# Towards Historical Analysis of Riverscape Development Utilizing Semantic Segmentation

**Saeid Shamsaliei[1,*], Odd Erik Gundersen[1,2], Knut Alfredsen[1], Jo Halvard Halleraker[1,3]**

[1] Norwegian University of Science and Technology, Trondheim, Norway
[2] Aneo AS, Trondheim, Norway
[3] Norwegian Environment Agency, Trondheim, Norway
[*] Corresponding author: saeid.shamsaliei@ntnu.no

## Abstract

Rivers are under pressure from human development, which impacts river ecosystems severely. As part of the UN Decade of Ecosystem Restoration, UN has decided to take actions to restore ecosystems. Fresh water ecosystems have been considered particularly degraded. Making effective policies for how to restore river ecosystems is practically impossible without quantitative data that takes historical development into consideration. We present a system that semantically segments historical aerial images of riverscapes into six different classes and an analysis framework for large-scale temporal analysis of riverscape development that will utilize the segmentation classes when completed. The best performing semantic segmentation model achieves an average MIoU 74.1% and utilizes both model-centric and data-centric methods. A qualitative error analysis shows that this performance is satisfactory for temporal analysis of riverscape development given the requirements of both local and global analyses.

## Introduction

The changes of the rivers and adjacent landscapes have been accelerating since start of the 20th century (Piégay et al. 2020). Pressures caused by Human activities such as development of hydropower sites, flood protection construction, gravel mining and urbanization are the main factors of such drastic change (Gilvear and Bryant 2016; Grill et al. 2019). These pressures impact the ecosystem provided by rivers and led to reduction of biodiversity and habitat loss (Wohl 2019). Therefore, it is crucial to provide restoration for the degraded rivers and to ensure that future decisions will consider the sustainability of the rivers and their ecosystems. The UN declared the decade from 2021 to 2030 as the UN Decade of Restoration to emphasize the importance of revival of ecosystems all around the world. The European Union has adopted the Water Framework Directive (WFD) with the goal of ensuring the sustainable use of river ecosystems and if necessary addressing required mitigation. However, in order to find potential restorations and develop sustainable policies, it is essential to understand the effect of previous decisions on the development of rivers. As different encroachments, like regulations for hydropower, have been applied to different rivers, studying the evolution of rivers will help to understand their impacts. This study should include multiple states of rivers from historical time, with as few anthropogenic influence as possible, to the recent time. However, the bottleneck of this study is the acquisition of data which represents the historical state of the river and such process should start with automatic semantic segmentation of historical images (Åström et al. 2017). Satellite images are not suitable, since before 1990 satellite images were not taken on a regular basis. However, aerial photography started to appear around the 1930s and are taken with high resolution, which make them a great source for data acquisition (Marchese et al. 2017; Arnaud et al. 2015). Even though such aerial images are available, these images should be mapped into desired habitats in order to use them for the assessment process (Piégay et al. 2020).

Several works aimed to study the evolution of limited rivers through time by using historical maps (Gurnell, Downward, and Jones 1994), historical topographical maps (García, Dunesme, and Piégay 2020), aerial images (Gurnell 1997), and combination of historical maps and aeiral images (Zanoni et al. 2008). New architectures have been proposed for semantic segmentation of land cover classification datasets such as DeepGlobe (Demir et al. 2018) and LandCover.ai (Boguszewski et al. 2021). Architectures such as NU-Net (Samy et al. 2018), FPN (Seferbekov et al. 2018a) with ResNet50 as backbone (He et al. 2016) and spatial dropout (Seferbekov et al. 2018b), DIResUNet (Priyanka et al. 2022), GLNet (Chen et al. 2019) and MagNet (Huynh et al. 2021). MagNet is state-of-the-art in DeepGlobe land cover classification. Moreover, stochastic weight averaging (SWA) (Izmailov et al. 2018) and Lovasz-Softmax loss function (Berman and Blaschko 2017) are used to improve the performance (Rakhlin, Davydow, and Nikolenko 2018).

In land cover classification of riverscapes, a lot of research is done on analysis of remote sensing at sub-metric resolutions (Marcus and Fonstad 2010; Carbonneau et al. 2012; Piégay et al. 2012, 2020). High resolution images lead to better understanding of river ecology (Vannote et al. 1980; Fausch et al. 2002; Carbonneau and Piégay 2012) and an abundance of high resolution images are available. Satellite images have been used in several studies such as work of (Bhatpuria et al. 2022). To our knowledge, the only other example from literature that utilizes deep learning for semantic segmentation of historical aerial images of riverscapes is Al-

fredsen et al. (2021). We reproduce their work and consider the reproduced model as **baseline**.

