Journal of Data-centric Machine Learning Research (2024) Submitted 1/24; Revised 3/24; Published 4/24

Highlighting Challenges of State-of-the-Art Semantic Segmentation with HAIR - A Dataset of Historical Aerial Images

Saeid Shamsaliei SAEID.SHAMSALIEI@NTNU.NO
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway

Odd Erik Gundersen ODDERIK@NTNU.NO
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway

Knut Alfredsen KNUT.ALFREDSEN@NTNU.NO
Norwegian University of Science and Technology
Trondheim, Norway

Jo H. Halleraker JO.HALLERAKER@NTNU.NO
Norwegian University of Science and Technology
Trondheim, Norway

Anders Foldvik ANDERS.FOLDVIK@NINA.NO
Norwegian Institute for Nature Research Trondheim, Norway

Reviewed on OpenReview: *https://openreview.net/forum?id=kPyc520hj2*

Editor: Joaquin Vanschoren

Abstract

We present HAIR, the first dataset of expert-annotated historical aerial images covering different spatial regions spanning several decades. Historical aerial images are a treasure trove of insights into how the world has changed over the last hundred years. Understanding this change is especially important for investigating, among others, the impact of human development on biodiversity. The knowledge contained in these images, however, has not yet been fully unlocked, as this requires semantic segmentation models that are optimized for this type of data. Current models are developed for modern color images, and they do not perform well in historical data that is typically in grayscale. Furthermore, there is no benchmark historical grayscale aerial data that can be used to develop specific segmentation models for it. We here assess the issues of using semantic segmentation models designed for modern color images in historic grayscale data, and introduce HAIR as the first benchmark dataset of large-scale historical aerial grayscale images. HAIR contains $\approx 9 \times 10^9$ pixels of high-resolution aerial land images covering the years within the period 1947 - 1998, with detailed annotations performed by domain experts. By using HAIR, we show that pre-training on modern satellite images converted to grayscale does not improve the performance compared to training only on historic aerial grayscale data, stressing the relevance of using actual historical and grayscale aerial data for these studies. We further

show that state-of-the-art models underperform when trained on grayscale data compared to using the same data in color, and discuss the challenges faced by these models when applied directly to aerial grayscale data. Overall, HAIR appears as a powerful tool to aid in developing segmentation models that are able to extract the rich and valuable information from historical grayscale images.

Keywords: Historical Aerial Images, Semantic Segmentation, Grayscale Imagery, Dataset Benchmarking

1 Introduction

Historical aerial images is a treasure trove of insights into how the world has changed over the last hundred years, as they have been captured systematically since the early 1900s (The U.S. National Archives and Records Administration, 2023; Historic Environment Scotland, 2023). For example, since then the human population has grown from 1.65 Billion to 7.9 Billion in 2022 (Roser et al., 2013). Population growth has led to an in creasing pressure on all ecosystems on the Earth. The pressure is now so severe that UN has declared the decade from 2021 to 2030 as the UN Decade of Restoration. The initiative is described as *"a rallying call for the protection and revival of ecosystems all around the world"*. Furthermore, in December 2022, 188 governments signed the Kunming Montreal Global Biodiversity Framework (Stephens, 2023). It states: *"Raising awareness on the critical role of science, technology and innovation to strengthen scientific and technical capacities to monitor biodiversity, address knowledge gaps and develop innovative solutions to improve the conservation and sustainable use of biodiversity."* This work is one of hopefully many initiatives to address the knowledge gaps and support the development of innovative solutions for addressing efficient restoration of ecosystems. Understanding the changes in the landscape is crucial for sustainable development (Mercuri and Florenzano, 2019). For example, rivers and their surrounding landscape have changed dramatically over the years (Grill et al., 2019; Belletti et al., 2020), which has consequences for the riparian and aquatic ecosystems and, consequently, ecosystem services provided by rivers (Petsch et al., 2023). Understanding how rivers have evolved over time is important for the planning of river restoration (Wohl, 2019) with a goal to recover the functions of river systems lost through anthropogenic activity. Imagery can be used in the assessment of historic conditions (Morgan et al., 2010), important in defining a baseline for the evaluation of environmental changes and for developing working restoration measures (Guzelj et al., 2020). Channel narrowing and vegetation encroachment on rivers have been seen as responses to flow regulation and river training works (Liébault and Piégay, 2002), and this can influence natural habitats for fish and invertebrates. Loss of gravel bars and a larger containment of the river channel can lead to loss of species with specific habitat requirements (Åström et al., 2017). The development of floodplains and increased vegetation can influence flood levels and economic losses during high flood events. Aerial images have played a major role in capturing the aforementioned population growth and its effects, and they are extremely valuable for many other studies spanning into the past. In contrast, satellite images have only been captured systematically since the 1990s (Loveland et al., 2000).

The data found in historical aerial images is important for reconstructing river development over time but is still mainly based on time-consuming manual delineation of features (Piégay et al., 2020). Therefore, automatic mapping would advance the utilization and

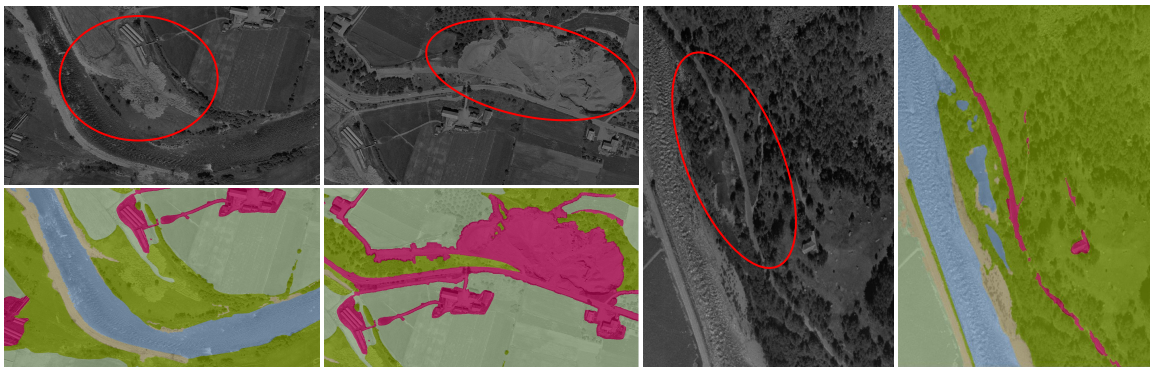


Figure 1: Ambiguous areas are marked in red. *Left* and *middle* show cases where gravel and human construction can be confused while *right* is an example where vegetation and water could be confused.

potential of using such images. Image recognition, specifically semantic segmentation, can be utilized for this mapping, but this technique can encounter a number of problems when applied to historical aerial images: (i) First, semantic segmentation relies on the availability of a large dataset of images annotated into desired habitats, but there are currently no large and carefully annotated datasets of historical aerial images. (ii) State-of-the-art segmentation models are designed for color images, but the historical data is typically in grayscale. Only panchromatic (grayscale) historic photographs are available for aerial images captured before the 2000s (Gergel and Turner, 2017). (iii) Modern segmentation models do not account for possible effects related to changes in camera technology over time, which are abundant in past data. A majority of the historical aerial images are captured using an analog film-based camera. The captured films are then scanned and converted into a digital format (Buller, 2023).

