset

We propose the novel TAS-NIR dataset to investigate the relationship of VIS+NIR images for a fine-grained vegetation and ground surface segmentation. The TAS-NIR dataset is unique due to its fine-grained semantic segmentation of the driving scenes in unstructured outdoor environments and the inclusion of NIR images in addition to the VIS images.

The dataset was recorded using our survey vehicle while driving in different unstructured outdoor environments during spring, summer and autumn. The VIS images and NIR images were recorded with two cameras mounted on a camera platform [24]. The NIR camera is mounted directly under the VIS camera, and both cameras share the same orientation. The region of interest, where both the VIS and NIR images overlap, is shown in Fig. 2. To record the NIR image we use a Basler acA1300-60gmNIR with a built-in EV76C661 CMOS sensor. A longpass filter is attached onto the lens of the NIR camera to prevent any light under $765\,\text{nm}$ to pass through the lens.

We transform the perspective of the NIR image to match that of the VIS image by applying a homography matrix $H_{\text{NIR}\to\text{VIS}}$. The homography matrix $H_{\text{NIR}\to\text{VIS}}$ is constructed assuming a flat ground plane visible in both cameras and knowing the height of the cameras over the ground plane. The homogra-

Dataset	No. Scenes	Resolution	Scene Type	Annotation Type
EPFL RGB-NIR Scene Dataset [22]	477	1024 × 768 px	Indoor & Outdoor Photography	Image Classification
EPFL Semantic Segmentation Dataset [14]	370†	1024 × 768 px	Outdoor Photography	Semantic Segmentation
HyKo2 [23]	78‡	214 × 417 px	Outdoor Driving	Semantic Segmentation
Freiburg Forest [5]	366	1024 × 768 px	Outdoor Driving	Semantic Segmentation
TAS-NIR (ours)	209	1200 × 480 px	Outdoor Driving	Fine-grained Semantic Segmentation

TABLE I: A comparison of known VIS+NIR datasets in terms of their size, scene type and annotation type. †: The outdoor scenes from the EPFL Semantic Segmentation Dataset are only compared here. ‡: Only scenes taken with the MQ022HG NIR camera are considered in the HyKo2 dataset.

phy is only an approximation of the true geometry.

$$H_{\text{NIR}\to\text{VIS}} = K_{\text{VIS}} \left(R_{\text{NIR}} + \frac{t_{\text{NIR}\to\text{VIS}}\ n_{\text{NIR}}^T}{d_{\text{NIR}}} \right) K_{\text{NIR}}^{-1} \tag{1}$$

Fig. 3 shows the camera setup in the vehicle. The TAS-NIR dataset consists of 209 VIS+NIR image pairs with a fine-grained semantic segmentation. Similar to the WildDash 2 benchmark [25], the TAS-NIR dataset does not provide enough images to train algorithms by itself and should primarily be used for validation and testing. The available data from the TAS500 dataset [3] uses the same labeling policy, but does not provide a NIR image. The TAS500 dataset is therefore used to train a semantic segmentation model solely on VIS images. In Tab. I, we compare the TAS-NIR dataset with other available VIS+NIR datasets.

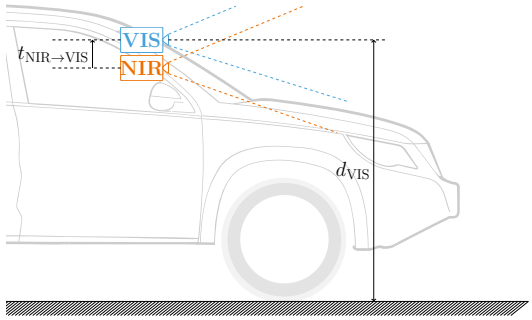

Fig. 3: The NIR camera is positioned directly under the VIS camera. The shared orientation and the known height of the vehicle over a flat ground surface allows the construction of a homography matrix $H_{\text{NIR}\to\text{VIS}}$.

