Multi-modal Land Cover Classification of Historical Aerial Images and Topographic Maps Exploiting Attention-based Feature Fusion

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, Attention-based Fusion, Semantic Segmentation, Historical Geodata, Remote Sensing Imagery, Topographic Maps

Abstract

Knowledge about past and present land cover is of interest for the assessment of the current status of our environment and, thus, for proper planning of the future. Information on past land cover is exclusively contained in an implicit way in historic remote sensing imagery and historic topographic maps. To make this information explicit, pixel-wise classification methods based on neural networks can be used. The method proposed in this paper aims to automatically predict land cover based on historic aerial imagery and scanned topographic maps. The proposed deep learning-based classifier extracts features at different scales from both modalities and fuses the most complex topographic map features of the smallest scale to enrich the ones derived from the aerial images. Both, the multi-modal features and those of the aerial images at larger scales, are mapped to pixel-wise predictions by means of a decoder. Comprehensive experiments show that the result of the proposed multi-modal classifier are superior compared to those of a uni-modal aerial image classifier; the multi-modal mIOU of 82.3% is 1.4% larger than the one of uni-modal classifier. This demonstrates that aerial image classification can benefit from additional information contained in topographic maps.

1. Introduction

Land cover information is highly relevant in the context of various applications, e,g, for studies related to ecological questions or urban development. In this context, not only the current status of the Earth's surface is of interest, but also trends for the change of the landscape. 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: 10.02.2025). One aim of the project is to obtain knowledge about historic land cover. Here, the term *historical* can refer to any point in the past; in this paper, it refers to any point in time after the early 1950s, which is the time when aerial imagery became available at a regional scale. Available regionwide sources of information for historic land cover are remote sensing imagery and topographic maps. In these sources, the desired information is typically contained in an implicit way, requiring an interpretation of the pixels to make the knowledge explicitly available for computer-aided analysis. As a manual inspection is infeasible at a large temporal and spatial scale, supervised pixel-wise classification techniques play an important role in automatically deriving land cover predictions.

In the context of land cover classification, the input for a classifier typically consists of data from a single modality, e.g. aerial or satellite images, or, in case of historic data, topographic maps, e.g. (Wu et al., 2023). Nevertheless, it is assumed that the joint consideration of multiple data sources as input is to be preferred to benefit from interdependencies and complementary knowledge contained in the data. For the classification of the current state of the landscape, data from various sources are available, and they can be combined in multi-modal classification techniques, e.g. (Garnot et al., 2022; Li et al., 2022; Wang et al., 2023), aiming to achieve more correct predictions compared to uni-modal techniques. For past epochs, mostly imagery, topographic maps and partly elevation models are available. Nevertheless, most research aiming to derive past land cover does so in a uni-modal way, e.g. by exploiting historic aerial images (Mboga et al., 2020; Sertel et al., 2023). Only very few works investigate multi-modal land cover classification from available data sources; one of the few exceptions is (Le Bris et al., 2020), where height information is considered jointly with historic aerial images. Particularly, no work could be identified that successfully considered aerial images and topographic maps for that purpose. As such data are the most important sources of information for past land cover and because they are complementary in terms of representing land cover classes, it is of special interest to investigate methods for multi-modal classification. For instance, class boundaries are clearer and intra-class variability is much smaller in maps, whereas object colour and texture are better represented in images, and the shapes of the objects reflect reality without generalisation effects. Thus, it is assumed that a classifier can benefit from using both modalities simultaneously.

In this context, it is not yet clear what is the optimal way to fuse the data from these modalities, both in terms of the actual fusion methodology and the stage of a classifier at which the information from the different modalities should be fused. This paper proposes a new approach for multi-modal land cover classification based on historical maps and historical aerial orthophotos with a focus on the optimal way of fusing the two modalities. In our previous work (Dorozynski et al., 2024), maps and images were treated as equally important at all stages of the approach, which turned out not to be beneficial: the multi-modal classifier

performed worse than the one solely based on digital orthophotos (DOPs). The proposed new approach combines the modalities more selectively, while utilizing a more advanced fusion scheme based on attention mechanisms. The scientific contributions of this work can be summarized as follows:

- We propose a new approach for fusing fine-grained aerial image features and high-level multi-modal features derived from scanned maps and images for historic land cover classification.
- In this context, we adapt an attention mechanism (Song et al., 2022) from the field of image retrieval for fusion of the two modalities for pixel-wise classification.
- We investigate whether convolutions or attention mechanisms are to be preferred for the extraction of spatial features from the two modalities.
- We provide a comprehensive analysis of the impact of the multi-modal features on classification compared to a uni-modal classification.

2. Related Work

This section provides an overview about existing literature related to pixel-wise land cover classification from remote sensing imagery and topographic maps. Besides uni-modal classification approaches based on these sources, techniques for combining multi-modal data are also discussed.

Classification of remote sensing imagery: Predicting land cover from remotely sensed images such as aerial images and satellite images is a classical task in Photogrammetry and Remote Sensing. Standard methods exploit fully convolutional networks (Long et al., 2015), encoder-decoder networks such as UNets (Ronneberger et al., 2015) or variants thereof, e.g. UPerNet (Xiao et al., 2018), to map the input to pixel-wise predictions for land cover. To do so, the input is first processed to generate feature maps by an encoder, frequently a convolutional neural network (CNN) (Krizhevsky et al., 2012), though current approaches tend do use attention-based architectures (Vaswani et al., 2017) such as Swin Transformers (Liu et al., 2022). Nevertheless, there are indications that convolutions are to be preferred for the extraction of spatial features, also because attention-based methods require patchification, which restricts the accuracy near object boundaries (Voelsen et al., 2024). Recently there has been a growing interest in using historic image-based data sources for the prediction of the past land cover, e.g. (Mboga et al., 2020, 2021; Van den Broeck et al., 2022; Sertel et al., 2023). While a pixel-wise classifier for the prediction of historic land cover is trained in a fully supervised way in (Mboga et al., 2020; Sertel et al., 2023), Mboga et al. (2021) and Van den Broeck et al. (2022) focus on domain adaptation to transfer knowledge about current land cover to past epochs. These works focus on pixel-wise classification (semantic segmentation in Computer Vision) of satellite or aerial images only and, thus, represent uni-modal approaches, neglecting other sources of information for land cover.

Classification of topographic maps: historical topographic maps provide an alternative source of information about past land cover (Uhl et al., 2021). In recent years, various approaches have been developed to use the potential of deep learning to identify map objects, e.g. text (Kim et al., 2023) or building footprints (Heitzler and Hurni, 2020). As in many deep learning domains, a major problem is the provision of ground truth, for which various approaches have been proposed. Jiao et al. (2022) suggest creating so-called imitation maps, i.e. using old symbols to create historical-looking maps from current digital landscape models. Wu et al. (2023) take advantage of the fact that despite the changes of topographic objects over the years, there is still a high probability that some objects (or parts of them) have not changed their position; this is considered in a domain adaptation framework. Nevertheless, these works also concentrate on a single modality only.

