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

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

HAIR is a dataset of historical aerial images of riverscapes with high-quality annotations made by experts that can be used for semantic segmentation.

This dataset fills several gaps:

1) Greyscale semantic segmentation: Most semantic segmentation algorithms are made for RGB color images.
2) Longitudinal analyses of human development: Improvement of greyscale semantic segmentation algorithms will enable longitudinal studies covering large parts of the previous century.
3) High quality annotations made by experts.

The decade from 2021 to 2030 is declared the UN Decade of *Ecosystem Restoration*, and freshwater ecosystems have been judged as particularly degraded by UN. Hence, this dataset could provide important insights to help make policies for how to maintain and restore the fragile ecosystems around rivers. HAIR has some issues of which some are not typical for other land cover datasets captured close in time or by satellites: 1) camera technology has developed immensely since the first aerial images were taken, and therefore the quality of the images in the dataset is diverse, 2) lighting conditions are affected by time of day and the path of the airplane, so spatially close images might differ in lighting, 3) most of the historic aerial land cover images are grayscale, which makes it easy to confuse very different types of areas that could easily be differentiated based on color information, and finally 4) the dataset is highly biased, as the most important of the five classes, gravel, is small compared to the others.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

This dataset has not been used before. HAIR is shared as part of the paper HAIR: A Dataset of Historic Aerial Images of Riverscapes for Semantic Segmentation submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

C. What (other) tasks could the dataset be used for?

None.

D. Who funded the creation dataset?

The Norwegian University of Science and Technology, Norwegian Institute for Nature Research and Norwegian Research Council through the OFFPHD program, Grant No: 289725 – Ecosystem-based management.

E. Any other comment?

None.

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

All images and annotations are shared as both TIFF and PNG images. Each pixel of the image is labeled as one class. The TIFF files contain georeference information of images, which is useful when, for example, studying the evolution of landscapes. The PNG files are added for convenience as they are more suitable to be used for training and evaluating semantic segmentation models.

B. How many instances are there in total (of each type, if appropriate)?

There are in total 178 images and 178 images of annotations for each of the two file formats.

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

For each land cover image there is a corresponding annotation image.

D. Is there a label or target associated with each instance? If so, please provide a description.

We annotated the images with six different classes that can help understand the human pressure on river biodiversity and hydromorphology.

- **Water (W)**: Water covered areas (not restricted to river).
- **Gravel (G)**: Gravel bars and point bars in the river - vegetation free.

- **Vegetation (V)**: Forest and other vegetated areas in the riparian corridor.
- **Farmland (F)**: Farmland and cultivated land in the river corridor.
- **Anthropogenic (A)**: Anthropogenic structures like houses and roads.
- **Unknown (U)**: Areas that do not contain any aerial images.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No information is missing. Areas that do not contain any aerial images are labeled as "unknown". All the other areas are labeled with one of the five land cover type classes.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

For each land cover image there is a corresponding annotation image.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

HAIR is comprised of 178 annotated images with resolutions of 8000×6000, 6400×4800 or 16000×12000 pixels, from the five different rivers Nea, Orkla, Surna, Gaula and Lærdal in Norway. The images are taken in the years 1947, 1962, 1963, 1976 and 1998, and the pixel resolution is 20 cm per pixel.

The date of the acquisition of the images from each river is as follows:

- Gaula: 1947-08-18, 1963-07-01, 1998-05-12
- Nea: 1962-07-10
- Orkla: 1962-07-10
- Surna: 1963-07-14
- Lærdal: 1976-07-05

Out of the 178 images, 20 are in the in-distribution (ID) test set and 9 are in the out-of-distribution (OOD) test set. The out of distribution test set contains images from one river, Orkla, that is not in the training set and and another one, Gaula, taken in the year 1998, which is 51 years of camera improvement compared to the images from 1947 found in the in-distribution training set.

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

One training set and two test sets are defined. Both test sets are stored in the same test set folder.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

HAIR has some issues of which some are not typical for other land cover datasets captured close in time or by satellites: 1) camera technology has developed immensely since the first aerial images were taken, and therefore the quality of the images in the dataset is diverse, 2) lighting conditions are affected by time of day and the path of the airplane, so spatially close images might differ in lighting, 3) most of the historic aerial land cover images are grayscale, which makes it easy to confuse very different types of areas that could easily be differentiated based on color information, and finally 4) the dataset is highly biased, as the most important of the five classes, gravel, is small compared to the others.

Only issue 1 could be interpreted as noise.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained. New versions will be versioned and publicly shared.

Any other comments?