Despite the importance of data, it has been the undervalued part of AI ecosystem (Aroyo et al. 2021). However, recent attention to data-centric [1] approaches (Whang and Lee 2020) led to promissing achievements in domains such as image segmentation (Motamedi, Sakharnykh, and Kaldewey 2021), object detection (Terzi, Azginoglu, and Terzi 2021) and semantic segmentation (Roth, Wüstefeld, and Weichert 2021).

Our main contributions are threefold. We present a deep learning model that semantically segment historical aerial images of riverscapes into six habitat types of riverscapes that improve over state-of-the-art. Second, we show through an ablation study that improved label quality and data augmentation are responsible for increasing the performance of the baseline from 64.6% to 72.5%. Changing the encoder from VGG16 to ResNet50 is responsible for the additional performance. Finally, we show how the semantic segmentation model can be utilized in analyses of how riverscapes develop over time.

## Challenges of Historical Aerial Images

We focus on historical aerial images of riverscapes from the 1960s to 1990s. Working with historical aerial images have unique inherent challenges: 1) Images are grayscale. Lack of color information make it more difficult to understand the underlying land type of each area. 2) Images are taken over a period of 40 years and changes in camera technology led to different contrast and brightness for different images. 3) Land types are not uniformly distributed so the data suffers from class imbalance. 4) The land type of some areas are hard to determine even for an expert, due to the low quality of images.

## Method and Experiments

We evaluate different deep learning methods for semantic segmentation along with other techniques including improved label quality. Models must semantically segment riverscapes into the habitat types water (W), gravel (G), vegetation (V), farmland (F), anthropogenic (A), and unknown (U). **Reproducing baseline:** The baseline is made by reproducing the work of (Alfredsen et al. 2021). The model used in this work is a U-Net with VGG16 (Simonyan and Zisserman 2014) encoder. The dataset, D0, contains (512×512 px) images of rivers Gaula 1963, Surna 1963 and Lærdal 1978 and the test set contains areas from river Gaula 1963, 1998 and Nea 1962. We tested if data augmentation would improve the performance of the baseline. **Improving labels:** We investigated if improving the data was a viable way of improving the performance of the baseline model. An informal inspection of labels indicated that the dataset had two main issues: 1) Errors and inconsistencies in labels and 2) Class imbalance. To improve the label quality, a set of domain experts re-annotated the data similar to (Terzi, Azginoglu, and Terzi 2021) and a set of strict guidelines were followed by the domain experts. To have detailed annotations, they were made by drawing with a pen on a tablet instead of typical polygon tools in GIS which is more typical (Carbonneau et al. 2020). To mitigate the class imbalance, a data augmentation method that rotates and samples images (RSA) was implemented. RSA is described in the appendix. In addition, an overview of dataset is described in the table A1. **Model comparison:** A model comparison study was conducted to identify whether the performance could be further improved. Several recent deep learning models developed for semantic segmentation as well as other techniques were evaluated. To improve the robustness and account for different camera quality, two different pipelines for online image augmentation (OA1 and OA2) were tested which are described in the appendix. To mitigate the class imbalance, weighted categorical cross entropy (WCE) (Samy et al. 2018) was used.

Two experiments are conducted. The first experiment evaluates the effect on improving data and the second experiment evaluates the effect of using other deep learning architectures and online data augmentation techniques.

**Experiment 1: Data improvements** In this experiment, we evaluate the impact of label quality and data augmentation on the performance. The baseline is trained with and without RSA on D0 and D1. Additionally, we test the more advanced architectures FPN, DeepLabV3+ and U-Net with ResNet50 as encoder on D0 to evaluate whether the more advanced architectures handle the noise D0 data better.

**Experiment 2: Model comparison** Four deep learning architectures, MagNet, FPN, DeepLabV3+ (Chen et al. 2018) and U-Net with ResNet50 as encoder were selected, in addition to SWA, RSA, WCE and online augmentation. In order to test these alternatives, a model comparison study was conducted. D1 was used as training data due to its performance in the previous experiment. MagNet is pretrained on DeepGlobe dataset. The detail of the experiment is described in the appendix.

## Results

**Experiment 1:** Table 2 shows the results of Experiment 1. Surprisingly the performance on the baseline did not improve by though data augmentation (RSA) when trained on D0. Also, none of the new architectures achieved any improvement when trained on D0. This can be explained by noise and inconsistencies in the annotation of D0. When training the baseline model on D1, the results improved by 4.86% and applying RSA led to a 2.1% additional improvement of average MIoU. Improved label quality and data augmentation have a significant effect on performance. **Experiment 2:** Table 1 shows the results of Experiment 2 for all combinations of the baseline but only the best performing combination for the other methods. For all the architectures SWA and RSA led to improvements. Less intensive online augmentation (OA2) performed better than OA1. Given that the data is grayscale, the model needs to solely rely on the texture of the image. Therefore some spatial augmentations might lead to confusion The full result is provided at the appendix.