Multi-modal classification: Combining several data sources in multi-modal classification, aiming to solve the task in a better way compared to uni-modal techniques only relying on a single data source, is an ongoing field of research. In the context of land cover classification, several data sources were jointly considered as inputs: images of two satellites (Stocker and Le Bris, 2020), aerial images and satellite images (Heidarianbaei et al., 2024), elevation data and remote sensing images (Chen et al., 2018; Le Bris et al., 2020), radar and optical images (Garnot et al., 2022; Wang et al., 2024), optionally also considering height (Li et al., 2022), and optical and LiDAR data (Wang et al., 2023). There are different strategies for fusing the two modalities in multi-modal methods (Hong et al., 2020; Garnot et al., 2022). In *data-level fusion*, also called *early fusion* (Chen et al., 2018), the input data are combined before being presented to the classification network. Such a fusion scheme requires both modalities to have the same spatial resolution and to provide a similar level of detail. Alternatively, in case of less synergistic modalities, a fusion of extracted features is to be preferred, where the degree of correspondence of the modalities determines whether to fuse at earlier, e.g. (Wang et al., 2024), or at later stages, e.g. (Stocker and Le Bris, 2020; Le Bris et al., 2020; Li et al., 2022). In *late fusion* approaches, e.g. (Wang et al., 2023), the feature maps of the decoder are combined, while the predictions of two classifiers trained independently for each modality can be fused in *decision level fusion* approaches.

Nevertheless, none of the works cited so far combine topographic maps and image data in a multi-modal classifier and, in particular, there has not yet been any investigation to find out which scheme is most suitable for fusing these modalities. Such a fusion is of interest because both modalities play a crucial role in understanding the evolution of landscapes (Liu et al., 2018; Van den Broeck et al., 2022; Ettehadi Osgouei et al., 2022; Minervino Amodio et al., 2023). While orthomosaics of different epochs are classified in (Van den Broeck et al., 2022), historic maps for older epochs and orthomosaics for more recent epochs are classified in (Liu et al., 2018; Ettehadi Osgouei et al., 2022; Minervino Amodio et al., 2023), respectively. It can be assumed that, if available, the combination of these sources could improve the classification results by leveraging the complementary information contained in the two data sources. To the best of our knowledge, the only work that combines these modalities for land cover classification is our previous one (Dorozynski et al., 2024). However, the early fusion approach proposed in that study does not outperform the uni-modal counterparts of the proposed classifier, demonstrating the need for further research on the fusion of these two modalities.

Discussion: The success of multi-modal classification methods over uni-modal ones, e.g. (Garnot et al., 2022), as well as the requirement to classify historical geodata, such as historical orthophotos (Mboga et al., 2020; Sertel et al., 2023) and maps (Chiang et al., 2023) to obtain information about past land

cover, e.g. (Ettehadi Osgouei et al., 2022; Minervino Amodio et al., 2023), lead to the conclusion that there is the need for multi-modal methods that successfully combine these data to obtain the most reliable predictions for historic states of landscapes. Our previous work (Dorozynski et al., 2024) has shown that early fusion by feature concatenation does not help to exploit the full potential of combining the two sources. We believe that this might be caused by the fact that the maps are affected by errors due to changes in the landscape not considered in the map, but also due to generalization and different underlying modelling schemes. Therefore, another fusion scheme is required to address these problems. This paper aims to take a further step towards closing this research gap for the successful fusion of maps and orthoimages for pixel-wise classification.

3. Methodology

In this section, we describe our new deep learning-based multi-modal classification network. The main idea is to fuse the features derived from the two input modalities, DOPs and scanned topographic maps, at the stage of the network that is expected to be most suitable in terms of the properties of the features derived from the individual modalities, exploiting a fusion strategy that will focus on the features that are most relevant for classification. This cannot be achieved by early fusion. Thus, our method extracts features in two separate encoder branches (one per modality) before fusing them using trainable attention weights to allow for focusing on the most relevant information. For that purpose, we adapt the attention-based strategy of Song et al. (2022), originally designed for image retrieval, which combines global and local attention weights both in the spatial and the feature dimensions and is thus designed to consider all relevant aspects of the multi-dimensional multi-modal features. The extracted features are mapped to pixel-wise predictions, utilizing a UPerNet decoder (Xiao et al., 2018). The network architecture is presented in Section 3.1, while the method used for training the network is outlined in Section 3.2.

3.1 Network Architecture

The network architecture of the proposed network is shown in Figure 1. The input consists of a DOP $\mathbf{x}_{ae}$ and a co-registered scanned map $\mathbf{x}_{tm}$. These inputs are presented to two separate branches of the encoder E (cf. Section 3.1.1). The resultant features are combined in a multi-modal fusion module (cf. Section 3.1.2), yielding the final combined encoder output. Finally, the resultant features are presented to the decoder D to obtain pixel-wise predictions of land cover(cf. Section 3.1.3).

3.1.1 Encoder: The encoder E consists of two branches, E_{ae} with weights $\mathbf{w}_{ae}^{E}$ and E_{tm} with weights $\mathbf{w}_{tm}^{E}$. In principle, any backbone architecture can be used, as long as it consists of multiple stages, each followed by a downsampling step (in our experiments we compare a CNN and a transformer-based architecture; cf. Section 5.1). We denote the number of stages by S; the set of all weights of E is denoted by $\mathbf{w}^{E}$. The branch E_{ae} takes the DOP $\mathbf{x}_{ae} \in \mathbb{R}^{C^{(0)}\times H^{(0)}\times W^{(0)}}$ as its input, while E_{tm} processes the map $\mathbf{x}_{tm}\mathbb{R}^{C^{(0)}\times H^{(0)}\times W^{(0)}}$. $C^{(0)}, H^{(0)}$ and $W^{(0)}$ are identical for both modalities, which implies that the two grids have to be co-registered. The two encoder branches E_{ae} and E_{tm} produce uni-modal features $\mathbf{F}_{ae}^{(s)}$, $\mathbf{F}_{tm}^{(s)} \in \mathbb{R}^{C^{(s)}\times H^{(s)}\times W^{(s)}}$ with $s \in \{1, ..., S\}$ at all S stages.