In this paper, we will assess in detail the difficulties faced by state of the art semantic segmentation models when applied to historical grayscale data through a number of experiments. Furthermore, to allow these and future analysis with historical aerial grayscale images, we introduce HAIR, the first benchmark dataset of historical aerial grayscale images. HAIR only contains grayscale images, as historical images are only available as panchromatic photographs. For semantic segmentation models to perform reliably on historical aerial images, they should extract information from these grayscale images. As the first benchmark dataset of historical aerial grayscale images, HAIR enables more research on semantic segmentation of historical aerial images. It is worth mentioning that in this work, we do not focus on a dataset that can directly be used for the analysis of changes in the landscape over time; instead, we focus on semantic segmentation of historical aerial images captured before the 2000s that plays as a bottleneck in studying the changes in the landscape. HAIR is the first annotated large-scale dataset for semantic segmentation of historical aerial imagery of land cover, and it is characterized by four aspects that may challenge segmentation models when applied to it: 1) as mentioned above, camera technology has advanced significantly over time, which results in HAIR presenting a varying image quality based on when images are captured, 2) the images are also affected by lightning conditions that are influenced

by factors such as the time of the day and the airplane's direction during the capture, 3) HAIR also shows class imbalance, as some classes are underrepresented due to the nature of the aerial images, and finally 4) the images are in grayscale, which means that they carry less information than satellite images including RGB, as well as sometimes infra red, channels. In addition to experiments comparing models and learning strategies to improve model performance, we will discuss the effect of these four characteristics when training the the segmentation models with HAIR images.

Contributions:

- We show the inability of using current satellite color image datasets for training segmentation models targeting aerial grayscale historic data analysis. This highlights the value of having an actual dataset for extracting useful information from historical aerial grayscale images.
- We release HAIR[1], the first dataset of high-resolution, historical aerial images with high-quality annotations of riverscapes made by experts. The dataset will be released under the CC BY-SA 4.0 license[2] and contains roughly 8.72 billion annotated pixels surpassing widely recognized land cover datasets, such as DeepGlobe (Demir et al., 2018), and is on par with Inria (Maggiori et al., 2017) . HAIR covers the years between 1947 - 1998, which is the largest time span covered by a high-resolution historical dataset to date. Images in HAIR are in grayscale. All these features make HAIR the first state-of-the-art dataset for historical aerial image data analysis [3].
- We introduce an out-of-distribution (OOD) test set to evaluate how segmentation models generalize on spatial and temporal changes in the data.
- We compare various state-of-the-art segmentation models on HAIR to provide future established references. Also, we assess how four main properties specific to historical aerial grayscale data impact the performance of the models and how these can be addressed for improvement in future studies.

2 Problem Description

Historical aerial images cover a spatial region captured at a given time. We define $\mathcal{X}$ to be the set of all historical aerial images of landscapes, such that:

$$x \in \mathcal{X}, x \in \mathbb{R}^{N \times W \times H} \tag{1}$$

where N is the number of channels, which would be 3 for RGB and 1 for grayscale images, W is the image width in pixels, and H is the image height in pixels. Every image x is a snapshot from a spatial region $s \in S$, and is captured at a time point $t \in T$, so that $x_i^{s,t}$ is image i of region s at time t, and $\mathcal{X}^{S,T}$ is the set of all images covering all regions for any possible time periods. A region could be specified very concretely as, for example, GPS

1. Because of its size (20GB), HAIR is not uploaded in supplementary materials.
2. https://creativecommons.org/about/cclicenses/
3. The dataset and benchmarks are publicly available at `https://riverscapes.ai`.

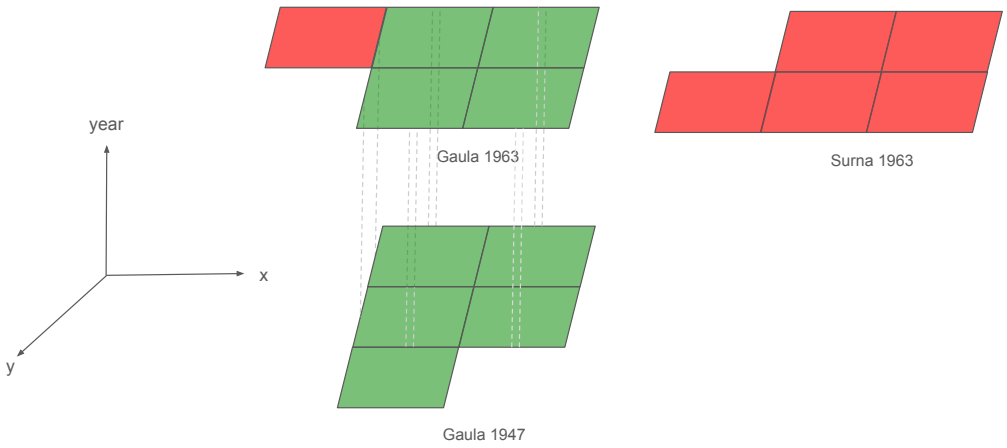


Figure 2: A simplified depiction of how images in our dataset are distributed in space and time. Each rectangle in the x-y plane represents an aerial image.

coordinates or more loosely as a name of an area, such as New York city or the Mississippi river. Similarly, time could be defined using any unit of time, for example the year of capture. Therefore, $\mathcal{X}^{NY,1963}$ is the set of all historical aerial images of New York city from 1963, while $x_1^{NY,1963}$ is the specific image $i = 1$ from the set of New York city images from 1963.

To analyze how a region of land has changed over time, each pixel p of an image, x_i, must be mapped into a desired class $c \in C$. The result of this mapping can be defined as $y_i = C \times W \times H$, and Y represents the set of mappings for all historical aerial images of landscapes that have been semantically segmented. The function $f(x) = y$ is a semantic segmentation function that performs this mapping so that each image $x_i^{s,t}$ has a corresponding mapping $y_i^{s,t}$. The function f can be estimated through supervised learning on a subset $A^{S,T} \subset \mathcal{X}^{S,T}$ of all images for which we already have the mapping. To evaluate how well the estimated function f generalize, we set aside a subset of the data, typically 10-20%, for testing $I^{S,T} \subset A^{S,T}$. The test set $I^{S,T}$ is an in-distribution dataset (IID), which if sampled randomly should reflect the distribution of the training data.

However, given that the goal is to perform semantic segmentation on other images in $\mathcal{X}^{S,T}$ than those in $A^{S,T}$, f must generalize beyond the training data. For example, one could envision that one would use a mapping function, f, that is trained on aerial images of New York in 1963, $\mathcal{X}^{NY,1963}$, to map aerial images of New York from 1973, $\mathcal{X}^{NY,1973}$, or to map aerial images of Philadelphia from 1963, $\mathcal{X}^{P,1963}$. Both sets are out-of-distribution compared to the one that the mapping function was trained on. The set $\mathcal{X}^{NY,1973}$ is *temporally different* while the set $\mathcal{X}^{P,1963}$ is *spatially different*. Generally, one would expect that aerial images that are captured close both temporally and spatially would be more similar than ones that are captured further away along one or two of the dimensions. Both Philadelphia and New York are big cities in USA, so we would expect them to be more similar than i.e. aerial images from the desert of Sahara. While aerial images from different years of the same region are expect to be somewhat similar as landscapes, usually, changes slowly,

the camera technology will also change over time, so that the aerial images will change subtly even though the landscapes are quite similar. Hence, images of New York from 1973 are expected to be more similar to images from 1963 than images from 2023, as camera technology has improved so much over time. To evaluate the generalization ability of the function f, we design a set $O^{S,T}$, containing images that are either spatially or temporally different than $A^{S,T}$. The test set $O^{S,T}$ is an out-of-distribution dataset (OOD).

In this work we focus on riverscapes as a subset of all the possible landscapes. In other words, we consider a selection of landscapes of riverscapes as the $A^{S,T}$ set. Figure 2 illustrates a simplified overview of how images in our dataset are distributed in space and time. Images of rivers $A^{Gaula,1947}$ and $A^{Gaula,1963}$ are temporally different, while images of rivers $A^{Gaula,1963}$ and $A^{Surna,1963}$ have the same temporal characteristics and are spatially different.