[1] See: https://datacentricai.org

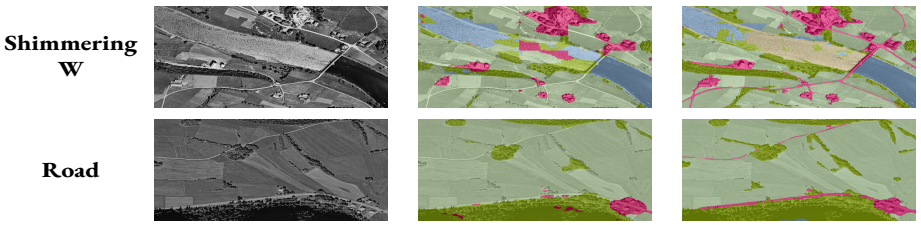


Figure 1: Examples of error cases observed in the baseline (middle column) as well as prediction of the improved model (right column). The improved model still struggles with shimmering water while the issue with roads are resolved.

| Model architecture | Encoder | SWA | RSA | WCE | OA1 | OA2 | Test sets (MIoU) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Gaula 63 | Nea 62 | Gaula 98 | Average |
| U-Net | VGG16 | | | | | | 78.16 | 68.4 | 61.82 | 69.46 |
| | | ✓ | | | | | 79.29 | 71.20 | 61.34 | 70.61 |
| | | ✓ | ✓ | | | | 79.13 | 73.43 | 64.75 | 72.43 |
| | | ✓ | ✓ | ✓ | | | **80.55** | **74.27** | 60.8 | 71.88 |
| | | ✓ | ✓ | | ✓ | | 74.76 | 65.23 | 69.95 | 69.98 |
| | | ✓ | ✓ | ✓ | | ✓ | 79.56 | 72.96 | 64.76 | 72.46 |
| | ResNet50 | ✓ | ✓ | ✓ | | ✓ | 77.59 | 71.25 | **73.55** | **74.13** |
| FPN | ResNet50 | ✓ | ✓ | ✓ | | ✓ | 79.30 | 72.10 | 67.61 | 73.00 |
| DeepLabV3+ | ResNet50 | ✓ | ✓ | | | | 79.45 | 71.98 | 63.35 | 71.59 |
| MagNet | FPN-ResNet50 | ✓ | ✓ | ✓ | | ✓ | 79.36 | 72.01 | 63.14 | 71.50 |

Table 1: MIoU of the model-centric experiments. Except for U-Net VGG16, only the best performing combination of model-centric methods are illustrated.

| Model architecture | Encoder | RSA | D0 | D1 | Test sets (MIoU) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | Gaula 63 | Nea 62 | Gaula 98 | Average |
| U-Net | VGG16 | | ✓ | | 69.12 | 69.80 | 54.18 | 64.60 |
| | | ✓ | ✓ | | 63.18 | 59.46 | 47.88 | 56.84 |
| | | | | ✓ | 78.16 | 68.4 | 61.82 | 69.46 |
| | | ✓ | | ✓ | **79.23** | **73.11** | **62.62** | **71.65** |
| | ResNet50 | | ✓ | | 70.10 | 67.94 | **55.73** | 64.59 |
| FPN | ResNet50 | | ✓ | | **73.25** | **70.99** | 40.06 | 61.43 |
| DeepLabV3+ | ResNet50 | | ✓ | | 68.26 | 67.46 | 54.32 | 63.35 |

Table 2: Results in form of mean intersection of union (MIoU) for Experiment 1: Data improvements. First row is the reproduction of the baseline and rows are colored based on the version of the data used for training.

## Qualitative Error Analysis

In order to have an insight into the reliability of the baseline, we perform the analysis of the baseline and improved model. The prediction of models on all test sets were manually inspected and most pervasive errors were grouped together as *error case*. Some error cases are described as (correct label:prediction error). The overview of the cases are shown in the Table 3 and Figure 1 provides visual examples of error cases. More detailed description of the cases are presented in the appendix.

