# Multi-modal Land Cover Classification of Historical Aerial Images and Topographic Maps: A Comparative Study

Mareike Dorozynski[1], Franz Rottensteiner[1], Frank Thiemann[2], Monika Sester[2], Thorsten Dahms[3], Michael Hovenbitzer[3]

[1] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
{dorozynski, rottensteiner}@ipi.uni-hannover.de
[2] Institute of Cartography and Geoinformatics, Leibniz Universität Hannover, Germany
{frank.thiemann, monika.sester}@ikg.uni-hannover.de
[3] Federal Agency for Cartography and Geodesy, Frankfurt am Main, Germany
{thorsten.dahms, michael.hovenbitzer}@bkg.bund.de

**Keywords:** Multi-modal Classification, Historical Geodata, Aerial Images, Topographic Maps, Semantic Segmentation, Land Cover.

## Abstract

Knowledge about land cover is relevant for many different applications such as updating topographic information systems, monitoring the environment, and planning future land cover. Particularly for monitoring, it is of interest to be not only aware of current land cover but of past land cover at different epochs, too. To allow for efficient, computer-aided spatio-temporal analysis, digital land cover information is required explicitly. In this context, historic aerial orthophotos and scanned historic topographic maps can serve as sources of information, in which land cover information is contained implicitly. The present work aims to automatically extract land cover from this data using classification. Thus, a deep learning-based multi-modal classifier is proposed to exploit information from aerial imagery and maps simultaneously for land cover prediction. Two variants of the classifier are trained, utilizing a supervised training strategy, for building segmentation and vegetation segmentation, respectively. Both classifiers are evaluated on independent test sets and compared to their respective two uni-modal counterparts, i.e. an aerial image classifier and a map classifier. Thus, a mean F1-score of 62.2% for multi-modal building segmentation and a mean F1-score of 83.7% for multi-modal vegetation segmentation can be achieved. Detailed analysis of quantitative and qualitative results gives hints for promising directions for future research of multi-modal classifiers to further improve the performance of the multi-modal classifier.

## 1. Introduction

Obtaining knowledge about past and present land cover has become an increasingly important topic. Such knowledge is not only highly relevant in updating topographic maps, but also allows for an analysis of former land cover, and thus, monitoring, change detection, and identifying trends in the spatial distribution of specific object types on the Earth’s surface. Against this background, the German Federal Agency for Cartography and Geodesy (BKG) has established the Gauss Centre (`https://www.gausszentrum.uni-hannover.de/en/`, accessed: 17.05.2024). The overall aim of the project is to derive land cover information for different points in time and to enable computer-aided analysis of time series of the same. Sources of information, particularly for past land cover, are historic remote sensing imagery, as well as historic topographic maps. Nevertheless, information about the coverage of the Earth is only implicitly contained in such data. To enable automated analysis of land cover, the original raster data, i.e. image data or scanned map data, must first be segmented according to a predefined object type catalog and thus be digitized. To obtain such explicit representations from implicit representations, classification methods can be exploited, in which raster data is presented to a land cover classifier that provides pixel-wise predictions for different land cover classes.

Typically, in training, raster data of a single modality, i.e. image data or map data, with a reference for land cover is used to learn a classifier to map the respective input data to high-quality predictions. To do so, recent approaches rely on deep learning, e.g. to extract building information from maps (Heitzler and Hurni, 2020) or to semantically segment aerial images (Mboga et al., 2020). Instead of utilizing a single source of information, multi-modal approaches incorporate different types of input data to exploit complementary information contained in the individual sources, e.g. historic aerial imagery and height information (Le Bris et al., 2020). Nevertheless, there does not yet seem to be any approach that combines aerial imagery with topographic map data. As these two kinds of data are often the only regionwide sources of information for historic land cover, it is of special interest to investigate the potential of semantic segmentation techniques to derive land cover from both sources simultaneously.

Accordingly, the goal of the present work is to consider image and map data jointly as inputs for land cover classification. The scientific contributions of the present work are the following:

- A multi-modal land cover classification approach is presented that takes aerial orthophotos and topographic maps to automatically derive land cover predictions. This is the first approach that combines these two input modalities.
- The multi-modal approach is compared to land cover predictions achieved by using the individual modalities as inputs for respective uni-modal classifiers.
- Based on a comprehensive analysis of the effects of exploiting multiple modalities for semantic segmentation on the classification accuracies, promising directions for future research of multi-modal classifiers are identified.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-107-2024

## 2. Related Work

Deriving information about land cover from different types of geodata is a classical task in geodesy, cartography, photogrammetry, and remote sensing. In recent years, the interest in historical geodata has become more and more relevant (Uhl et al., 2021; Farella et al., 2022; Dahle et al., 2024), e.g. in historical aerial photographs (Farella et al., 2022; Dahle et al., 2024), as well as in historical topographic maps (Uhl et al., 2021). For land cover classification, each of the two data sources just mentioned comes with its strengths and weaknesses, where it is of interest to investigate a combination of them for land cover classification compared to utilizing these sources independently from each other to derive land cover. In the following paragraphs, relevant existing works, as well as research gaps are described concerning the segmentation of historic imagery, semantic segmentation of historic maps, and approaches relying on both modalities.

**Image Classification:** Recent approaches for pixel-wise image classification, i.e. semantic segmentation, utilize fully convolutional networks (Long et al., 2015) or encoder-decoder networks such as UNets (Ronneberger et al., 2015) to extract relevant image features with an encoder, e.g. based on convolutional neural networks (LeCun et al., 1989; Krizhevsky et al., 2012), and subsequently processing the features to come to one prediction per pixel of the input image. Such approaches also have been applied to historical imagery, aiming to semantically segment them. While a UNet predicts six classes, including ice, snow, and sky, for all pixels of individual historic oblique aerial images in (Dahle et al., 2024), orthomosaics are classified by a UNet in (Le Bris et al., 2020; Mboga et al., 2020). Five land cover classes are differentiated in (Le Bris et al., 2020), where predictions are made for two different epochs, i.e. 1981 and 2001. Mboga et al. (2021) differentiates three and six land cover classes, respectively. UNets are also considered for semantic segmentation of current aerial imagery. In this context, recent works exploit variants of Vision Transformers (Dosovitskiy et al., 2021) as encoders, e.g. (He et al., 2022). Nevertheless, all these works rely on (historical) images only to predict land cover and do not consider topographic map data.