Estimating a function that can generalize well on historical aerial images of riverscapes is challenging, as these images cover various areas and are captured on different times. For evaluation of the estimated function, we designed two test sets, IID and OOD, in the same way as described above. To make the IID test set, we randomly sample approximately 10% of the $A^{S,T}$. The rest of 90% of $A^{S,T}$ can be used for training and validation of the function we aim to estimate. To evaluate the generalization ability of the estimated functions, we designed an OOD test set. The OOD test set consists of two separate sets, a set of temporally different images, and another set of spatially different images.

The OOD test set takes into account the shift in geographical location, time of the capture, and camera technologies used for capturing of the images. Our definition of OOD is similar to the one provided by Hendrycks et al. (2021), which considers images with different geographical locations, time of capture, and camera technology used for capturing the image, etc., to be the OOD set. The section 4 describes our dataset, HAIR, in detail.

3 Related work

Since the introduction of FCN (Long et al., 2015), 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 (Ronneberger et al., 2015), ParseNet (Liu et al., 2015), PSPNet (Zhao et al., 2017), DeepLab (Chen et al., 2014, 2017a,b, 2018), and HRNet (Wang et al., 2020). Recently, there has been a growing interest in pure transformer-based architectures inspired by the success of Visual Transformers (Dosovitskiy et al., 2020). For instance, Segmentor (Strudel et al., 2021), Segformer (Xie et al., 2021), and Swin-Unet (Cao et al., 2021) are pure transformer-based architectures. Additionally, some studies propose models combining transformers and CNNs, like TransUnet (Chen et al., 2021). Minaee et al. (2021), provide an overview of the segmentation models. In land cover classification, several models have been developed for datasets like DeepGlobe (Demir et al., 2018) and LandCover.ai (Boguszewski et al., 2021). FPN with ResNet50 backbone (Seferbekov et al., 2018), and U-Net with Lovasz-Softmax (Berman and Blaschko, 2017) loss function (Rakhlin et al., 2018) were used for DeepGlobe. Another group of studies, modified the well-known architectures or proposed new architectures, specific for remote sensing data. NU-Net (Samy et al., 2018) modified U-Net to capture more global information and was used for DeepGlobe dataset. DiresUNet (Priyanka et al., 2022) utilized dense global spatial

Table 1: Semantic segmentation datasets of land cover classification.

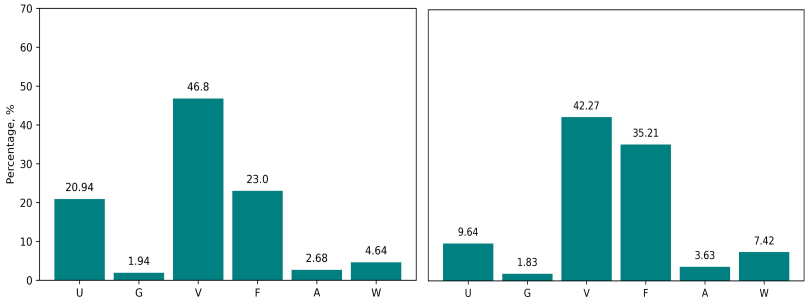
Dataset	#Classes	Resolution	#Channels	#Pixel	Image size
DeepGlobe Land Cover (Demir et al., 2018)	7	0.5	RGB	$6.87 * 10^9$	2448x2448
LandCover.ai (Boguszewski et al., 2021)	3	0.25,0.5	RGB	$2.98 * 10^9$	9000x9500;4200x4700
Agriculture-Vision (Chiu et al., 2020)	9	0.1,0.15,0.2	RGB+NIR	$2.26 * 10^{10}$	512x512
Extended Agriculture-Vision (Wu et al., 2023)	9	0.1,0.15,0.2	RGB+NIR	$810.0 * 10^{10}$	15000x15000
HAIR	6	0.2	Grayscale	$8.72 * 10^9$	8000x6000;6400x4800;16000x12000

pyramid pooling module (DGSPP) to help with the global context information extraction and was used for landcover.ai.

The size of aerial and satellite images are usually very large, e.g. Inria Aerial (Maggiori et al., 2017) contains 5000×5000 images and introduces challenges with the memory capacity on GPU's. Therefore images are typically downsampled or patched in order to fit in GPU memory. GLNet (Chen et al., 2019) used two branches of both downsampled and patched input and combine them in order to take advantage of global context.MagNet (Huynh et al., 2021) starts with segmenting the smallest scale of input image, and then progressively refines the segmentation output by increasing the input scale. 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.

Semantic segmentation of historical images is a particularly challenging task since the data varies in spatial and temporal dimensions. When dealing with spatial data, the underlying pattern of the data varies across the Earth's surface (Anselin, 1989). This means that a model trained on this dataset may fit better in some areas than others (Goodchild and Li, 2021). Semantic segmentation becomes more difficult for data that varies in the temporal dimension (Digra et al., 2022). Current semantic segmentation models are not able to generalize well on multi-temporal data. This is because the characteristics of data vary with respect to the time of capture (Yang et al., 2022b). For example, different post-processing methods could be used for preparing the data, or different camera technologies could be used to capture the images from the exact geographical location.

There has been several works proposing new datasets for semantic segmentation of landscapes. 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. (2020) give an overview of many datasets. LandCover.ai is a semantic segmentation dataset of aerial images from areas across Poland with resolutions between 25 to 50 cm per pixel and contains five classes. Agriculture-Vision (Chiu et al., 2020) is an aerial image dataset for pattern analysis of agricultural lands in the US with nine classes and 10 cm per pixel resolution. In a later study, this dataset is extended and improved. While the original version is limited to small images of 512×512 pixels, the updated version includes 3600 large, high-resolution images (Wu et al., 2023). DeepGlobe is a Satellite 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 six classes. Like DeepGlobe, satellite images (e.g., Landsat, Sentinel-1, Sentinel-2) were used to create many datasets in remote sensing (Wu et al., 2023). Two notable examples are Biome (Foga et al., 2017), designed for cloud detection, and Inria, which is a semantic segmentation dataset of urban settlements with two classes of building and not building. Ratajczak et al.



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 3: 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).

(2019) introduce a historical aerial image dataset of grayscale images. However, similar to the EuroSAT dataset (Helber et al., 2019), the problem is formulated as a classification task where smaller image patches are classified. However, the low resolution limits the usefulness of the data for studying land use evolution (Carbonneau and Piégay, 2012; Fausch et al., 2002). Other studies focus on semantic segmentation of land cover images captured by unmanned aerial vehicles (UAVs). For example, Wang and Mahmoudian (2023) present a UAV dataset of aerial fluvial images with multiple camera perspectives. Images are from the fall of 2017 and are labeled into river, boat, bridge, sky, forest vegetation, dry sediment, drone, and fluvial classes.

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), HAIR images are exclusively grayscale, as this is the nature of historical aerial images. Figure 7 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. Having high resolution images are important for many important applications, such as understanding of river habitat Carbonneau et al. (2020) and their evolution through time Piégay et al. (2020), ice morphology Alfredsen et al. (2018) and fish ecology Vannote et al. (1980), 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 (Barlaup et al., 2008; Hauer et al., 2016). Finally, the annotations in HAIR are finer than previous works, and these detailed annotations are made by experts. As shown in Table 1, the size of HAIR exceeds that of many widely recognized datasets in the field.

