The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B1-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission I, 4–11 July 2026, Toronto, Canada

# Zero-shot multi-class semantic segmentation of remote sensing images using SAM 2 with prior database information

Paula L. Lippmann, Mareike Dorozynski, Franz Rottensteiner, Christian Heipke

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
{lippmann, dorozynski, rottensteiner, heipke}@ipi.uni-hannover.de

**Keywords:** Semantic Segmentation, Visual Foundation Models, Segment Anything Model

**Abstract**

Remote sensing images (RSI) play a central role in the regular update of land cover datasets. A first step in this process is RSI semantic segmentation, which is mainly solved by deep learning today. Vision foundation models (VFM) have gained increasing importance in this context. Having been trained on large datasets, VFM for segmentation can yield good results on data from various domains without further training. We present a new method for using the VFM Segment Anything Model 2 (SAM 2) for multi-class semantic segmentation of Sentinel-2 images that does not require training data. Our method is based on a prompt engineering approach, using SAM 2 in its pre-trained form and generating different prompt types on the basis of existing topographic data. We also propose a post-processing step for merging the output of SAM 2 to obtain a multi-class label image. The results of our experiments show that our method achieves an overall accuracy (OA) of up to 93% at pixel-level using polygon mask prompts, while using point and box prompts leads to an OA of 81% and a mF1-score of 72%. Experiments with other Sentinel-2 composite images do not show significantly different results compared to R-G-B images. Incorporating data from different time steps for map updating shows good results, but remains inconclusive due to the small amount of change in the dataset.

## 1. Introduction

Land cover data is stored in topographic databases and need to be updated regularly because of frequent land cover changes. Nowadays, this is still mainly done manually. For instance, in the German federal state of Lower Saxony, the topographic data (data of the Authorative Topographic-Cartographic Information System (ATKIS) (AdV, 2025)) is currently updated regularly on the basis of aerial images taken every three years; also satellite images play a certain role. A first step towards achieving automated updates is the semantic segmentation of these remote sensing images (RSI). Semantic segmentation, also called pixel-wise classification, is the assignment of a class to every pixel. This information can then be used for change detection and map updating. This paper only deals with the first step, semantic segmentation.

Today, semantic segmentation is mainly solved by deep learning (DL). Traditional DL models include convolutional neural networks (CNN), transformer models and hybrid models, specifically designed for a selected task. These models achieve good results, but face the challenge of scarcity of the needed labelled data for training. For some years, a growing number of visual foundation models (VFM) has been published and gained increasing importance in this context. Having been trained on large datasets, VFM can yield good inference results on data from various domains without further training; this ability is called a good ”zero-shot performance”. In particular, VFM specialise in specific generic tasks and can be transferred to various other applications. The Segment Anything Model (SAM) (Kirillov et al., 2023) is a large VFM designed for image segmentation based on user-provided prompts. These prompts contain hints from the user regarding the approximate position of area of interest to be segmented. Its extension SAM 2 (Ravi et al., 2025) can perform prompt-based segmentation of both images and videos. To overcome the problem of label scarcity in remote sensing, this VFM could be applied to RSI directly, without further training. However, SAM and SAM 2 do not predict class labels. To apply this model for database updates, class labels are necessary, so extensions are necessary to use this model for semantic segmentation.

There is some research on transferring VFM, trained for general segmentation using natural images, to other domains. The objective is to utilize the general features and knowledge acquired through the VFM for a different task. Specifically, there has been some research on using SAM and SAM 2 for semantic segmentation and change detection based on RSI (e.g. Ren et al., 2024a; Sun et al., 2024; Qiu et al., 2025; Sultan et al., 2025). Wan et al. (2025) identifies three main strategies for integrating VFM, such as SAM or SAM 2, into a corresponding workflow. Firstly, SAM can be used in its original pre-trained form (zero-shot). Thus, on the one hand it can serve as an auxiliary tool for other networks where the output of SAM is used to enhance specialised models (Liu and Zhou, 2024; Li et al., 2024; Ma et al., 2024). SAM relies on input prompts to segment specific regions of the image. Therefore, on the other hand the zero-shot performance can be improved through prompt engineering to enhance the output of the pre-trained SAM. For example Ren et al. (2024b) and Sultan et al. (2025) use another network to generate task-specific prompts for SAM. Secondly, parts of SAM can be fine-tuned, as done in Chen et al. (2024), and finally trainable adapters can be integrated into SAM (Zhang et al., 2025). There are also combinations of these strategies (Luo et al., 2024). In any case, the second and third strategies require labelled training data.

In this work, we present a new method for using the VFM SAM 2 for multi-class semantic segmentation of Sentinel-2 images. Our method can be seen as a prompt engineering approach combined with a post-processing step. While it uses

SAM 2 in its pre-trained (frozen) form, the prompts are generated based on existing topographic data, enabling our method to be used for database updates. This aspect has not been investigated in detail in this paper. Instead we focus on the multi-class semantic segmentation using SAM 2. To the best of the authors' knowledge, there is no approach for multi-class semantic segmentation using a pre-trained VFM, that makes direct use of topographic database information. Thus, the main contributions of this work are as follows:

1. We propose a new method for multi-class semantic segmentation of RSI based on the frozen SAM 2 that does not require any training data.
2. We propose an automated prompt generation strategy that uses information available in a topographic database.
3. We propose a new post-processing method for merging the outputs of the VFM to obtain a multi-class label image.
4. As a small contribution, we investigate which types of prompts are most useful in the context of our proposed method.

Although this paper focuses on the VFM SAM 2, the proposed method should also be transferable to other VFM that work with prompting and can output logits.

## 2. Related Work

In this section we discuss related work for semantic segmentation and prompt engineering. We set the focus mainly to the application of SAM or SAM 2 for RSI, but do present general work as well. SAM (Kirillov et al., 2023) is a large VFM designed for image segmentation based on user-provided prompts. Its extension SAM 2 (Ravi et al., 2025) can perform prompt based segmentation of both images and videos (several single frames).

