HED-UNET: A MULTI-SCALE FRAMEWORK FOR SIMULTANEOUS SEGMENTATION AND EDGE DETECTION

Konrad Heidler [1,2], *Lichao Mou* [1,2], *Celia Baumhoer* [3], *Andreas Dietz* [3], *Xiao Xiang Zhu* [1,2]

[1] Data Science in Earth Observation, Technical University of Munich (TUM), Munich, Germany
[2] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany
[3] Remote Sensing Data Center (DFD), German Aerospace Center (DLR), Wessling, Germany

ABSTRACT

Segmentation models for remote sensing imagery are usually trained on the segmentation task alone. However, for many applications, the class boundaries carry semantic value. To account for this, we propose a new approach that unites both tasks within a single deep learning model. The proposed network architecture follows the successful encoder-decoder approach, and is improved by employing deep supervision at multiple resolution levels, as well as merging these resolution levels into a final prediction using a hierarchical attention mechanism. This framework is trained to detect the coastline in Sentinel-1 images of the Antarctic coastline. Its performance is then compared to conventional single-task approaches, and shown to outperform these methods. The code is available at `https://github.com/khdlr/HED-UNet`.

Index Terms— Semantic segmentation, edge detection, Antarctica, glacier front

1. INTRODUCTION

Many tasks in remote sensing are based around the segmentation of imagery. Be it the extraction of building footprints, land cover mapping, or coastline detection, all these tasks require the pixels of an image to be separated into multiple classes [2].

Unlike with other computer vision tasks, the boundary between regions oftentimes has some semantic value in remote sensing. Looking at a photograph, the boundary between a person and the background does not have any special importance by itself. This is different in remote sensing where, for example, the boundary between water and land in an image represents the *coastline*. Therefore, the application of edge detection approaches has very high value in this field (e. g. building boundaries, road extraction, biomass estimation, coastline mapping, etc.). Making this observation, it is only natural to ask the question "can we exploit the relationship between segmentation and edge detection?"

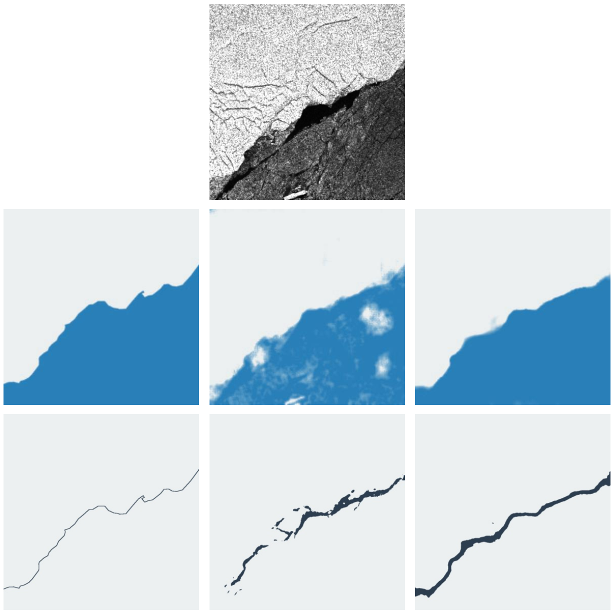

Fig. 1. Combining segmentation and edge detection greatly improves performance. Top: Input SAR image (HH polarization). Middle (left to right): Segmentation ground truth, UNet prediction, HED-UNet prediction. Bottom (left to right): Edge detection ground truth, HED prediction, HED-UNet prediction. The displayed tile measures about $30\,\mathrm{km} \times 30\,\mathrm{km}$.

Given the fact that semantic segmentation models often produce blurry results close to the class borders (cf. fig. 1), the idea of introducing edge detection into segmentation frameworks is not new.

The combination of the tasks can be done in a sequential manner, where the edge detection is done first, and the predicted edges are then used as an additional input channel for the segmentation model [3].

Another simple method of combining the two is to augment a deep segmentation architecture by adding an auxiliary output for edge detection, without any additional changes to the architecture. This can already improve the segmentation

An extended version of this conference report is available at [1].

