HAIR: A Dataset of Historic Aerial Images of Riverscapes for Semantic Segmentation

Saeid Shamsaliei, Odd Erik Gundersen,[*] Jo H. Halleraker,[†] Knut Alfredsen
Norwegian University of Science and Technology, Trondheim, Norway
{saeid.shamsaliei,odderik,knut.alfredsen,jo.halleraker}@ntnu.no

Anders Foldvik
Norwegian Institute for Nature Research, Trondheim, Norway
anders.foldvik@nina.no

Abstract

Accurate and reliable semantic segmentation of historical aerial images of landscapes is crucial for tracking, analyzing, and understanding land use over time. Development of such models is challenging due to the lack of annotated datasets of historical images. We introduce HAIR, the first dataset for semantic segmentation of historical aerial imagery for land cover to address this issue. The dataset contains high resolution, grayscale images of riverscapes with a resolution of 20 cm per pixel, captured from 1947 to 1998. The images of this large-scale dataset are annotated into six land types in meticulous detail by domain experts. We benchmark state-of-the-art semantic segmentation models and present both quantitative and qualitative results on in-distribution and out-of-distribution test sets. Our baseline experiments show that pre-training on a recent high-resolution satellite image dataset that is converted to grayscale does not improve performance. They also show that state-of-the-art models do not generalize well on out-of-distribution data. Finally, we characterize four challenges facing the segmentation of historical aerial images, including HAIR, and by this hope to spur interest in developing models that generalize well on historical images to support temporal analysis of land use.

1 Introduction

Rivers are central for the human condition. Early civilizations were build in river valleys, and today they remain important, i.e. for fresh drinking water and livelihood, such as fishing, agriculture and power production. However, this importance comes at a cost. Riverscapes are under pressure from human development, and this challenges the biodiversity and hydromorphology around the rivers. Given that rivers are home to some of the most diverse and endangered wildlife on Earth, the problem is especially severe. Since 1900 the human population has grown from 1.65 Billion to 7.9 Billion in 2022 [51]. The growth has lead to an increasing pressure on all ecosystems on the Earth, including riverscapes. UN has declared the decade from 2021 to 2030 as the UN Decade of Restoration, and it is described as "a rallying call for the protection and revival of ecosystems all around the world".

[*]Aneo AI Research, Trondheim, Norway
[†]Norwegian Environment Agency, Trondheim, Norway

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

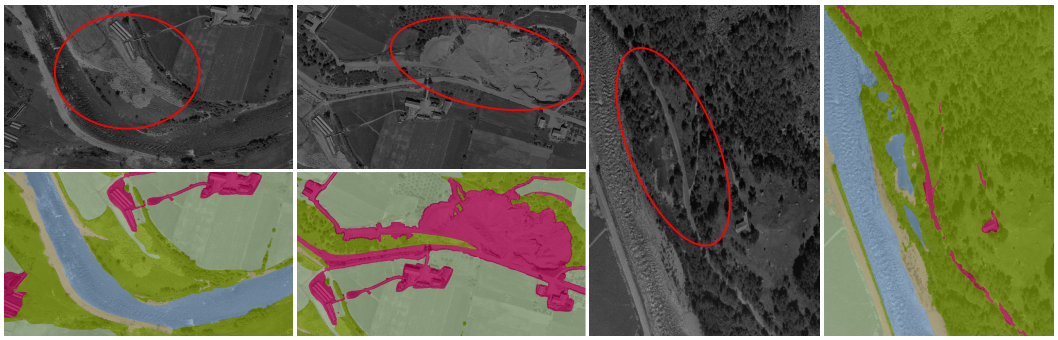


Figure 1: Ambiguous areas are marked in red. *Left* and *middle* show cases where gravel and human construction can be confused while *right* is an example where vegetation and water could be confused.

To understand the impact of human development in a region, it is crucial to understand the state of the landscape from a time when the world's population was significantly smaller. While satellite data has facilitated the production of land use and land cover maps since the 1990s [40], aerial images have been systematically captured since the early 1900s in some parts of the world. However, the majority of these aerial images were captured using an analog film-based camera before 2005. The captured films were then scanned and converted into a digital format [9]. Additionally, for aerial images captured before the 2000s, only panchromatic (grayscale) historic photographs are available [27]. Historical aerial images have the potential to be used as a valuable data source for understanding, monitoring, and analyzing land use over time. However, using aerial images for these purposes requires automatic and reliable mapping of the aerial grayscale images into desired habitats [44]. Image recognition, specifically semantic segmentation, can be utilized for this mapping but this relies on the availability of a large dataset of historical images annotated into desired the habitats.

Datasets of historical aerial images have four characteristics that impacts semantic segmentation: 1) camera technology has advanced significantly over time, which results in varying image quality based on when images are captured, 2) lightning conditions that are influenced by factors such as the time of the day and the airplane's direction during the capture, 3) class imbalance, as some classes are underrepresented due to the nature of the aerial images, and 4) grayscale, which means that they carry less information than satellite images that include RGB channels and sometimes additional infra-red channels as well.

Contributions: Our contributions are threefold. First, we release HAIR, the first dataset of high-resolution, historical aerial images with high-quality annotations of riverscapes made by experts. The dataset is released under the CC BY-SA 4.0 license[3] and contains roughly 8.72 billion annotated pixels surpassing widely recognized land cover datasets, such as DeepGlobe [22] and Inria [41]. Second, we present a benchmark of state-of-the-art semantic segmentation models to provide as baselines for future work. Third, experiments show that pre-training on high-resolution satellite image that are converted to grayscale does not improve performance and that state-of-the-art models do not generalize well to out-of-distribution data.

2 Related work

The datasets, LandCover.ai, DeepGlobe and Agriculture-Vision, which are summarized in Table 1 are most similar to HAIR given that they all contain images of natural landscapes with high resolution. Long et al.[39] give an overview of many other datasets. LandCover.ai is a semantic segmentation dataset of aerial images from rural areas across Poland with resolutions between 25 to 50 cm per pixel and contains 5 classes. Agriculture-Vision [21] is an aerial image dataset for pattern analysis of agricultural lands in US with nine classes and 10 cm per pixel resolution. DeepGlobe is a Satellite

[3] https://creativecommons.org/about/cclicenses/

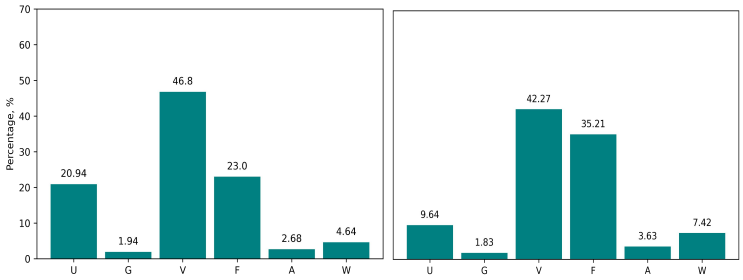
Table 1: Comparison of HAIR and similar natural landscapes datasets for semantic segmentation. The resolution unit is (meters per pixel).