The features generated in the last $S_f \leq S$ encoder stages (S_f is a hyperparameter) are fused to generate multi-modal features $\mathbf{F}^{(s)} \in \mathbb{R}^{2\cdot C^{(s)}\times H^{(s)}\times W^{(s)}}$ in the way described in Section 3.1.2. These multi-modal features are combined with the uni-modal features $\mathbf{F}_{ae}^{(s)}$ generated by the encoder branch processing the DOP in the first $(S - S_f)$ stages, yielding a combined set of features $\mathbf{F}_{ae}^{(1)}, ..., \mathbf{F}_{ae}^{(S-S_f)}, \mathbf{F}^{(S_f)}, ..., \mathbf{F}^{(S)}$ which provides the input for the decoder D. The uni-modal features $\mathbf{F}_{tm}^{(s)}$ generated by the first $(S - S_f)$ stages of E_{tm} are discarded. This strategy, using features of (relatively) high resolution only when derived from the DOP and considering the information from maps only at a relatively coarse resolution, is designed to overcome the problems of early fusion we identified in (Dorozynski et al., 2024): In this way, small deviations of objects in the map do not mislead the classifier. Geometrically detailed information about land cover is only provided by the image features at higher geometrical resolutions. Nevertheless, the topographic map features at coarser geometric resolutions are assumed to provide information about larger structures, which we expect to support the classification. We believe that this strategy will allow the classification to benefit from the strengths of both modalities.

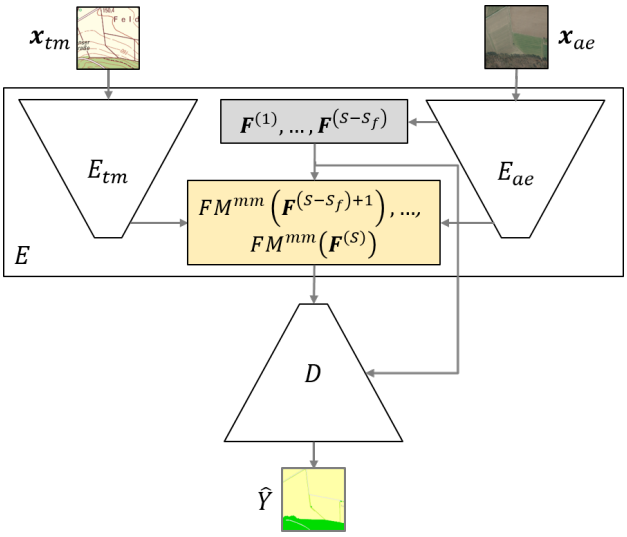

Figure 1. Multi-modal UPerNet with attention-based feature fusion. A topographic map $\mathbf{x}_{tm}$ and a co-registered DOP $\mathbf{x}_{ae}$ are presented to a uni-modal map encoder E_{tm} and a uni-modal DOP encoder E_{ae}, respectively, generating features at S stages. The uni-modal features of the last S_f stages are fused to multi-modal features by means of a fusion module FM^{mm}. The resulting features (yellow box) and the aerial image features of the first $S - S_f$ stages (gray box) are processed by a decoder D to obtain pixel-wise predictions $\hat{\mathbf{Y}}$.

3.1.2 Multi-modal Fusion: The fusion of the uni-modal image and map features to multi-modal features is realized in a multi-modal fusion module (FM^{mm} in Figure 1) with learnable weights $\mathbf{w}^{mmf}$ using global-local attention mechanisms (GLAM) (Song et al., 2022). In this module, GLAM is applied to each of the last S_f stages individually. At each of these stages s, the features $\mathbf{F}_{ae}^{(s)}$ and $\mathbf{F}_{tm}^{(s)}$ of that stage are concatenated to form a tensor $[\mathbf{F}_{ae}^{(s)}, \mathbf{F}_{tm}^{(s)}] =: \mathbf{F}_{in}^{(s)} \in \mathbb{R}^{2\cdot C(s)\times H(s)\times W(s)}$. In order to extract locally relevant context, convolutions are applied both along the $2 \cdot C(s)$ channels and in the spatial dimensions $H^{(s)} \times W^{(s)}$. This is referred to as the local attention (LA) mapping function in (Song et al., 2022) and results in features $\mathbf{F}_{lo}^{(s)}$. Furthermore, to also consider global context, self-

attention (Vaswani et al., 2017) is applied both in the channel and spatial dimensions, which is referred to as global attention (GA) mapping function and results in features $\mathbf{F}_{gl}^{(s)}$. The features $\mathbf{F}_{lo}^{(s)}$ and $\mathbf{F}_{gl}^{(s)}$ are combined with $\mathbf{F}_{in}^{(s)}$ by computing the weighted average of all three features using trainable scalars w_{lo}, w_{gl}, w_{in}, resulting in the output $\mathbf{F}^{(s)}$ for a stage s of the multi-modal fusion module. This can be formalized as follows:

$$\begin{aligned}\mathbf{F}^{(s)} &= GLAM([\mathbf{F}_{ae}^{(s)}, \mathbf{F}_{tm}^{(s)}]) = GLAM(\mathbf{F}_{in}^{(s)}) \\ &= w_{lo} \cdot LA(\mathbf{F}_{in}^{(s)}) + w_{gl} \cdot GA(\mathbf{F}_{in}^{(s)}) + w_{in} \cdot \mathbf{F}_{in}^{(s)} \\ &= w_{lo} \cdot \mathbf{F}_{lo}^{(s)} + w_{gl} \cdot \mathbf{F}_{gl}^{(s)} + w_{in} \cdot \mathbf{F}_{in}^{(s)}.\end{aligned} \quad (1)$$

For more details about the two mapping functions LA and GA, the reader is referred to (Song et al., 2022).

3.1.3 Decoder: After the encoding of the inputs $\mathbf{x}_{ae}$ and $\mathbf{x}_{tm}$ to features $\mathbf{F}^{(1)}, ..., \mathbf{F}^{(s)}, ..., \mathbf{F}^{(S)}$, i.e. $S - S_f$ unimodal features $\mathbf{F}_{ae}^{(1)}, ..., \mathbf{F}_{ae}^{(S-S_f)}$ and S_f multi-modal features $\mathbf{F}^{((S-S_f)+1)}, ..., \mathbf{F}^{(S)}$, the resultant S features are mapped to a pixel-wise class-score map $\mathbf{S} \in \mathbb{R}^{K \times H^{(0)} \times W^{(0)}}$ by a UPerNet decoder (Xiao et al., 2018) with trainable weights $\mathbf{w}^D$. The class score map has the same spatial extent as the input $\mathbf{x}$. At each spatial position $(h^{(0)}, w^{(0)})$ with $h^{(0)} \in \{1, ..., H^{(0)}\}$, $w^{(0)} \in \{1, ..., W^{(0)}\}$, $\mathbf{S}$ consists of a K-dimensional vector of softmax scores, where K is the number of classes to be differentiated. Each vector with K class scores $\mathbf{sm}_{h^{(0)},w^{(0)}}$ depends on the network weights $\mathbf{w} := [\mathbf{w}_{ae}^E, \mathbf{w}_{tm}^E, \mathbf{w}^{mmf}, \mathbf{w}^D]$ and inputs $(\mathbf{x}_{ae}, \mathbf{x}_{tm})$ defined in the preceding sections. Its elements are interpreted as the probabilities for the pixel $\mathbf{x}(h^{(0)}, w^{(0)})$ to belong to class $k \in \{1, ..., K\}$. The network's prediction $\hat{\mathbf{Y}}(h^{(0)}, w^{(0)})$ for $\mathbf{x}(h^{(0)}, w^{(0)})$ is the class C_k belonging to the largest softmax score $sm_{h^{(0)},w^{(0)},k} \in \mathbf{sm}_{h^{(0)},w^{(0)}}$.