SAM and SAM 2 are VFM trained for the segmentation of natural images without providing class information. In order to adapt and improve SAM for other applications related to natural images, different strategies are proposed, i.e. prompt engineering, adapter insertion, or fine-tuning.. For example, to perform semantic segmentation, Ren et al. (2024b) introduce Grounded SAM for natural images. The semantic class for the segmentation is obtained via an object detector model, Grounding DINO (Liu et al., 2025). The object detector model predicts bounding boxes of specific objects whose class is prompted via text. These boxes are then used as box prompts into SAM to perform segmentation of these specific objects. This approach focusses on prompt engineering, an approach to redesign and adjust the prompts with respect to the task at hand. In Chen et al. (2023), the focus is set on tasks like shadow detection or camouflaged object detection in natural images. Therefore an adapter is inserted into SAM (SAM-Adapter), to improve the segmentation result. By inserting trainable adapters into a pre-trained network, the effort required to retrain the model is reduced because only the adapters will be trained while the other parameters of the model remain unchanged (frozen). Nevertheless, training data is necessary for this task. A third strategy involves fine-tuning the pre-trained model. Fine-tuning methods aim to optimize some or all parameters of a model for an appropriate solution of the new problem (Goodfellow et al., 2016).

There has also been some research on VFM that focused on their capabilities for tasks related to RSI. This involves VFM trained on natural images (Wan et al., 2025) and VFM trained explicitly on RSI (Lu et al., 2025). Analogous to the proposed adaptation strategies for natural images, semantic segmentation is also performed on RSI using SAM. In the field of adapter insertion and fine-tuning, Zhang et al. (2025) propose an approach to improve the segmentation result for RSI segmentation. The authors insert two adapters into SAM for the extraction of remote sensing-related information. These adapters incorporate information from different scales of RSI and can extract task specific features. Annotated training data is necessary to train the adapters. In Sultan et al. (2025) the mask decoder is fine-tuned to be able to output multi-channel segmentation maps, instead of binary segmentations. Fine-tuning the model also requires training data, which is not always available in sufficient amounts in the RSI field. For this reason we focus on a new approach which does not need training data. Our approach thus uses a prompt engineering technique.

Other methods for prompt engineering have also been proposed. These methods either work directly with available data or other pre-trained models or use training to generate prompts. In Diab et al. (2025), the authors propose a pipeline for aerial image segmentation that is promptable via natural language (text) and is refined using pre- and post-processing steps. The pipeline uses user defined text prompts, which assumes knowledge about the visible data, in combination with aerial images for the generation of bounding boxes via Grounding DINO, similar to Ren et al. (2024b). The resultant bounding boxes are filtered to eliminate outliers and subsequently used as prompts for the frozen SAM. By merging all segmentation masks a binary segmentation is obtained. The works of Osco et al. (2023) and Ren et al. (2024a) explore SAM's segmentation capabilities in zero-shot applications. Instead of a pre-trained model for prompt generation, the authors use available data (polygons or raster data), to generate box and point prompts. These approaches require labelled data to be applicable. On the other hand, in (Zhang et al., 2025), the box prompts are derived from high-frequency components via Fast-Fourier transform of one of the trainable adapters (based on hand crafted features). In Luo et al. (2024) a new object detector is trained to automatically obtain object categories and bounding boxes from the multi-scale features of the image encoder for box prompting. In Chen et al. (2024) a new module (*prompter*) is developed. This *prompter* is trained to create category-related prompt embeddings directly from the multi-scale image features of the image encoder. Sultan et al. (2025) use a pre-connected network (here a U-Net model, which produces a multi-class segmentation map and which needs to be trained) to generate pseudo labels, which are then used to extract point prompts. These prompt engineering approaches all rely on specifically collected training data. In contrast, we do not use labelled data, but we propose to use data from a topographic database to generate the prompts; we do not need to perform any training.

Another aspect is the output type of the above-mentioned models. One of the main applications of SAM is semantic segmentation (Wan et al., 2025). SAM performs class-independent segmentation, which means that the output is just a binary segmentation based on the input prompt. Therefore, it is necessary to add further steps in order to perform multi-class semantic segmentation. Most of the approaches mentioned above only focus on binary segmentation. In other words, the prompts usually represent only a single class, for which the semantic segmentation is performed. In Ren et al. (2024a), the authors also propose a method called RAM-Grounded-SAM, which uses the image tagging model Recognize Anything (RAM) (Zhang et

al., 2024) to recognize class categories in natural images. These categories are then used as text prompts into the object detector (Grounding DINO) to generate bounding boxes. Afterwards, the detected boxes are fed into SAM, and the resulting masks are combined to generate dense annotations of the images. This can be seen as a multi-class semantic segmentation of natural images, although the authors do not clarify how they combine the binary segmentations into dense segmentations. With respect to RSI, in (Zhang et al., 2025) it is possible to generate high frequency components for each class separately out of the inserted adapter. But this requires individual training of the model for each class. The final class for each pixel can then be determined by selecting the class with the highest probability. An approach that goes one step further is proposed by Luo et al. (2024). Here, the focus is set on the segmentation of instances, i.e. individual objects of a class. The model is trained and used with multi-class data and can output all instances of all classes simultaneously. Sultan et al. (2025) combine logit values from SAM to predict multi-class label images. Unlike our method, however, the SAM mask decoder is retrained here, which needs training data. To the best of the authors' knowledge, there is no method for multi-class semantic segmentation using SAM without further training. To overcome the problem of scarcity of available training data we propose a post-processing step, instead of training, to obtain multi-class label images from the raw outputs of the VFM.

One open question that remains is which prompts work best. SAM 2 permits the input of three different types of prompts. In Osco et al. (2023) and Ren et al. (2024a) these different prompt types are tested. The results achieved are contradictory in these two works, which further motivates our analysis of different prompt types for the proposed method in this paper.

In summary, to the best of our knowledge, there is no method using SAM for multi-class semantic segmentation of RSI without training data. Existing models either use a prompt engineering approach that requires training, or combine prompt engineering directly with approaches such as adapter insertion or fine-tuning, which also need training. Using fine-tuning and adapter modules to employ a VFM for downstream tasks can be computationally expensive and memory-intensive. Furthermore, as mentioned before, only a limited amount of labelled data for remote sensing applications exists. For this reason, we present a new method that does not require any training. Another finding is that most of the existing works focus on the segmentation of individual classes, resulting in binary segmentation. However, as a first step for future map updates, comprehensive segmentation is necessary. Therefore, we propose a method that combines the raw data of the VFM in a post-processing step, to obtain a multi-class label image. In order to take full advantage of this new method, various prompt types and their combinations are tested. Our method enables multi-class semantic segmentation of RSI, with topographic database information without training of the VFM.