Additional information about the images:

- Gaula 1947: Type=ortofoto 20, Panchromatic, Coverage number=WF-0265, Color depth=8 bit/px, Recording method=analogue camera
- Gaula 1963: Type=ortofoto 20, Panchromatic, Coverage number=WF-1396, Color depth=8 bit/px, Recording method=analogue camera
- Gaula 1998: Type=ortofoto 10 Panchromatic, Coverage number=FN-98077, Color depth=8 bit/px, Recording method=analogue camera
- Surna 1963: Type=ortofoto 20, Panchromatic, Coverage number=WF-1428, Color depth=8 bit/px, Recording method=analogue camera
- Orkla 1962: Type=ortofoto 10, Panchromatic, Coverage number=WF-1280, color depth=8 bit/px, Recording method=analogue camera
- Lærdal 1976 Type=ortofoto 20, Panchromatic, Coverage number=NF-1654, Color depth=8 bit/px, Recording

method=analogue camera

- Nea 1962 Type=ortofoto 20, Panchromatic, Coverage number=WF-1295, Color depth=8 bit/px, Recording method=analogue camera

III. Collection Process

A. *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?*

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.

B. *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

Annotations are made manually by the experts using Adobe Photoshop on an iPad using a pen as it enables detailed annotations. The most common annotation tools for this purpose is GIS software with polygon editing tools, such as QGIS. Each large image was loaded as a layer to Adobe Photoshop. For the annotation, five distinct colors were selected so that each color represents one class. Then, the five classes were colored by the experts using the corresponding color of the class. The annotations were added as a layer on top of the source image. 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 Marching Ants algorithm was used to modify the edges of the objects.

Each large image was assigned to one of the experts, The experts followed a common procedure. Areas that were considered ambiguous by individual experts have been discussed in the group to reduce the noise in the annotation. Thus, the annotation has been taken very seriously and is considered to be of high quality, although ambiguous examples certainly can be found in such a large dataset. Due to a clear procedure and definitions of the land type classes, disagreement did not occur very often, and a discussion among the experts would lead to a unanimous decision.

The distinction between anthropogenic and gravel, water and vegetation, and vegetation with no tree coverage and farmland, were some of the main confusion for the experts during the annotations.

C. *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

The dataset contains images from 1947, 1962, 1963, 1978, and 1998 from rivers Gaula, Nea, Orkla, Surna, and Lærdal. The time and space represented by the images are selected so that the HAIR dataset contains both spatial and temporal overlaps. This allows for evaluating how well models generalize over images that cover different regions in the same time period and over images that capture the same region in different time periods.

D. *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

Experts (PhD students, a professor, a researcher) paid by the financing institutions.

E. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

The images were taken in the years 1947, 1962, 1963, 1976, and 1998.

Annotations were made in 2021, 2022 and 2023.

IV. Data Preprocessing

A. *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

No preprosessing was done.

B. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

The raw data is shared in the dataset.

C. *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

Software used to label is Adobe Photoshop for iPad.

D. *Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?*

Not applicable.

E. Any other comments

The following procedure was followed by all the annotators:

1) All the images should be annotated and areas with no image should be labeled as “Unknown”.
2) Borders of classes should be as fine as possible.
3) Only what is visible is considered to be the true state of the map. For example if a road disappeared in the forest. It is not considered a road.
4) Dark shadows are considered to belong to the class that makes the shadow.
5) Main uncertainty challenges and how to solve them:
   a) Confusion between human construction and gravel class: Due to the fact that images are historical, some roads, mines or even building constructions are very similar to gravel. In order to fix that, the current map should be checked and if there is a road or building in that place, it should be labeled as a human construction class.
   b) Some areas that are not forest and not clearly farmland. These areas should be classified as vegetation.
6) In case of any uncertainty in the class, recent land cover maps of the uncertain area needs to be inspected to achieve more information about the area.
7) If recent maps did not help, cases should be reported to the domain-expert to resolve the uncertainty.

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

zenodo.org and a website promoting the dataset.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

Creative Commons 4.0 BY-SA.

C. Are there any copyrights on the data?

Statkart-Geovekst holds the copyright.

D. Are there any fees or access/export restrictions?

No.

E. Any other comments?

The source of the raw images that are released as part of HAIR is the Norwegian Mapping Authority who shares these images on their website . The following information is provided on their website [1] (it is translated from Norwegian to English by us):

For the non-commercial use of the data: Contact one of the rights holders for permission to publish in, for example, reports, historical articles etc.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The authors of the paper mentioned above.

B. Will the dataset be updated? If so, how often and by whom?

It might be updated, and if so irregularly by the authors of the paper.

C. How will updates be communicated? (e.g., mailing list, GitHub)

If new versions of the dataset are released, they will be published referenced on the website as well.

D. If the dataset becomes obsolete how will this be communicated?

Not applicable.

E. Is there a repository to link to any/all papers/systems that use this dataset?

A web page will be made if the paper is accepted at NeurIPS 2023.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

If someone would like to contribute directly to the HAIR, we recommend emailing `saeid.shamsaliei@ntnu.no`. We will refer to all papers (we are made aware of) that use the dataset on the website `https://riverscapes.ai`. If new versions of the dataset are released, they will be published referenced on the website as well.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

None.

[1] https://norgeibilder.no

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

Not applicable. All source images are shared on https://norgeibilder.no

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

Not applicable.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Only the six classes mentioned above.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

No.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

It was obtained through a third party: `https://norgeibilder.no`

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Not applicable.

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Not applicable.

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

M. Any other comments?

None.