Dataset	#Classes	#Images	Resolution	#Channels	#Pixel	Size
DeepGlobe Land Cover [22]	7	1147	0.5	RGB	$6.87 * 10^9$	2448x2448
LandCover.ai [5]	3	41	0.25,0.5	RGB	$2.98 * 10^9$	9000x9500;4200x4700
Agriculture-Vision [21]	9	94986	0.1,0.15,0.2	RGB+NIR	$2.49 * 10^{10}$	512x512
HAIR	6	178	0.2	Grayscale	$8.72 * 10^9$	8000x6000;6400x4800;16000x12000

Image Understanding Challenge with the three challenges: road extraction, building detection and land cover classification. DeepGlobe has high resolution images of 50 cm per pixel from India, Indonesia and Thailand and has 6 classes. Ratajczak et al. [49] introduce a historical aerial image dataset with grayscale images. They formulate the problem as a classification task where smaller image patches are given one class labels. However, the low resolution limits the usefulness of the data for studying land use evolution [13, 25].

HAIR has some key differences from other datasets. First, unlike most datasets that consist of recent aerial and satellite images with at least three channels (RGB), images are exclusively grayscale, as this is the nature of historical aerial images. Color information could help the segmentation model to distinguish two otherwise similar classes. In the absence of color information, models must solely rely on the texture and context of images. However, capturing the global context becomes challenging due to the high resolution of HAIR images and the limited memory capacity of current GPUs. Having high resolution images might increase the complexity of semantic segmentation and hinder the effectiveness of pretraining on common low resolution landscape datasets. However, for many important applications, such as understanding of river habitat [12] and their evolution through time [44], ice morphology [3] and fish ecology [59], it is crucial to have high resolution images. By providing high resolution images, we hope to direct further research into the challenges related to the context of historical grayscale images. Figure 8 in the appendix provides a visual comparison of HAIR aerial images and Sentinel-1 SAR images. As it can be seen from the figure, these SAR images have low resolution which make them undesirable for studying the evolution of the rivers. Additionally, Sentinel-1 images are only available after 2014 and cannot be used to study the historical state of the landscape. Second, to the best of our knowledge, HAIR is the only dataset with *gravel* class. This class is critical in analyzing the evolution of riverscapes and their ecosystems [4, 30]. Gravel serves as the exclusive habitat for certain insect and plant species, making it crucial to monitor changes in this class for biodiversity tracking [45]. Finally, even though grayscale images are more difficult for humans to annotate, as we rely on color information when interpreting images, annotations in HAIR are finer than in previous works. For example, small vegetation on top of gravel bars is annotated. These details help to develop models that are useful in ecological applications. Furthermore, the detailed annotations are made by experts. As illustrated by Table 1, the size of HAIR exceeds that of many widely recognized datasets in the field.

Since the introduction of FCN [38], the first end-to-end deep learning semantic segmentation model, numerous studies have proposed CNN-based architectures to enhance performance. These architectures include U-Net [50], ParseNet [36], PSPNet [67], DeepLab [15, 16, 17, 18], and HRNet [61]. Recently, there has been a growing interest in pure transformer-based architectures inspired by the success of Visual Transformers [24]. For instance, Segmentor [57], Segformer [62], and Swin-Unet [11] are pure transformer-based architectures. Additionally, some studies propose models combining transformers and CNNs, like TransUnet [14]. Minaee et al. [42], provide an overview of the segmentation models. In remote sensing, many models have been developed for datasets like DeepGlobe and LandCover.ai. Examples include FPN [54] with ResNet50 [31] as encoder and spatial dropout, NU-Ne [53], and DIResUNet [46]. Some proposed models, such as GLNet [20] and MagNet [32], leverage both downsampled and patched input images to capture the global context. MagNet is currently state-of-the-art in DeepGlobe land cover classification. The development of models in the field has predominantly focused on datasets containing images with three or more channels, so the research of semantic segmentation for grayscale images have been relatively limited.



River	Year	Total	ID	OOD
Gaula	1947	1.15	0.14	-
Gaula	1963	3.60	0.43	-
Gaula	1998	0.29	-	0.29
Surna	1963	0.48	0.10	-
Lærdal	1976	2.35	0.24	-
Orkla	1962	0.51	-	0.51
Nea	1962	0.34	0.03	-
Total	**-**	**8.72**	**0.94**	**0.80**

Figure 2: Diagrams: the distribution of classes in the dataset (left) and the OOD test set. Table: overview of the number of pixels taken from the different rivers (Rivers) in the different years (Year) and how many of these are in the ID and OOD test sets (in billion).

3 Dataset description

HAIR is comprised of 8.72 billion pixels spread over 178 annotated images from the five rivers Nea, Orkla, Surna, Gaula and Lærdal in Norway. The images come in one of three different sizes (8000×6000, 6400×4800 or 16000×12000 pixels), with a resolution of 20 cm per pixel. High quality annotations are made by experts, which limits the number of images that can be annotated; quality annotations take long time to make and they are made by experts that is a limited resource. This effort is inspired by the data-centric movement[4].

Figure 2 provides an overview of the dataset. The dataset contains images from 1947, 1962, 1963, 1978 and 1998 redand therefore, all of them are panchromatic (grayscale) since aerial images before the 2000 were captured in grayscale [27]. These images are selected so that we have both spatial and temporal overlap in the dataset. By spatial overlap, we mean when two images capture the same geographical area in different years, and by temporal overlap, we mean when two images are captured the same year (but not necessarily over the same geographical area). The images of river Gaula represent the spatial overlap, and the river Surna represents the temporal overlap with Gaula images from 1963.

The test sets have been designed to test whether both spatial and temporal characteristics are learned and can be generalized. Approximately 10% of the images from each river in the training set were chosen randomly for the ID test set, and the rest of images were used for training and validation. The OOD test set consists of two different rivers Gaula and Orkla. The OOD images of river Gaula are captured in 1998, which is a different time period than those found in the training set and translates to 51 years of camera improvement compared to the images from 1947 found in the training set. The OOD pictures of Orkla are captured in 1962, which are close in time to the Nea 1962, Surna 1963 and Gaula 1963. However, the Orkla images cover a completely different spatial area. Out of the 178 large images, 20 are in the ID test set and nine in the OOD test set. Note that there is a spatial overlap for areas in different time periods but not for the same period. All images from the same time period are of different spatial areas.