IV. Methodology

A. Local Temperature Scaling

Neural networks for semantic segmentation tasks make use of the softmax function σ_{SM} on the network logits $\mathbf{z}(x)$ of each output pixel location x to produce the confidence score $\hat{p}(x)$. The network logits $\mathbf{z}(x)$ is a vector of length k with a logit for each possible semantic class.

$$\sigma_{\text{SM}}(\mathbf{z}(x))^{(k)} = \frac{\exp(z(x)^{(k)})}{\sum_{j=1}^{K} \exp(z(x)^{(j)})} \tag{2}$$

Traditionally, the confidence score was taken to be the largest activation from the softmax function.

$$\hat{p}(x) = \max_k \sigma_{\text{SM}}(\mathbf{z}(x))^{(k)} \tag{3}$$

Neural networks, especially recent models trained with batch normalization, have shown to produce overconfident output probabilities during later stages of training [17], [26]. This effect is less of a problem for applications, where the network output is the final stage of the perception process. For our probabilistic model, where we fuse the network output probability of a network based solely on VIS images with hand-crafted vegetation indices from VIS+NIR image pairs, we want the network output probability to resemble the empirically observable segmentation errors. The process of adjusting the network output probabilities is called calibration.

The temperature concept, which applies a Platt scaling with a single scale parameter $T > 0$ for all classes can be used to calibrate the network outputs [17]. The temperature T can raise the output entropy of the network by softening the softmax function for $T > 1$. T is optimized with respect to the NLL on the validation set to produce a calibrated confidence value $\hat{q}$.

$$\hat{q}(x) = \max_k \sigma_{\text{SM}}(\mathbf{z}(x)/T)^{(k)} \tag{4}$$

Temperature scaling was originally intended for image classification tasks, where the prediction on the input image corresponds to a network output vector of logits.

For semantic segmentation tasks, where the network produces output logits $\mathbf{z}(\mathbf{x})$ for each pixel location x in the input image, the temperature concept was extended in Local Temperature Scaling (LTS) [21]. The calibration is now extended from calibrated predictions $\hat{q}(x)$ being adjusted by the same scalar temperature value T to an image n and location x dependent probability map $\hat{\mathbb{Q}}_n(x, T_n(x))$ and temperature map $T_n(x)$.

$$\hat{\mathbb{Q}}_n(x, T_n(x)) = \max_{k\in K} \sigma_{\text{SM}}\big(\mathbf{z}_n(x)/T_n(x)\big)^{(k)} \tag{5}$$

where $T_n(x) \in \mathbb{R}^+$ is image and location dependent.

The image and location dependence is modeled by a small neural network $\mathscr{H}$ which takes the input image I and network output logits map $\mathbf{Z}$. So $\mathscr{H}$ learns the mapping $(I, \mathbf{Z}) \to \hat{\mathbb{Q}}$. $\mathscr{H}$ uses a tree-like convolutional neural network and optimizes over the NLL of the x pixel locations on the N_{val} images in the hold-out validation dataset [27], [28].

$$\theta^* = \arg\min_\theta - \sum_{n=1}^{N_{\text{val}}} \sum_{x\in\Omega} \log\left(\sigma_{\text{SM}}\Big(\frac{\mathbf{z}_n(x)}{\mathscr{H}(\theta, \mathbf{Z}_n, I_n, x)} \Big) \right) \tag{6}$$

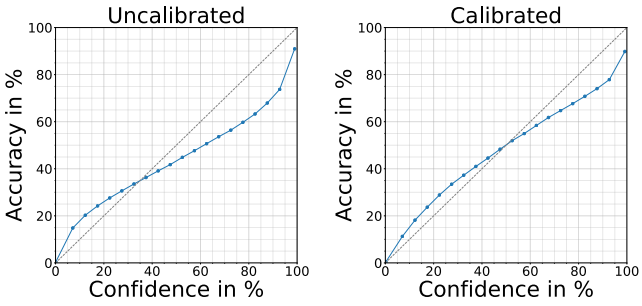

Fig. 4: The reliability diagrams [29] for an uncalibrated and calibrated DeepLabV3+ network on the TAS-NIR dataset test split. A reliability diagram can give insight on the calibration of a predictive model. Here we compare the confidence $\hat{q}(x)$ scores for each pixel with the accuracy reported when comparing it to the ground truth of the test split. A well calibrated model would have its confidence values close to its eventual accuracy (e.g. only 40% of all pixels with a confidence score of 40% are correctly classified). The ideal calibration is signified by the gray dashed line.