**Classification of maps:** The automatic interpretation of historic maps has seen a growing interest in recent years. Those maps contain past states of the Earth’s surface and land cover or land use and allow to inspect and analyze the temporal evolution of those states back in time (Uhl et al., 2021). Different approaches have been developed to exploit the potential of Deep Learning methods to identify the map objects, e.g. building footprints (Heitzler and Hurni, 2020). An essential problem is providing ground truth, for which different approaches have been proposed. (Jiao et al., 2022) propose to create so-called Imitation maps, i.e. use old symbology to create maps from current vector data. (Wu et al., 2023) exploit the fact that despite changes in topographic objects over the years, there is still a high likelihood that some objects (or parts thereof) did not change their position; this co-occurrence is implemented in a domain adaptation framework. Even though these works aim at the automatic interpretation of historical maps, none of them investigated exploiting further modalities to potentially improve the results.

**Multi-modal classification:** While aerial imagery and topographic maps were interpreted in uni-modal frameworks in the works cited so far, there is a growing interest in simultaneously exploiting multiple input modalities in the context of classification. Many works aim to improve land cover classification by combining remote sensing imagery with other types of geodata: For instance, Wang et al. (2023) combine image data with LiDAR data, optical data is combined with radar data in (Garnot et al., 2022) and with height information in (Le Bris et al., 2020), respectively, and optical data from multiple satellites are jointly considered in (Li et al., 2022). All these works demonstrate that the consideration of multiple sources of information can improve the quality of semantic segmentation. The only works that consider both, historical aerial images and historic topographic maps, e.g. (Liu et al., 2018), exploit ancient maps as a source of information for older epochs and aerial imagery as such for younger epochs; due to the combination, a longer period can be considered for analyzing land cover changes. No work could be identified, that combines topographic maps and aerial (ortho-) images of the same epoch for multi-modal land cover segmentation.

**Discussion:** It has been shown that both, historic images, e.g. (Le Bris et al., 2020), as well as scanned historic maps, e.g. (Heitzler and Hurni, 2020), can be utilized to extract information about land cover. Furthermore, many different approaches have demonstrated that the consideration of multiple modalities can improve classification performance, e.g. (Wang et al., 2023). To the best of the authors’ knowledge, there are not yet any approaches investigating multi-modal semantic segmentation of topographic maps and aerial imagery. Thus, inspired by the success of the respective uni-modal approaches, this work aims to investigate the potential of a first multi-modal classification considering topographic maps and aerial imagery of (approximately) the same epoch.

## 3. Methodology

The goal of the proposed method is to exploit historical topographic maps and historical aerial images simultaneously to predict land cover at the epoch of the historic input data. For this purpose, an aerial image $\mathbf{x}_{ae}$ and a scanned topographic map $\mathbf{x}_m$ showing the same area at approximately the same point in time and with the same spatial resolution are jointly presented to the proposed multi-modal classifier. The proposed classifier is based on a UPerNet (Xiao et al., 2018) (see Figure 1) and consists of one Swin Transformer encoder (Liu et al., 2022) per input modality and one decoder that exploits features at different stages of both input modalities. As output, one of the $K$ land cover classes $c_1, ..., c_k, ..., c_K$ of interest is predicted per pixel. The label map thus produced has the same extent as the two inputs $\mathbf{x}_{ae}$ and $\mathbf{x}_m$, respectively. As the predictions provided by the proposed multi-modal classifier are based on both, information contained in maps, as well as information contained in aerial images, they allow for an analysis of the potential of such a multi-modal classifier compared the to respective uni-modal classifiers.

### 3.1 Network Architecture

The main objective of the proposed multi-modal UPerNet is to allow for exploiting knowledge about past land cover contained in multiple modalities, namely in historical orthophotos $\mathbf{x}_{ae}$ originating from aerial flights and in scanned historical topographic maps $\mathbf{x}_m$, to provide pixel-wise predictions for $K$ land cover classes of interest. Accordingly, the inputs of the network are two input images $\mathbf{x}_{ae}$, $\mathbf{x}_m$ of the size $H$ x $W$ pixels; $H$ is the height of the image and $W$ is the width of the image with

https://doi.org/10.5194/isprs-annals-X-4-2024-107-2024 | © Author(s) 2024. CC BY 4.0 License.

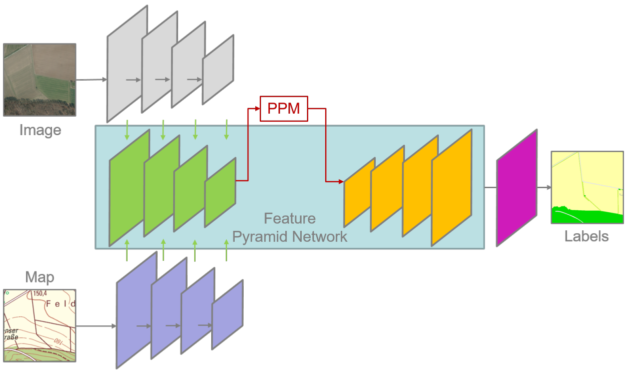


Figure 1. Multi-modal UPerNet. The encoder feature maps of the image branch (gray) and the map branch (purple) are combined (green arrows). The resulting combined feature maps (green) are presented to a Feature Pyramid Network (FPN; blue) with Pyramid Pooling Module (PPM; red). The output of the decoder (orange) is presented to a classification head (pink) the output of which is a label map with pixel-wise predictions.

$H = W$. In the present work, both images are RGB images, but the presented method is adaptable to deal with another number of input channels, e.g. a single grayscale channel as available for historic data before roughly 2000.

Each of the images $\mathbf{x}_{ae}$, $\mathbf{x}_m$ is presented to another encoder $E_{ae}(\mathbf{x}_{ae}, \mathbf{w}^E_{ae})$, $E_m(\mathbf{x}_m, \mathbf{w}^E_m)$ with an individual set of trainable weights $\mathbf{w}^E_{ae}$, $\mathbf{w}^E_m$, where all weights in the encoder are denoted as $\mathbf{w}^E := [\mathbf{w}^E_{ae}, \mathbf{w}^E_m]^T$. Both encoders are tiny variants of Swin Transformers (Liu et al., 2022), requiring a partitioning of the input images into patches of $P$ x $P$ pixels. The embeddings of all patches are presented to four subsequent Transformer blocks, delivering feature maps $f$ for the respective input modality at four different down-sampling stages $s$, i.e. $\{f^{(s)}_{ae}(\mathbf{x}_{ae}, \mathbf{w}^E_{ae}), f^{(s)}_m(\mathbf{x}_m, \mathbf{w}^E_m)\}^4_{s=1}$. To get multi-modal feature maps, the uni-modal feature at each stage $s$ are concatenated so that in case of having $d^{(s)}$ uni-modal feature maps per modality at stage $s$ the resulting multi-modal features consist of $2 \cdot d^{(s)}$ feature maps.