Data Acquisition: HAIR consists of images from historical aerial photos used to develop the digital orthophoto covering the whole of Norway. The images can be accessed through a database of aerial imagery of the mainland of Norway (www.norgeibilder.no) covering both recent and historic photos that is maintained by the Norwegian mapping authority. The earliest images are from 1937 and the most recent are from 2022. A large backlog of historic aerial photos of the whole of Norway are being digitized and will be shared in the database continuously after being completed. All images are georeferenced in the database. The images in HAIR are projected into EUREF89-UTM33N.

The aerial photos are taken at different times in the summer season when there is no snow, but also during varying periods of the vegetation season. The dataset includes a wide variety of optical

[4]Workshop on Data-centric AI at NeurIPS 2021: https://datacentricai.org

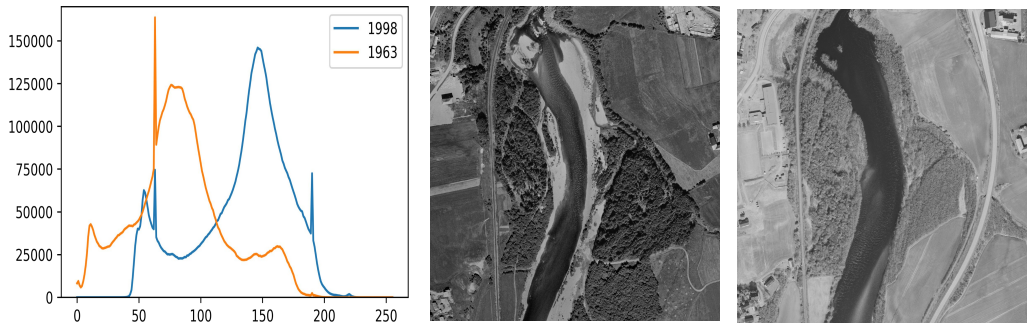


Figure 3: Intensity distribution of images covering the same section of Gaula from 1963 (left) and 1998 (right). It is calculated as the number of pixels with a given color (256 gray levels).

conditions including different shadow lengths, angels of sunlight and saturation. As mentioned by Boguszewski et al. [5], one would expect that this would make the dataset robust. However, the results on the OOD test data shows that it is not exactly the case.

Annotation: Traditionally, the most common annotation tools for this purpose is GIS software with polygon editing, such as QGIS [47]. However, polygon editing lacks the precision required for the high quality labels. For HAIR, annotations were made manually by the experts using Adobe Photoshop on an iPad using a pen, as this enables detailed annotations. Each large image of mainly 8000×6000 pixels was loaded as a layer to Adobe Photoshop. Annotations were done on top of the source image in a layer of its own. A specific color were assigned to each class, and the domain experts colored the six classes using the corresponding colors. Adobe Photoshop provided many tools that help facilitate the annotation. Magic Wand was, for example, used as a selection tool for most of the roads, and the Marching Ants algorithm [60] was used to modify the edges of the objects.

The experts followed a common procedure. Areas that were considered ambiguous by individual experts were discussed by the group to reduce the noise in the annotation. The annotation has been taken very seriously and is considered to be of high quality, although ambiguous examples can without doubt be found in such a large dataset. Figure 1 illustrates three out of the many different cases that were discussed by the experts.

Classes: We annotated the images with six different classes that can help understand the human pressure on river biodiversity and hydromorphology. These six classes are chosen pragmatically based on ease of manual annotation versus value of analyzing their change over time. More classes could have been added with high value, but these classes would have been even smaller than the gravel class with an even higher potential for misclassification.

Water (W): Water covered areas (not restricted to river).

Gravel (G): Gravel bars and point bars in the river – vegetation free.

Farmland (F): Farmland and cultivated land in the river corridor.

Vegetation (V): Forest and other vegetated areas in the riparian corridor.

Anthropogenic (A): Anthropogenic structures like houses and roads.

Unknown (U): Only areas that do not contain any aerial images are labeled as “unknown”.

Statistics for HAIR: The distribution of different classes in HAIR is shown in Figure 2 and indicates an imbalanced dataset where the two classes gravel and anthropogenic are underrepresented. The most common classes, vegetation and forest, cover 69.8% of the images. Gravel, which is the most important class besides water for the analysis, covers only 1.9% of the images while water covers 4.6%. Figure 3 shows a comparison of the the intensity distribution of images covering the same section of river Gaula from the years 1963 and 1998. The intensity distribution is calculated as the number of pixels with a given pixel color, where pixel color is one out of 256 gray levels.

4 Two Benchmark Tasks

We select U-Net [48], FPN [55], DeepLabV3+ [19], HRNet [61], Magnet [32] and Swin-Unet [11] as baselines to get a broad set of models for the benchmark. The selected models are evaluated on both the ID and OOD test sets.

4.1 Method

The segmentation of grayscale images is under-studied primarily due to the lack of large-scale datasets. In this benchmark, we try to shed some lights into greyscale segmentation of aerial images by testing a set of approaches to provide insights into these. One promising approach is to employ transfer learning, where a common dataset of colored images is converted to grayscale and used for pre-training selected models. To investigate that, we convert the widely used DeepGlobe dataset to grayscale and employ it to pre-train the baselines. We evaluate two different scenarios for all models: 1) all models are trained on HAIR only, and 2) models are first pre-trained on grayscale DeepGlobe dataset and then fine-tuned on HAIR.

To convert the images of the DeepGlobe dataset into grayscale, we use the luminance method [33], which is widely used in computer vision [6] and is implemented in several image processing software and libraries such as OpenCV [8]. In luminance method, the function $\mathcal{G}_{luminance}$ is defined as: $\mathcal{G}_{luminance} \leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$, where R, G and B represents the red, green and blue channels of the image[5]. Additionally for all the trainings, to leverage the strong feature extraction of encoders pre-trained on RGB Imagenet, we replicate the grayscale channel into the R,G, and B channels, transforming our single-channel input into a standard three-channel image format. ResNet50 pre-trianed on ImageNet is used as the encoder for most models, except for HRNet and Swin-Unet that are not pre-tained.

The hyperparameters used for training MagNet are the same as in [32]. For other models, we used the Hyperband algorithm [35] with a maximum of 100 epochs and 11 hybrid iteration to find the hyperparameters. Also, validation accuracy was used as the objective. More detail of the hyperparameters for each model is provided in the appendix. The code used to run the experiments are available in `https://github.com/SaeidShamsaliei/HAIR`.