3.2 Training

The network weights $\mathbf{w}$ are determined by minimizing a loss function $\mathcal{L}$ that measures the discrepancies between the predictions $\hat{\mathbf{Y}}$ and the known reference values for the correct class $\mathbf{Y} \in \mathbb{R}^{K \times H^{(0)} \times W^{(0)}}$. We use the categorical softmax cross-entropy loss for that purpose:

$$\begin{aligned}&\mathcal{L}\left(\hat{\mathbf{Y}}\left(\mathbf{x}_{ae}, \mathbf{x}_{tm}, \mathbf{w}\right), \mathbf{Y}\right) \\ &= -\sum_{h^{(0)}=1}^{H^{(0)}} \left(\sum_{w^{(0)}=1}^{W^{(0)}} \left(\sum_{k=1}^{K} \delta_{h^{(0)},w^{(0)},k} \cdot sm_{h^{(0)},w^{(0)},k} \right) \right).\end{aligned} \quad (2)$$

It is calculated based on the predictions $\hat{\mathbf{Y}}$ of the decoder D. The scalar $\delta_{h^{(0)},w^{(0)},k} \in \hat{\mathbf{Y}}$ is a binary indicator variable with $\delta_{h^{(0)},w^{(0)},k} = 1$ in case the pixel at position $(h^{(0)}, w^{(0)})$ belongs to the k^{th} class and $\delta_{h^{(0)},w^{(0)},k} = 0$ in all other cases. Equation 2 makes clear that the predictions $\hat{\mathbf{Y}}$ of the decoder depend on the weights $\mathbf{w}^D$ of that decoder, those of the multi-modal fusion module $\mathbf{w}^{mmf}$, as well as those of the two uni-modal encoder branches, i.e. $\mathbf{w}_{ae}^E$ and $\mathbf{w}_{tm}^E$. Minimizing the loss $\mathcal{L}$ affects the values of all these network weights so that multi-modal dependencies are learned during training.

4. Datasets

The datasets for the evaluation of the multi-modal classifiers are based on scanned topographic maps at a scale of 1:25.000 (TK25) from 2011 and a DOP obtained from aerial photographs from 2010. Both data sources show the city of Hamelin (Germany) and its surroundings. Both modalities were transformed to a joint coordinate system (ETRS89 / UTM zone 32N; EPSG: 25832). Originally, the two modalities were available at different spatial resolutions. As our method requires aligned grids for the two input modalities, they were re-sampled to a joint ground sampling distance (GSD) of 1 m using bi-linear interpolation. This is a compromise between the information content of the TK25, which corresponds to a GSD of 2.5 m in the best case, and the GSD of 20 cm of the original DOP. The reference for land cover was generated by manual digitization based on a visual inspection of the DOP and considering the TK25. The digitized polygons were rasterized at the same GSD as the input using nearest neighbour interpolation. The reference was generated in two different test areas of different land cover characteristics in the dataset just described and using two different class structures, resulting in two different datasets to be used for evaluation that are described in the subsequent sections. Section 4.3 discusses the properties of the used topographic maps.

4.1 Multi-modal Building Dataset

The first dataset covers an area of 8.7 km^2 and consists of 33 tiles of 512 x 512 pixels each with the GSD of 1 m. In this case, only two classes are differentiated in the reference: *Building* and *No Building*, hence we refer to it as the *building dataset*. The dataset covers the centre of Hamelin, suburban areas, alloted settlements and industrial areas.

The dataset is split into three disjoint subsets for training, validation and testing, respectively. For splitting, the type of building was considered in a visual inspection such that all three subsets are representative and contain all kinds of buildings. The number of tiles per subset and the class distributions in the three subsets can be found in Table 1. The class *No Building* is dominant in all subsets (around 75%-85% of the pixels).

Class name	Frequencies [%]		
	Train	Validation	Test
No Building	86.3	75.6	74.9
Building	13.7	24.4	25.1

Table 1. Statistics for the building dataset. *Class name*: Name of the class; *Frequencies [%]*: Percentage of pixels belonging to the respective class in the respective subset.

4.2 Multi-modal Vegetation Dataset

This dataset, to which we refer as the *vegetation dataset*, covers an area of 4 km^2 and consists of four tiles of 500 x 500 pixels each with a GSD of 1 m. In total, 9 land cover classes can be differentiated (Figure 2), but due to the low frequency of five of them, these classes are summarized in a joint background class *Other*, which is differentiated along with three foreground classes *Crop*, *Deciduous trees* and *Coniferous trees*.

The dataset is split into three disjoint subsets for training, validation and testing, respectively. Splitting was conducted based on the tiles such that every class is contained in every subset with roughly the same relative frequency. As a result, the subdivision with 12 tiles for training, 2 tiles for validation and 2 tiles for testing as visualized in Figure 2 was achieved. The corresponding class distributions can be found in Table 2. *Deciduous trees* cover more than half of the area of each subset (54%-70%), while the classes *Other* and *Coniferous trees* are underrepresented (1.7% and 9.8%, respectively).

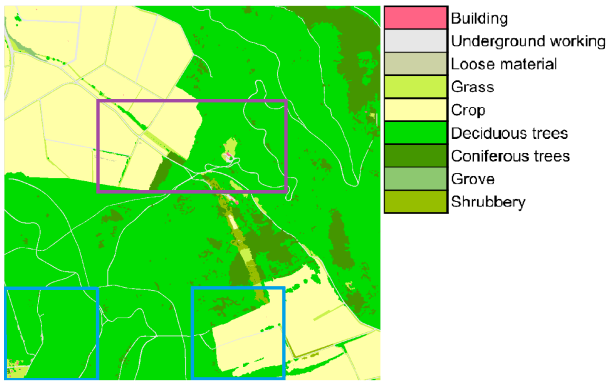

Figure 2. Reference of the vegetation dataset, covering 4 km^2. The blue and purple rectangles indicate the validation and test subsets, the remaining areas are used for training.

Class name	Frequencies [%]		
	Train	Validation	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 vegetation dataset. *Class name*: Name of the class; *Frequencies [%]*: Percentage of pixels belonging to the respective class in the respective subset.