$$s.t. \quad \mathscr{H}(\theta, \mathbf{Z}_n, I_n, x) > 0$$

Practically speaking, LTS adds an additional post-processing stage to the overall training process, where we optimize the weights of the temperature mapping network $\mathscr{H}$. This is alleviated by the few parameters in $\mathscr{H}$ and the relatively small size of the hold-out validation dataset. The effects of network output calibration can be observed in the reliability diagrams in Fig. 4.

B. Vegetation Indices

The appearance of spatial content in NIR imagery differs in a few ways from their appearance in the visible spectrum (see Fig. 5). For vegetation in particular, where the chlorophyll in the vegetation appears transparent in the NIR spectrum, allowing the light in the NIR spectrum to reflect on the water contained in the vegetation [30]. The intensity of the NIR reflection depends on the season and the type of plants, which is why vegetation indices tend to combine the NIR response with other image properties from the VIS spectrum. For instance, the common Normalized Difference Vegetation Index (NDVI) metric makes use of the low reflectivity of vegetation in the red channel $\mathrm{R_{VIS}}$ of the debayered VIS image in combination with the high reflectivity in the NIR image.

$$\mathrm{NDVI} = \begin{cases} \dfrac{\mathrm{NIR} - \mathrm{R_{VIS}}}{\mathrm{NIR} + \mathrm{R_{VIS}}} & \text{otw.} \\ 0 & \text{if } \mathrm{NIR} = \mathrm{R_{VIS}} = 0 \end{cases} \tag{7}$$

Both the NIR and VIS image require the same bit depth for the NDVI calculation. The result NDVI value is in $[-1, 1]$ range. We can observe that NDVI ≥ 0 for pixels, where the reflectance in the NIR spectrum is higher than the reflectance in the R channel and vice versa. As observed in remote sensing applications, this translates to negative NDVI values for bodies of water and NDVI values close to zero for surface types like rocks, sands and concrete surfaces. Positive NDVI values come up for vegetation like crops, shrubs, grasses and forests [31].

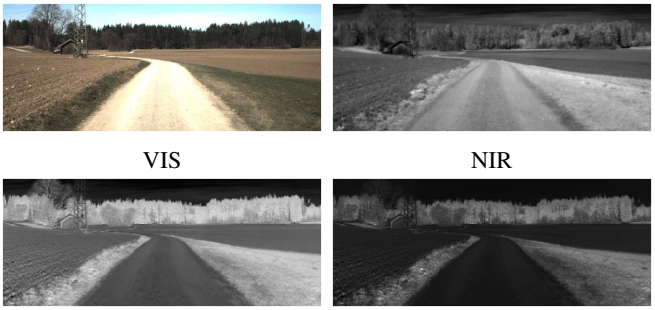

Fig. 5: Both vegetation indices NDVI and EVI make use of the high reflectivity of vegetation in the near infrared spectrum and the low reflectivity of vegetation in the visible red spectrum.

In our experiments, we will also take a closer look at the Enhanced Vegetation Index. The EVI was introduced to compensate soil and atmospheric effects and incorporates the light from the visible blue channel $\mathrm{B_{VIS}}$ as well. We had troubles finding a consistent formula of the EVI index, we therefore rely on the following formula [10].

$$\mathrm{EVI} = \begin{cases} \dfrac{2 \cdot (\mathrm{NIR} - \mathrm{R_{VIS}})}{\mathrm{NIR} + C_1 \cdot \mathrm{R_{VIS}} + C_2 \cdot \mathrm{B_{VIS}}} & \text{otw.} \\ 0 & \text{if } \mathrm{NIR} = \mathrm{R_{VIS}} = \mathrm{B_{VIS}} = 0 \end{cases} \tag{8}$$

$$\text{where } C_1 = 6.0 \text{ and } C_2 = 7.5$$

The EVI index is in $[-\frac{2}{C_1}, 2]$ range. The weights C_1 and C_2 adjust the use of the blue channel in aerosol correction of the red channel. The mentioned atmospheric effects do not relate to the image acquisition process from the ground with a camera system, but we will use the Eq. (8) for the sake of consistency. In Fig. 5 we present the NDVI and EVI image for a scene from the TAS-NIR dataset.