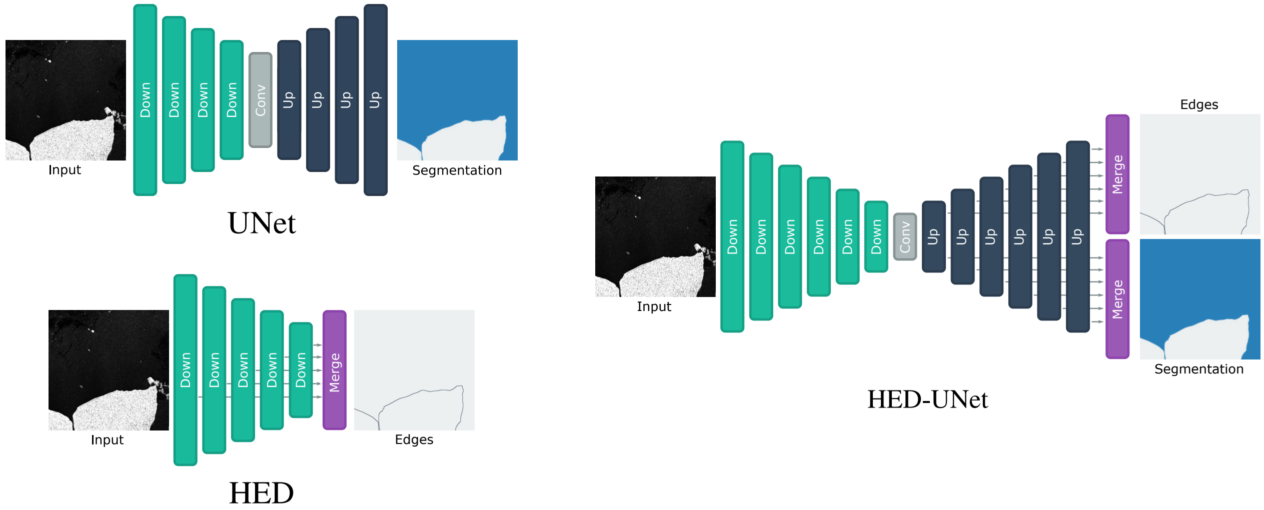

Fig. 2. Conceptual amalgamation of UNet and HED (left) into HED-UNet (right). While HED merges the outputs of "Down" blocks, HED-UNet uses the outputs of "Up" blocks, allowing for larger receptive fields and deeper decision paths.

results compared to pure segmentation in some cases, like in building footprint extraction [4] and coastline detection [5].

Finally, it is also possible to use predicted edges to guide a label propagation algorithm on an initial segmentation map, which also helps reduce blurry edges [6].

In contrast to these existing methods, we introduce a framework based on a unified theory of segmentation and edge detection, where both tasks are treated interchangeably. This means that both the edge detection and the semantic segmentation are predicted in a parallel, cooperative manner, allowing the model to fully exploit the bidirectional synergies arising from the connection of the tasks.

2. METHOD

As starting points for our framework, we choose the segmentation architecture UNet [7], and the edge detector HED [8]. Both have been extensively applied within their respective fields and can be considered as established methods.

The central idea in combining the two architecture is to exploit the fact that both are based on multi-scale processing of the imagery. While the UNet follows an encoder-decoder architecture, HED follows an encode-and-merge architecture, which aggregates features on multiple scale levels and merges the result into a final prediction. Our combined HED-UNet employs all three of these stages, as can be seen in fig. 2. First, a pyramid of feature maps at different resolutions is built in the encoder stage. Then, in the decoder stage, feature maps of increasing resolution are assembled by combining the signal with skip connections from the encoder stage. Finally, a merging head is employed to combine predictions at different stages into the model's final prediction. To allow for larger receptive fields, we increase the depth of the model to 6 down- and upsampling blocks instead of 4 (UNet) and 5 (HED).

Simultaneous segmentation and edge detection is then achieved by using two merging heads, one for segmentation and one for edge detection. This framework can be easily extended to additional tasks by adding new merging heads.

2.1. Hierarchical Attention Merging Heads

Choosing the right structure for the merging heads is vital to the performance of the model. In order to predict a segmentation or edge map from a pyramid of features, the merging head first computes coarse predictions for each resolution level. These are then resampled to the full output resolution and merged into the final prediction.