4.3 Generalization and Representation of Land Use in Topographic Maps (TK 25)

The topographic map series TK 25 represents generalized geographical information. In contrast to the land cover, which can be observed in aerial images and digitized from those images, the topographic map primarily shows aggregated areas of land use. Land use areas may contain various land cover types, and conversely, the same land cover type can be found within different land uses. Not all objects are included in the map, as the selection relies on minimal capture size thresholds based on object types, e.g. 1 hectare (ha) for agricultural land, 0.1 ha for forests and lakes or 50 m^2 for buildings. Water bodies wider than 12 m are shown as areas, otherwise, they are depicted as lines. Forests are generalized into deciduous, coniferous, or mixed types, represented by symbols spaced about 5 mm apart on the map, equating to a distance of 125 m in reality, with no clear boundaries demarcating different forest types. This makes it very difficult to distinguish between the tree types in mixed forests on the topographic map. It is also important to note that there is no guarantee of temporal alignment between map data and image data for individual objects. Update cycles and priorities vary, potentially leading to discrepancies in data up-to-date-ness. All of this leads to both semantic and geometric discrepancies between aerial images and topographic maps. This discrepancies motivate the design of the fusion process in the network architecture.

5. Experiments

The goal of the experiments is to assess the performance of the proposed multi-modal classifier. For this purpose, the method presented in Section 3 is applied to the data described in Section 4. Section 5.1 provides an overview about the setup of all conducted experiments as well as the evaluation protocol. In Section 5.2, the results are presented, analysed, and discussed.

5.1 Experimental Setup

We conducted experiments using two variants of the network, differing by the backbone used in the encoder. In both variants, the encoder has $S = 4$ stages. One variant uses a ResNet18 (He et al., 2016) backbone, whereas the other one uses the tiny variant of the Swin Transformer (Liu et al., 2022) with a patch size of $P = 4$ pixels for the patchification of the input and a window size $W = 16$. The different backbones used in the models are indicated with a corresponding superscript R or S in the experiments. Preliminary experiments not reported here for lack of space have shown that the consideration of TK25 features of the last stage only (i.e., $S_f = 1$; cf. Section 3.1.1) is to be preferred in terms of the classification performance compared to other options. The reduction of the spatial resolution of the inputs follows (Xiao et al., 2018), so that the field of view after the fourth down-sampling stage amounts to 32 pixels in case of a ResNet-based encoder; in case of a Swin encoder, the entire input patch of 256 $\times$ 256 pixels contributes to the value of a node in the latest stage.

For all experiments, before being presented to the classifier, the channels of both input modalities, DOP and TK25, are individually normalized to have a zero mean and a standard deviation of one for all subsets. In training, patches of a size of 256 x 256 pixels are randomly cropped from these tiles, i.e. $H^{(0)} = W^{(0)} = 256$ and $C^{(0)} = 3$. Data augmentation is applied in training, using a random rotation by an angle between $0°$ and $360°$, random flipping in horizontal and vertical directions, as well as a potential transposition. All random components in the generation of the input patches are identical for each modality, so that each pixel in each modality refers to the same location on the ground. In each training iteration, a batch of 8 such patches is presented to the classification network. The network weights $\mathbf{w}$ are randomly initialized using variance scaling (He et al., 2015), except for the weights $\mathbf{w}_{ae}^{E}$, $\mathbf{w}_{tm}^{E}$ of the respective variant of the encoder that are initialized with weights obtained in a pre-training on ImageNet (Russakovsky et al., 2015). All weights are optimized using ADAM (Kingma and Ba, 2015) with standard parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\hat{\epsilon} = 1 \cdot 10^{-8}$) and a learning rate of $1 \cdot 10^{-2}$. Training is proceeded until no improvement in the mean F1 Score on the validation set can be observed for 30 epochs, where an epoch is defined to consist of 1000 training iterations.

The trained model is applied to the test set tiles, where the predictions for evaluation are determined in a sliding window approach with a horizontal and vertical shift of 128 pixels; in overlap regions, the class scores are averaged. The predictions thus obtained are compared to the land cover reference and per experiment the mean F1 score (mF1), the mean Intersection over Union (mIOU) and the Overall Accuracy (OA) are determined. The class-wise IoU and F1 scores are also analysed. Each experiment is conducted three times. We report mean values and standard deviations of the quality metrics.

All conducted experiments follow the general setup just described and are listed in Table 3. To allow for an analysis of the performance of the multi-modal classifier, the classifier is applied both to the building dataset (B_{ae+tm}^{R}, B_{ae+tm}^{S}) and the vegetation dataset (V_{ae+tm}^{R}, V_{ae+tm}^{S}). Furthermore, it of interest to asses the quality of a uni-modal aerial DOP classifier to be able to analyse the impact of the map features in the multi-modal classifier. Thus, a classifier per dataset is trained on aerial DOP only (B_{ae}^{R}, B_{ae}^{S} and V_{tm}^{R}, V_{tm}^{S}), which is realized by presenting the DOPs to the encoder E_{ae} (Figure 1) and setting

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-G-2025-221-2025 | © Author(s) 2025. CC BY 4.0 License.

$S_f = 0$, i.e. feature fusion is realized at none of the stages such that exclusively aerial features are considered in the decoder D. To be able to get an impression of which land cover information can be extracted from the topographic maps, uni-modal map classifiers are trained additionally (B^R_{tm}, B^S_{tm} and V^R_{tm}, V^S_{tm}), though this is not the main goal of this paper. For training the uni-modal map classifiers, the topographic maps are presented to E_{ae} (which is identical to E_{tm}), and again S_f is set to $S_f = 0$. A comparison of the experiments with ResNet18 encoders E_{ae} and E_{tm} to those with Swin-based encoders allows for an analysis of the suitability of convolutions versus self-attention for feature extraction.

Name	Dataset	Modality		Encoder
		DOP	TK25	
B^R_{ae+tm}	Building	yes	yes	ResNet18
B^R_{ae}	Building	yes	no	ResNet18
B^R_{tm}	Building	no	yes	ResNet18
B^S_{ae+tm}	Building	yes	yes	Swin
B^S_{ae}	Building	yes	no	Swin
B^S_{tm}	Building	no	yes	Swin
V^R_{ae+tm}	Vegetation	yes	yes	ResNet18
V^R_{ae}	Vegetation	yes	no	ResNet18
V^R_{tm}	Vegetation	no	yes	ResNet18
V^S_{ae+tm}	Vegetation	yes	yes	Swin
V^S_{ae}	Vegetation	yes	no	Swin
V^S_{tm}	Vegetation	no	yes	Swin

Table 3. List of Experiments. *Name*: Name of the experiment; *Dataset*: either the building (Section 4.1) or the vegetation dataset (Section 4.2); *Modality*: Input modality presented to the network (DOP: $\mathbf{x}_{ae}$, TK25: $\mathbf{x}_{tm}$); *Encoder*: Encoder backbone for E_{ae}, E_{tm}, superscript R for ResNet and S for Swin.