4.2 Results

We calculate performance using MIoU of the experiments in the same way as Huynh et al. [32]. Results are presented in Tables 2 and 3 as mean $\pm$ standard deviation of model performance. Variation is introduced through training five seed replicates for each model. A seed replicate is a model trained using different seeds for the pseudo-random generator [7]. The performance is reported as the mean and standard deviation of the MIoU for the set of five seed replicates. As can be seen, DeepLabV3+ achieves the best performance on both datasets. However, it does not perform considerably better than the other models. Gravel is the class that all models struggle the most to identify, and forest is the class for which all models achieve their highest score. For the OOD test set, gravel is still the hardest class to predict for all models, while it is relatively easier for the models to predict the vegetation and farmland.

Pre-training on the converted DeepGlobe dataset led to a slight decrease in performance for all models except HRNet and Swin-Unet, which do not perform well in general. Although the limited size of the DeepGlobe dataset could be a factor, the results highlight the need for developing large scale historical grayscale aerial imagery datasets. Another factor that might be a cause of the comparatively poorer performance is the luminance method for converting DeepGlobe to grayscale. Kanan and Cottrell [33] report that the conversion method affects the performance of downstream tasks, which could easily be the case for deep learning as well as pre-processing is known to affect reproducibility of deep learning methods [26, 28].

[5]The method is used in many papers such as [6], however, in these works the name of method is not mentioned. We got the name from [33]. The coefficients are the central component of the method.

Table 2: The MIoU prediction of different semantic segmentation architectures on the ID test set. MIoU is mean ± standard deviation over five seed replicates runs

Models	Backbone	Not pre-trained on grayscale DeepGlobe						Pre-trained on grayscale DeepGlobe					
		IoU per class					MIoU	IoU per class					MIoU
		G	V	F	A	W		G	V	F	A	W	
U-Net	ResNet50	60.81	**88.02**	85.31	**65.70**	85.54	77.07 ± 00.48	58.19	**87.85**	85.01	**64.74**	84.67	76.10 ± 00.72
FPN	ResNet50	60.83	87.84	85.11	65.59	82.78	76.43 ± 01.84	59.40	87.40	84.75	64.70	84.05	76.06 ± 00.57
DeepLabV3+	ResNet50	**61.07**	87.92	**86.23**	65.43	**86.19**	**77.37 ± 01.02**	**60.63**	87.54	**85.14**	63.37	**84.76**	**76.28 ± 01.26**
HRNet	HRNetV2-W18	52.56	87.17	82.20	62.97	77.70	72.52 ± 00.97	56.08	87.04	82.15	64.34	77.87	73.50 ± 00.79
Swin-Unet	Swin Transformer	43.86	83.50	76.32	47.03	60.35	62.21 ± 00.40	44.41	83.43	76.31	49.79	61.38	63.06 ± 01.52
MagNet	FPN	51.75	77.55	76.55	57.75	62.51	65.22 ± 01.39	44.21	72.53	70.54	44.96	49.36	56.32 ± 01.12

Table 3: The MIoU prediction of different semantic segmentation architectures on the OOD test set. MIoU is mean ± standard deviation over five seed replicates.

Models	Backbone	Not pre-trained on grayscale DeepGlobe						Pre-trained on grayscale DeepGlobe					
		IoU per class					MIoU	IoU per class					MIoU
		G	V	F	A	W		G	V	F	A	W	
U-Net	ResNet50	**32.09**	**78.38**	**75.98**	61.20	73.20	64.17 ± 00.80	30.19	**76.49**	66.35	51.58	**75.86**	60.10 ± 01.03
FPN	ResNet50	29.24	72.43	72.73	56.39	69.78	60.11 ± 01.51	32.02	76.29	69.42	57.22	63.05	59.60 ± 01.64
DeepLabV3+	ResNet50	30.55	76.93	73.83	**61.57**	**78.57**	**64.29 ± 02.66**	**32.27**	72.93	**75.43**	**57.43**	75.34	**62.68 ± 02.81**
HRNet	HRNetV2-W18	22.27	67.03	60.27	50.00	68.17	53.44 ± 02.46	22.09	66.84	55.55	50.95	62.65	51.61 ± 02.63
Swin-Unet	Swin Transformer	14.29	62.43	50.99	35.12	50.86	42.74 ± 00.84	14.40	62.71	51.02	35.58	52.09	43.16 ± 01.82
MagNet	FPN	31.65	64.33	74.00	54.28	61.01	57.06 ± 03.84	21.01	53.59	54.31	38.88	40.09	41.58 ± 04.51

Table 4: The MIoU of the DeepLabV3+ on the two different rivers in the OOD test set and Gaula 1963 in ID test set.

River	Year	MIoU	
		Not pre-trained	Pre-trained
Orkla	1962	69.64 ± 03.06	69.17 ± 01.58
Gaula	1998	61.62 ± 03.21	59.43 ± 03.70
Gaula	1963	79.83 ± 01.19	80.25 ± 00.26

The performance of MagNet experiences a relatively larger decline, compared to other models, when pre-trained on the grayscale DeepGlobe dataset. The combination of model's complex architecture and relatively small scale of DeepGlobe used for pre-training, seems to affect the performance to a larger extent. The backbones of HRNet and Swin-Unet were not pre-trained on any large-scale dataset to help with extracting the low level features, like ImageNet. This could be a contributing factor to their relatively lower performance.

All models perform substantially worse on the OOD test set compared to the ID test set. DeepLabV3+ generalizes better than the other models, but only slightly better than U-Net when not pre-trained. Table 4 shows how DeepLabV3+ performs on the two different types of OOD data and that it generalizes better to data from the same time-period where the same or a similar camera technology is used than for the images captured more than 35 years later. The better performance of DeepLabV3+ on OOD might be explained by how Atrous Spatial Pyramid Pooling (ASPP) mechanism fuses the extracted features from the input image. Since ASPP increases the field of view of the model, it can efficiently access a larger context of the input image.

4.3 Discussion

Here we present a short discussion of how the four characteristics of historical aerial images can affect the results.