## 3. Segment Anything Model

The general structure of SAM and SAM 2 consists of three main parts. In the following we restrict the description to SAM 2, as this is the model used in our work. SAM 2 takes an image and prompts indicating locations of the area of interest as input. The output consists of a binary segmentation mask. It is also possible to output the raw scores per pixel, called logits. These are the values before applying the sigmoid function, which is used for binary segmentation. The logits indicate how certain SAM 2 is to assign a pixel to the foreground segment based on the input prompt. The prompts are a mandatory input into SAM 2. They are designed to give hints about the location of the area in the image to segment. Possible prompt types are points, boxes and masks. Point prompts can be positive and negative prompts indicating a location to be part of a segment (positive) or explicitly not to be part of a segment (negative). The box prompts are axis-aligned bounding boxes that surround the area of interest. The mask prompts are binary images with $1/4$ of the resolution of the input images. Prompts per area of interest are processed individually. Nevertheless, an area can be prompted with a combination of prompt types. Also, several image parts can be prompted simultaneously using batches, i.e. several prompts together with one image as input. Neither SAM nor SAM 2 contain any class information for the segments. The overall structure of SAM is depicted in Figure 1, where the additional parts of SAM 2, designed for video segmentation, are marked in grey; these components are not used in this work.

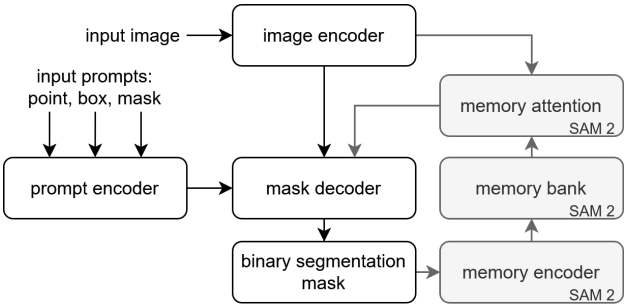



Figure 1. Structure of SAM and SAM 2. The memory mechanisms, marked in grey, are only available in SAM 2 for the segmentation of videos. Thus they can be skipped when using single images as input.

First, an image encoder processes the input image of size $1024\ px \times 1024\ px \times 3$ channels to obtain image embeddings of size $64 \times 64 \times 256$. The input image may have a different height and width, but will then be resized via bilinear interpolation to fit the size defined by SAM 2. The image encoder of SAM 2 is a hierarchical vision transformer (Ryali et al., 2023; Bolya et al., 2024) pre-trained as a masked autoencoder (He et al., 2022). This hierarchical encoder enables the use of multi-scale features in the mask decoder. A prompt encoder processes the prompts to generate prompt embeddings. Both, image and prompt embeddings, are then combined in the mask decoder to predict a binary segmentation mask with the same resolution as the input image. SAM 2 allows the user to choose whether to generate a single or three alternative segmentation masks that the model considers plausible for each input prompt. In this work, we have chosen to output a single segmentation mask. This is always the mask that receives the highest internal score from SAM 2. The mask decoder uses transformers to combine prompt and image embeddings into a segmentation mask. Output masks are generated for each prompt. More details about SAM 2 can be found in Ravi et al. (2025).

The SAM 2 was jointly trained on a very large dataset of images and videos. The training dataset contained images with mask annotations without any class label for the images or mask. The mask annotations range from large annotations like buildings (not top-view) to fine grained annotations like door handles. The focus is set on *natural images* without domain specific data

like medical images or RSI, which is why further developments are necessary to transfer the model to these domains.

## 4. Methodology

We propose a new method for multi-class semantic segmentation of RSI using SAM 2 with frozen parameters. Our method takes information from a topographic database for prompt generation and RSI as input. The topographic database must contain polygons with corresponding class information. The RSI need to be georeferenced and contain three channels with a radiometric resolution of 8 bit each. The output of our method is a multi-class label image with the same resolution as the input image. An overview of the workflow is shown in Figure 2.

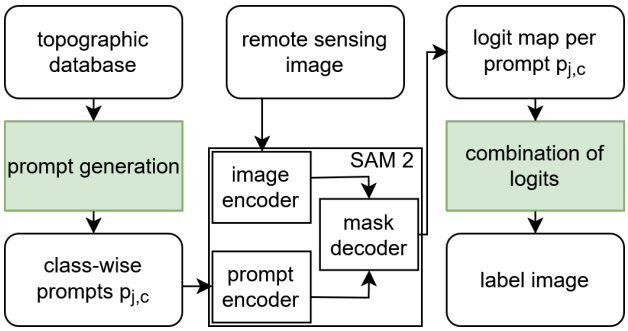



Figure 2. Proposed method for zero-shot semantic segmentation of RSI using SAM 2. The developed modules are marked in green. $p_{j,c}$ is the prompt for polygon $j$ of class $c$.

Our method adds two modules to SAM 2. The first is the prompt generation and the second is a post-processing step for multi-class semantic segmentation. The content of the topographic database is used to generate prompts. The prompt generation is done separately for each class $c$ and polygon $j$ to obtain class-wise prompts $p_{j,c}$ (sec. 4.1). All prompts $p_{j,c}$ of one class $c$ for one image form a batch of prompts and are fed into the prompt encoder of SAM 2, while simultaneously using the corresponding RSI as input into SAM 2's image encoder. The embeddings are then processed in the mask decoder. The output is a logit map per prompt $p_{j,c}$ with the resolution of the input image. These logit maps are used in post-processing, where all logit maps of one class $c$ are combined to generate one single class logit per pixel for each class. The multi-class label image of the entire RSI is then generated by identifying the class with the maximum logit among all class logits on a per pixel basis (sec. 4.2). The developed modules of our method are describe in the following subsections.

### 4.1 Prompt generation

All three possible prompt types, points, boxes and masks, can be generated based on the information contained in the database. The class label associated with a prompt is derived from the class of a polygon in the database. To generate prompts, the position, shape, size and class information of the polygons in the database are used.