5.2 Results and Discussion

5.2.1 Average Quality Metrics: The quality metrics per experiment averaged over three runs, as well as the corresponding standard deviations are listed in Table 4. For all four experimental series, the multi-modal variants of the classifier (B^R_{ae+tm}, B^S_{ae+tm}, V^R_{ae+tm}, V^S_{ae+tm}) perform best for all three quality metrics, i.e. better than their uni-modal counterparts. The only exception is the classification of the vegetation dataset with a ResNet-based encoder, where the OA of the uni-modal aerial classifier (V^R_{ae}) performs slightly but not significantly better than the multi-modal classifier (V^R_{ae+tm}) in that series. This shows that the consideration of both modalities compared to considering only one modality is to be preferred. The higher mF1 and higher mIOU, respectively, demonstrate that particularly individual classes can be differentiated in a better way under consideration of both modalities. Comparing all classification results per dataset, i.e. all ResNet-based and all Swin-based classifiers, the best performing classifier on the building dataset extracts features based on convolutions (B^R_{ae+tm}) whereas the best performing classifier on the vegetation dataset does so based on attentions (V^S_{ae+tm}). The behaviour on the building dataset can be explained by the requirement to identify small details to make correct building predictions at a GSD of 1 m, which is possible with the ResNet encoder that takes individual pixels as an input. In contrast, the objects in the vegetation dataset are larger compared to buildings, such that a larger global context is more of interest than details. The Swin-based encoder enables to consider not only local but also global context, and the patchification (P = 4 pixels) might lead to features that can represent larger structures in a better way compared to convolutions. Be that as it may, the average metrics demonstrate that the classifier can benefit from learning from both modalities, where, depending on the granularity of the objects, convolutions or attentions are to be preferred for feature extraction.

Name	Quality metric [%]		
	mF1	mIOU	OA
B^R_{ae+tm}	**90.1** $\pm$ 0.25	**82.3** $\pm$ 0.39	**92.5** $\pm$ 0.24
B^R_{ae}	89.2 $\pm$ 0.49	80.9 $\pm$ 0.73	91.8 $\pm$ 0.37
B^R_{tm}	84.6 $\pm$ 0.24	74.2 $\pm$ 0.37	88.7 $\pm$ 0.29
B^S_{ae+tm}	**86.9** $\pm$ 0.29	**77.4** $\pm$ 0.42	**90.1** $\pm$ 0.14
B^S_{ae}	86.2 $\pm$ 0.76	76.4 $\pm$ 1.07	89.7 $\pm$ 0.43
B^S_{tm}	84.0 $\pm$ 0.42	73.3 $\pm$ 0.56	88.2 $\pm$ 0.12
V^R_{ae+tm}	**82.2** $\pm$ 0.21	**72.1** $\pm$ 0.25	91.5 $\pm$ 0.33
V^R_{ae}	81.6 $\pm$ 1.34	71.3 $\pm$ 1.71	**91.6** $\pm$ 0.22
V^R_{tm}	58.7 $\pm$ 0.29	51.4 $\pm$ 0.40	87.2 $\pm$ 0.40
V^S_{ae+tm}	**83.0** $\pm$ 0.42	**73.0** $\pm$ 0.59	**91.6** $\pm$ 0.17
V^S_{ae}	82.1 $\pm$ 0.22	71.8 $\pm$ 0.31	91.3 $\pm$ 0.05
V^S_{tm}	54.0 $\pm$ 2.41	46.9 $\pm$ 1.99	84.8 $\pm$ 0.45

Table 4. Quality metrics achieved in all experiments. The numbers are mean and standard deviations achieved in three runs per experiment. *Name*: Name of the experiment (cf. Table 3). The best results per series is highlighted in bold font.

5.2.2 Class-specific Quality Metrics: The class-specific quality metrics for the two experimental series on the building dataset are shown in Table 5 and those for the vegetation dataset in Table 6, respectively. Generally, the class-wise IOU and F1 scores are in line with the average metrics.

The highest metrics per class per experimental series on the **building dataset** (Table 5) can be achieved for the multi-modal variant of the classifier, where the ResNet-based encoder is to be preferred over the Swin-based one. A closer look shows that particularly the class of interest, i.e. *Building*, can benefit from the fusion of the modalities. While the IOU for *No Building* is improved by 0.7% (B^R_{ae+tm} compared to B^R_{ae}) and 4.3% (B^R_{ae+tm} compared to B^R_{tm}), respectively, the IOU for *Building* is improved by 2.0% (B^R_{ae+tm} compared to B^R_{ae}) and 11.9% (B^R_{ae+tm} compared to B^R_{tm}), respectively. Similarly, the improvements in the F1 scores are much larger for the class *Building* compared to the class *No Building*. A visual inspection of the predictions (see Figure 3) makes clear that both of the modalities come along with different strengths and the multi-modal classifier benefits from both strengths: the map-based classifiers (B^R_{tm}, B^S_{tm}) allow for the prediction of the basic building structures but fail to predict some buildings that are not contained in the map, as well as fine-grained building parts (which is to be expected from the contents in the provided topographic map); the aerial image classifiers (B^R_{ae}, B^S_{ae}) allow for the prediction of details, but partly fail to correctly predict inner parts of building footprints. The multi-modal classifiers (B^R_{ae+tm}, B^S_{ae+tm}) tend to correctly predict both details and all parts of a building object, which becomes particularly clear from the building at the bottom right in the ResNet-based classifier variants (B_{ae}, B_{tm}, B_{ae+tm} in Figure 3).

In general, the class-wise metrics on the **vegetation dataset** (Table 6) show a similar behaviour as the average ones: the highest metrics are achieved for the multi-modal variants of the classifier. Exceptions from this general observation are the classes *Crop* and *Deciduous* in the ResNet-based series of experiments; while for the class *Crop* both, IOU and F1 score, are (nearly) identical for all three classifiers (V^R_{tm}, V^R_{ae}, V^R_{ae+tm}), the IOU and F1 score of *Deciduous* are on par for the multi-modal classifier (V^R_{ae+tm}) and the uni-modal orthoimage

Class-specific **IOU** [%]			
Class name	Experiment		
	B^R_{ae+tm}	B^R_{ae}	B^R_{tm}
No Building	**90.4** ± 0.28	89.7 ± 0.45	86.1 ± 0.41
Building	**74.1** ± 0.49	72.1 ± 1.10	62.2 ± 0.33
Class name	Experiment		
	B^S_{ae+tm}	B^S_{ae}	B^S_{tm}
No Building	**87.6** ± 0.22	87.2 ± 0.50	85.6 ± 0.12
Building	**67.1** ± 0.74	65.7 ± 1.67	61.1 ± 1.02
Class-specific **F1 scores** [%]			
Class name	Experiment		
	B^R_{ae+tm}	B^R_{ae}	B^R_{tm}
No Building	**94.9** ± 0.19	94.6 ± 0.25	92.6 ± 0.21
Building	**85.1** ± 0.33	83.8 ± 0.74	76.7 ± 0.24
Class name	Experiment		
	B^S_{ae+tm}	B^S_{ae}	B^S_{tm}
No Building	**93.4** ± 0.12	93.1 ± 0.25	92.2 ± 0.05
Building	**80.3** ± 0.49	79.3 ± 1.25	75.8 ± 0.80