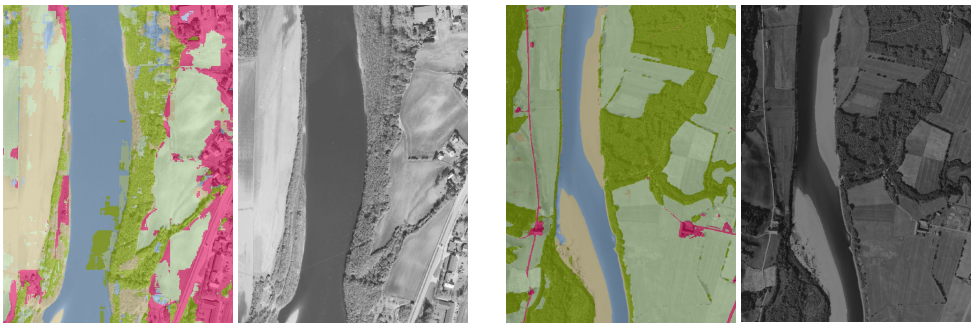


Figure 4: Comparison of predictions for Gaula 1998 (left) and 1963 (right).

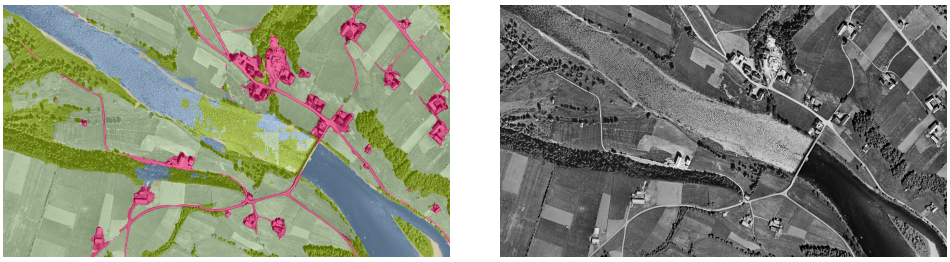


Figure 5: Different light conditions make accurate prediction difficult.

Characteristic #1 - Camera technology: The results on the OOD test set indicate that camera technology is an issue. The performance of the models drop considerably for the images of Gaula 1998 compared to Orkla 1962. The Gaula 1998 were captured 35 years after the Orkla 1962 images. However, DeepLabV3+ performs much better when tested on the ID images of Gaula from 1963, as can be seen in Table 4. This indicates that it is not river Gaula itself that is challenging, but that it is the development of camera technology that is the cause of the performance drop. While intensity differs between the river in the ID test set compared to the older images in the OOD dataset, as seen in Figures 3 and 4, light intensity alone should not be a problem as contrast and brightness were changed randomly as part of the training, as described in methodology. More research on data augmentation methods designed specifically for grayscale aerial images might alleviate this issue. Examples include modification of brightness and contrast during the training [52, 64], blurring filters, histogram transformations, special optical and lens distortion [37], and generative models [56].

Characteristic #2 - Lighting conditions: Light conditions make the prediction more challenging. In Figure 5, the reflections in the shimmery section represent an unusual visual effect on the water. Because this effect happens only when the light from the sun hits the water and is reflected in the camera, there are relatively few examples of such effects in the training data and little data for the model to learn the pattern. Challenge #1 covers this by suggesting that more research on online augmentations and generative models, designed specifically for historical grayscale aerial images, could mitigate this issue.

Characteristic #3 - Class imbalance: Figure 2 shows the imbalanced class distribution, and the results presented in Table 2 and Table 3 indicate that the models have the worst performance on smallest classes. The class distribution clearly poses a problem. More research on effective methods to mitigate the class imbalance of high resolution aerial images, such as more efficient sampling methods [58] and more effective loss functions [23] could help overcoming this challenge.

Characteristic #4 - Grayscale: Texture and context become the main sources of information for grayscale images. Land covers that are easily distinguishable when having color information become hard to distinguish in grayscale. Figure 6 shows several cases where texture alone makes

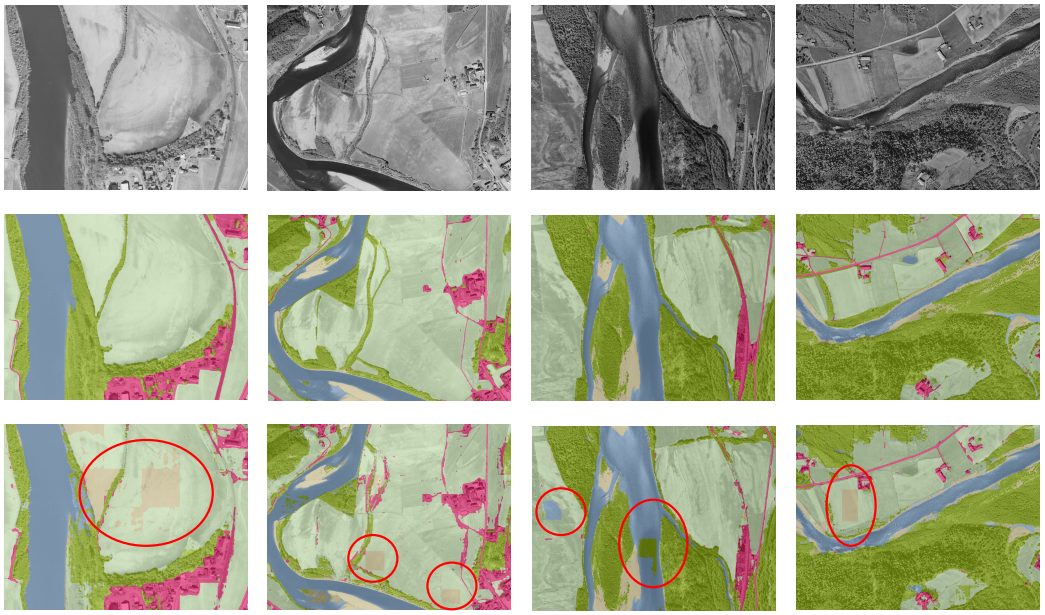


Figure 6: Source images (top row), labels (middle-row) and best predictions (bottom row).

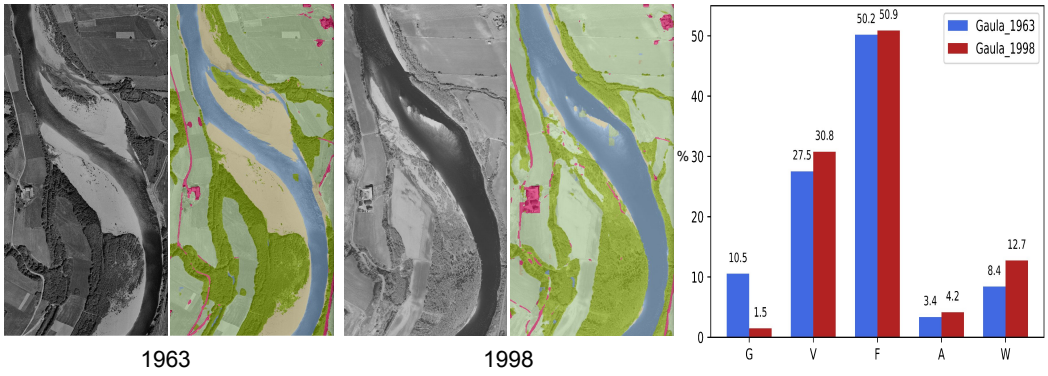


Figure 7: A section of River Gaula from 1963 and 1998 with distribution of classes in percentage.

the prediction hard. The deep learning segmentation methods typically divide the larger images into smaller patches and run these through the deep nets. Context is the underlying issue when the method miss-classifies two patches that clearly are of the same class. Examples of this can easily be spotted in Figure 6 where patches of the wrong class can be seen very clearly in larger regions that are mostly predicted correctly. These errors happen when there are examples of patches in the test set that have similar texture but different class than patches in the training data. It is easy to imagine that color information would help solve these cases. We believe that continued development of architectures similar to MagNet and GLNet that use different scales of images to provide the context to smaller patches could mitigate this issue.