Initial experiments conducted with learned weights for the merging exhibited unstable performance, as the optimal merging coefficients change dramatically between different scenes. In order to remedy this behavior, we introduce a dynamic merging scheme based on attention [9, 10]. Instead of a sequential or spatial attention, the merging is done using hierarchical attention. This mechanism allows the model to take into account the local confidence for each prediction level, and therefore *attend to different resolution levels*, depending on where in the image a prediction is performed.

Further, the intermediate multiresolution predictions are also optimized to match an adequately re-scaled version of the ground truth. The resulting *Deep Supervision* [11] improves training convergence and encourages the model to widen its effective receptive field.

2.2. Loss Function

To allow for a unified theory of segmentation and edge detection, the loss function needs to be adapted.

For semantic segmentation, the *cross-entropy* loss is usually used. Given an image with height H and width W, a

Method		Segmentation		Edge Detection		
	Avg. Deviation	Accuracy	mIoU	F_1 ODS	F_1 OIS	Runtime
UNet [7]	271 m	0.892	0.806	—	—	283 ms
HED [8]	341 m	—	—	0.384	0.410	71 ms
HED-UNet	**222 m**	**0.920**	**0.849**	**0.397**	**0.416**	334 ms

Table 1. Numerical validation results for the trained models. It can be seen that the combination of the tasks leads to improved performance. Especially the average deviation between actual coastline and predicted coastline is reduced greatly.

target segmentation into K classes can be represented as a one-hot encoded target map $(T_{ijk}) \in \{0,1\}^{H\times W\times K}$. For class probabilities (P_{ijk}) of the same shape, predicted by a model, the cross-entropy loss is calculated as

$$\mathcal{L}(T,P) = -\frac{1}{HW}\sum_{i,j,k}\log(P_{ijk})T_{ijk}\,. \quad (1)$$

While this loss function is suitable for segmentation, it does not perform well for edge detection, because the classes are extremely unbalanced – almost all pixels belong to the "not edge" class, with only very few pixels actually lying on the edge.

In an attempt to close to the original cross-entropy loss, this can be fixed by adaptively weighting the loss per class, according to the number of pixels in that class:

$$\mathcal{L}(T,P) = -\sum_{k}\frac{1}{K\cdot\sum_{i,j}T_{ijk}}\cdot\sum_{i,j}\log(P_{ijk})T_{ijk} \quad (2)$$

The adaptive class-weighting factor $1/(K\cdot\sum_{i,j}T_{ijk})$ ensures that no matter the amount of pixels per class, each class contributes equally to the overall loss for the scene. For perfectly balanced classes, it is equivalent to Eq. 1.

These two properties make the loss function from Eq. 2 a good drop-in replacement for regular cross-entropy when both segmentation and edges are to be predicted.

3. EXPERIMENTAL SETUP

To assess the relative performance of this model compared to baseline approaches, we evaluate it on a Sentinel-1 dataset of the Antarctic Coastline. Coastline detection in the Antarctic is notoriously difficult, especially for SAR imagery. This is both due to the general challenges in SAR data like speckle and large dynamic range [12], as well as challenges specific to this location, like snow melt, sea ice and icebergs, or dry-snow facies of the higher ice sheet that are almost impossible to distinguish from open water from SAR imagery alone [13]. The dataset consists of 16 cropped Sentinel-1 GRD scenes of Antarctica's coastline acquired between June 2017 and December 2018, and covers a combined area of around $730\,000\,\mathrm{km}^2$. The imagery has a spatial resolution of $40\,\mathrm{m}$ in dual polarization. Four of the 16 scenes were reserved as validation scenes in a way so that training and testing scenes are spatially distinct.

On this dataset, we train and evaluate the baseline single-task models HED [8] for edge detection and UNet [7] for semantic segmentation. Finally, we do the same for our proposed HED-UNet model, but train it to perform both tasks at once. The scenes were split into patches of 768×768 pixels, allowing for large convolutional receptive fields.

All models are trained for 15 epochs on the dataset, and then validated on the unseen, spatially distinct test scenes.