Subsequently, these multi-modal features are presented to a Pyramid Parsing Module (Zhao et al., 2017), parameterized by the weights $\mathbf{w}^{PPM}$, and a Feature Pyramid Network (Lin et al., 2017), parameterized by the weights $\mathbf{w}^{FPN}$. The outputs of the Feature Pyramid Network are feature maps with identical spatial dimensions as $f^{(1)}_m(\mathbf{x}_m, \mathbf{w}_m)$ and $f^{(1)}_{ae}(\mathbf{x}_{ae}, \mathbf{w}_{ae})$, respectively, where the number of feature maps is $2 \cdot d^{(1)}$.

These feature maps are presented to a classification head that reduces the $2 \cdot d^{(1)}$ feature maps to $K$ feature maps based on convolutions with weights $\mathbf{w}^{head}$, where the entire set of weights of the decoder is denoted by $\mathbf{w}^D := [\mathbf{w}^{FPN}, \mathbf{w}^{head}]^T$. Afterwards, the feature maps are bilinearly up-sampled to the input image size, i.e. $H$ x $W$ pixels, and the resulting $H \cdot W$ $K$-dimensional class scores $\vec{a}_{h,w}$ that are normalized using the Softmax function. The class $\hat{c}_k$ with the highest Softmax activation for the $K$ features at position $(h, w)$, $h \in [1, ..., H]$, $w \in [1, ..., W]$ is predicted for the corresponding pixel in the input images. The output of the classifier is the label map containing the predictions for all pixels.

### 3.2 Training

The proposed multi-modal UPerNet is trained by iteratively updating the network's weights $\mathbf{w} := [\mathbf{w}^E, \mathbf{w}^E, \mathbf{w}^{PPM}, \mathbf{w}^D]^T$ so that the loss function $\mathcal{L}$ becomes minimal. In the present work, the Softmax cross-entropy is used to measure the network's ability to correctly predict the reference class $c_k$ for the pixel at position $(h, w)$, indicated by a binary indicator variable $t_{h,w,k} = 1$ ($t_{h,w,k} = 0$ for all other classes). Accordingly, the loss for a single training sample, consisting of an aerial image $\mathbf{x}_{ae}$, the corresponding scanned map $\mathbf{x}_m$, and reference information $\mathbf{t} := \{t_{h,w,k}\ h = 1, ..., H \wedge w = 1, ..., W \wedge k = 1, ...K\}$, becomes

$$\mathcal{L}(\mathbf{x}_{ae}, \mathbf{x}_m, \mathbf{t}, \mathbf{w}) = \sum_{(h,w)} \sum_k t_{h,w,k} \cdot sm(\vec{a}_{h,w}, \mathbf{w}). \quad (1)$$

Due to the dependencies of the Softmax activation $sm$ in equation 1 on both input modalities and thus, the dependency of the loss $\mathcal{L}$ on the same, the network is forced to learn weights $\mathbf{w}$ such that the most meaningful information is extracted from the two input modalities.

## 4. Dataset

The data used to evaluate the multi-modal UPerNet is based on digital orthophotos (DOPs) and scanned topographic maps with a scale of 1:25 000 (TK25), and manually created reference data for land cover. All data represent areas in the region of the German city Hamelin, where a DOP from 2010 is available, and the first TK25 of that region (map sheet 3822) produced after this image acquisition year is selected, i.e. a map from 2011. The land cover reference was created by manually digitizing one vector layer per land cover class of interest based on the visual interpretation of the DOP, where the polygons of all layers belonging to one epoch are disjoint in space. Below, section 4.1 provides details about the general workflow for generating a multi-modal reference required by the method described in section 3. Based on available reference polygons two datasets are thus generated: One dataset for binary building classification belonging to the inner city of Hamelin (section 4.2), and one dataset for multi-class classification of different vegetation types belonging to a rural area in the north of Hamelin (section 4.3).

### 4.1 Preparation of the input data

The goal of the data preparation is to obtain aligned raster representations of the DOPs, the TK25, and the reference polygons, where corresponding pixels of these three data sources have to represent identical areas in object space, i.e. on the ground. Thus, in the first step, a joint coordinate reference system is selected, which is here ETRS89 / UTM zone 32N (EPSG: 25832), i.e. the reference system of the georeferenced DOP and thus, of the reference polygons, too. In the second step, the scanned TK25 maps are georeferenced by measuring the four corner points of the map sheet, assigning UTM coordinates to them, and applying an affine transformation to the scanned map. The resulting georeferenced TK25 raster then is available with a possibly deviation of an average of 1 m in the ground sampling distance in horizontal and vertical direction. At this point, all three data sources are available in the same coordinate reference system. As the classification method requires aligned rasters with quadratic cells/pixels of all data, a joint raster is defined in a third step: As the TK25 map sheet are given in a scale of

https://doi.org/10.5194/isprs-annals-X-4-2024-107-2024 | © Author(s) 2024. CC BY 4.0 License.

1:25 000, objects in the map are represented with an accuracy of about 1.25 m, assuming a line width of approximately 0.05 mm and an ideal map accuracy. The DOPs of the two time steps are available with a ground sampling distance of 20 cm. Thus, a joint raster with a ground sampling distance of 1 m is defined to have a compromise between maintaining the detailed information in the DOPs and still having a meaningful difference between neighbouring TK25 pixels. Furthermore, the 1 m raster is defined such that the upper left corner of a pixel belongs to integer values of Easting and Northing of the UTM coordinates. In the next step, the values of all data sources are interpolated to that joint raster, where bilinear interpolation is used for the DOPs and the TK25. Nearest neighbor interpolation is used to obtain the reference rasters, where the reference raster contains values between 0 and $K$, having $K$ classes of interest in a region. In the final step, the raster is partitioned into tiles of 500 x 500 pixels, where individual tiles are assigned to training, validation, or testing, respectively, and from which the input patches for network training are randomly drawn.