5 Future Work and Conclusions

Manual analyses of aerial images of riverscapes is the standard method for understanding the development of individual rivers [29, 66], and they include spatio-temporal dynamics of parameters such as river width, bank cover, sinousity and other geomorphological features. Figure 7 shows raw images and segmentations of the years 1963 and 1998 as well as a bar chart that enables us to perform a basic version of such an analysis on the current dataset. The images show that the reduction in gravel is substantial and so is the increase in water. The bar chart confirms not only this but that all classes except for gravel increase. The analysis could indicate that the conditions for the plants and insects living in gravel has worsened and that their existence is threatened. However, as these images could have been taken at different seasons we cannot draw this conclusion so easily. Hence, we need to semantically segment images of River Gaula for more years and the findings for one river must be compared to findings in other rivers to draw more certain conclusions. This is an example for a large scale analysis. However, such large scale analyses covering multiple rivers have not been done before. To achieve this, it is essential to automate the land cover classification of historical aerial images [2], and the segmentation model must generalize well to OOD data. Given that such land area cover maps made by segmentation models exist for multiple rivers over a long period of time (1940's through 2020's), a completely new type of analysis can be done.

In this paper, we present a free and open dataset for advancing semantic segmentation of historic, high resolution, grayscale aerial images of land cover. We provide baselines for selected state-of-the-art deep learning models on the two benchmark tasks and show several issues with applying state-of-the-art models developed for color images on grayscale images. It is clear that the training is not robust and that the models do not generalize well beyond the training data. Pre-training on a recent dataset of converted satellite images is shown to not improve the performance of the models, which highlights the importance of generating large scale datasets of historical aerial images. Additionally, more research can be done to improve the effectiveness of pre-training on low resolution RGB datasets to improve semantic segmentation of high resolution grayscale images. For example, one can explore the impact of converting the aerial resolution of HAIR images (e.g., from 20 cm per pixel to 50 cm per pixel as in DeepGlobe images) to determine whether the conversion would improve when using models pre-trained on lower resolution datasets. Finally, we identify and show how four characteristics of historical aerial images negatively affect the performance of the tested models.

Grayscale aerial images are a potentially important data source that has not received much attention. Our hope is that bringing attention to this data source will support the development of generalizable models that can overcome the challenges of the dataset and be used to perform large scale land-mapping of historical grayscale aerial images accurately. This will enable longitudinal studies and quantified analyses of land use, and support large scale investigations into which practices are sustainable and which are not. Such analyses require far too much manual work of experts to be feasible without automatic solutions.

Acknowledgments and Disclosure of Funding

Thanks to Statkart-Geovekst for supporting the sharing of the enhanced dataset publicly under the Creative Commons 4.0 SA-BY license. J. H. H. was funded by the Norwegian Research Council through the OFFPHD program, Grant No: 289725 – Ecosystem-based management.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng.

TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Knut Alfredsen, Arild Dalsgård, Saeid Shamsaliei, Jo Halvard Halleraker, and Odd Erik Gundersen. Towards an automatic characterization of riverscape development by deep learning. *River Research and Applications*, 38(4):810–816, December 2021.

[3] Knut Alfredsen, Christian Haas, Jeffrey A Tuhtan, and Peggy Zinke. Brief communication: Mapping river ice using drones and structure from motion. *The Cryosphere*, 12(2):627–633, 2018.

[4] Bjørn T Barlaup, Sven Erik Gabrielsen, Helge Skoglund, and Tore Wiers. Addition of spawning gravel—a means to restore spawning habitat of atlantic salmon (salmo salar l.), and anadromous and resident brown trout (salmo trutta l.) in regulated rivers. *River Research and Applications*, 24(5):543–550, 2008.

[5] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. 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*, pages 1102–1110, June 2021.

[6] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. Ieee, 2007.

[7] Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pages 725–734. PMLR, 2019.

[8] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[9] Hardy Buller. personal communication with Norwegian Mapping Authority, Aug. 14 2023.

[10] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020.

[11] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[12] Patrice E Carbonneau, Stephen J Dugdale, Toby P Breckon, James T Dietrich, Mark A Fonstad, Hitoshi Miyamoto, and Amy S Woodget. Adopting deep learning methods for airborne rgb fluvial scene classification. *Remote Sensing of Environment*, 251:112107, 2020.

[13] Patrice E. Carbonneau and Hervé Piégay. Introduction: The growing use of imagery in fundamental and applied river sciences, August 2012.

[14] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021.

[15] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017.

[17] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[18] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[19] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[20] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. 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, June 2019.

[21] Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander G. Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier, Greg Rose, David Wilson, Adrian Tudor, Naira Hovakimyan, Thomas S. Huang, and Honghui Shi. Agriculture-vision: A[jenssen] large aerial image database for agricultural pattern analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2825–2835, 2020.

[22] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. 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, June 2018.

[23] Rongsheng Dong, Xiaoquan Pan, and Fengying Li. Denseu-net-based semantic segmentation of small objects in urban remote sensing images. *IEEE Access*, 7:65347–65356, 2019.

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[25] Kurt D Fausch, Christian E Torgersen, Colden V Baxter, and Hiram W Li. 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, 2002.

[26] Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–49, 2021.

[27] Sarah E Gergel and Monica G Turner. *Learning landscape ecology: a practical guide to concepts and techniques*. Springer, 2017.

[28] Odd Erik Gundersen, Kevin Coakley, Christine Kirkpatrick, and Yolanda Gil. Sources of irreproducibility in machine learning: A review, 2023.

[29] Angela M Gurnell. 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, 1997.