C. Late-fusion of VIS+NIR Predictions

Many instances of the same semantic class share similar values on both vegetation indices. The vegetation classes with a similar visual appearance in the VIS image can be discriminated among the semantic class in the NIR image. A lack of semantically segmented VIS+NIR image pairs prevents us from training a multimodal network for this domain. We therefore suggest adding a post-processing stage to the VIS-only neural network output by calibrating its outputs and adjusting the predictions based on the vegetation index values for each pixel position of the prediction. The prediction using the vegetation index is based on the accumulated image histograms of each semantic class across the validation set. The image histogram for each semantic class in the NDVI image is clustered into 16 bins, while the histograms for the EVI images consist of 20 bins. The weight of a bin is defined

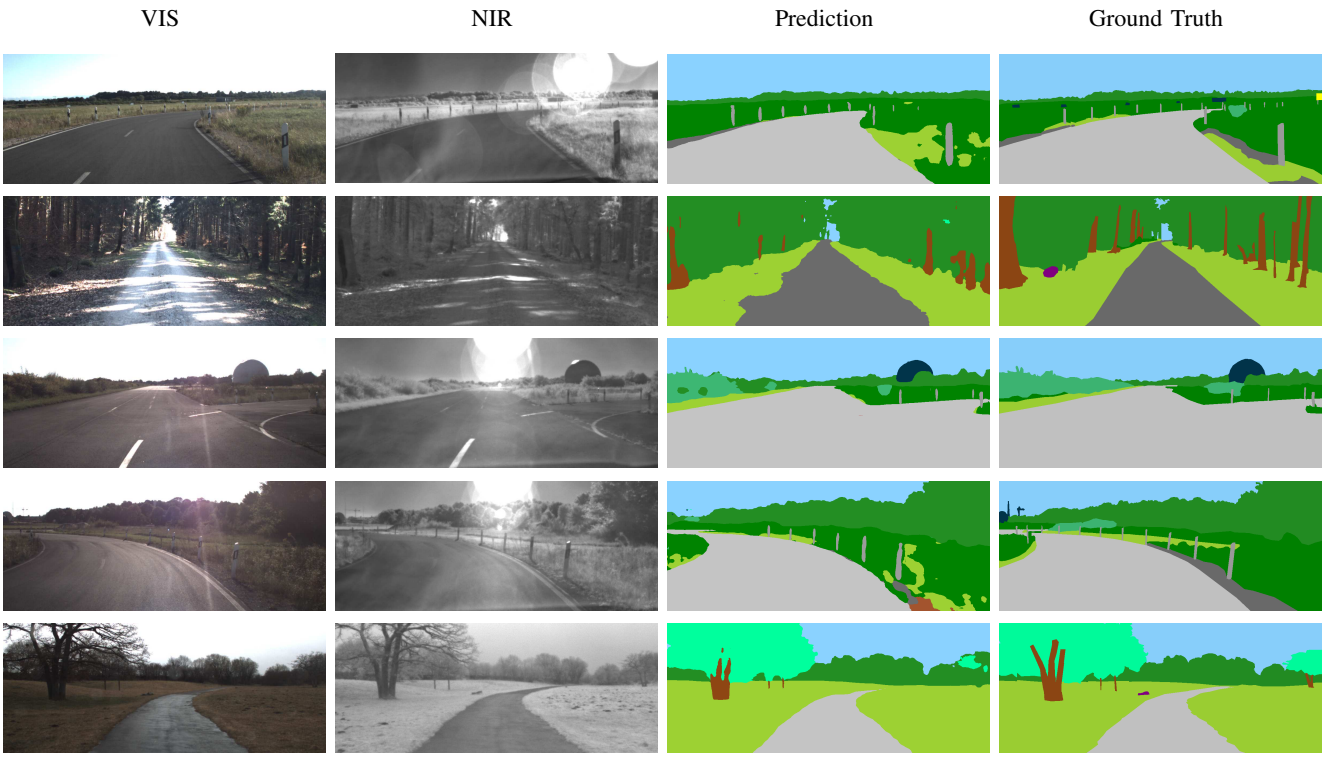

Fig. 6: A qualitative evaluation of the DeepLabV3+ prediction performance with different types of vegetation, such as a tree trunk, tree crown, low grass, high grass, forest and bush. The presented images also show the pole and obstacle class, which we didn't quantitatively evaluate in Tab. II.

TABLE II: The Intersection over Union (IoU) in percent for each semantic class of interest and the mean Intersection over Union (mIoU) over all semantic classes of interest. We use local temperature scaling (LTS) to calibrate the output of DeepLabV3+. The addition of a image histogram-based approach alone ($\beta = 0.75$) shows no significant improvement. We observe a significant improvement from the fully-connected conditional random field (CRF with $\theta_\alpha = 10$, $\theta_\beta = 13$). $\checkmark_x$ notes, that the CRF uses image modality x as structural input during inference.

network	LTS	Histogram	CRF	mIoU	asphalt	gravel	soil	low grass	high grass	bush	tree crown	tree trunk	forest
DeepLabv3+	✗	✗	✗	30.61	50.24	34.14	2.55	56.86	30.98	10.14	7.43	10.80	72.38
	✓	NDVI	✗	30.56	50.08	34.17	2.50	56.75	30.80	10.26	7.42	10.74	72.39