Point and box prompts are generated individually for each polygon from the database. To generate a point prompt, a representative point is selected within the polygon. In this work we only use positive point prompts. These prompts indicate one point inside the segment being searched for. In order to find such a point, we use an algorithm where a horizontal line which does not intersect any polygon corner and which lies near the midpoint of the polygon's vertical extent is used. The longest section of the line that lies within the polygon is determined, and the corresponding centre point is selected as the representative point[1]. This ensures that the point lies within the polygon.

To generate a box prompt, the minimum and maximum values of the coordinates of an individual polygon $j$ are chosen to generate the axis-aligned minimum bounding rectangle. The coordinates of the upper left and the lower right corner of the bounding box are used to define the box prompt.

Mask prompts are binary images as described in sec. 3. There are two options to generate these prompts. Either single polygons can be used, or all polygons of one class can be processed together. When using every single polygon separately, there will be as many masks per image as there are polygons. These masks are called *polygon masks* in this paper. Each mask indicates for every pixel whether it belongs to the corresponding polygon (pixel value 1) or not (0). When using all polygons of one class jointly, the process results in one mask per class and image. This mask is called *combined mask* in this paper and each mask indicates for every pixel whether it belongs to the corresponding class or not.

Except in the case of the combined mask, the number of prompts for class $c$ corresponds to the number of polygons $j$ of that class in the database. It is possible to combine several prompts and different prompt types in a batch and to feed the batch into SAM 2. A batch is generated by using the prompts of all polygons $j$ of one class $c$ inside an RSI, resulting in one batch per image. Depending on the size of the batch and the system's processing power, the batches can also be divided into sub-batches.

### 4.2 Multi-class semantic segmentation

In the post-processing step, the logit maps are combined to obtain a multi-class label image. The process is visualized in Figure 3. For each prompt $p_{j,c}$, the resulting logit map contains logits $l_{j,c}$ per pixel. These logits are associated with the corresponding class $c$ of the prompt. Our proposed approach involves evaluating the maximum logit values on a per-pixel basis for each class separately.

When employing point, box or polygon mask prompts, we first need to aggregate the logits of all prompts of a class by selecting the maximum logit value per pixel. This is necessary, because with these prompt types, one prompt per polygon $j$ is generated and combined into a batch of prompts for one class to be processed. After aggregation, one logit per pixel for this specific class (here called class logit $l_c$) is obtained and stored. This aggregation step is then repeated for each class to obtain one class logit for each pixel $l_c$ for each class separately. Subsequently, the final class of a pixel is determined by selecting the class with the maximum class logit $l_c$ among all class logits.

For combined mask prompts, the first step does not need be performed, as there is only one input prompt per class, i.e. only one logit per pixel per class $l_c$ from the outset. These class logits $l_c$ can then directly be used for the second step, the selection of the class with the maximum class logit.

[1] `https://locationtech.github.io/jts/javadoc/org/locationtech/jts/algorithm/InteriorPointArea.html`

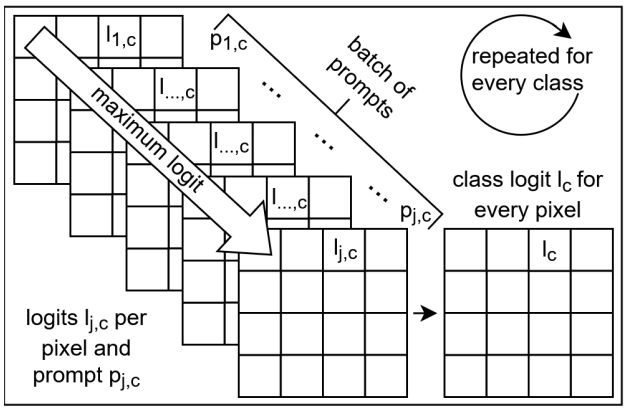



Figure 3. Combination of logits per pixel and prompt $l_{j,c}$ into class logits $l_c$ for every pixel. This step is performed for every class separately.

## 5. Experiments

### 5.1 Dataset

The test area we use covers the whole area of the Federal State of Lower Saxony in Germany, approx. 47 700 m$^2$ in size. The RSI are Sentinel-2 images acquired on or nearest to 31/03/2019 and 31/03/2020. The images are split into tiles of $8 \times 8$ $km^2$ and filtered for a cloud coverage smaller than 5% to avoid disturbances. We use the precomputed true colour images (TCI) as well as the georeferenced bottom-of-atmosphere reflectance data. TCI are R-G-B images precomputed from the Red (R), Green (G), and Blue (B) bands[2].

To obtain the information for prompt generation, data from the official German landscape model (ATKIS) are used (AdV, 2025). ATKIS is updated every three months based on specified standards for the geotopography. The basis for the updates are aerial images taken every three years and in-situ surveys. The topographic data are stored as polygon, line or point geometries with corresponding attributes. For our experiments, we use the ATKIS data from 31/03/2019 and 31/03/2020 with the corresponding Sentinel-2 images acquired nearest to these dates.

The available ATKIS geometries are classified according to the ATKIS object type catalogue (AdV, 2008). In this work, we use the available polygon and line geometries of ATKIS in order to derive *land cover classes*. The line geometries represent the centre line of an object with a specific width. This width is used to create polygons from the line geometries using buffering. As the data includes more than 100 classes, the original object types are aggregated into 7 *land cover classes* (cf. Voelsen et al., 2024): *settlement (stl.), barren land (bar.), vegetation (veg.), sealed area (sld.), agriculture (agr.), forest (for.)* and *water (wat.)*. In addition, the class *other* is used for areas outside the border of Lower Saxony, which are available in the Sentinel-2 images but not in the database. This class is not taken into account for the evaluation.

### 5.2 Experimental setup

The method is tested with different combinations of prompt types: only points, only boxes, only polygon masks and only combined masks. Also, other combinations are tested: points with boxes, points with polygon masks or combined masks, boxes with polygon masks or combined masks, and all three prompt types combined. The prompts are generated using the prompt generation method proposed in section 4.1. These prompt types together with the Sentinel-2 images are tested with two different versions of SAM 2; the large version with 224.4 million parameters and the tiny version with 38.9 million parameters (Ravi et al., 2025). These experiments are designed to show how effectively the proposed prompt generation works for multi-class semantic segmentation. To do this, we first use the Sentinel-2 images (TCI) from end of March 2019 together with the prompts from 31/03/2019.