[30] F Richard Hauer, Harvey Locke, Victoria J Dreitz, Mark Hebblewhite, Winsor H Lowe, Clint C Muhlfeld, Cara R Nelson, Michael F Proctor, and Stewart B Rood. Gravel-bed river floodplains are the ecological nexus of glaciated mountain landscapes. *Science Advances*, 2(6):e1600026, 2016.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16755–16764. Computer Vision Foundation / IEEE, 2021.

[33] Christopher Kanan and Garrison W Cottrell. Color-to-grayscale: does the method matter in image recognition? *PloS one*, 7(1):e29740, 2012.

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[35] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

[36] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv preprint arXiv:1506.04579*, 2015.

[37] Yan Liu, Qirui Ren, Jiahui Geng, Meng Ding, and Jiangyun Li. Efficient patch-wise semantic segmentation for large-scale remote sensing images. *Sensors*, 18(10):3232, 2018.

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2015.

[39] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. Dirs: On creating benchmark datasets for remote sensing image interpretation. *CoRR*, abs/2006.12485, 2020.

[40] Thomas R Loveland, Bradley C Reed, Jesslyn F Brown, Donald O Ohlen, Zhiliang Zhu, LWMJ Yang, and James W Merchant. Development of a global land cover characteristics database and igbp discover from 1 km avhrr data. *International journal of remote sensing*, 21(6-7):1303–1330, 2000.

[41] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017.

[42] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[44] Hervé Piégay, Fanny Arnaud, Barbara Belletti, Mélanie Bertrand, Simone Bizzi, Patrice Carbonneau, Simon Dufour, Frédéric Liébault, Virginia Ruiz-Villanueva, and Louise Slater. Remotely sensed rivers in the anthropocene: State of the art and prospects. *Earth Surface Processes and Landforms*, 45(1):157–188, 2020.

[45] Francesca Pilotto, Christer Nilsson, Lina E Polvi, and Brendan G McKie. First signs of macroinvertebrate recovery following enhanced restoration of boreal streams used for timber floating. *Ecological Applications*, 28(2):587–597, 2018.

[46] Priyanka, Sravya N, Shyam Lal, J Nalini, Chintala Sudhakar Reddy, and Fabio Dell'Acqua. DIResUNet: Architecture for multiclass semantic segmentation of high resolution remote sensing imagery data. *Applied Intelligence*, March 2022.

[47] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2009.

[48] Alexander Rakhlin, Alex Davydow, and Sergey Nikolenko. 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*, pages 262–266, 2018.

[49] Rémi Ratajczak, Carlos Fernando Crispim-Junior, Elodie Faure, Béatrice Fervers, and Laure Tougne. Automatic land cover reconstruction from historical aerial images: An evaluation of features extraction and classification algorithms. *IEEE Transactions on Image Processing*, 28(7):3357–3371, 2019.

[50] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[51] Max Roser, Hannah Ritchie, and Esteban Ortiz-Ospina. World population growth, 2013. https://ourworldindata.org/world-population-growth.

[52] Dimitrios Sakkos, Hubert PH Shum, and Edmond SL Ho. Illumination-based data augmentation for robust background subtraction. In *2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pages 1–8. IEEE, 2019.

[53] Mohamed Samy, Karim Amer, Kareem Eissa, Mahmoud Shaker, and Mohamed ElHelw. 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, June 2018.

[54] Selim Seferbekov, Vladimir Iglovikov, Alexander Buslaev, and Alexey Shvets. Feature pyramid network for multi-class land segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 272–2723, 2018.

[55] Selim Seferbekov, Vladimir Iglovikov, Alexander Buslaev, and Alexey Shvets. Feature pyramid network for multi-class land segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 272–275, 2018.

[56] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[57] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, October 2021.

[58] Antonio Tavera, Edoardo Arnaudo, Carlo Masone, and Barbara Caputo. Augmentation invariance and adaptive sampling in semantic segmentation of agricultural aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1656–1665, 2022.

[59] Robin L. Vannote, G. Wayne Minshall, Kenneth W. Cummins, James R. Sedell, and Colbert E. Cushing. The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences*, 37(1):130–137, January 1980.

[60] Alberto Viseras, Rafael Ortiz Losada, and Luis Merino. Planning with ants. *International Journal of Advanced Robotic Systems*, 13(5):172988141666407, September 2016.

[61] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[62] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12077–12090, 2021.

[63] Pavel Yakubovskiy. Segmentation models. `https://github.com/qubvel/segmentation_models`, 2019.

[64] Bo Yang, Kaiyong Xu, Hengjun Wang, and Hengwei Zhang. Random transformation of image brightness for adversarial attack. *Journal of Intelligent & Fuzzy Systems*, (Preprint):1–12, 2022.

[65] Yuhui Yuan, Xiaokang Chen, Xilin Chen, and Jingdong Wang. Segmentation transformer: Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019.

[66] Luca Zanoni, Angela Gurnell, Nick Drake, and Nicola Surian. Island dynamics in a braided river from analysis of historical maps and air photographs. *River Research and Applications*, 24(8):1141–1159, 2008.

[67] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Limitations to size of dataset and benchmarking methodology were discussed in the text.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We mentioned that we do not see any negative social impacts to our work. On the contrary we see only potential in positive change of ecosystem restoration policies.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code, data and datasheet can be found here: https://folk.idi.ntnu.no/odderik/HAIR/ All above will be shared on a website and through Zenodo if accepted with a DOI.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We did only run experiments with five random seeds. Results only for comparison purposes, not main deliverable.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This is shared in the appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] All models used for benchmarking have been cited.

(b) Did you mention the license of the assets? [Yes] We share them using CC 4.0 BY-SA

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] These are land area images, which do not contain such information.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

Include extra information in the appendix. This section will often be part of the supplemental material. Please see the call on the NeurIPS website for links to additional guides on dataset publication.