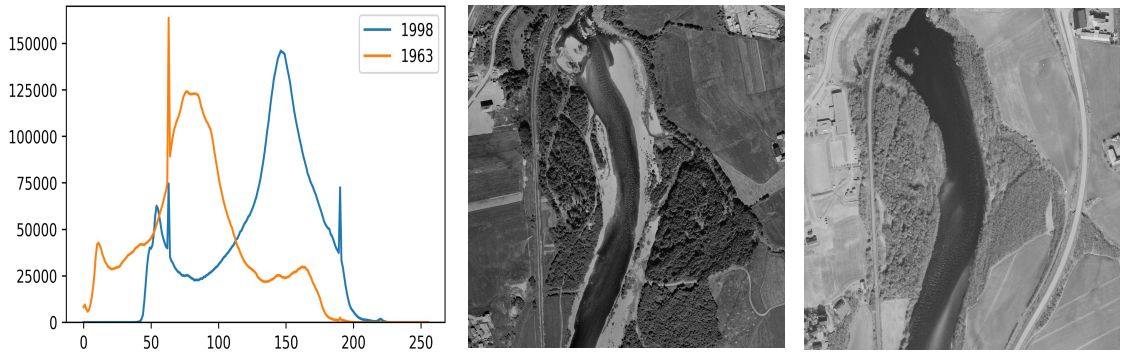


Figure 4: Intensity distribution of images covering the same section of Gaula from 1963 (left) and 1998 (right). It is calculated as the number of pixels with a given color (256 gray levels).

4 HAIR

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] in which found that improved annotation quality could improve performance more than model optimization.

Figure 3 provides an overview of the dataset. The dataset contains images from 1947, 1962, 1963, 1978 and 1998 and therefore, all of them are panchromatic (grayscale) since aerial images before the 2000 were captured in grayscale (Gergel and Turner, 2017). These images are selected so that we have both spatial and temporal overlap in the dataset. The $A^{S,T}$ set of HAIR contains images of riverscapes $A^{Gaula,1947}$, $A^{Gaula,1963}$, $A^{Surna,1963}$, $A^{Surna,1963}$, $A^{Nea,1962}$, $A^{Lærdal,1976}$.

The test sets have been designed to enable evaluation of whether both spatial and temporal characteristics are learned and can be generalized. Approximately 10% of images of each of the riverscapes in $A^{S,T}$ were chosen randomly for the ID test set. The remainder is split into a training set (80%) and a validation set (20%). The OOD test set, $O^{S,T}$, consists of two different rivers, namely 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. This means that, $O^{S,T} = \{ O^{Orkla,1962}, O^{Gaula,1998}\}$. Out of the 178 large images in HAIR, 20 are in the ID test set and nine in the OOD test set.

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

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 2023. 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, which means that vegetation differs quite a lot. The dataset includes a wide variety of optical conditions including different shadow lengths, angels of sunlight and saturation.

Annotation: Traditionally, the most common annotation tools for this purpose is GIS software with polygon editing, such as QGIS (QGIS Development Team, 2009). 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 iPads using pens, 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 (Viseras et al., 2016) was used to modify the edges of 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.
- Gravel (G): Vegetation free gravel bars in the river.
- Farmland (F): Farmland and cultivated land.
- Vegetation (V): Forest and other vegetated areas.
- Anthropogenic (A): Human-built structures.
- Unknown (U): Areas with no aerial images.

Statistics for HAIR: The distribution of different classes in HAIR is shown in Figure 3 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 an analysis of the impact of human development on riverscapes, covers only 1.9% of the images while water covers 4.6%. Figure 4 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.

5 Experiments

Here, we present two experiments of which one is an investigation of how removing color information affects the performance of a state of the art semantic segmentation model and and the other is a benchmark of state-of-the-art semantic segmentation models on HAIR.

5.1 Materials

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 grayscale 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 refer to the DeepGlobe converted to grayscale as DeepGlobe-G, and the original DeepGlobe as DeepGlobe-RGB. We evaluate two different scenarios for all models: 1) all models are trained on HAIR only, and 2) models are first pre-trained on DeepGlobe-G dataset and then fine-tuned on HAIR.

To convert the images of the DeepGlobe dataset into grayscale, we use the luminance method (Kanan and Cottrell, 2012), which is widely used in computer vision (Bosch et al., 2007) and is implemented in several image processing software and libraries such as OpenCV (Bradski, 2000). 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.

5.2 Removing Color Information

To investigate how removing color information from aerial landscape images affect the performance of state of the art semantic segmentation models, we trained MagNet on DeepGlobe-G and compared its performance to MagNet trained and evaluated on the DeepGlobe-RGB. MagNet is state-of-the-art on DeepGlobe-RGB (Huynh et al., 2021), and

5. The method is used in many papers such as (Bosch et al., 2007), however, in these works the name of the method is not mentioned. We got the name from (Kanan and Cottrell, 2012). The coefficients are the central component of the method.

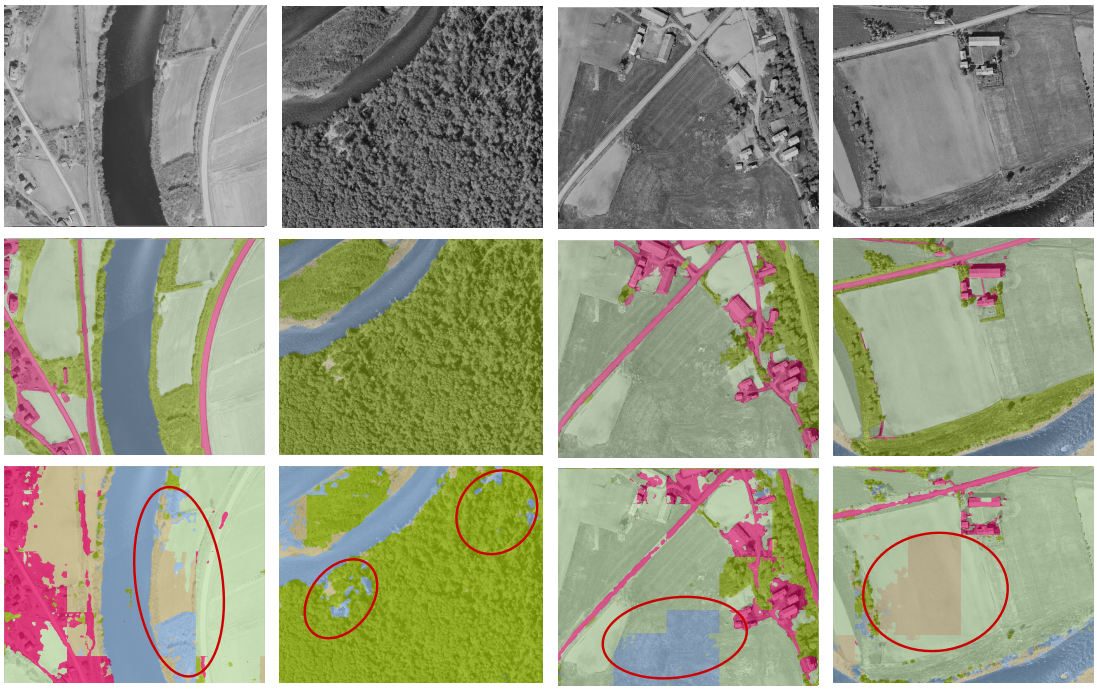


Figure 5: Source images (top row), labels (middle-row) and best predictions (bottom row).

we followed the exact same procedure as they used to train a MagNet model on DeepGlobe-RGB when we trained a MagNet model on DeepGlobe-G.

Results: The resulting MIoU of MagNet trained on DeepGlobe-G was 55.92 % compared to 72.96%, which is the result reported by Huynh et al. (2021) for MagnNet on DeepGLobe-RGB. This means that removing color information from the dataset decreases the performance of MagNet by 17.04%, which clearly shows how lack of color information negatively affects the performance.

5.3 Two Benchmark Tasks

We select U-Net, FPN, DeepLabV3+ (Chen et al., 2018), HRNet, Magnet and Swin-Unet 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.