	✓	EVI	✗	30.62	**50.31**	34.13	2.54	56.88	30.97	10.13	7.44	10.76	72.40
	✓	NDVI	$\checkmark_{\text{NDVI}}$	47.97	50.16	61.02	23.11	**62.48**	46.93	37.78	46.34	30.65	**73.12**
	✓	EVI	$\checkmark_{\text{EVI}}$	50.16	47.27	62.08	30.40	62.17	**48.43**	38.59	56.98	**32.64**	72.88
	✓	✗	$\checkmark_{\text{VIS}}$	**52.19**	45.31	70.50	**43.10**	60.09	47.99	**41.08**	**59.89**	29.23	72.60

as the ratio of the semantic pixels in a bin compared to all semantic pixels in the histogram of a semantic class.
The calibrated prediction for each pixel is supplemented with the histogram-based prediction of the respective NDVI and EVI pixel by adding the normalized bin weights $\omega(x)$ for each semantic class k.

$$\hat{\mathbb{Q}}_n(x, T_n, \omega) = \max_{k \in K} \beta\ \omega(x)^{(k)} + \sigma_{\text{SM}}(\frac{\mathbf{z}_n(x)}{T_n(x)})^{(k)} \quad (9)$$

The hyperparameter β weights the influence of the histogram-based predictions $\omega(x)$. $\beta = 0.75$ is set for both the NDVI and EVI image predictions and β has been optimized on the validation split.

To smooth the fused predictions, we use a fully connected conditional random field [32]. The possible labelings of an image conditioned over the input image pixel intensities are characterized by a graph and its cliques. The cliques induce a potential, which assign a cost to assigning labels to neighboring pixels [33]. The potentials consists of two kernels.

The first is a so-called smoothness kernel, that removes small isolated regions in the labeling [34]. The second kernel is the appearance kernel, which penalizes if nearby pixels have different semantic labels (*nearness*) and penalizes if pixels with similar pixel intensities in the input image have

different semantic labels in the prediction (*similarity*).

A higher correlation between the pixel intensities in the vegetation indices to the semantic classes of different surface and vegetation types can be observed. We investigate the effect of passing the vegetation indices or the NIR image to the CRF as input image improves the final semantic segmentation.