Secondly, an experiment is performed with different 3-channel input images. SAM 2 is restricted on an input of 3-channel 8-bit images, but RSI usually contain more than 3 channels. In order to be able to use the information content of different channels, we create new 3-channel composite images out of the georeferenced bottom-of-atmosphere reflectance data of the Sentinel-2 images, instead of using the preprocessed TCI. The bottom-of-atmosphere reflectance data consists of 16-bit images. For the composition of the 3-channel images, we need to convert the data into 8-bit. For this purpose, we calculate the 1%-quantile ($q1$) and 99%- quantile ($q99$) of the bands 2, 3, 4 and 8 (blue, green, red and near infrared (NIR)) to determine a new interval for the bottom-of-atmosphere reflectance and to eliminate possible outliers. Instead of using individual quantile values per band, we use the mean value of the quantile values ($\mu_{q1}, \mu_{q99}$) of the used bands per $8 \times 8$ $km^2$ tile. Subsequently, we normalize the data. For the normalisation, we use $g' = (g - \mu_{q1})/(\mu_{q99} - \mu_{q1}) \cdot 255$ and clip the values to $g' \in [0, 255]$. Here, $g'$ are the new grey values per band and $g$ are the raw bottom-of-atmosphere reflectance values. We generate NIR-R-G and R-G-B images by stacking the new values per band from the three corresponding channels into an image of size $H \times W \times 3$. We perform experiments with both NIR-R-G images and R-G-B images. In doing so, we intend to investigate, if SAM 2 can also produce reasonable results when fed with images that consist of channels other than its training data (namely R-G-B). Especially for RSI, these different channels contain additional useful information. Although the TCI already consist of R-G-B channels, in this experiment we generate R-G-B images from the 16-bit bottom-of-atmosphere reflectance values in order to establish comparability with the NIR-R-G images.

The third experiment is related to map updating. In this case we cannot select images and prompts from the same time epoch. Rather, we perform further experiments with the best versions of the previous experiments, using the prompts from 31/03/2019 with Sentinel-2 images from a year later, end of March 2020.

For all experiments we use the frozen SAM 2 for inference. As no random components are involved, each experiment is carried out only once. The evaluation of the experiments is performed on a per pixel level using (rasterised) ATKIS data. As the database is not fully updated every three month, there is no fully correct ground truth available. The used reference may therefore contain label noise. This can limit the interpretability of the results. The performance of our proposed semantic segmentation is measured by evaluating the prediction accuracy compared to the reference. We use the class-specific F1-score ($F1_c$), mean F1-score (mF1) and overall accuracy (OA) as evaluation metrics. The F1-score for a class $c$ (eq. 1) is based on the number

[2] https://sentiwiki.copernicus.eu/web/glossary#Glossary-TrueColourImages(TCI)

of true positive ($TP_c$), false positive ($FP_c$) and false negative ($FN_c$) predictions for that class:

$$F1_c = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c}. \quad (1)$$

The mF1 is the mean of the F1-score over all classes, and the OA is the percentage of pixels with correctly predicted class labels.

The evaluation of the first and second experiment is carried out with the rasterised ATKIS data from 31/03/2019, while for the third experiment, the rasterised ATKIS data from 31/03/2020 is used for the calculation of the evaluation metrics.

### 5.3 Analysis of different prompt types and tiny and large model versions

The results achieved when using TCI and different prompting strategies with the tiny and large versions of SAM 2 are shown in Tables 1 and 2, respectively. When using the tiny version, the two types of mask prompts perform best in OA (93.8% and 90.1%) and mF1-score (83.9% and 80.0%). The results of the approach of using a single mask per polygon are better than using one mask for all polygons of a class. This is probably due to the higher intra-class variance that occurs if all polygons of one class are considered as a single entity in the mask, which makes it more difficult for SAM 2 to determine the true segment. The results for single point prompts are worst, with an OA of 69.6% and mF1-score of 55.3%. Combining mask prompts with point prompts increases the OA and mF1-score by approximately 9% compared to just using point prompts. The results for box prompts and any combination of prompts with box prompts are third best.

Performing the same prompting types with the large version of SAM 2 generally leads to worse results for both types of mask prompts. For the point and box prompt combination a similar result is achieved, while the combination of all three prompt types together leads to slightly better results. The exact reason remains unclear. The best result is reached again with the polygon mask prompt (OA of 84.1%). The results for point prompts are worst, for box prompts they are better, and combinations of points and boxes again increase the results slightly compared to single box or single point prompts.

Considering the individual F1-scores of the classes, it can be observed that the results for the classes *barren land* and *sealed area* are particularly poor in comparison to the results of all other classes. This might be due to the shape of the polygons. *Sealed area* contains ATKIS classes like roads and railways, and *barren land* contains narrow objects along rivers and streets. These classes tend to be elongated polygons in the database. When prompting boxes for these polygons, the bounding boxes can become very large and cover a significantly larger area than the actual polygon. This makes it difficult for the model to predict just the intended elongated polygon by segmentation. Also for point prompts, one single point might not be enough to correctly represent such a long polygon. Furthermore, narrow polygons such as roads partially disappear in the downsampled mask prompts because they are too narrow. It therefore seems necessary to adjust the prompt generation strategy for mask prompts. Because *sealed area* contains more line-like polygons than *barren land*, the results are worse for this class.

Qualitative examples for the different prompting results with the tiny version can be seen in Figure 4. When using point prompts, the segmented regions seem to become larger than the reference region. This is especially visible for the *water* areas, which are enlarged in Figure 4 (c) when using point prompts. In Figures 4 (c)-(f) the inaccuracy with segmenting elongated areas is clearly visible. These elongated areas occur, for example, between agricultural areas and are barely recognisable in the reference. In the segmented images, however, the classes *sealed area* and *barren land* appear as larger areas. When using polygon masks prompts (Figure 4 (g)) these areas seem to disappear. The label image looks more irregular in general than it does for the other prompting methods. This result is probably obtained because the mask prompt is an image with resolution four times coarser than the original image. This illustrates the strong influence of the prior information obtained via the mask prompt. The location of the mask prompt has a major influence on the final segmentation result.