The optimization and image augmentation methods are selected based on (Shamsaliei et al., 2023). The hyperparameters used for training MagNet are the same as in (Huynh et al., 2021). For other models, we used the Hyperband algorithm (Li et al., 2018) 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://zenodo.org/doi/10.5281/zenodo.10512593`.

Results: We calculate performance using MIoU of the experiments in the same way as Huynh et al. (2021). Results are presented in Tables 3 as mean ± 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 (Bouthillier et al., 2019). 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 DeepGlobe-G dataset led to a decrease in performance for all models except HRNet and Swin-Unet. The reason for the reduced performance might be that the DeepGlobe dataset has a relatively small scale. The small scale especially affects MagNet as it has a comparatively more complex architecture than the other models. HRNet and Swin-Unet on the other hand do not perform well in general, pre-training or no pre-training. A contributing factor to their relatively lower performance could be that their backbones were not pre-trained on any large-scale dataset, such as ImageNet. Another factor that might cause reduced performance is the method for converting DeepGlobe to grayscale. Kanan and Cottrell (2012) 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 (Ferrari Dacrema et al., 2021; Gundersen et al., 2023).

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 3 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. As described in the Section 4 a variety of optical conditions such as different angles of sunlight and length of shadows, are included in the dataset. As mentioned by Boguszewski et al. (2021), one would expect that this would make the dataset robust. However, the results on the OOD test data shows that it is not the case.

Table 2 shows the variance of performance of DeepLabV3+ with respect to the year of collection of the images in both ID and OOD test sets. The table suggests that the model generally finds more recent images to be more challenging. We observe a substantial drop in performance when the model is tested on the images from Lærdal, captured in 1976, despite this river being part of the ID test set. This drop in performance could be attributed to the temporal differences since most of the images in the training set are from the 60s. However, images from 1947 and 1976 have approximately similar temporal differences to those in the 60s, yet we do not observe any drop in performance in the images from 1947. It could possibly be due to the similarity between images from the 40s and 60s. When examining

Table 2: The MIoU of the DeepLabV3+ on test sets, categorized by the year.

year	river	OOD	MIOU	
			Not pre-trained	Pre-trained
1947	Gaula		89.12 ± 02.49	88.20 ± 00.74
1962	Nea		77.25 ± 00.86	73.70 ± 01.33
	Orkla	✓	69.64 ± 03.06	69.17 ± 01.58
1963	Gaula		89.12 ± 02.49	88.20 ± 00.74
	Surna		78.03 ± 05.32	77.98 ± 00.23
1976	Lærdal		66.31 ± 03.81	66.47 ± 03.53
1998	Gaula	✓	61.62 ± 03.21	59.43 ± 03.70

Table 3: The MIoU prediction of different semantic segmentation architectures on the ID and OOD test set. MIoU is mean ± standard deviation over five seed replicates. The Pre-trained, means the model is pre-trained on DeepGlobe-G.

Model	ID test set		OOD test set	
	Pre-trained	Not Pre-trained	Pre-trained	Not Pre-trained
U-Net	77.07 ± 00.48	76.10 ± 00.72	64.17 ± 00.8	60.10 ± 01.03
FPN	76.43 ± 01.84	76.06 ± 00.57	60.11 ± 01.51	59.60 ± 01.64
DeepLabV3+	**77.37 ± 01.02**	**76.28 ± 01.26**	**64.29 ± 02.66**	**62.68 ± 02.81**
HRNet	72.52 ± 00.97	73.50 ± 00.79	53.44 ± 02.46	51.61 ± 02.63
Swin-Unet	62.21 ± 00.40	63.06 ± 01.52	42.74 ± 00.84	43.16 ± 01.82
MagNet	65.22 ± 01.39	56.32 ± 01.12	57.06 ± 03.84	41.58 ± 04.51

the images, it is evident that images from 1947 bear a closer resemblance to those from the 1960s than those from 1976. For instance, in Figure 8 in the appendix, which depicts images from 1947, 1963, 1976, and 1998, it is noticeable that images from 1947 and 1963 share more similarities than those between 1962 and 1976. When looking at the performance of the pre-trained models, we observe that there is a larger variance when models are applied to more recent images. Table 5 in the appendix presents the variability in model performance across all benchmark models relative to the years of river data.

Table 4: Comparison of complexity of semantic segmentation models when applied to HAIR.

Metric	U-Net	FPN	DeepLabV3+	HRNet	Swin-Unet	MagNet	
						Back-bone	Refiner
Trainable parameters (m)	32.51	26.86	11.82	9.51	28.08	28.06	0.09
Training Time (s/step)	0.68	1.15	1.21	0.95	1.19	0.72	0.615
Inference Time (s/image)	11.77	13.64	10.34	13.63	14.95	15.04	

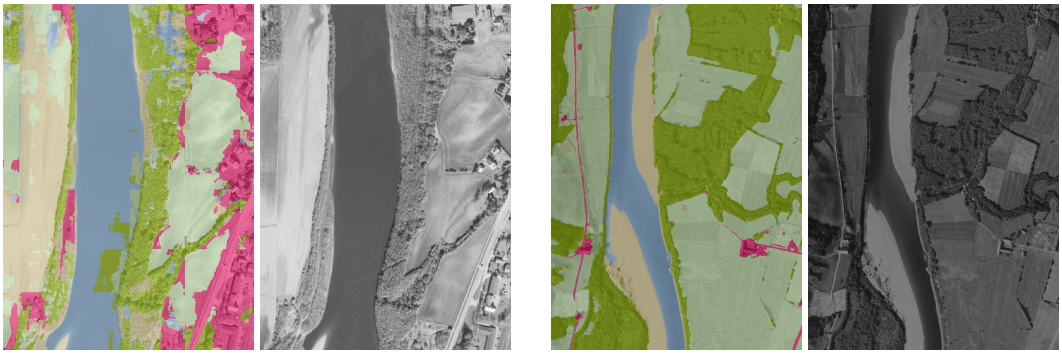


Figure 6: Comparison of predictions for Gaula 1998 (left) and 1963 (right).

To evaluate the complexity of the models, we compared three metrics: number of trainable parameters, training time for each step, and inference time of one 6000 × 8000 pixels image. This evaluation methodology is similar to the approach used in (Yi et al., 2019). Table 4 shows these metrics for all the models used for the two benchmark tasks.

Models with fewer trainable parameters are typically smaller in size, and we observe that HRNet and DeepLabV3+ have fewer trainable parameters compared to the other models. However, it is important to mention that the number of parameters is not the only factor in GPU memory consumption. For example, while HRNet has comparably fewer trainable parameters than DeepLabV3+, it requires more GPU memory to train. HRNet keeps the high-resolution representations of the input throughout the feed-forward process. It starts with high resolution and gradually adds lower resolutions, connecting the multi-resolution streams in parallel. This design choice leads to higher consumption of GPU memory. On the other hand, DeepLabV3+ incorporates dilated convolutions and reduces the spatial resolution of the input at an early stage, requiring less GPU memory.

To compare the training time, since each model has different batch sizes, which affect the number of steps of each epoch, we compared the time it takes for one step in the model. To compare the inference time, we selected five images from the test set with the size of 6000 × 8000 pixels and calculated the average inference time for these images. All the experiments are conducted on the same type of hardware and software. We observe that, comparably, U-Net requires less time to train, which is consistent with the simple design of the architecture. However, DeepLabV3+ has a slightly faster inference time compared to other models. This can be due to having fewer parameters and employing Atrous Separable Convolutions, which reduce the computational cost. On the flip side, MagNet has two phases on the inference and applies two separate networks to one image, making the inference more time-consuming. In practice, we train the model once and utilize it for inference several times, which makes inference time a more important factor.