### 4.2 Multi-modal building dataset

For binary building classification, the procedure described in section 4.1 is applied to DOPs and building reference polygons from 2010 showing the inner city of Hamelin, as well as the corresponding TK25 from 2011. This results in a multi-modal dataset, where the region of interest is visualized based on the binary reference information in Figure 2.

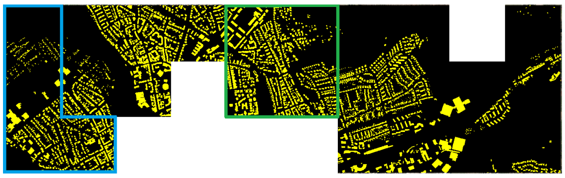


Figure 2. Reference of the binary building classification dataset, covering 6 $km^2$. Yellow belongs to the class *Building* and black to the class *No Building*. The blue area is used for validation, the green area for testing, and the remaining areas for training.

In Figure 2, all 24 tiles of 500 x 500 pixels are shown; the extent covers 4 tiles in height and 10 tiles in width. Only tiles with reference information for all buildings that are visible in the corresponding DOP tile were selected to contribute to the dataset. The tiles are assigned to a training, a validation, and a test subset, respectively, such that each subset contains all types of buildings in the dataset. The distribution of the data over the subsets and the corresponding relative class frequencies are shown in Table 1. In general, most of the pixels belong to the background class *No Building*. The amount of building pixels is around 15% in the training and validation sets. In the test set, around 23% of the pixels belong to buildings.

| Set | #Tiles | Frequencies [%] | |
|---|---|---|---|
| | | No Building | Building |
| Train | 16 | 86.0 | 14.0 |
| Val | 4 | 84.6 | 15.4 |
| Test | 4 | 76.8 | 23.2 |

Table 1. Statistics for the binary building dataset. *Set*: Name of the subset; *#Tiles*: Number of tiles in that subset; *Frequencies [%]*: Percentage of pixels belonging to the respective class in the respective subset.

### 4.3 Multi-modal vegetation dataset

Similarly, a multi-modal multi-class dataset is generated from DOPs, the TK25 map sheet, and the reference polygons (see Figure 3). The test area consists of 16 tiles of 500 x 500 pixels with a ground sampling distance of 1m, where 9 classes are contained in the reference (see the legend in Figure 3). Due to the rare frequency of all classes except for *Crop*, *Deciduous trees* and *Coniferous trees*, they are merged and considered as *Other* class so that in total 4 classes (the 3 just mentioned and a *Other* class) are differentiated for that area. The tiles are partitioned into 12 tiles for training and 2 tiles each for validation and testing (see Figure 3). Note that considering all of the nine classes would not have allowed to split the tiles into these three subsets, such that each class is contained in each subset. Thus, the statistics about relative class frequencies in Table 2 can be obtained. The most dominant class is *Deciduous trees*, followed by the class *Crop*, and the least frequent class of interest is *Coniferous trees*. The *Other* class, consisting of several object types, makes up in total between 4.7% and 8.6% of the subsets.

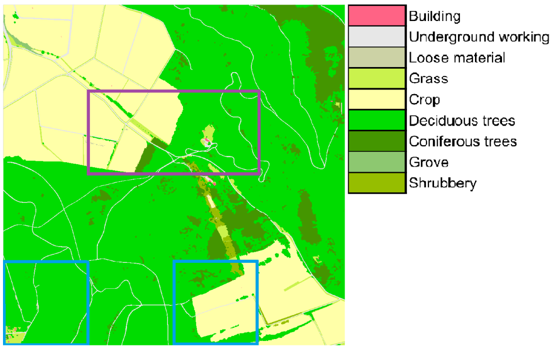


Figure 3. Reference of the multi-class vegetation classification dataset, covering 4 $km^2$. The blue area is used for validation, the purple area for testing, and the remaining areas are used for training.

| Class name | Frequencies [%] | | |
|---|---|---|---|
| | Train | Val | Test |
| Crop | 30.5 | 24.4 | 36.5 |
| Deciduous trees | 54.2 | 69.2 | 51.4 |
| Coniferous trees | 9.8 | 1.7 | 3.5 |
| Other | 5.5 | 4.7 | 8.6 |

Table 2. Statistics for the multi-class vegetation dataset. *Class name*: Name of the class; *Frequencies [%]*: Percentage of pixels belonging to the respective class in the respective subset.

## 5. Experiments

The experiments conducted in this work aim to evaluate the impact of multi-modal classification, utilizing the method presented in section 3, compared to the two respective uni-modal classifiers based on the two multi-modal datasets described in section 4. First of all, an overview of the experimental setup with training specifications, the conducted experiments, and the utilized evaluation protocol is presented in section 5.1. Afterward, the results thus obtained are presented, described, and discussed in section 5.2.

### 5.1 Experimental Setup

The overall goal of the experiments is to evaluate the quality of land cover predictions produced by the multi-modal classification approach and to compare the achieved qualities to those of the uni-modal counterparts of the classifier. Thus, the multi-modal approach is compared to uni-modal classification approaches, where a single modality, i.e. either map or aerial image, is presented to the network and a single encoder extracts feature maps for the PPM and the subsequent FPN with classification head. This is realized based on the building dataset and the vegetation dataset (section 4) by training the network architecture presented in section 3.1 based on the loss function presented in section 3.2. For training, images of $H = W = 256$ pixels are randomly extracted from the respective training tiles and partitioned into patches with P=4 pixels (see section 3.1 for a definition), where additional data augmentation (random rotation, transposition, horizontal and vertical flipping) is applied to artificially enlarge the training dataset; the augmentations for both input modalities are the same. Furthermore, the RGB input data is normalized to zero mean and a standard deviation of one for all channels of the two modalities independently. Thus, training batches with a batch size of 8 samples are produced, and presented to the network in each training iteration. The network weights $\mathbf{w}$ are initialized randomly using variance scaling (He et al., 2015), except for the encoder weights $\mathbf{w}_{ae}^{E}$ and $\mathbf{w}_{m}^{E}$; each set of encoder weights is initialized using pre-trained weights achieved on ImageNet (Russakovsky et al., 2015). During training, all weights are updated using mini-batch stochastic gradient descent with adaptive moments, i.e. Adam (Kingma and Ba, 2015). Preliminary experiments showed that a learning rate of $1 \cdot 10^{-2}$ is optimal in combination with the standard parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\hat{\epsilon} = 1 \cdot 10^{-8}$). Training is terminated using early stopping, where the stopping criterion is reached after 10 epochs of training without any further improvement of the validation (mean F1) score. To get an impression of the stability of the conducted experiments, all experiments are conducted three times, each with another random seed for the random components, i.e. weight initialization, and mini-batch sampling. For evaluation of the experiments, the Overall Accuracy (OA), the mean F1 score (mF1), and the mean Intersection Over Union (mIOU) are calculated on the independent test set. The presented mean values and the corresponding standard deviations of those three metrics are calculated out of the metrics achieved in the three runs per experiment.