## An Example Of Changes Over Time

An analysis of how Gaula has changed from 1963 to 1998 is presented to illustrate how the segmentation model can be used. Hydropower is not allowed in River Gaula. However, it is not immune to other human influences. Figure 2 shows two images of the same geographical location of River Gaula in 1963. The left image highlights the changes in the river and the middle image illustrates how gravel has changed. These classes are the output of the semantic segmentation model with quick inspection of domain experts. In addition, the percentage of frequency of each class is illustrated in the bar chart. Figure shows the gravel bars in this area completely disappeared from 1963 to 1998 The decrease in gravel can be due to human pressures or different discharge of the river or a combination of both. The decrease of gravel in River Gaula plays an integral role in the biodiversity of the surrounding environment. The proportion of the anthropogenic class increased and indicates the potential effect the river and gravel bars. An increase in the frequency of water and vegetation can also be seen.

## Temporal-analysis Framework

The analysis of development of riverscapes can be done in two levels: 1) global and 2) local. The first step of both analyses is to segment the aerial images of rivers in multiple time points into the six desired classes. Segmented images will be used as input for both analyses. The local analysis focuses on the development of one river or a section of it (e.g. 100 km length). Local analysis investigates the detailed changes of the morphology of a river over time. Such changes include development of side channels, gravel bars, width of the river and change in the meander ratio (Rosgen 1994). By comparing the development of one river regulated for i.e., hydropower with an unregulated river in the same region, it is possible to disentangle the impact of the regulation from other anthropogenic and natural influences. Local analyses is more detailed compared to global analyses. Therefore, the output can be slightly corrected by experts. However, global analyses investigates the development of multiple rivers on a large scale. The segmentation map of many rivers will be stored in a geospatial database. The output of semantic segmentation can be used for computing large scale morphological and land use indices, like Connectivity Status Index (Grill et al. 2020), and to compare the development of rivers at a large scale. For example, comparing the development of all the regulated rivers in an area to the unregulated ones is

| Test sets | Gaula 1963 | | | Gaula 1998 | | | | Nea 1962 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Error Cases | W:F | Road | Noisy A | W:F | V:W | W:V | W:G | W:F | Road | Shimmering W | F:W |
| Baseline | | | | | | | | | | | |
| U-Net ResNet50 | | | | | | | | | | | |

Table 3: An overview of error cases observed in the baseline and the improved model. The red, yellow and green respectively indicate that there is a problem, the problem is partially solved and the problem is solved. It shows that shimmering waters still remain an issue for the improved model and Water:Farmland and Water:Gravel issues are not completely solved.

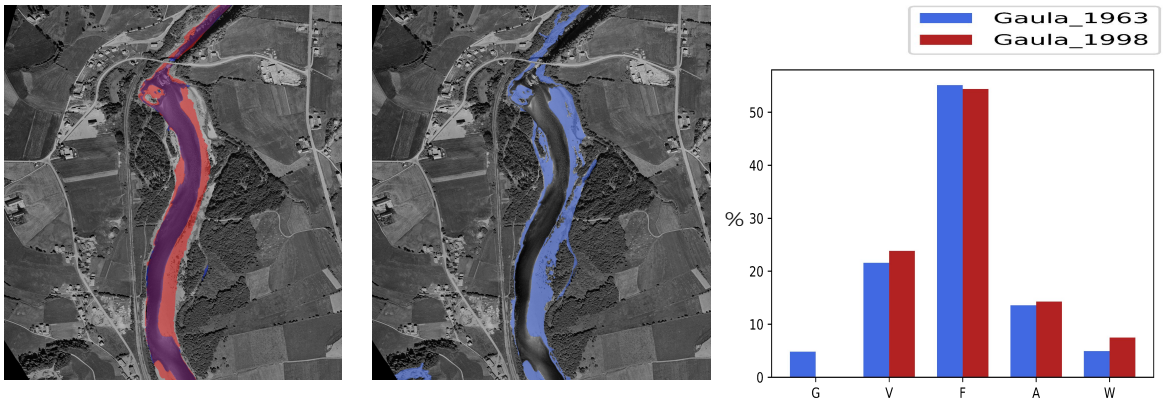


Figure 2: Comparison of land types of a section of River Gaula in two years of 1963 (blue) and 1998 (red). The left image is the aerial image masked with water class, The middle image is masked with gravel The right diagram shows the frequency of each class.