As mask prompts represent essential prior information and their placement has a major impact on the results using the proposed method, they may not be suitable for every application which uses prior knowledge from a database. This depends on how accurate the information in the database actually is. In contrast, data for the generation of point or box prompts can be less accurate within the database to still deliver acceptable results. Therefore we will perform the following experiments with both types of mask prompts, as well as with a combination of points with boxes, in which we explicitly exclude the mask information. As the performance of the tiny version is better than the one of the large version, we will use the tiny model for further experiments.

### 5.4 Analysis of different input channels

For the newly created 3-channel images, the results in Table 3 show only slightly differences between the NIR-R-G and the generated R-G-B images. The OA and mF1-scores are also similar to the ones achieved with the preprocessed TCI (Table 1). This shows that it almost does not make any difference whether the TCI or bottom of atmosphere reflectance values are used. Furthermore, it also demonstrates that SAM 2 can process 3-channel input images that differ from the ones used in the training data. However, using this further information from NIR-R-G does not lead to an information gain for RSI.

### 5.5 Semantic segmentation for database updates

The results achieved when using different time steps for prompts and images are shown in Table 4. Using polygon mask prompts leads to an OA of 93.2% with a mF1-score of 82.7%. The combined mask prompts are slightly worse with an OA of 89.1%. Using combined points and box prompts with the tiny version of SAM 2 leads to an OA of 81.5% and mF1 of 71.5%. These results are only slightly worse than those for 2019 in Table 1. This indicates that a direct transfer for updating databases might be possible. However, it must be noted that only 0.8% of the data used has undergone any changes at all. This makes it difficult to provide a clear statement on the actual detection of land cover changes. Furthermore, the current classification is pixel-based, which would need to be transferred to objects, i.e. polygons, for the actual update of topographic databases.

| | F1-scores [%] | | | | | | | [%] | |
|---|---|---|---|---|---|---|---|---|---|
| prompt types | *stl.* | *bar.* | *veg.* | *sld.* | *agr.* | *for.* | *wat.* | mF1 | OA |
| points | 65.2 | 15.3 | 67.9 | 11.2 | 76.9 | 78.7 | 72.1 | 55.3 | 69.6 |
| boxes | 85.0 | 55.9 | 81.9 | 5.6 | 86.1 | 88.0 | 93.1 | 70.8 | 79.7 |
| polygon masks | **88.9** | **76.9** | **92.6** | **42.3** | **95.9** | **94.9** | **95.8** | **83.9** | **93.8** |
| combined mask | 86.3 | 70.9 | 89.0 | 37.2 | 92.6 | 89.6 | 94.6 | 80.0 | 90.1 |
| points and boxes | 84.4 | 58.2 | 83.4 | 6.7 | 87.0 | 89.6 | 94.0 | 71.9 | 81.0 |
| points and polygon masks | 72.1 | 30.3 | 76.3 | 18.5 | 81.8 | 85.1 | 85.4 | 64.2 | 78.7 |
| points and combined mask | 73.6 | 31.3 | 77.8 | 19.2 | 82.5 | 85.8 | 86.1 | 65.2 | 79.7 |
| boxes and polygon masks | 86.0 | 60.0 | 85.3 | 8.6 | 91.1 | 90.7 | 93.7 | 73.6 | 85.6 |
| boxes and combined mask | 86.2 | 57.9 | 85.2 | 8.2 | 90.2 | 90.7 | 93.7 | 73.2 | 85.1 |
| points, boxes and polygon masks | 84.0 | 58.7 | 83.5 | 6.9 | 87.5 | 89.4 | 94.0 | 72.0 | 81.3 |
| points, boxes and combined mask | 84.3 | 57.8 | 83.6 | 7.0 | 87.4 | 89.8 | 94.1 | 72.0 | 81.5 |

Table 1. Results for zero-shot multi-class semantic segmentation using Sentinel-2 **TCI** images and SAM 2 **tiny**. The Sentinel-2 images and the database information for prompting are taken from end of March 2019. Best result marked in **bold**, second best result are underlined.

| | F1-scores [%] | | | | | | | [%] | |
|---|---|---|---|---|---|---|---|---|---|
| prompt types | *stl.* | *bar.* | *veg.* | *sld.* | *agr.* | *for.* | *wat.* | mF1 | OA |
| points | 56.1 | 15.1 | 63.7 | 10.5 | 73.4 | 73.9 | 70.3 | 51.8 | 65.5 |
| boxes | 81.1 | 56.5 | 80.5 | 5.7 | 87.0 | 85.9 | 93.0 | 69.9 | 78.0 |
| polygon masks | 74.1 | 57.2 | 82.0 | **30.4** | 87.8 | 85.3 | 83.7 | 71.5 | **84.1** |
| combined mask | 64.6 | 41.0 | 74.3 | 20.1 | 86.2 | 81.2 | 79.4 | 63.8 | 79.4 |
| points and boxes | 83.8 | 58.3 | 83.0 | 7.0 | 88.5 | 88.7 | 93.9 | 71.9 | 81.4 |
| points and polygon masks | 64.8 | 23.6 | 70.5 | 15.1 | 75.9 | 80.8 | 81.5 | 58.9 | 72.9 |
| points and combined mask | 63.8 | 21.9 | 69.9 | 15.2 | 77.1 | 79.9 | 80.4 | 58.3 | 72.4 |
| boxes and polygon masks | 82.3 | 59.6 | 82.2 | 6.5 | 88.5 | 87.3 | 93.0 | 71.3 | 80.4 |
| boxes and combined mask | 81.9 | 59.8 | 83.0 | 6.9 | 89.4 | 88.1 | 93.3 | 71.8 | 81.5 |
| points, boxes and polygon masks | **84.2** | **60.4** | 83.7 | 7.5 | 88.8 | 89.5 | 93.9 | 72.6 | 82.3 |
| points, boxes and combined mask | 84.1 | 60.3 | **84.3** | 7.9 | **89.5** | **89.8** | **94.1** | **72.8** | 83.1 |