In addition to the models described so far, we fine-tuned the Segment Anything Model (SAM) (Kirillov et al., 2023). SAM is a promotable semantic segmentation model based on Transformer vision models (Dosovitskiy et al., 2020). SAM has been trained on a large dataset and has demonstrated impressive performance in zero-shot tasks. The fine-tuning is described in Section A.5 of the appendix. Our findings indicate that further research is needed to enhance SAM's performance on the HAIR dataset. Despite the balanced

performance on ID and OOD test sets, the fine-tuned SAM model did not yield satisfactory results in either of the test sets. Interestingly, the model performed slightly better on the OOD test set than the ID test set. This indicates that additional studies on fine-tuning large pre-trained models like SAM could result in models exhibiting enhanced generalization performance when utilized on the HAIR dataset.

6 The Four Challenges

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

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 2. 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 4 and 6, light intensity alone should not be a problem as contrast and brightness were changed randomly as part of the training, as described in Appendix A. 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 (Sakkos et al., 2019; Yang et al., 2022a), blurring filters, histogram transformations, special optical and lens distortion (Liu et al., 2018), and generative models (Shorten and Khoshgoftaar, 2019).

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 3 shows the imbalanced class distribution, and the results presented in 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 (Tavera et al., 2022) and more effective loss functions (Dong et al., 2019) 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 5 shows several cases where texture alone makes the prediction hard. The MIoU of MagNet on DeepGlobe-G was 55.92 % which is 17.04% less than the performance reported in (Huynh et al., 2021) for DeepGlobe-RGB. 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. 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. Further development is needed to better support grayscale landscape images.

7 Future Work and Conclusions

Historical grayscale aerial images are a potentially important data source. However, they have not received much attention. These images have the capacity to offer valuable insight into the evolution of the landscape over the past century, shedding light on the role of human development in these transformations. However, current state-of-the-art semantic segmentation models used to access this information do not produce satisfactory results because they are not tailored for this type of data.

We show that the current satellite dataset containing RGB images is not suitable for training segmentation models aimed at analyzing historical aerial images. This finding shows the value of having an annotated dataset containing the actual historical aerial images. Additionally, we present HAIR, the first dataset of high resolution historical aerial images, annotated by domain experts. HAIR contains 8.72 billion annotated pixels and covers the years between 1947 and 1998. As the source aerial images are collected by the Norwegian Mapping Authority, our dataset is currently limited to Norwegian rivers. But our dataset encompasses a diverse range of rivers within Norway covering a variety of river characteristics, such as different sizes, morphological features, and riparian landscapes. We show that the performance of semantic segmentation models drops when applied to grayscale images compared to the same images in RGB. Furthermore, we provide a baseline for future studies by benchmarking the HAIR on selected state-of-the-art semantic segmentation models. The experiments show that pre-training on a recent dataset of converted satellite images does not improve the performance of the models, which further highlights the importance of generating large scale datasets of historical aerial images. Finally, we evaluated the impact of the four challenges associated with historical aerial images on the performance of the models and discussed how they can be mitigated in future studies.

Broader Impact Statement

Automatic semantic segmentation of historical aerial images of the landscape is a bottleneck in studying the changes in the landscape over time. This study is vital for understanding the impact of human development on the environment, finding potential restoration policies for landscapes and protecting biodiversity. However, semantic segmentation of historical aerial images is not well-studied due to the lack of a labeled dataset. In this work, we illustrate the shortcoming of current state-of-the-art semantic segmentation methods when applied to historical grayscale aerial images. To address this shortcoming, we present HAIR, the first dataset of historical aerial images of landscapes covering different regions across several decades. This dataset enables research on semantic segmentation of historical aerial images, which is key to analyzing landscape evolution over time.

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

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. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

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.

Luc Anselin. What is special about spatial data? alternative perspectives on spatial data analysis (89-4). 1989.

Jens Åström, Frode Ødegaard, Oddvar Hanssen, and Sandra Åström. Endring i leveområder for elvesandjeger og stor elvebreddedderkopp ved gaula. forekomst og dynamikk av elveører fra 1947 til 2014. 2017.

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.

Barbara Belletti, Carlos Garcia de Leaniz, Joshua Jones, Simone Bizzi, Luca Börger, Gilles Segura, Andrea Castelletti, Wouter Van de Bund, Kim Aarestrup, James Barry, et al. More than one million barriers fragment europe's rivers. *Nature*, 588(7838):436–441, 2020.

Maxim Berman and Matthew B. Blaschko. Optimization of the jaccard index for image segmentation with the lovász hinge. *CoRR*, abs/1705.08790, 2017. URL `http://arxiv.org/abs/1705.08790`.

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.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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.

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

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

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

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. ISSN 2078-2489. doi: 10.3390/info11020125. URL `https://www.mdpi.com/2078-2489/11/2/125`.

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.

Patrice E. Carbonneau and Hervé Piégay. Introduction: The growing use of imagery in fundamental and applied river sciences, August 2012. URL `https://doi.org/10.1002/9781119940791.ch1`.

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.

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. URL `https://arxiv.org/abs/2102.04306`.

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.

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, 2017a.

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

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.

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. doi: 10.1109/cvpr.2019.00913. URL `https://doi.org/10.1109/cvpr.2019.00913`.

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. doi: 10.1109/CVPR42600.2020.00290.

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. doi: 10.1109/cvprw.2018.00031. URL `https://doi.org/10.1109/cvprw.2018.00031`.

Monia Digra, Renu Dhir, and Nonita Sharma. Land use land cover classification of remote sensing images based on the deep learning approaches: a statistical analysis and review. *Arabian Journal of Geosciences*, 15(10):1003, 2022.

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.

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.

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.

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.

Steve Foga, Pat L Scaramuzza, Song Guo, Zhe Zhu, Ronald D Dilley Jr, Tim Beckmann, Gail L Schmidt, John L Dwyer, M Joseph Hughes, and Brady Laue. Cloud detection algorithm comparison and validation for operational landsat data products. *Remote sensing of environment*, 194:379–390, 2017.

Vitor Fortes Rey, Dominique Nshimyimana, and Paul Lukowicz. Don't freeze: Finetune encoders for better self-supervised har. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, pages 195–196, 2023.

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

Michael F Goodchild and Wenwen Li. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118(35):e2015759118, 2021.

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

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

Martin Guzelj, Christoph Hauer, and Gregory Egger. The third dimension in river restoration: how anthropogenic disturbance changes boundary conditions for ecological mitigation. *Scientific reports*, 10(1):13106, 2020.

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.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

Historic Environment Scotland. The national collection of aerial photography. `https://ncap.org.uk`, 2023. Online; accessed 31-aug-2023.

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. URL https://openaccess.thecvf.com/content/CVPR2021/html/Huynh_Progressive_Semantic_Segmentation_CVPR_2021_paper.html.

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

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. URL http://arxiv.org/abs/1412.6980.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

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. URL http://jmlr.org/papers/v18/16-558.html.

Frédéric Liébault and Hervé Piégay. Causes of 20th century channel narrowing in mountain and piedmont rivers of southeastern france. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group*, 27(4):425–444, 2002.

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

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.

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. doi: 10.1109/cvpr.2015.7298965. URL https://doi.org/10.1109/cvpr.2015.7298965.

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. URL https://arxiv.org/abs/2006.12485.

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.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3226–3229, 2017. doi: 10.1109/IGARSS.2017.8127684.

Anna Maria Mercuri and Assunta Florenzano. The long-term perspective of human impact on landscape for environmental change (lotec) and sustainability: from botany to the interdisciplinary approach, 2019.

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.