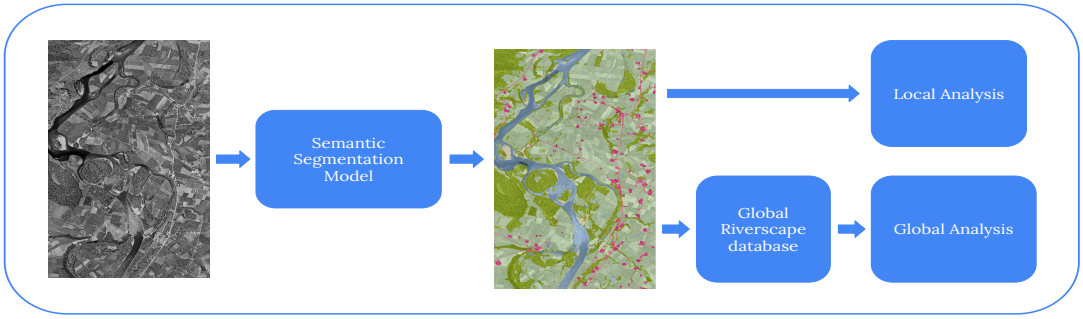


Figure 3: Framework for analysis of development of riverscapes through time. The input is a historical aerial image of a riverscape which is segmented into six habitats using semantic segmentation model. The output of the model is used for local analysis which focuses on one river, as well as global analysis which is large scale analysis of multiple rivers over time.

a possibility. The error rate of semantic segmentation on test sets of this work can be used as an input to a global analysis to account for potential errors. However, this solution might introduce some bias to the analysis. Historical aerial images have high variety and the error of the test set of this work might not represent the performance of the model on a new river. One alternative is to use the predictive uncertainty of the model using methods such as Monte-Carlo Dropout (Gal and Ghahramani 2016) similar to (Dechesne, Lassalle, and Lefèvre 2021). It is shown that the incorrect areas tend to have higher predictive uncertainty compared to the correct ones (Czolbe et al. 2021; Wickstrøm, Kampffmeyer, and Jenssen 2020). However, Neural Networks can be extremely confident in wrong predictions which make it challenging to use the estimation of uncertainty as the input to the analysis framework. Figure 3 illustrate the analysis framework for analysing the development riverscapes over time.

## References


Alfredsen, K.; Dalsgård, A.; Shamsaliei, S.; Halleraker, J. H.; and Gundersen, O. E. 2021. Towards an automatic characterization of riverscape development by deep learning. *River Research and Applications*, 38(4): 810–816.

Arnaud, F.; Piégay, H.; Schmitt, L.; Rollet, A.; Ferrier, V.; and Béal, D. 2015. Historical geomorphic analysis (1932–2011) of a by-passed river reach in process-based restoration perspectives: The Old Rhine downstream of the Kembs diversion dam (France, Germany). *Geomorphology*, 236: 163–177.

Aroyo, L.; Lease, M.; Paritosh, P. K.; and Schaekermann, M. 2021. Data Excellence for AI: Why Should You Care. *CoRR*, abs/2111.10391.

Åström, J.; Ødegaard, F.; Hanssen, O.; and Åström, S. 2017. Endring i leveområder for elvesandjeger og stor elvebreddedderkopp ved Gaula. Forekomst og dynamikk av elveører fra 1947 til 2014.

Berman, M.; and Blaschko, M. B. 2017. Optimization of the Jaccard index for image segmentation with the Lovász hinge. *CoRR*, abs/1705.08790.

Bhatpuria, D.; Matheswaran, K.; Piman, T.; Tha, T.; and Towashiraporn, P. 2022. Assessment of Large-Scale Seasonal River Morphological Changes in Ayeyarwady River Using Optical Remote Sensing Data. *Remote Sensing*, 14(14): 3393.

Boguszewski, A.; Batorski, D.; Ziemba-Jankowska, N.; Dziedzic, T.; and Zambrzycka, A. 2021. LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1102–1110.

Carbonneau, P.; Fonstad, M. A.; Marcus, W. A.; and Dugdale, S. J. 2012. Making riverscapes real. *Geomorphology*, 137(1): 74–86. Geospatial Technologies and Geomorphological Mapping Proceedings of the 41st Annual Binghamton Geomorphology Symposium.

Carbonneau, P. E.; Belletti, B.; Micotti, M.; Lastoria, B.; Casaioli, M.; Mariani, S.; Marchetti, G.; and Bizzi, S. 2020. UAV-based training for fully fuzzy classification of Sentinel-2 fluvial scenes. *Earth Surface Processes and Landforms*, 45(13): 3120–3140.

Carbonneau, P. E.; and Piégay, H. 2012. Introduction: The Growing Use of Imagery in Fundamental and Applied River Sciences.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, W.; Jiang, Z.; Wang, Z.; Cui, K.; and Qian, X. 2019. Collaborative Global-Local Networks for Memory-Efficient Segmentation of Ultra-High Resolution Images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Czolbe, S.; Arnavaz, K.; Krause, O.; and Feragen, A. 2021. Is Segmentation Uncertainty Useful? In Feragen, A.; Sommer, S.; Schnabel, J.; and Nielsen, M., eds., *Information Processing in Medical Imaging*, 715–726. Cham: Springer International Publishing. ISBN 978-3-030-78191-0.