Table 2. Results for zero-shot multi-class semantic segmentation using Sentinel-2 **TCI** images and SAM 2 **large**. The Sentinel-2 images and the database information for prompting are taken from end of March 2019. Best result marked in **bold**, second best result are underlined.

| | [%] | |
|---|---|---|
| image and prompt type | mF1 | OA |
| NIR-R-G, polygon masks | 83.9 | 93.8 |
| NIR-R-G, combined masks | 79.4 | 89.0 |
| NIR-R-G, points and boxes | 72.0 | 81.5 |
| R-G-B, polygon masks | 83.9 | 93.8 |
| R-G-B, combined masks | 80.2 | 90.3 |
| R-G-B, points and boxes | 72.2 | 81.8 |

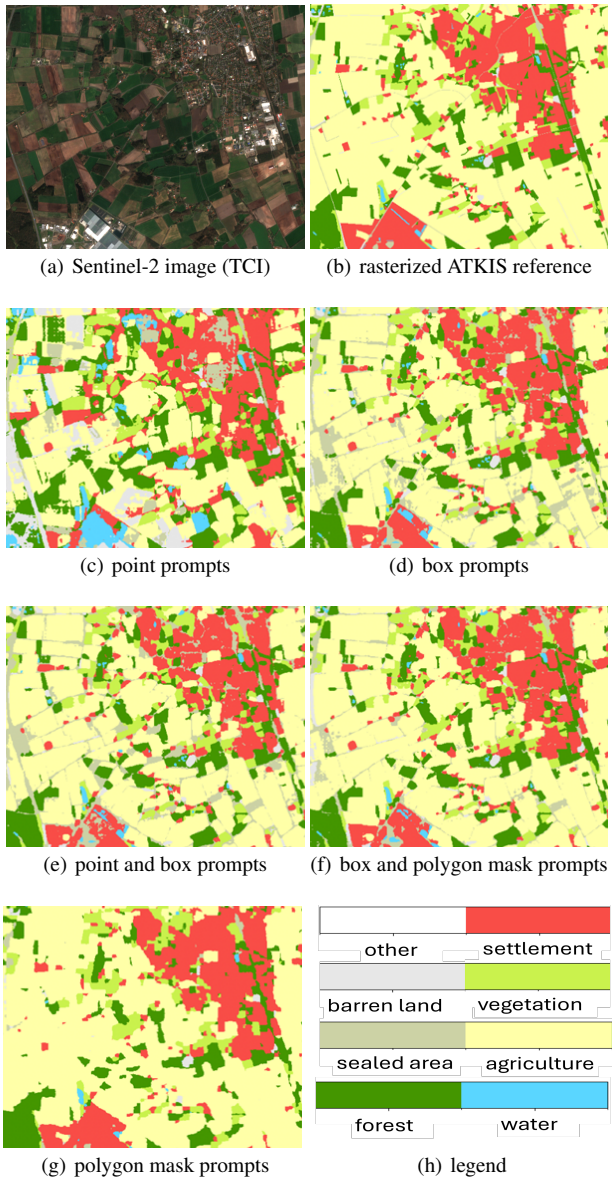
Table 3. Results for zero-shot multi-class semantic segmentation using newly created 3-channel composite images and SAM 2 **tiny** with different prompt types. The images and database information for prompting are from 31/03/2019.

## 6. Conclusion and Outlook

This work presents a method to perform multi-class semantic segmentation using SAM 2 without fine-tuning or training and without requiring any training data. Instead, prior knowledge from a topographic database is used to generate prompts for a VFM (here SAM 2), followed by a post-processing step that combines logits per class for every pixel into a multi-class label image. The method is tested on two different versions of SAM 2 and with different prompt types. Furthermore, experiments with different Sentinel-2 images (TCI and 3-channel composite images) and for varying time steps are performed. The proposed method achieves an overall accuracy of up to 93% at pixel-level when using polygon masks as prompts (strong prior information). With weaker preliminary information, such as bounding boxes and point prompts, an OA of around 81% and a mF1-score of 71.9% is achieved. Changing the input images into different 3-channel composite images does not influence the results for OA and mF1-score significantly. Also different results are achieved for different classes. Especially the classes *barren land* and *sealed area* contain many elongated areas, for which the proposed approach for prompt generation is not useful. The segmentation for different time steps also shows good results. But the very small amount of changed pixels in the dataset makes it difficult to draw conclusions about areas or datasets with more changes. Overall, the experiments show that using polygon mask prompts seem to be best suited when using SAM 2 with a combination of (ATKIS) database data and Sentinel-2 images. If masks must not be used, the combination

| | F1-scores [%] | | | | | | | [%] | |
|---|---|---|---|---|---|---|---|---|---|
| image and prompt type | *stl.* | *bar.* | *veg.* | *sld.* | *agr.* | *for.* | *wat.* | mF1 | OA |
| TCI, polygon masks | 88.4 | 71.2 | 91.6 | 41.5 | 95.5 | 94.6 | 95.9 | 82.7 | 93.2 |
| TCI, combined masks | 85.3 | 66.2 | 88.0 | 35.6 | 92.0 | 88.3 | 94.3 | 78.5 | 89.1 |
| TCI, points and boxes | 84.5 | 54.4 | 83.3 | 7.0 | 87.6 | 89.6 | 94.1 | 71.5 | 81.5 |

Table 4. Results for zero-shot multi-class semantic segmentation using different Sentinel-2 images and SAM 2 **tiny**. The Sentinel-2 images where taken end of March 2020, while the database information for prompting is taken from 31/03/2019. For evaluation the database information from 31/03/2020 was used.




(a) Sentinel-2 image (TCI) (b) rasterized ATKIS reference
(c) point prompts (d) box prompts
(e) point and box prompts (f) box and polygon mask prompts
(g) polygon mask prompts (h) legend

Figure 4. Exemplary prediction results using different prompt types and the tiny version of SAM 2 on a Sentinel-2 tile. (a) - (b) Sentinel-2 TCI image and rasterized ATKIS reference, (c) - (g) results for different prompt types. (h) Legend.

of box and point prompts performs best in our scenario. For the application of database updates, further investigations are necessary.