Following this general experimental setup, the experiments listed in Table 3 are conducted. For each of the two datasets, a multi-modal experiment, as well as two uni-modal experiments, i.e. a uni-modal aerial image classification and a uni-modal map classification, are conducted. Thus, the performance of the multi-modal classifications ($B_{ae+m}$ and $V_{ae+m}$, respectively) can be compared to those of the uni-modal experiments ($B_{ae}$, $B_m$ and $V_{ae}$, $V_m$, respectively). Furthermore, a comparison of experiments with identical input modalities evaluated on datasets with different object types can be conducted (e.g. $B_{ae+m}$ and $V_{ae+m}$), allowing for an analysis of the ability of a certain modality to represent a certain object type such that the proposed classifier can be trained to correctly predict it.

### 5.2 Results and Discussion

The average quality metrics of all conducted experiments are presented in Table 4. For both of the datasets, all metrics are highest in the case of uni-modal classification based on aerial digital orthophotos ($B_{ae}$ and $V_{ae}$, respectively). While the difference in performance to the other experiments conducted on the same dataset is significant by a large margin on the building dataset, i.e. all metrics obtained in $B_{ae}$ are around 5.4% to 11.3% higher than those of the second best experiment $B_m$, the difference in all metrics is relatively small between the best experiment $V_{ae}$ and the second best experiment $V_{ae+m}$ (1.9% - 2.9%). This behavior is somewhat unexpected, having assumed that learning from multiple modalities would support the classifier in distinguishing different land cover classes. A more detailed analysis of the results, as well as potential reasons for the observed results, will be presented below, where Tables 5 and 6 show the class-specific F1-scores and the class-specific IOUs.

| Name | Dataset | Modality | |
|---|---|---|---|
| | | DOP | TK25 |
| $B_{ae+m}$ | Building | yes | yes |
| $B_{ae}$ | Building | yes | No |
| $B_m$ | Building | No | yes |
| $V_{ae+m}$ | Vegetation | yes | yes |
| $V_{ae}$ | Vegetation | yes | No |
| $V_m$ | Vegetation | No | yes |

Table 3. List of Experiments. *Name*: Name of the experiment; *Dataset*: either Building dataset (section 4.2) or Vegetation dataset (section 4.3); *Modality*: Input modality presented to the network (DOP for $\mathbf{x}_{ae}$ and TK25 for $\mathbf{x}_m$).

**Building dataset:** As already observed in the average quality metrics in Table 4, the uni-modal experiment with aerial DOPs also results in the highest class-specific quality metrics (see Table 5). In Table 5, the class of interest, i.e. *Building*, achieves the lowest scores in all of the three experiments, where the metrics in the multi-modal experiment $B_{ae+m}$ are lower than those obtained for both of the uni-modal experiments, i.e. $B_{ae}$ and $B_m$. There are several potential reasons for this: An analysis of the input modalities and the corresponding reference label maps shows that there are different kinds of discrepancies between the input data, i.e. DOP, TK25, and reference label map (see Figure 4), which are mainly due to generalization effects:

- Both of the examples show that due to **generalization** of the buildings in the map, the map does not contain all details of the building outlines that are visible in the DOP and the reference, respectively.

- The first example (upper row) shows a **displacement** between some buildings in the DOP and thus, also the reference compared to the TK25.

- Furthermore, the two examples in Figure 4 show that not all **buildings in the TK25** are not contained in the reference. Specifically, the second example (bottom row) shows that there are building parts contained in the TK25 that are not considered in the reference. A closer look at the corresponding DOP indeed shows that it is often quite hard to distinguish between gray roofs, e.g. from garages, and roads, which is likely to be the reason for the discrepancies in the data.

A solution to reduce such effects could be to consider both, DOPs and the TK25, for labelling.

Still, a closer look at the predictions of the three classifiers in Figure 5 shows that the two uni-modal classifiers $B_{ae}$ and $B_m$ predict relatively meaningful results, which fit to the quality metrics in Tables 4 and 5. While the boundaries of the two classifiers $B_{ae}$ and $B_m$ tend to be too roundish, $B_{ae}$ tends to predict finer details of the buildings, and $B_m$ does not provide such details. To produce shapes which have adequate building characteristics (e.g. rectangularity, parallelism) can be either

| Name | Quality metric [%] | | |
|---|---|---|---|
| | mF1 | mIOU | OA |
| $B_{ae+m}$ | 61.5 ± 0.61 | 45.6 ± 0.45 | 65.4 ± 0.74 |
| $B_{ae}$ | **89.1** ± 0.12 | **80.8** ± 0.26 | **92.2** ± 0.20 |
| $B_m$ | 81.1 ± 0.21 | 69.5 ± 0.21 | 86.8 ± 0.12 |
| $V_{ae+m}$ | 82.3 ± 0.64 | 72.0 ± 1.03 | 91.0 ± 0.61 |
| $V_{ae}$ | **84.4** ± 0.75 | **74.9** ± 1.00 | **92.6** ± 0.29 |
| $V_m$ | 59.3 ± 0.69 | 52.6 ± 0.56 | 87.0 ± 0.50 |

Table 4. Average experimental results with mean and standard deviation achieved in three runs per experiment. *Name*: Name of the experiment; *Quality metric*: respective average quality metric achieved in an experiment. The best results per dataset is highlighted in bold font.

| **F1 Score** | | | |
|---|---|---|---|
| Class name | Experiment | | |
| | $B_{ae+m}$ | $B_{ae}$ | $B_m$ |
| No Building | 73.6 ± 1.32 | **94.9** ± 0.20 | 91.5 ± 0.12 |
| Building | 49.4 ± 2.32 | **83.2** ± 0.15 | 70.8 ± 0.45 |
| **IOU** | | | |
| Class name | Experiment | | |
| | $B_{ae+m}$ | $B_{ae}$ | $B_m$ |
| No Building | 58.2 ± 1.64 | **90.3** ± 0.30 | 84.3 ± 0.21 |
| Building | 32.8 ± 2.02 | **71.3** ± 0.26 | 54.7 ± 0.50 |

Table 5. Class-specific results on the building dataset. *Class name*: Name of the class; *Experiment*: Name of the experiment. Per metric and class, the experiment with the highest metric is highlighted in bold font.