Jessica L Morgan, Sarah E Gergel, and Nicholas C Coops. Aerial photography: a rapidly evolving tool for ecological management. *BioScience*, 60(1):47–59, 2010.

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. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Danielle Katharine Petsch, Vivian de Mello Cionek, Sidinei Magela Thomaz, and Natalia Carneiro Lacerda Dos Santos. Ecosystem services provided by river-floodplain ecosystems. *Hydrobiologia*, 850(12):2563–2584, 2023.

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.

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. doi: 10.1007/s10489-022-03310-z. URL `https://doi.org/10.1007/s10489-022-03310-z`.

QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2009. URL `http://qgis.osgeo.org`.

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.

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. doi: 10.1109/TIP.2019.2896492.

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.

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

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.

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. doi: 10.1109/cvprw.2018.00050. URL `https://doi.org/10.1109/cvprw.2018.00050`.

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.

Saeid Shamsaliei, Odd Erik Gundersen, Knut Alfredsen, and Jo Halvard Halleraker. Towards historical analysis of riverscape development utilizing semantic segmentation. Presented at The AAAI 2023 Workshop on Artificial Intelligence for Social Good, AI4SG-23, 2023.

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

Tim Stephens. The kunming–montreal global biodiversity framework. *International Legal Materials*, page 1–20, 2023. doi: 10.1017/ilm.2023.16.

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. doi: 10.1109/iccv48922.2021.00717. URL `https://doi.org/10.1109/iccv48922.2021.00717`.

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.

The U.S. National Archives and Records Administration. National archieves of aerial photography. `https://www.archives.gov/research/cartographic/aerial-photography`, 2023. Online; accessed 31-aug-2023.

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. doi: 10.1139/f80-017. URL `https://doi.org/10.1139/f80-017`.

Alberto Viseras, Rafael Ortiz Losada, and Luis Merino. Planning with ants. *International Journal of Advanced Robotic Systems*, 13(5):172988141666407, September 2016. doi: 10.1177/1729881416664078. URL https://doi.org/10.1177/1729881416664078.

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.

Zihan Wang and Nina Mahmoudian. Aerial fluvial image dataset for deep semantic segmentation neural networks and its benchmarks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:4755–4766, 2023. doi: 10.1109/JSTARS.2023.3275068.

Ellen Wohl. Forgotten legacies: understanding and mitigating historical human alterations of river corridors. *Water Resources Research*, 55(7):5181–5201, 2019.

Jing Wu, David Pichler, Daniel Marley, David Wilson, Naira Hovakimyan, and Jennifer Hobbs. Extended agriculture-vision: An extension of a large aerial image dataset for agricultural pattern analysis, 2023.

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. URL https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html.

Pavel Yakubovskiy. Segmentation models. https://github.com/qubvel/segmentation_models, 2019.

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, 2022a.

Xuan Yang, Bing Zhang, Zhengchao Chen, Yongqing Bai, and Pan Chen. A multi-temporal network for improving semantic segmentation of large-scale landsat imagery. *Remote Sensing*, 14(19):5062, 2022b.

Yaning Yi, Zhijie Zhang, Wanchang Zhang, Chuanrong Zhang, Weidong Li, and Tian Zhao. Semantic segmentation of urban buildings from vhr remote sensing imagery using a deep convolutional neural network. *Remote sensing*, 11(15):1774, 2019.

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

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.

Appendix A.

A.1 Runtime Environment

GPU: NVIDIA Tesla V100

CPU: Intel(R) Xeon-Gold 6240

Number of Cores: 18 cores @ 2.6 Ghz

RAM: 32 GiB

Models were implemented using Tensorflow (Abadi et al., 2015) and Pytorch (Paszke et al., 2019) packages. In addition, Albumentations library (Buslaev et al., 2020) was used for Online Augmentation, and SegmentModel (Yakubovskiy, 2019) library was used in some of the implementations. In order to train the MagNet (Huynh et al., 2021), the script provided by the paper's github page was used.

A.2 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 (Kingma and Ba, 2015) 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, we perform a rotate and sample augmentation (RSA) method, where 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 (Yuan et al., 2019).

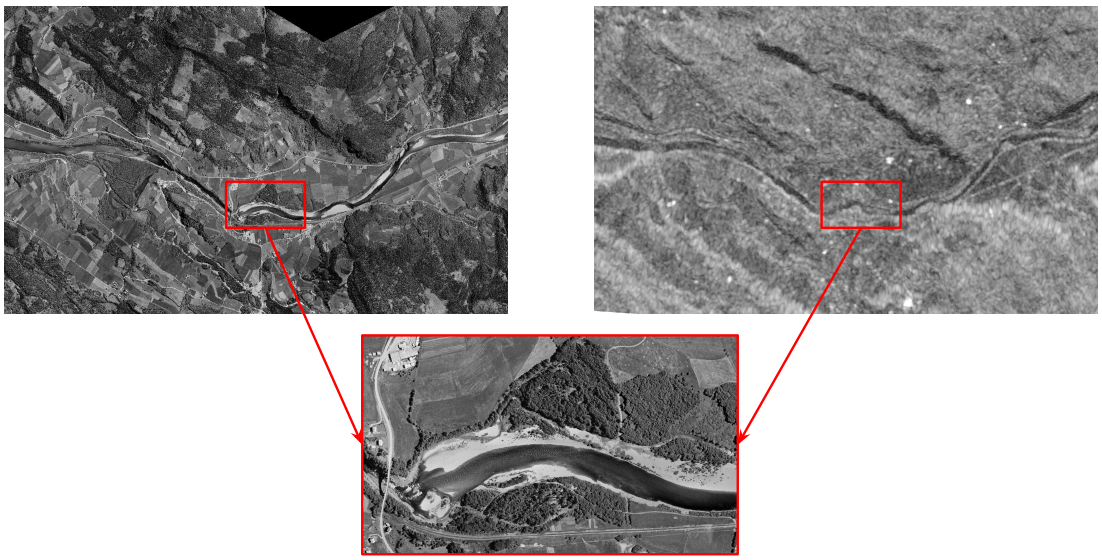


Figure 7: A visual comparison of HAIR images of Gaula captured in 1963 (top left) and SAT image of Gaula captured by Sentinel-1 in 2015 (top right). A subsection of HAIR images of Gaula 1963 is magnified to illustrate the high resolution of the HAIR images.

A.3 Visualization of the Synthetic Aperture Radar (SAR) images

As illustrated in the Figure 7, 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.

A.4 Variation of segmentation quality over time

Table 5 states the variance in the performance of semantic segmentation models in the benchmark with respect to the year of the rivers in the test sets.

A.5 Fine-tuning Segment Anything

Recently, there has been a large focus on developing foundation models in computer vision. Foundation models can adapt to tasks and data distributions that extend beyond their initial training data (Bommasani et al., 2021). In this work, we select the Segment Anything Model (SAM) as a foundation model and fine-tune and evaluate its performance on the HAIR dataset. This model has three components:
Image encoder: Encodes the input image into image embeddings.
Prompt encoder: Encodes additional prompts such as points, boxes, text, and input masks.
Mask decoder: Decodes the encoded image embeddings and prompts embeddings to create the desired output masks.