V. Experiment

We quantitatively evaluate our method by comparing the prediction performance of a semantic segmentation model, that only makes use of the VIS images from the test split, to our combined approach of both calibrated neural network predictions on VIS data and hand-crafted VIS+NIR vegetation indices.

For our experiment we have trained a DeepLabv3+ network [35] with the 439 VIS images of the TAS500 [3] dataset. The images in the training dataset are cropped to $1200 \times 480\,\mathrm{px}$ and are matched to the same region of interest as in the TAS-NIR dataset. The neural network has been trained for 80.000 iterations in total.

In Tab. II we compare the general semantic segmentation performance of all methods and compare the performance on a per-class level. The low IoU scores for specific semantic classes like bush, soil, tree trunk and tree crown in Tab. II is caused by only very few occurrences in the TAS-NIR dataset test split. This leads to misclassifications of these few instances to more heavily effect the IoU score. In Fig. 6 we present some predictions of our proposed method.

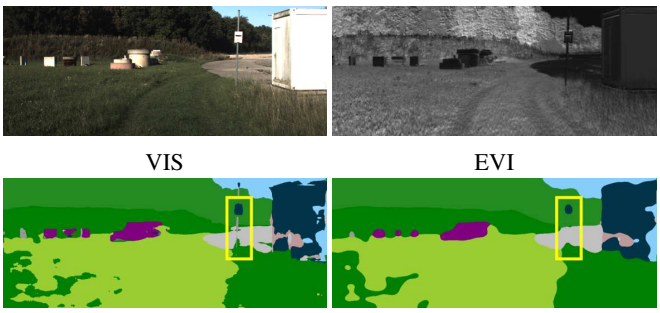

Fig. 7: For the conditional random field (CRF) we can observe how the smoothness kernel can lead to thin structures like the pole in the yellow box (■) to disappear in the final segmentation.

VI. Conclusions

In this work, we introduced the novel TAS-NIR dataset for the fine-grained semantic segmentation of ground surface and vegetation types in unstructured outdoor environments from VIS+NIR image pairs. The image histogram-based approach to supplement the semantic segmentation coming from the VIS image has shown only very little influence on improving the overall semantic segmentation performance. The use of a CRF improves the performance significantly. The CRF can fill small semantic regions in the final prediction. This comes at the cost of missing thin obstacles like the pole in Fig. 7.

References

[1] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A RUGD Dataset for Autonomous Navigation and Visual Perception in Unstructured Outdoor Environments," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, 2019.

[2] P. Jiang, P. Osteen, M. Wigness, and S. Saripalli, "RELLIS-3D Dataset: Data, Benchmarks and Analysis," 2020.

[3] K. A. Metzger, P. Mortimer, and H.-J. Wuensche, "A Fine-Grained Dataset and its Efficient Semantic Segmentation for Unstructured Driving Scenarios," in *Int. Conf. Pattern Recognition (ICPR)*, Milano, Italy (Virtual Conference), Jan. 2021.

[4] D. Maturana, P.-W. Chou, M. Uenoyama, and S. Scherer, "Real-time Semantic Mapping for Autonomous Off-Road Navigation," in *Field and Service Robotics*. Springer, 2018, pp. 335–350.

[5] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "AdapNet: Adaptive Semantic Segmentation in Adverse Environmental Conditions," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4644–4651.

[6] R. W. Wood, "Photography by Invisible Rays," *Photographic Journal*, pp. 329–338, 1910.

[7] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination Invariant Face Recognition Using Near-Infrared Images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, pp. 627–639, 2007.

[8] J. W. R. Jr., R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the Great Plains with ERTS," vol. 1, 1974, p. 309.

[9] F. J. Kriegler, W. A. Malila, R. F. Nalepka, and W. Richardson, "Preprocessing Transformations and Their Effects on Multispectral Recognition," vol. 2, Oct. 1969, pp. 13–16.

[10] A. R. Huete, K. Didan, W. J. D. van Leeuwen, A. Jacobson, R. Solanos, and T. D. Laing, "MODIS Vegetation Index (MOD 13) Algorithm Theoretical Basis Document (ATBD)," 1999, p. 129.