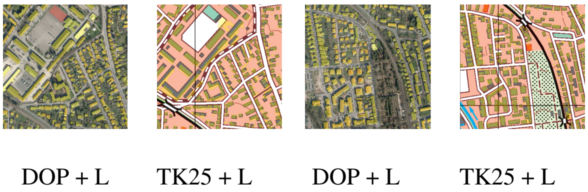


Figure 4. Examples for test tiles with reference overlay. DOP + L: Aerial image with yellow reference overlay. TK25 + L: topographic map with yellow reference overlay.

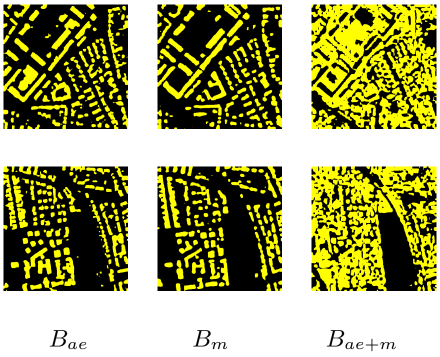


Figure 5. Examples for predictions achieved on two test tiles. Each row shows one of the test tile areas. The columns belong to the three conducted experiments (see Table 3).

achieved in post processing (Feng et al. (2020), Neidhart and Sester (2008), or by including it in the segmentation process (Marmanis et al. (2018)).

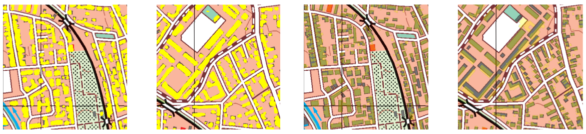
Beyond that, it is an interesting finding that the predictions of the uni-modal map classifier fit relatively well with the map contents: Figure 6 presents qualitative results of $B_m$, i.e. the building predictions and the corresponding part of the map. It can be seen that the predictions of $B_m$ fit well with the map content, particularly compared to the alignment of the reference and the map in Figure 4. Accordingly, the classifier learned in a relatively good way to predict buildings, even though the training data had contained shifts. As the evaluation was done with regards to the reference data, this leads to large discrepancies and thus the quantitative numbers in Table 5 are the lowest for the class *Building* in the experiment $B_m$.

The predictions of the multi-modal classifier $B_{ae+m}$ (see Figure 5, right) show that the classifier predicts buildings not only for regions that belong to buildings but also for many other regions. This is likely to be caused by the displacements between the TK25 and the DOP (and thus the reference), which becomes clear when, e.g. looking at the first example in Figure 4: as the yellow reference is aligned with the buildings in the DOP but not with those in the TK25, buildings in the TK25 and the DOP are not aligned. To achieve better multi-modal predictions, another fusion scheme, i.e. late fusion, could mitigate the problems of the current classifier by exploiting the already relatively well uni-modal predictions. Furthermore, auxiliary supervision, e.g. (Garnot et al., 2022), forcing a network to produce correct predictions based on all input modalities in addition to the multi-modal-based predictions might be helpful to overcome this issue, too.

**Vegetation dataset:** As already observed in the average quality metrics in Table 4, the uni-modal experiment with aerial DOPs results in the highest class-specific quality metrics (see Table 6) on the vegetation dataset, too. In contrast to the quality metrics achieved on the building dataset, the multi-modal classifier ($V_{ae+m}$) can predict all classes nearly as good as the best performing uni-modal aerial image classifier ($V_{ae}$). Accordingly, it is assumed that the fusion of all encoder features is suitable for predicting land cover in rural areas, in contrast to urban areas. This could be caused by fewer occurrences of object boundaries in rural areas compared to urban areas such that small geometric discrepancies between the DOP and the map have a lower impact on the classification performance.

A second interesting observation is that two of the three foreground classes in the vegetation dataset, i.e. the classes *Crop* and *Deciduous trees*, are predicted relatively well by the uni-modal map classifier ($V_m$): The F1-score and the IOU for the class *Crop* are en par for all of the three classifiers (93.7% - 95.2%) and the class *Deciduous trees* is also predicted relatively well based on maps ($V_m$) with an F1-score of 90.4% and an IOU of 82.6%. While correctly predicting *Other* is much more difficult for all of the three classifiers compared to the two foreground classes just mentioned, the map-based classifier $V_m$ tends to fully fail in predicting *Coniferous trees*. A qualitative analysis of the results provides some potential reasons for the class-specific quality metrics in Table 6.

Figure 7 shows the input data for both modalities and the reference label maps, whereas the predictions of all three vegetation classifiers for the two test tiles in the vegetation dataset are presented in Figure 8. There, it can be seen that the class *Other* is mostly confused with the class *Crop*. While a part of the agricultural area in the upper example in Figure 8 is predicted as *Other* in $V_{ae+m}$ and partly $V_{ae}$, a larger part of the class *Other* in the lower example in Figure 8 is predicted as *Crop* both by $V_{ae}$ and $V_{ae+m}$. This might be the case because these classes are relatively heterogeneous in their appearance and in particular, both classes, *Crop* and *Other*, contain areas with low green vegetation, which explains the confusion between these

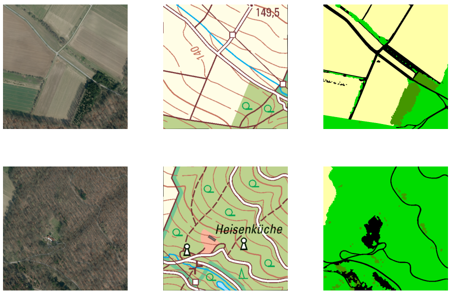


DOP + $B_m$ TK25 + $B_m$ DOP + L TK25 + L

Figure 6. Examples for test tiles with uni-modal map predictions ($B_m$). The test tiles in this figure are identical to those in Figure 4 (for an easier comparison, the original overlay of the reference data is replicated here again; DOP + L, TK25 + L).