Dechesne, C.; Lassalle, P.; and Lefèvre, S. 2021. Bayesian U-Net: Estimating Uncertainty in Semantic Segmentation of Earth Observation Images. *Remote Sensing*, 13(19).

Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; and Raskar, R. 2018. DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE.

Fausch, K. D.; Torgersen, C. E.; Baxter, C. V.; and Li, H. W. 2002. Landscapes to riverscapes: bridging the gap between research and conservation of stream fishes: a continuous view of the river is needed to understand how processes interacting among scales set the context for stream fishes and their habitat. *BioScience*, 52(6): 483–498.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

García, J. H.; Dunesme, S.; and Piégay, H. 2020. Can we characterize river corridor evolution at a continental scale from historical topographic maps? A first assessment from the comparison of four countries. *River Research and Applications*, 36(6): 934–946.

Gilvear, D.; and Bryant, R. 2016. Analysis of remotely sensed data for fluvial geomorphology and river science.

Grill, G.; Lehner, B.; Thieme, M.; Geenen, B.; Tickner, D.; Antonelli, F.; Babu, S.; Borrelli, P.; Cheng, L.; Crochetiere, H.; et al. 2019. Mapping the world's free-flowing rivers. *Nature*, 569(7755): 215–221.

Grill, G.; Lehner, B.; Tieme, M.; Ticker, D.; and Geenen, B. 2020. Mapping the world's free-flowing rivers using the Connectivity Status Index (CSI). In *EGU General Assembly Conference Abstracts*, 22487.

Gurnell, A.; Downward, S.; and Jones, R. 1994. Channel planform change on the River Dee meanders, 1876–1992. *Regulated rivers: research & management*, 9(4): 187–204.

Gurnell, A. M. 1997. Channel change on the River Dee meanders, 1946–1992, from the analysis of air photographs. *Regulated Rivers: Research & Management: An International Journal Devoted to River Research and Management*, 13(1): 13–26.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huynh, C.; Tran, A. T.; Luu, K.; and Hoai, M. 2021. Progressive Semantic Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 16755–16764. Computer Vision Foundation / IEEE.

Izmailov, P.; Podoprikhin, D.; Garipov, T.; Vetrov, D. P.; and Wilson, A. G. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. In *Proceedings of Uncertainty in AI (UAI 2018)*.

Li, L.; Jamieson, K.; DeSalvo, G.; Rostamizadeh, A.; and Talwalkar, A. 2018. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185): 1–52.

Marchese, E.; Scorpio, V.; Fuller, I.; McColl, S.; and Comiti, F. 2017. Morphological changes in Alpine rivers following the end of the Little Ice Age. *Geomorphology*, 295: 811–826.

Marcus, W. A.; and Fonstad, M. A. 2010. Remote sensing of rivers: the emergence of a subdiscipline in the river sciences. *Earth Surface Processes and Landforms*, 35(15): 1867–1872.

Motamedi, M.; Sakharnykh, N.; and Kaldewey, T. 2021. A data-centric approach for training deep neural networks with less data. *arXiv preprint arXiv:2110.03613*.

Piégay, H.; Alber, A.; Lauer, J. W.; Rollet, A.-J.; and Wiederkehr, E. 2012. Biophysical Characterisation of Fluvial Corridors at Reach to Network Scales.

Piégay, H.; Arnaud, F.; Belletti, B.; Bertrand, M.; Bizzi, S.; Carbonneau, P.; Dufour, S.; Liébault, F.; Ruiz-Villanueva, V.; and Slater, L. 2020. Remotely sensed rivers in the Anthropocene: state of the art and prospects. *Earth Surface Processes and Landforms*, 45(1): 157–188.

Priyanka; N, S.; Lal, S.; Nalini, J.; Reddy, C. S.; and Dell'Acqua, F. 2022. DIResUNet: Architecture for multiclass semantic segmentation of high resolution remote sensing imagery data. *Applied Intelligence*.

Rakhlin, A.; Davydow, A.; and Nikolenko, S. 2018. Land cover classification from satellite imagery with u-net and lovász-softmax loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 262–266.

Rosgen, D. L. 1994. A classification of natural rivers. *Catena*, 22(3): 169–199.

Roth, A.; Wüstefeld, K.; and Weichert, F. 2021. A Data-Centric Augmentation Approach for Disturbed Sensor Image Segmentation. *Journal of Imaging*, 7(10).

Samy, M.; Amer, K.; Eissa, K.; Shaker, M.; and ElHelw, M. 2018. NU-Net: Deep Residual Wide Field of View Convolutional Neural Network for Semantic Segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE.