4. RESULTS & DISCUSSION

Table 1 quantifies the performance of the evaluated models on the withheld validation dataset. Here, *mIoU* denotes the "mean Intersection over Union", which is generally considered a more informative metric for segmentation than the plain pixel-wise accuracy. Perhaps the most significant metric for the task of coastline detection is the *average deviation*, which denotes the average distance between the true coastline and the predicted one. Finally, F_1 ODS and F_1 OIS are edge detection metrics that denote the F_1 score obtained when thresholding the edge predictions at an optimal threshold for the entire dataset (ODS) or the optimal threshold for each separate image (OIS). Finally, we also record the average time that each model takes to predict the coastline for a $100\,\mathrm{km}\times 100\,\mathrm{km}$ scene on a NVIDIA V100 GPU.

The validation metrics show that the proposed HED-UNet improves greatly upon the baseline models, which we attribute mostly to the following three factors:

1. The synergy between the two tasks greatly helps the model learn useful representations. Plain semantic segmentation does not teach the model the concept of boundaries, while plain edge detection does not understand the difference between the classes. The combined model therefore has a better understanding of the entire scenery.

2. Deep supervision and the added down-/upsampling steps encourage a larger receptive field. Through the deep supervision and merging procedure, HED-UNet is forced to encode meaningful features into its deeper layers, which in turn means that a larger contextual

window is taken into account. Its final merged predictions are therefore based on more contextual information, which helps with the correct classification.

3. Attention merging allows the model to focus on intricate details in regions where they are needed, like rugged coastlines. Farther from the edge, the robust, coarse predictions can be given a higher weight, resulting in less noise in these regions.

5. CONCLUSION

Simultaneous segmentation and edge detection entails great synergies that can be easily exploited. The model proposed here does so by combining ideas from the widely used UNet segmentation architecture and the HED edge detection architecture. Further, the introduction of hierarchical attention merging heads allows for the adaptive merging of the information present in the multiresolution feature maps. We show that this approach is indeed highly beneficial for the task of coastline detection in Antarctica.

Compared to the regular UNet segmentation architecture, our model only needs marginal additional computational resources. We are convinced that it can be of great use for other tasks where edges have a special significance, like the mapping of building footprints, roads or bodies of water. Finally, through the plug-and-play nature of the attention merging heads, it is easily extendable to incorporate other dense prediction tasks like keypoint detection.

6. REFERENCES

[1] Konrad Heidler, Lichao Mou, Celia Baumhoer, Andreas Dietz, and Xiao Xiang Zhu, "HED-UNet: Combined segmentation and edge detection for monitoring the antarctic coastline," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2021.

[2] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[3] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, Jan. 2018.

[4] Zhongze Jiang, Zhong Chen, Kaixiang Ji, and Jian Yang, "Semantic segmentation network combined with edge detection for building extraction in remote sensing images," in *MIPPR 2019: Pattern Recognition and Computer Vision*, Nong Sang, Jayaram K. Udupa, Yuehuan Wang, and Zhenbing Liu, Eds. International Society for Optics and Photonics, 2020, vol. 11430, pp. 60 – 65, SPIE.

[5] Dongcai Cheng, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "FusionNet: Edge aware deep convolutional networks for semantic segmentation of remote sensing harbor images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 10, no. 12, pp. 5769–5783, Dec. 2017.

[6] Liang-Chieh Chen, Jonathan T. Barron, George Papandreou, Kevin Murphy, and Alan L. Yuille, "Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016, pp. 4545–4554.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, Eds., Oct. 2015, pp. 234–241.

[8] Saining Xie and Zhuowen Tu, "Holistically-nested edge detection," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 11, 2017.

[10] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," *Proc. 2018 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, p. 10, 2018.

[11] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu, "Deeply-supervised nets," in *Int. Conf. Artif. Intell. Stat. (AISTATS)*, Guy Lebanon and S. V. N. Vishwanathan, Eds., May 2015, vol. 38, pp. 562–570.

[12] Xiaoxiang Zhu, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, and Richard Bamler, "Deep learning meets SAR: Concepts, Models, Pitfalls, and Perspectives," 2021.

[13] Celia A. Baumhoer, Andreas J. Dietz, C. Kneisel, and C. Kuenzer, "Automated extraction of antarctic glacier and ice shelf fronts from sentinel-1 imagery using deep learning," *Remote Sens.*, vol. 11, no. 21, pp. 2529, Jan. 2019.