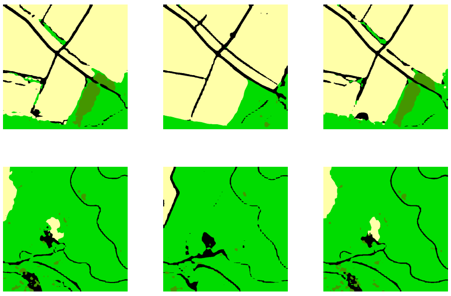
| **F1 Score** | | | |
|---|---|---|---|
| Class name | Experiment | | |
| | $V_{ae+m}$ | $V_{ae}$ | $V_m$ |
| Crop | 93.7 ± 0.71 | **95.2** ± 0.32 | 94.5 ± 0.15 |
| Deciduous | 94.1 ± 0.44 | **95.2** ± 0.15 | 90.4 ± 0.45 |
| Coniferous | 79.8 ± 2.40 | **81.6** ± 0.95 | 3.5 ± 1.31 |
| Other | 61.4 ± 1.85 | **65.8** ± 2.15 | 48.7 ± 1.61 |
| **IOU** | | | |
| Class name | Experiment | | |
| | $V_{ae+m}$ | $V_{ae}$ | $V_m$ |
| Crop | 88.2 ± 1.27 | **90.9** ± 0.59 | 89.6 ± 0.25 |
| Deciduous | 88.8 ± 0.84 | **90.8** ± 0.25 | 82.6 ± 0.80 |
| Coniferous | 66.4 ± 3.27 | **68.9** ± 1.35 | 1.8 ± 0.66 |
| Other | 44.3 ± 1.95 | **49.0** ± 2.44 | 32.2 ± 1.34 |

Table 6. Class-specific results on the vegetation dataset. *Class name*: Name of the class; *Experiment*: Name of the experiment. Per metric and class, the experiment with the highest metric is highlighted in bold font.



DOP TK25 Reference

Figure 7. Test tiles for the vegetation dataset. Each row shows one of the test tile areas. DOP: Aerial image. TK25: topographic map. Reference: Reference label map. For the legend see Figure 3 and black indicates *Other*.



$V_{ae}$ $V_m$ $V_{ae+m}$

Figure 8. Predictions achieved on the two test tiles. Each row shows one of the test tile areas. The columns belong to the three conducted experiments (see Table 3).

two classes. In contrast, the map in the lower examples in Figure 7 has predominantly a homogeneous red signature in the area, where $V_{ae}$ and $V_{ae+m}$ confuse *Crop* and *Other*. Accordingly, $V_m$ is better in correctly predicting *Other* in that region, which can be seen in Figure 8. Ideally, the multi-modal classifier would have taken advantage of the ability to correctly predict the class *Other* based on the map. It is assumed that auxiliary supervision in future work could support the multi-modal classifier in exploiting this strength of the map-based classification.

There is another interesting observation that supports the idea of learning land cover not only from aerial images but also from maps. In Figure 7, there are water bodies in the areas of both of the two test tiles according to the topographic map. Because of vegetation like trees and bushes that are visible in the orthophotos, the reference label maps, that were produced from these images, do not contain this information. As concluded in the context of the building dataset, both input modalities are recommended to be used for the generation of future reference label maps to allow for the full exploitation of the information in both of the modalities.

Other than that, there is another interesting observation concerning the prediction of the class *Coniferous trees*. A comparison of the topographic map (TK25) and the reference labels (L) in the upper example in Figure 7 shows that the area with *Coniferous trees* in the reference is marked to be *Deciduous trees* in the map; only a small part of the map content in the bottom example is marked to be an area with *Coniferous trees*. Thus, it can be explained that the quality metrics for *Coniferous trees* are extremely low for the map-based classifier $V_m$, i.e. an F1 score of 3.5% and an IOU of 1.8% are achieved (see Table 6). Indeed, the predictions of $V_m$ in the upper example in Figure 8 fit very well with these low-quality metrics, because $V_m$ predicts *Deciduous trees* for most of the forest area instead of *Coniferous trees*, which would be required by the reference. In the multi-modal classification scenario, $V_{ae+m}$, *Coniferous trees* is correctly predicted (see Figure 8), which is why it is assumed that the information provided by the digital orthophoto compensates for the misleading information that the classifier obtained from the map. Accordingly, *Coniferous trees* are similarly well predicted by $V_{ae+m}$ as by the aerial image classifier $V_{ae}$. It is concluded that a multi-modal setting with maps and aerial images is to be preferred over a uni-modal map setting in such cases to overcome issues with error-prone or less detailed information in maps (see the upper example in Figure 7) because the aerial images always show the actual land cover at the time of recording.

**Summary:** The conducted experiments showed that a multi-modal land cover classification of topographic maps and aerial imagery is possible in principle. Currently, the uni-modal aerial image classifier performs best on both, the urban and the rural datasets, where several reasons have been identified that explain this behavior. For both datasets, it has been found that future reference labels should be generated under consideration of both modalities instead of based on orthophotos only. Thus, for instance, buildings with a roof appearing similar to the surrounding underground workings can be detected more easily, leading to fewer conflicts between the inputs and the land cover reference. Further conflicts observed in urban areas are caused by spatial displacements of buildings in the maps compared to the other data and the generalization of the buildings. As the uni-modal classifiers performed much better on the building dataset, it is assumed that both, auxiliary supervision, as well as a modification of the fusion scheme to late fusion, might help to fully exploit the info in the individual modalities and to improve the current results. In the rural test area, the three foreground classes were predicted rather well by the multi-modal classifier. Only the heterogeneous *Other* class achieved lower quality metrics in the multi-modal classification. In a qualitative analysis, it was found that some of the areas with the class *Other* are relatively homogeneous in the map, which is why the uni-modal map classifier delivered better predictions in that area compared to the other two classifiers. Also in this context, auxiliary supervision is assumed to help a future multi-modal classifier to better exploit such strengths of an individual modality. It is particularly noteworthy that the multi-modal classifier was able to correctly predict coniferous trees in contrast to the uni-modal map classifier, even though the signature in the map is wrong. This demonstrates the general ability of the current multi-modal classifier to rely on the more informative input modality for making correct predictions for some of the classes. All in all, it was found that there is a great potential for learning to predict land cover utilizing both aerial imagery and topographic maps.