Seferbekov, S.; Iglovikov, V.; Buslaev, A.; and Shvets, A. 2018a. Feature Pyramid Network for Multi-class Land Segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 272–2723.

Seferbekov, S.; Iglovikov, V.; Buslaev, A.; and Shvets, A. 2018b. Feature pyramid network for multi-class land segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 272–275.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Terzi, R.; Azginoglu, N.; and Terzi, D. S. 2021. False positive repression: Data centric pipeline for object detection in brain MRI. *Concurrency and Computation: Practice and Experience*.

Vannote, R. L.; Minshall, G. W.; Cummins, K. W.; Sedell, J. R.; and Cushing, C. E. 1980. The River Continuum Concept. *Canadian Journal of Fisheries and Aquatic Sciences*, 37(1): 130–137.

Whang, S. E.; and Lee, J.-G. 2020. Data collection and quality challenges for deep learning. *Proceedings of the VLDB Endowment*, 13(12): 3429–3432.

Wickstrøm, K.; Kampffmeyer, M.; and Jenssen, R. 2020. Uncertainty and interpretability in convolutional neural networks for semantic segmentation of colorectal polyps. *Medical Image Analysis*, 60: 101619.

Wohl, E. 2019. Forgotten Legacies: Understanding and Mitigating Historical Human Alterations of River Corridors. *Water Resources Research*, 55(7): 5181–5201.

Zanoni, L.; Gurnell, A.; Drake, N.; and Surian, N. 2008. Island dynamics in a braided river from analysis of historical maps and air photographs. *River Research and Applications*, 24(8): 1141–1159.

# Appendix

## Dataset Details

| River | Dataset D0 | Dataset D1 |
|---|---|---|
| Gaula 1963 | 1086 | 8755 (57) |
| Lærdal 1978 | 609 | 3686 (24) |
| Surna 1963 | 4613 | 922 (6) |
| Sum | 6307 | 13363 (87) |

Table A1: Number of annotated 512×512 patches in D0 and D1 before data augmentation (8000×6000 in parentheses).

## RSA

RSA adds patches that are sampled from the rotated versions of the large images. Only patches that have center pixels from the class "Gravel" are added. This reduces the imbalance of both "Gravel" and "Water", which are underrepresented in both datasets. Patches that have large overlap with other patches and that contains large amount of the unknown class are filtered out. The algorithm of RSA is described in Algorithm 1.

**Algorithm 1:** Rotation Algorithm for one large image.

**Input :** $Image_{M*N}$ (Large image with MxN dimension)
**Input :** $P, Q$ (Dimension of sample images)
**Input :** $SelectionLowerBound$ (minimum number of sampling images)
**Input :** $SamplingClass$ (class of interest for sampling)
**Input :** $Angle$ (Rotation Angle)
**Input :** $MaxUnknownPercentage$
**Input :** $MaxOverlapPercentage$
**Output:** $ImageList$: list of rotated images

Initialize ImageList as empty list
Rotate $Image_{M*N}$ by $Angle^{\circ}$
$PotentialImageNumber \leftarrow (M * N)/(Q * N)$
$SelectionLowerBound \leftarrow max(SelectionLowerBound, PotentialImageNumber)$
**for** $i \leftarrow 0$ **to** $SelectionLowerBound$ **do**
    $SampleCandidate \leftarrow$ randomly sample an image with center of $SamplingClass$;
    **if** *less than* $MaxUnknownPercentage$ *of* $SampleCandidate$ *is UnknownClass* ***AND*** *less than* $MaxOverlapPercentage$ *of* $SampleCandidate$ *is already been sampled and added to* $ImageList$ **then**
        $ImageList$.append($SampleCandidate$)
    **end**
**end**

## Experiment setups

Hyperparameters are selected using the Hyperband algorithm (Li et al. 2018) with the objective of validation accuracy with maximum 20 epochs and 10 hybrid iterations. It is used to select the initial learning rate between (0.01, 0.001, 0.0001) and dropout rate between (0.0, 0.1, 0.2) for architectures which have dropout. The batch size for all models is 16, except for MagNet which is 12 for backbone and 8 for refinement modules. L2 regularization is used for convolutional layers of all models except MagNet. Stochastic gradient descent with momentum of 0.9 and weight decay of 5e-4, is used for optimizating MagNet's backbone. The refinement module of MagNet is trained as described in the original paper. For other models, Adam is used with ReduceLROnPlateau, to reduce the learning rate by a factor of 0.5, if value loss did not decrease for more than 5 epochs. SWA is used with the constant learning rate of 5e-5 and is activated after the convergence, training stopped when value loss does not decreased for 20 epochs. MagNet trained for 484 epochs. Then SWA is activated and learning continues for 100 more epochs.