Table 5: The MIoU score of all of the models with respect to the year of capturing the images in both ID and OOD test sets. The MIoU is provided as the average plus minus the standard deviation for each model.

model	**pre-trained**	***Year and river***						
		1947	**1962**		**1963**		**1976**	**1998**
		Gaula	**Nea**	**Orkla**	**Gaula**	**Surna**	**Lærdal**	**Gaula**
U-Net		88.34 ± 0.40	76.39 ± 1.25	69.35 ± 2.85	88.34 ± 0.40	80.66 ± 0.68	67.74 ± 1.40	61.58 ± 1.07
	✓	86.46 ± 1.56	73.53 ± 1.29	68.21 ± 1.85	86.46 ± 1.56	80.46 ± 1.16	66.37 ± 1.83	56.04 ± 1.49
FPN		85.92 ± 3.03	74.87 ± 1.96	64.60 ± 4.14	85.92 ± 3.03	78.55 ± 2.74	66.12 ± 2.96	57.87 ± 3.20
	✓	87.31 ± 0.67	74.00 ± 2.15	68.84 ± 0.63	87.31 ± 0.67	79.48 ± 0.57	66.83 ± 1.50	54.98 ± 2.98
DeepLabV3+		89.12 ± 2.49	77.25 ± 0.86	69.64 ± 3.06	89.12 ± 2.49	78.03 ± 5.32	66.31 ± 3.81	61.62 ± 3.21
	✓	88.20 ± 0.74	73.70 ± 1.33	69.17 ± 1.58	88.20 ± 0.74	77.98 ± 0.23	66.47 ± 3.53	59.43 ± 3.70
HRNet		83.08 ± 2.82	73.52 ± 1.58	58.02 ± 3.04	83.08 ± 2.82	77.78 ± 1.13	60.08 ± 1.25	51.31 ± 3.05
	✓	86.74 ± 0.93	74.74 ± 2.33	53.55 ± 3.70	86.74 ± 0.93	76.91 ± 1.62	62.13 ± 1.38	50.65 ± 2.74
Swin-Unet		67.19 ± 1.64	61.71 ± 1.13	44.06 ± 1.89	67.19 ± 1.64	67.90 ± 1.85	50.72 ± 0.62	42.08 ± 1.52
	✓	67.59 ± 1.69	62.60 ± 2.04	44.99 ± 1.71	67.59 ± 1.69	69.44 ± 2.41	51.34 ± 2.39	42.25 ± 2.33
MagNet		64.32 ± 2.12	73.54 ± 1.28	66.40 ± 4.71	64.32 ± 2.12	62.41 ± 9.69	45.34 ± 1.30	52.38 ± 4.85
	✓	54.16 ± 3.86	62.76 ± 2.69	51.07 ± 4.44	54.16 ± 3.86	43.52 ± 5.47	36.86 ± 3.81	36.83 ± 5.09

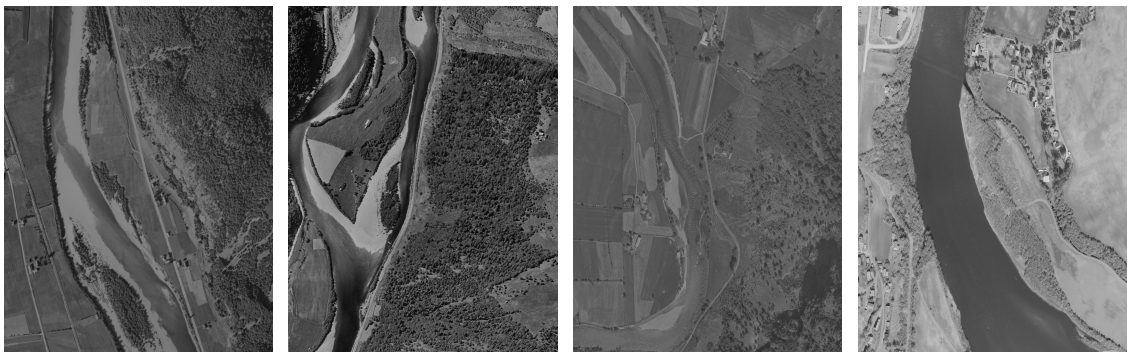


Figure 8: The images of rivers captured in 1947, 1963, 1976, and 1998 (from left to right). We observe that images from 1947 and 1963 are more similar than 1976 and 1963.

Table 6: The MIoU of the Fine-Tuned SAM on test sets.

Model	**Test sets**	
	ID	**OOD**
SAM	52.24	54.57

Table 7: The MIoU of the Fine-Tuned SAM on test sets categorized by year.

model	**Year and river**						
	1947	**1962**		**1963**		**1976**	**1998**
	Gaula	**Nea**	**Orkla**	**Gaula**	**Surna**	**Lærdal**	**Gaula**
SAM	44.30	54.40	57.64	44.30	52.64	38.23	53.04

To fine-tune SAM on the HAIR dataset, we initialize the SAM ViT-L model with the pre-trained weights. Then, we add a convolutional layer as the segmentation head to transform the output of SAM into the desired shape of the HAIR dataset. The segmentation head consists of a convolutional layer followed by upsampling. First, all the components are frozen except the segmentation head. We train the segmentation head during the first ten epochs. Afterward, we unfreeze the parameters of the mask decoder component and continue the training until convergence. This learning approach is similar to that used in (Fortes Rey et al., 2023). We used similar Hyperparameters as used in (Huynh et al., 2021). Figure 9 shows the process of fine-tuning SAM for the HAIR dataset.

Table 6 shows the performance of SAM for both ID and OOD test sets. Additionally, Table 7 states the performance of SAM on the test sets categorized by the year. The results imply that further research and optimization are required to enhance the performance of fine-tuning foundation models such as SAM on the HAIR dataset. Notably, the average time taken for each step during the fine-tuning of SAM is 2.77 seconds, which is comparatively longer than the other models in the benchmark. Table 8 shows the model complexity of SAM. We observe that the inference time of SAM is considerably higher than other models in the benchmark, which is consistent with the model's size.

Table 8: Comparison of complexity of semantic segmentation models when applied to HAIR.

Metric	**Fine-Tuned SAM**
Trainable parameters (m)	405.85
Training Time (s/step)	2.77
Inference Time (s/image)	51.81

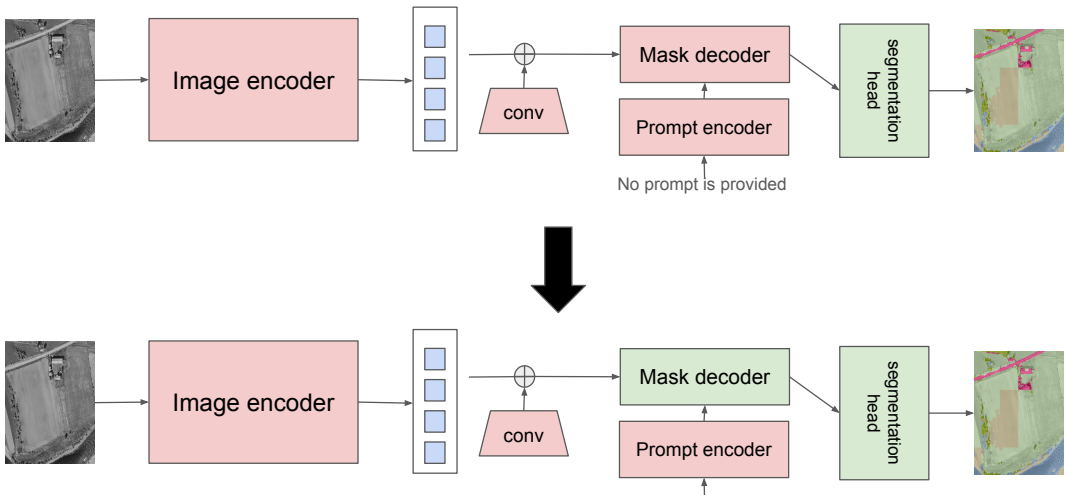


Figure 9: The process of fine-tuning SAM (Kirillov et al., 2023) for HAIR dataset. For the first ten epochs, we only train the segmentation head (top row). Afterward, we train the segmentation head and mask decoder together (bottom row). Components marked in red are not being trained, while those marked in green are undergoing training.