[11] A. Valada, G. L. Oliveira, T. Brox, and W. Burgard, "Deep Multispectral Semantic Scene Understanding of Forested Environments using Multimodal Fusion," in *Int. Symp. Experimental Robotics (ISER)*, 2016.

[12] L. Schaul, C. Fredembach, and S. Süsstrunk, "Color image dehazing using the near-infrared," *Proc. IEEE Int. Conf. Image Processing (ICIP)*, pp. 1629–1632, 2009.

[13] C. Fredembach and S. Süsstrunk, "Colouring the Near-Infrared," 2008, pp. 176–182.

[14] N. Salamati, D. Larlus, G. Csurka, and S. Süsstrunk, "Semantic Image Segmentation Using Visible and Near-Infrared Channels," in *Proc. European Conf. Comput. Vision (ECCV)*, 2012.

[15] A. Valada, R. Mohan, and W. Burgard, "Self-Supervised Model Adaptation for Multimodal Semantic Segmentation," *International Journal of Computer Vision (IJCV)*, Jul. 2019.

[16] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture," in *Asian Conf. on Comput. Vision*, Nov. 2016.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proc. Int. Conf. Mach. Learning (ICML)*, Aug. 2017, pp. 1321–1330.

[18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles," in *Conf. on Neural Inf. Processing*, 2017, p. 6405–6416.

[19] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Proc. Int. Conf. Mach. Learning (ICML)*, Jun. 2016, pp. 1050–1059.

[20] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," in *Advances in Large-Margin Classifiers*, 1999, pp. 61–74.

[21] Z. Ding, X. Han, P. Liu, and M. Niethammer, "Local Temperature Scaling for Probability Calibration," in *Proc. IEEE Int. Conf. Comput. Vision (ICCV)*, 2021, pp. 6889–6899.

[22] M. Brown and S. Süsstrunk, "Multispectral SIFT for Scene Category Recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2011, pp. 177–184.

[23] C. Winkens, F. Sattler, V. Adams, and D. Paulus, "HyKo: A Spectral Dataset for Scene Understanding," in *Proc. IEEE Int. Conf. Comput. Vision Workshops (ICCVW)*, 2017, pp. 254–261.

[24] A. Unterholzner and H.-J. Wuensche, "Hybrid Adaptive Control of a Multi-Focal Vision System," in *IEEE Trans. Intell. Veh.*, 2010, pp. 534–539.

[25] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. F. Dominguez, "WildDash - Creating Hazard-Aware Benchmarks," in *Proc. European Conf. Comput. Vision (ECCV)*, Sep. 2018.
[26] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania, "Calibrating Deep Neural Networks using Focal Loss," in *Conf. on Neural Inf. Processing*, 2020, pp. 15 288–15 299.
[27] O. Irsoy and E. Alpaydin, "Autoencoder Trees," in *Asian Conference on Machine Learning*, 2026, pp. 378–390.
[28] C.-Y. Lee, P. Gallagher, and Z. Tu, "Generalizing Pooling Functions in CNNs: Mixed, Gated, and Tree," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 863–875, 2018.
[29] A. Niculescu-Mizil and R. Caruana, "Predicting Good Probabilities with Supervised Learning," in *Proc. Int. Conf. Mach. Learning (ICML)*, 2005, p. 625–632.
[30] K. Nassau, *The Physics and Chemistry of Color: The Fifteen Causes of Color*, 2nd ed. Wiley, 2001.
[31] H. Jones and R. Vaughan, *Remote Sensing of Vegetation: Principles, Techniques, and Applications*. Oxford University Press, 2010.
[32] P. Krähenbühl and V. Koltun, "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials," in *Conf. on Neural Inf. Processing*, 2011.
[33] J. Lafferty, X. Zhu, and Y. Liu, "Kernel Conditional Random Fields: Representation and Clique Selection," in *Proc. Int. Conf. Mach. Learning (ICML)*, 2004.
[34] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context," *Int. J. Comput. Vision*, 2009.
[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proc. European Conf. Comput. Vision (ECCV)*, 2018.