The 'Gravel' and 'Water' classes have more importance when assessing the evolution of the river. Therefore, these two classes were given higher weight to encode this importance into the objective in WCE. Finally, stochastic weight averaging (SWA) (Izmailov et al. 2018) was used with the aim of improving the generalization of the model.

### Online Augmentations

- **OA1:** random flipping, transposition, random changing of brightness and contrast, random image compression, optical and grid distortion, blurring filters.
- **OA2:** random flipping, transposition, random changing of brightness and contrast.

**Experiment 1** This experiment is done using the smaller 512×512 patches and not the full size 8000×6000 images, and MagNet could not be evaluated in this experiment. D0 is trained with and without the RA, using the baseline model. As Table 2 shows, unexpectedly, the RA led to worse performance. This can be explained by noise and inconsistencies in the annotation of D0, which led to low quality augmented samples. These noisy samples aggravated the training and it led to worse performance. Additionally, new architectures, namely, FPN, DeepLabV3+ and U-Net with ResNet50 as encoder, were also used for training the D0. Surprisingly, all of these more advanced models performed worse than the baseline model on D0.

However, when D1 was used for training the baseline model, it outperformed the baseline which illustrates the effectiveness of improving the quality of the data, or more generally, the data-centric method. Moreover, when D1 is used for training, applying the RA led to further improvement of the result.

**Experiment 2** Since the DeepGlobe land cover classification dataset is relatively close to the dataset of this work, MagNet was pre-trained on DeepGlobe. The last layer was changed from 7 to 6 dimensions, to match the number of classes in our data. To train the MagNet, images were

patched into $2448 \times 2448$ px, similar to DeepGlobe. Additionally 2 refinement modules with the scales of $612 \times 612$ and $1224 \times 1224$ were trained with an FPN backbone similar to the original paper. To reduce the class imbalance and encode the importance of 'water' and 'gravel' classes, WCE used the weights U:01.72%, W:22.41%, G:22.41%, V:17.24%, F:17.24% and A:18.97%.

## Qualitative Error Analysis

### Water:Farmland

- **Baseline:** Large segments of farmland can be found in the water. This problem is more pervasive in Gaula 1963, but it was observed in all the test sets.
- **Improved:** The noisy segments of farmland on the water mostly disappeared. The noise remain in some small areas only. The improved labeling of D1 reduced noise. RSA and U-Net ResNet50 reduce it further.

### Issue with roads

- **Baseline:** Roads are frequently ignored in Gaula 1963 and Nea 1962 dataset.
- **Improved:** It is noticeable that the anthropogenic class has more detail when the improved model is used. Improved label quality is the main source of improvement.

### Noisy anthropogenic segments

- **Baseline:** This issue is mostly observed in Gaula 1963. Noisy segments of the anthropogenic class can be found in vegetation. When training annotation was checked, similar noises were found in the data.
- **Improved:** This issue was resolved by improved label quality.

### Shimmering water

- **Baseline:** Due changes in the light condition and surface structure of the river in Nea 1962, some areas contain shimmering water. This confuses models and led to strange prediction.
- **Improved:** When using the best performing improved model, this area is wrongfully predicted as gravel. However, other architectures such as MagNet and U-Net ResNet without OA do not have this issue.

### Vegetation:Water

- **Baseline:** Mostly in Gaula 1998, light forest is mislabeled as water. This issue can be traced back to the dataset.
- **Improved:** This issue is resolved by improved label quality.

### Water:Vegetation

- **Baseline:** Mostly in calm areas of the river in Gaula 1998, large noisy segments of vegetation can be observed. The issue can be related to the global context since two adjacent prediction windows have completely different predictions.
- **Improved:** This problem is resolved by improved label quality and SWA and OA.

### Water:Gravel

- **Baseline:** In Gaula 1998, some shallow areas of the rivers were mislabeled as gravel. However, when inspected by domain experts, these areas were considered to be vague since images are grayscale and old.
- **Improved:** Even though this issue is improved considerably when the improved model is used, still some segments of gravel can be seen in these areas. It might be due to the uncertain nature of the problem. Even domain experts have difficulties distinguishing the gravel bars in the middle of the river, with the shallow water.

### Farmland:Water

- **Baseline:** Mostly in Nea 1962 some farmlands are mislabeled as water. This problem might be related to the context since the error happened at the edge of prediction windows.
- **Improved:** This problem is resolved by improved label quality.

**New Issues:** The improved models introduces new issues. For example, some dark forest areas were classified as water. Even though it is an error that affects the overall MIoU, it will not affect the development studies relying on this method as these areas are far from the riverscapes and will not be part of the analyses.