Table 5. Class-specific quality metrics on the building dataset. Experiment: Name of the experiment. Best results per experimental series are highlighted in bold font.

classifier (V^R_{ae}). Furthermore, while the highest IOU (90.1%) and F1 score (94.8%) for the class *Crop* are achieved in the multi-modal classifier with attention-based features (V^S_{ae+tm}), the highest scores for the class *Deciduous* are achieved with the uni-modal aerial classifier relying on convolution-based features (V^R_{ae}). This might be caused by the requirement to predict not only large structures but also fine-grained details for these two classes, particularly for *Deciduous*, which becomes clear from Figure 4. This Figure also visualizes that the fine-grained *Deciduous* tree objects along the roads are not predicted by the map-based classifiers (V^R_{tm}, V^S_{tm}), which is reasonable given the provided input data (TK25 in Figure 4; cf. also Section 4.3) and explains the lower metrics for *Deciduous* trees achieved by the map-based classifiers (see Table 6). Similarly, the metrics in Table 6 and the predictions in Figure 4 for the class *Coniferous* trees by the classifiers V^R_{tm}, V^S_{tm} can be explained by the input TK25; as the map does not provide any information about coniferous trees in the forested area at the bottom right, the classifier can not correctly predict them, even though they are present according to the orthophoto (DOP in Figure 4). Nevertheless, the multi-modal classifiers (V^R_{ae+tm}, V^S_{ae+tm}) rely on the more informative DOP for predicting *Coniferous* trees and are even better than the aerial classifiers (V^R_{ae}, V^S_{ae}) in correctly predicting *Coniferous* trees, which is assumed to be caused by a more homogeneous representation of the other classes in the map, leading to less confusion with the class *Coniferous* trees in the multi-modal cases. To summarize, the ResNet-based classifier and particularly the Swin-based classifier benefit from the joint consideration of aerial images and maps in general, with fine-grained structures exclusively contained in the aerial image being slightly better predicted by the uni-modal aerial classifier relying on convolutions.

6. Conclusions & Outlook

In this paper, a multi-modal classification approach was proposed that combines historic aerial orthoimages and historic topographic maps for the pixel-wise prediction of land cover. To the best of the knowledge of the authors, this is the first classifier that successfully combines these two modalities for pixel-wise classification. The classifier benefits from both modalities by exploiting fine-grained information about the Earth's surface contained in the aerial images and exploiting high-level features

Class-specific **IOU** [%]			
Class name	Experiment		
	V^R_{ae+tm}	V^R_{ae}	V^R_{tm}
Crop	**89.5** ± 0.21	**89.6** ± 0.31	**89.6** ± 0.29
Deciduous	89.3 ± 0.94	**89.6** ± 0.29	83.2 ± 0.59
Coniferous	**65.8** ± 3.35	63.1 ± 5.84	0.3 ± 0.33
Other	**43.7** ± 2.41	42.9 ± 43.7	32.3 ± 1.44
Class name	Experiment		
	V^S_{ae+tm}	V^S_{ae}	V^S_{tm}
Crop	**90.1** ± 0.50	89.4 ± 0.25	85.2 ± 2.91
Deciduous	**89.5** ± 0.21	89.0 ± 0.16	81.7 ± 1.16
Coniferous	**66.7** ± 2.92	64.7 ± 1.50	0.2 ± 0.28
Other	**45.8** ± 1.51	44.1 ± 1.20	20.7 ± 6.10
Class-specific **F1 scores** [%]			
Class name	Experiment		
	V^R_{ae+tm}	V^R_{ae}	V^R_{tm}
Crop	**94.5** ± 0.12	**94.5** ± 0.17	**94.5** ± 0.17
Deciduous	94.4 ± 0.53	**94.6** ± 0.17	90.8 ± 0.36
Coniferous	**79.4** ± 2.45	77.2 ± 4.37	0.7 ± 0.59
Other	**60.8** ± 2.37	60.1 ± 0.91	48.8 ± 1.69
Class name	Experiment		
	V^S_{ae+tm}	V^S_{ae}	V^S_{tm}
Crop	**94.8** ± 0.28	94.4 ± 0.17	92.0 ± 1.69
Deciduous	**94.5** ± 0.12	94.2 ± 0.08	89.9 ± 0.68
Coniferous	**80.0** ± 2.12	78.6 ± 1.11	0.4 ± 0.52
Other	**62.8** ± 1.39	61.2 ± 1.12	33.8 ± 8.14

Table 6. Class-specific quality metrics on the vegetation dataset. *Deciduous* / *Coniferous*: *Deciduous trees* / *Coniferous trees*. Experiment: Name of the experiment. Best results per experimental series are highlighted in bold font.

of both, maps and aerial images. These high-level features are fused based on attentions, allowing to focus not only on relevant local and global contexts but also on relevant feature maps. A comprehensive evaluation demonstrates that the multi-modal classifier outperforms both uni-modal counterparts. This emphasizes that even a coarse representation in terms of maps can support the interpretation of the images. Conversely, it is clear that the classification of maps using the image-related ground-truth labels cannot yield an optimal solution, due to the discrepancies described above. Furthermore, the experiments indicate that convolution-based feature extraction is to be preferred for fine-grained objects, whereas larger objects are predicted in a better way utilizing attention-based features.

Future work can build on the present one in various ways. From a methodological point of view, it could be investigated how the multi-modal classifier could be further improved. For instance, auxiliary supervision, e.g. (Garnot et al., 2022), could explicitly force the classifier to extract as much information as possible from each modality. Furthermore, the method could be adapted such that it allows for inputs with different GSDs, avoiding the resampling of the input and consequently, the loss of information in case of downsampling. In addition, further sources of information could be considered, e.g. satellite images, e.g. (Garioud et al., 2023; Heidarianbaei et al., 2024), or height information derived from historic aerial stereo images (Le Bris et al., 2020). It might also be helpful to start from pre-trained weights obtained from topographic maps and aerial orthoimages by self-supervised learning (Wang et al., 2022), which would have to be adapted for the modalities of interest. From an application point of view, it would be interesting to investigate the generality of the approach in terms of its ability to perform on datasets of other regions as well as on datasets of other epochs. As it is assumed that this might be challenging, regional domain adaptation, e.g. (Wittich and Rottensteiner, 2021), and temporal domain adaptation, e.g. (Mboga et al., 2021; Van den Broeck et