## 6. Conclusions & Outlook

In this paper, a multi-modal classifier for predicting land cover from historical aerial orthophotos and historical topographic maps was proposed. The classifier takes photos and maps from approximately the same epoch, having the same ground sampling distance. For the supervised training approach, a pixel-wise land cover reference is required to update the network weights. The classifier is trained to extract representative uni-modal features from both input modalities and to predict land cover based on the combined features that are processed in a joint decoder. In comprehensive experiments, the multi-modal classifier is compared to a uni-modal aerial image classifier and a uni-modal map classifier, respectively. The results show that currently, uni-modal classification based on aerial imagery performs best for building classification, achieving a mean F1-score of 89.2%, and the results are in the same order of magnitude for such a uni-modal classifier and the multi-modal classifier for vegetation classification, i.e. around 84% as a mean F1-score. The main reason for the different result characteristics on the building dataset is assumed to be the realized fusion scheme; as both uni-modal classifiers perform better than the multi-modal classifier, late fusion should be investigated. Nevertheless, the results obtained in the context of classifying vegetation demonstrated that the multi-modal classifier can rely on the more informative input modality to come to a prediction, showing potential for future refined multi-modal classifiers.

Accordingly, there are many directions for investigations in future work. Concerning the reference information to be used for training, all input modalities are recommended to be used for creating (further) manual reference labels. Expanding the reference information is of special interest to be able to differentiate more land cover classes and to have a better representation of underrepresented classes in the training data. Furthermore, it would be very interesting to investigate the generality of the proposed approach by applying it to other datasets, even though, to the best of the knowledge of the authors, there is not yet any dataset publicly available coming along with topographic map data, aerial imagery, and a high-quality land cover reference. From a methodological point of view, realizing auxiliary supervision, e.g. (Garnot et al., 2022), is also assumed to help to overcome the problems in the context of building classification, because the two uni-modal building classifiers were found to perform better by a large margin than the multi-modal classifier. Furthermore, a more advanced fusion scheme, e.g. using attention mechanisms (Guo et al., 2022), has the potential to help the classifier to focus on relevant features. Beyond that, it would be interesting how well uni-modal and multi-modal classification of land cover can be realized in older epochs where typically grayscale aerial images and topographic maps with fewer colors are available. Having the required data for multiple epochs would also allow for an expansion of the classification approach to a multi-temporal classifier, e.g. by adapting the uni-modal, multi-temporal classifier in (Voelsen et al., 2023).

## References

Dahle, F., Lindenbergh, R., Wouters, B., 2024. Revisiting the Past: A comparative study for semantic segmentation of historical images of Adelaide Island using U-nets. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 11, 100056.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16words: Transformers for image recognition at scale. *ICLR 2021 - 9th International Conference on Learning Representations*.

Farella, E. M., Morelli, L., Remondino, F., Mills, J. P., Haala, N., Crompvoets, J., 2022. The EuroSDR time benchmark for historical aerial images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2022, 1175–1182.

Feng, Y., Yang, C., Sester, M., 2020. Multi-scale building maps from aerial imagery. *ISPRS Archives; 43, B3*, 43(B3), 41–47.

Garnot, V. S. F., Landrieu, L., Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187, 294–305.

Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., Hu, S. M., 2022. Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 8.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1026–1034.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y., 2022. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.

Heitzler, M., Hurni, L., 2020. Cartographic reconstruction of building footprints from historical maps: A study on the Swiss Siegfried map. *Transactions in GIS*, 24(2), 442–461.

Jiao, C., Heitzler, M., Hurni, L., 2022. A fast and effective deep learning approach for road extraction from historical maps by automatically generating training data with symbol reconstruction. *International Journal of Applied Earth Observation and Geoinformation*, 113, 102980.

Kingma, D. P., Ba, J., 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, Conference Track Proceedings*.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1, 1097–1105.

Le Bris, A., Giordano, S., Mallet, C., 2020. CNN semantic segmentation to retrieve past land cover out of historical orthoimages and DSM: first experiments. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020, 1013–1019.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., Jackel, L. D., 1989. Backpropagation applied to handwritten ZIP code recognition. *Neural Computation*, 1(4), 541–551.

Li, Y., Zhou, Y., Zhang, Y., Zhong, L., Wang, J., Chen, J., 2022. DKDFN: Domain Knowledge-Guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 186, 170–189.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, D., Toman, E., Fuller, Z., Chen, G., Londo, A., Zhang, X., Zhao, K., 2018. Integration of historical map and aerial imagery to characterize long-term land-use change and landscape dynamics: An object-based analysis via Random Forests. *Ecological indicators*, 95, 595–605.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L. et al., 2022. Swin transformer v2: Scaling up capacity and resolution. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3431–3440.

Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135, 158–172.

Mboga, N., D’aronco, S., Grippa, T., Pelletier, C., Georganos, S., Vanhuysse, S., Wolff, E., Smets, B., Dewitte, O., Lennert, M., Wegner, J. D., 2021. Domain adaptation for semantic segmentation of historical panchromatic orthomosaics in Central Africa. *ISPRS International Journal of Geo-Information*. 10(8), 523.

Mboga, N., Grippa, T., Georganos, S., Vanhuysse, S., Smets, B., Dewitte, O., Wolff, E., Lennert, M., 2020. Fully convolutional networks for land cover classification from historical panchromatic aerial photographs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167, 385–395.

Neidhart, H., Sester, M., 2008. Extraction of building ground plans from LiDAR data. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37, 405–410.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

Uhl, J. H., Leyk, S., Li, Z., Duan, W., Shbita, B., Chiang, Y.-Y., Knoblock, C. A., 2021. Combining remote-sensing-derived data and historical maps for long-term back-casting of urban extents. *Remote sensing*, 13(18), 3672.

Voelsen, M., Lauble, S., Rottensteiner, F., Heipke, C., 2023. Transformer models for multi-temporal land cover classification using remote sensing images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1/W1-2023, 981–990.

Wang, Y., Wan, Y., Zhang, Y., Zhang, B., Gao, Z., 2023. Imbalance knowledge-driven multi-modal network for land-cover semantic segmentation using aerial images and LiDAR point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202, 385–404.

Wu, S., Schindler, K., Heitzler, M., Hurni, L., 2023. Domain adaptation in segmenting historical maps: A weakly supervised approach through spatial co-occurrence. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197, 199–211.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding. *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.