1. Submission introducing new datasets must include the following in the supplementary materials:
   (a) Dataset documentation and intended uses. Recommended documentation frameworks include datasheets for datasets, dataset nutrition labels, data statements for NLP, and accountability frameworks.
   (b) URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers. *URL: https://https://folk.idi.ntnu.no/odderik/HAIR/hair.zip*
   (c) Author statement that they bear all responsibility in case of violation of rights, etc., and confirmation of the data license. *We confirm that we bear all responsibility in case of violations of rights. We confirm that the data will be shared under CC 4.0 BY-SA.*
   (d) Hosting, licensing, and maintenance plan. The choice of hosting platform is yours, as long as you ensure access to the data (possibly through a curated interface) and will provide the necessary maintenance. *Code is shared on GitHub and data will be shared through Zenodo. a website will be made to host all information about the data. The website will also link to new versions of the data set if new ones are made.*
2. To ensure accessibility, the supplementary materials for datasets must include the following:
   (a) Links to access the dataset and its metadata. This can be hidden upon submission if the dataset is not yet publicly available but must be added in the camera-ready version. In select cases, e.g when the data can only be released at a later date, this can be added afterward. Simulation environments should link to (open source) code repositories. *Data: https://folk.idi.ntnu.no/odderik/HAIR/hair.zip, code: https://github.com/SaeidShamsaliei/HAIR*
   (b) The dataset itself should ideally use an open and widely used data format. Provide a detailed explanation on how the dataset can be read. For simulation environments, use existing frameworks or explain how they can be used. *The image data and annotations are stored as both tiff and png files, which are standard formats that are supported by basically all machine larning libraries.*
   (c) Long-term preservation: It must be clear that the dataset will be available for a long time, either by uploading to a data repository or by explaining how the authors themselves will ensure this. *We share the data through Zenodo.org, which will store the data safely for the future in CERN's Data Centre for as long as CERN exists.*
   (d) Explicit license: Authors must choose a license, ideally a CC license for datasets, or an open source license for code (e.g. RL environments). *We have chosen Creative Commons 4.0 BY-SA.*
   (e) Add structured metadata to a dataset's meta-data page using Web standards (like schema.org and DCAT): This allows it to be discovered and organized by anyone. If you use an existing data repository, this is often done automatically. *This will be done.*
   (f) Highly recommended: a persistent dereferenceable identifier (e.g. a DOI minted by a data repository or a prefix on identifiers.org) for datasets, or a code repository (e.g. GitHub, GitLab,...) for code. If this is not possible or useful, please explain why. *DOIs are supported by zenodo.org and we will get one when the dataset is stored there.*
3. For benchmarks, the supplementary materials must ensure that all results are easily reproducible. Where possible, use a reproducibility framework such as the ML reproducibility checklist, or otherwise guarantee that all results can be easily reproduced, i.e. all necessary datasets, code, and evaluation procedures must be accessible and documented.
4. For papers introducing best practices in creating or curating datasets and benchmarks, the above supplementary materials are not required.

B Runtime Environment

All experiments were run on one node the Yoda cluster of Computer Science department of Norwegian University of Science and Technology. The information of this node is described below:

GPU: NVIDIA Tesla V100

CPU: Intel(R) Xeon-Gold 6240

Number of Cores: 18 cores @ 2.6 Ghz

RAM: 32 GiB

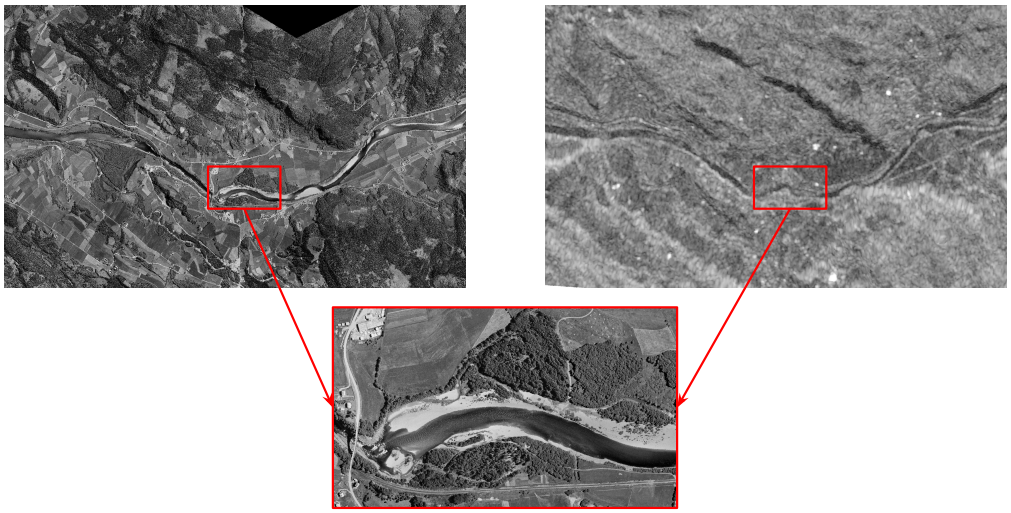


Figure 8: A visual comparison of HAIR images of Gaula captured in 1963 (top left) and SAT image of Gaula captured by Sentinel-1 in 2015 (top right). A subsection of HAIR images of Gaula 1963 is magnified to illustrate the high resolution of the HAIR images.

Models were implemented using Tensorflow [1] and Pytorch [43] packages. In addition, Albumentations library [10] was used for Online Augmentation, and SegmentModel [63] library was used in some of the implementations. In order to train the MagNet [32], the script provided by the paper's github page was used.

C Hyperparameters of benchmark models

The batch size for the models is 16, except for HRNet, Swin-Unet and backbone of MagNet, which is 12, and MagNet refinement module, which is 8. Models were trained using a weighted categorical cross entropy loss function to mitigate the class imbalance of the dataset. For training the MagNet, stochastic gradient decent (SGD) is used for with a weight decay of 0.9, For the other architectures the Adam optimizer [34] is used.

The ReduceLROnPlateau algorithm was applied in Adam to reduce the learning rate by a factor of 0.5 if value loss did not decrease for more than 5 epochs. Except for MagNet, L2 regularization is used for convolutional layers. The FPN backbone of the MagNet is pretrained on DeepGlobe dataset and the output layer is changed to have 6 outputs instead of 7. MagNet is trained for 484 epochs while the other architectures are trained until convergence .

Gradient clipping with clipping value of 0.5 is applied for training the Swin-Unet. During training, images were randomly flipped and transposed, and their contrast and brightness were randomly transformed with the changing factor set to 0.1, and the probability of applying the changes was set to 20%. For training the MagNet, images were sampled as 2448×2448 px patches, and the rest are sampled as 512×512 px patches. To sample the patches, each large image is first divided into smaller patches with no overlap. Afterwards, images were flipped and patches with center pixel of gravel and water were added to the training set. The input size of MagNet used is the same as in the original paper for DeepGlobe. For the rest of the models, 512×512 is used as the input size. For inference, large images were divided into patches with the same size used to train the corresponding model, and each of these patches was used for inference. Due to low GPU memory, we did not used object-contextual representation for HRNet [65].

D Visualization of the Synthetic Aperture Radar (SAR) images

As illustrated in the Figure 8, the resolution of the SAR images from Sentinel-1 are comparably lower than the aerial images in the HAIR dataset. Additionally, Sentinel-1 started the mission in 2014, so evolution of the landscape during previous years is not captured.