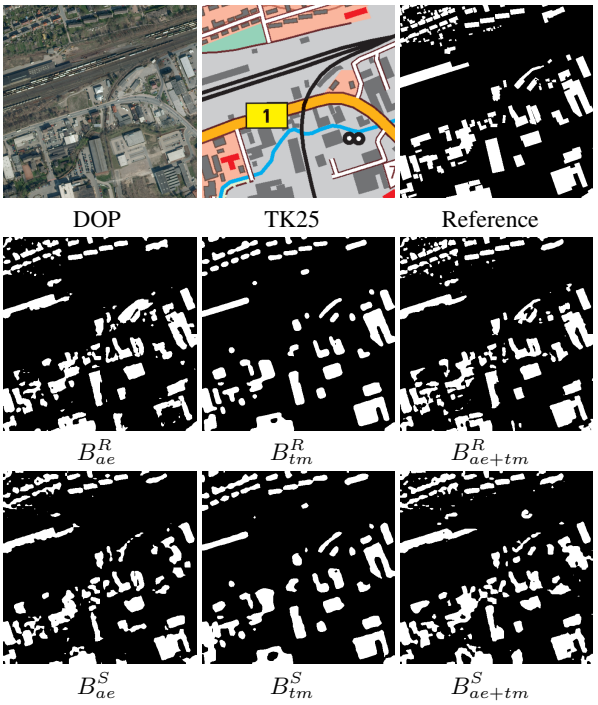

Figure 3. Test tile for the building dataset. The first row shows the available data (DOP: digital orthophoto; TK25: topographic map; Reference: Reference label map) and the second and third rows show the predictions of the classifiers (cf. Table 3).

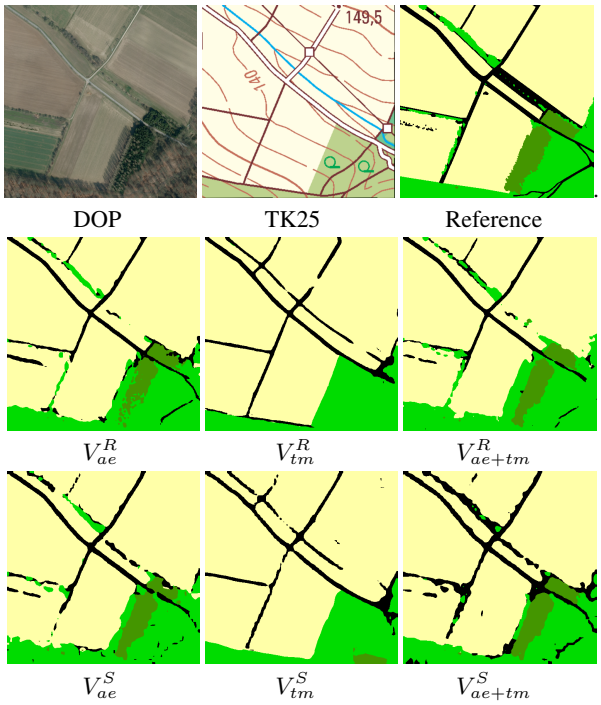

Figure 4. Test tile for the vegetation dataset. The first row shows the available data (DOP: digital orthophoto; TK25: topographic map; Reference: Reference label map) and the second and third rows show the predictions of the classifiers (cf. Table 3).

al., 2022), would be options to tackle such challenges. Finally, in order to be able to analyse the development of land cover over time, the method presented in this paper will be embedded into an approach for the classification of time series of historic multi-modal data.

References

Chen, K., Weinmann, M., Gao, X., Yan, M., Hinz, S., Jutzi, B., Weinmann, M., 2018. Residual shuffling convolutional neural networks for deep semantic image segmentation using multi-modal data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2, 65–72.

Chiang, Y.-Y., Chen, M., Duan, W., Kim, J., Knoblock, C. A., Leyk, S., Li, Z., Lin, Y., Namgung, M., Shbita, B. et al., 2023. Geoai for the digitization of historical maps. *Handbook of Geospatial Artificial Intelligence*, CRC Press, 217–247.

Dorozynski, M., Rottensteiner, F., Thiemann, F., Sester, M., Dahms, T., Hovenbitzer, M., 2024. Multi-modal land cover classification of historical aerial images and topographic maps: A comparative study. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4-2024, 107–115.

Ettehadi Osgouei, P., Sertel, E., Kabadayı, M. E., 2022. Integrated usage of historical geospatial data and modern satellite images reveal long-term land use/cover changes in Bursa/Turkey, 1858–2020. *Scientific Reports*, 12(1). 9077.

Garioud, A., Gonthier, N., Landrieu, L., De Wit, A., Valette, M., Poupée, M., Giordano, S., Wattrelos, b., 2023. FLAIR: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems (NIPS)*, 36, 16456–16482.

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.

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, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. *Proceedings of the IEEE/CVF European Conference on Computer Vision (ECCV)*, 630–645.

Heidarianbaei, M., Kanyamahanga, H., Dorozynski, M., 2024. Temporal ViT-U-Net Tandem Model: Enhancing Multi-Sensor Land Cover Classification Through Transformer-Based Utilization of Satellite Image Time Series. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-3-2024, 169–177.

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.

Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2020. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5), 4340–4354.

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.

Kim, J., Li, Z., Lin, Y., Namgung, M., Jang, L., Chiang, Y.-Y., 2023. The MapKurator system: A complete pipeline for extracting and linking text from historical maps. *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*.

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

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 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.

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.

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.

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).

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.

Minervino Amodio, A., Gioia, D., Danese, M., Masini, N., Sabia, C. A., 2023. Land-use change effects on soil erosion: the case of Roman “Via Herculia” (southern Italy) – Combining historical maps, aerial images and soil erosion model. *Sustainability*, 15(12), 9479.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015, part III 18*, 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.

Sertel, E., Avci, C., Kabadayi, M. E., 2023. Deep learning-based land use land cover segmentation of historical aerial images. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2622–2625.

Song, C. H., Han, H. J., Avrithis, Y., 2022. All the attention you need: Global-local, spatial-channel attention for image retrieval. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2754–2763.

Stocker, O., Le Bris, A., 2020. Can Spot-6/7 CNN semantic segmentation improve sentinel-2 based land cover products? Sensor assessment and fusion. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020, 557–564.

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.

Van den Broeck, W. A. J., Goedemé, T., Loopmans, M., 2022. Multiclass land cover mapping from historical orthophotos using domain adaptation and spatio-temporal transfer learning. *Remote Sensing*, 14(23).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, Polosukhin, I., 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30.

Voelsen, M., Rottensteiner, F., Heipke, C., 2024. Transformer models for land cover classification with satellite image time series. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92, 547–568.

Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., Zhu, X. X., 2022. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4), 213–247.

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.

Wang, Y., Zhang, W., Chen, W., Chen, C., Liang, Z., 2024. MFFnet: Multimodal feature fusion network for synthetic aperture radar and optical image land cover classification. *Remote Sensing*, 16(13), 2459.

Wittich, D., Rottensteiner, F., 2021. Appearance based deep domain adaptation for the classification of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180, 82–102.

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.