In summary, this work demonstrates a successful configuration of a method for using database information together with Sentinel-2 images for semantic segmentation without training, by using a pre-trained VFM.

However, there are some limitations and challenges that should be addressed in future work. Our analysis of the single classes shows that elongated areas are not segmented reliably using the proposed prompt generation strategy. To overcome this limitation, a different strategy to generate the prompts for this areas must be designed. For example, elongated areas (depending on the shape and size) could be split into smaller sub-areas, for which prompts are generated. In order to use the method for database updating, there are still various aspects that need to be taken into account. These include, on the one hand, the application of semantic segmentation to objects rather than pixels. The database information used is available as vector data, which means that the generated semantic label should also fit this format. Furthermore, the prompts predefine the locations of areas of a specific class. Regarding the detection of changes, possible class changes could be further analysed by using the feature space. For example, a loss for the difference or similarity of features depending on the class could be investigated. The proposed method shows good results, however, comparative experiments with alternative VFMs should be carried out to investigate its transferability. In addition, VFMs are usually very large models that require a lot of computational power. Approaches such as knowledge distillation could be investigated in order to take advantage of the capabilities of VFMs while saving computational power. These approaches could also tackle RSI specific characteristics, such as the use of additional information from all available bands of the RSI.

## References

AdV, 2008. GeoInfoDok 6.0 - Dokumente. `https://www.adv-online.de/GeoInfoDok/GeoInfoDok-6.0/Dokumente/`. Accessed 12/05/2026; AdV: Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland.

AdV, 2025. Authorative Topographic-Cartographic Information System (ATKIS®). `https://www.adv-online.de/Products/Geotopography/ATKIS/`. Accessed 12/05/2026; AdV: Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland.

Bolya, D., Ryali, C., Hoffman, J., Feichtenhofer, C., 2024. Window attention is bugged: How not to interpolate position embeddings. *International Conference on Learning Representations (ICLR)*.

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z., 2024. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62.

Chen, T., Zhu, L., Ding, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P., 2023. SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3359–3367.

Diab, M., Kolokoussis, P., Brovelli, M. A., 2025. Optimizing zero-shot text-based segmentation of remote sensing imagery using SAM and Grounding DINO. *Artificial Intelligence in Geosciences*, 6(1), 100105.

Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment Anything. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Li, K., Cao, X., Meng, D., 2024. A New Learning Paradigm for Foundation Model-Based Remote-Sensing Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62.

Liu, N., Xu, X., Su, Y., Zhang, H., Li, H.-C., 2025. PointSAM: Pointly-Supervised Segment Anything Model for Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 63.

Liu, Y., Zhou, G., 2024. Laddering vision foundation model for remote sensing image change detection. *Journal of Applied Remote Sensing (JRS)*, 18(3), 036503.

Lu, S., Guo, J., Zimmer-Dauphinee, J. R., Nieusma, J. M., Wang, X., vanValkenburgh, P., Wernke, S. A., Huo, Y., 2025. Vision Foundation Models in Remote Sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine (MGRS)*, 190–215.

Luo, M., Zhang, T., Wei, S., Ji, S., 2024. SAM-RSIS: Progressively Adapting SAM With Box Prompting to Remote Sensing Image Instance Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62.

Ma, X., Wu, Q., Zhao, X., Zhang, X., Pun, M.-O., Huang, B., 2024. SAM-Assisted Remote Sensing Imagery Semantic Segmentation With Object and Boundary Constraints. *IEEE Transactions on Geoscience and Remote Sensing*, 62.

Osco, L. P., Wu, Q., de Lemos, E. L., Gonçalves, W. N., Ramos, A. P. M., Li, J., Marcato, J., 2023. The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124, 103540.

Qiu, J., Liu, W., Zhang, X., Li, E., Zhang, L., Li, X., 2025. DED-SAM:Adapting Segment Anything Model 2 for Dual Encoder–Decoder Change Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)*, 18, 995–1006.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., 2025. SAM 2: Segment Anything in Images and Videos. *International Conference on Representation Learning (ICLR)*, 2025, 28085–28128.

Ren, S., Luzi, F., Lahrichi, S., Kassaw, K., Collins, L. M., Bradbury, K., Malof, J. M., 2024a. Segment anything, from space? *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 8355–8365.

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L., 2024b. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *International Conference on Computer Vision (ICCV) Demo Track*, arXiv preprint arXiv:2401.14159.

Ryali, C., Hu, Y.-T., Bolya, D., Wei, C., Fan, H., Huang, P.-Y., Aggarwal, V., Chowdhury, A., Poursaeed, O., Hoffman, J., Malik, J., Li, Y., Feichtenhofer, C., 2023. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles. A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (eds), *40th International Conference on Machine Learning (PMLR)*, 202, 29441–29454.

Sultan, R. I., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., Zhu, D., 2025. GeoSAM: Fine-tuning SAM with Multi-Modal Prompts for Mobility Infrastructure Segmentation. *European Conference on Artificial Intelligence (ECAI)*, Frontiers in Artificial Intelligence and Applications, 28, 501–508.

Sun, Z., Song, H., Zhang, K., Dong, G., Liang, L., Zhao, Y., 2024. Segment Anything Model Guided Semantic Knowledge Learning For Remote Sensing Change Detection. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5830–5834.

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(5), 547–568.

Wan, Z., Wang, S., Han, W., Wang, Y., Huang, X., Zhang, X., Chen, X., Chen, Y., 2025. A systematic survey and meta-analysis of the segment anything model in remote sensing image processing: Challenges, advances, applications, and opportunities. *ISPRS Journal of Photogrammetry and Remote Sensing*, 229, 436–466.

Zhang, J., Li, Y., Yang, X., Jiang, R., Zhang, L., 2025. RSAM-Seg: A SAM-Based Model with Prior Knowledge Integration for Remote Sensing Image Semantic Segmentation. *Remote Sensing*, 17(4).

Zhang, Y., Huang, X., Ma, J., Li, Z., Luo, Z., Xie, Y., Qin, Y., Luo, T., Li, Y., Liu, S., Guo, Y., Zhang, L., 2024. Recognize anything: A strong image tagging model. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1724–1732.