Guiding SAM 2 for semantic segmentation of satellite images

Phd colloquium Bonn

Paula Lippmann

Content

- Introduction
- Methodology
- Experiments
- Conclusion
- Future work and research direction

Motivation

- Topographic databases as a basis for planning and decision making
- Nowadays manual updates based on aerial images every few years
- High temporal resolution of satellite images available
- Neural networks for image analysis
- Foundation Models for a wide range of applications

Why Foundation models?

- Pretrained on large datasets
  - Ability to generalize across different domains
  - Vision Foundation models have strong visual recognition capabilities
- Pretraining enables possibility to save resources and energy for training
- Adaptation to specific tasks is possible

➢ Here: Segment Anything Model 2

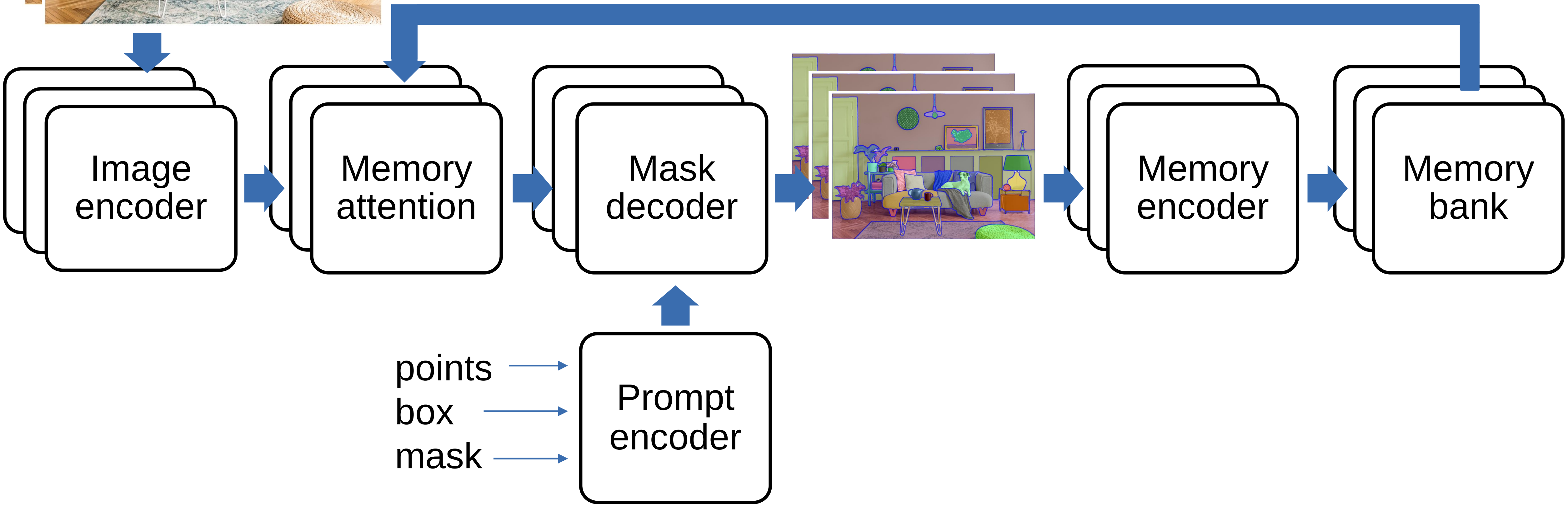
Segment Anything Model 2 (SAM 2) [Ravi2024]

- Promptable visual segmentation in images and videos
- Prompting: Location information about object necessary
- Video segmentation might be useful for satellite image time series

Segment-anything.com/demo#

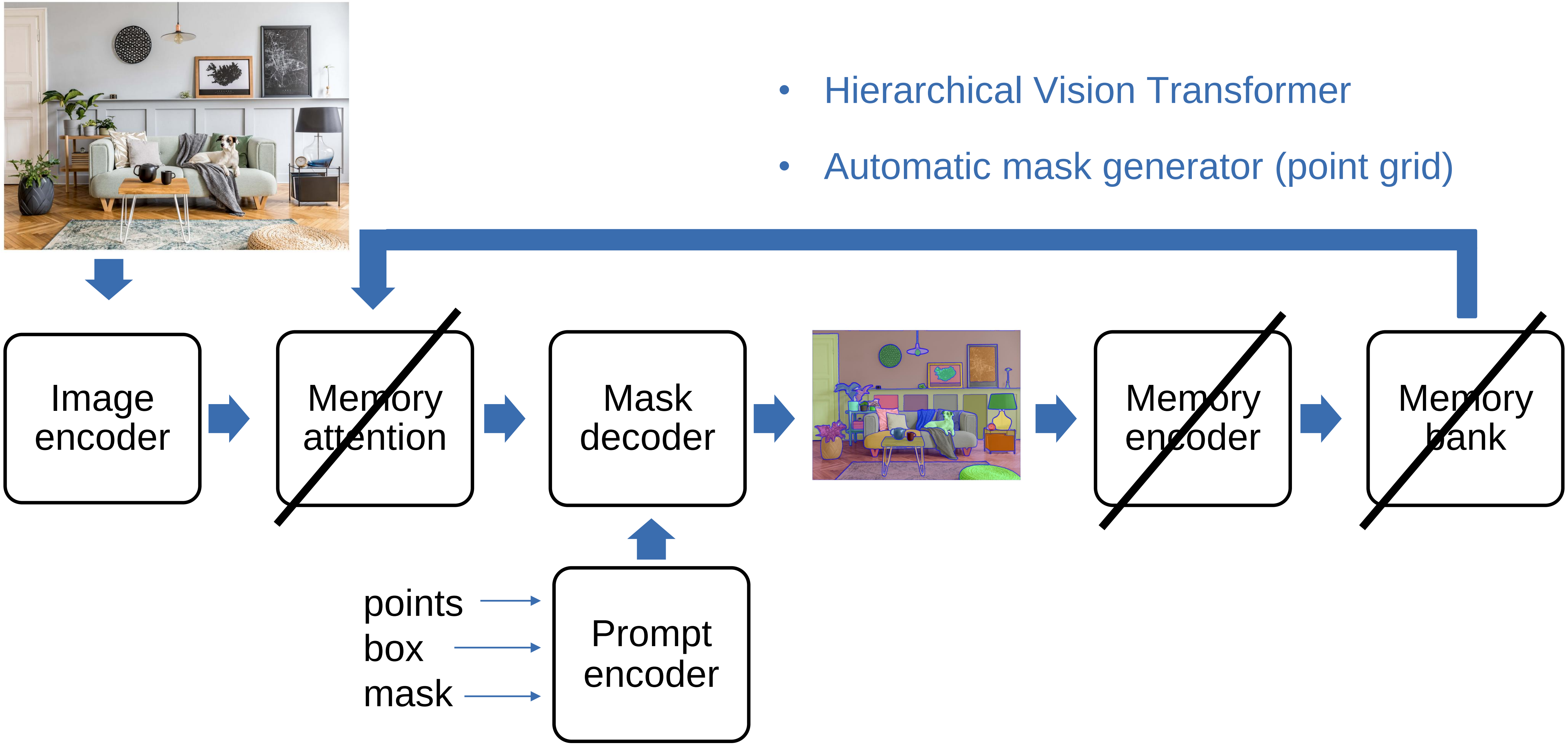
SAM 2

- Hierarchical Vision Transformer
- Automatic mask generator (point grid)



SAM 2

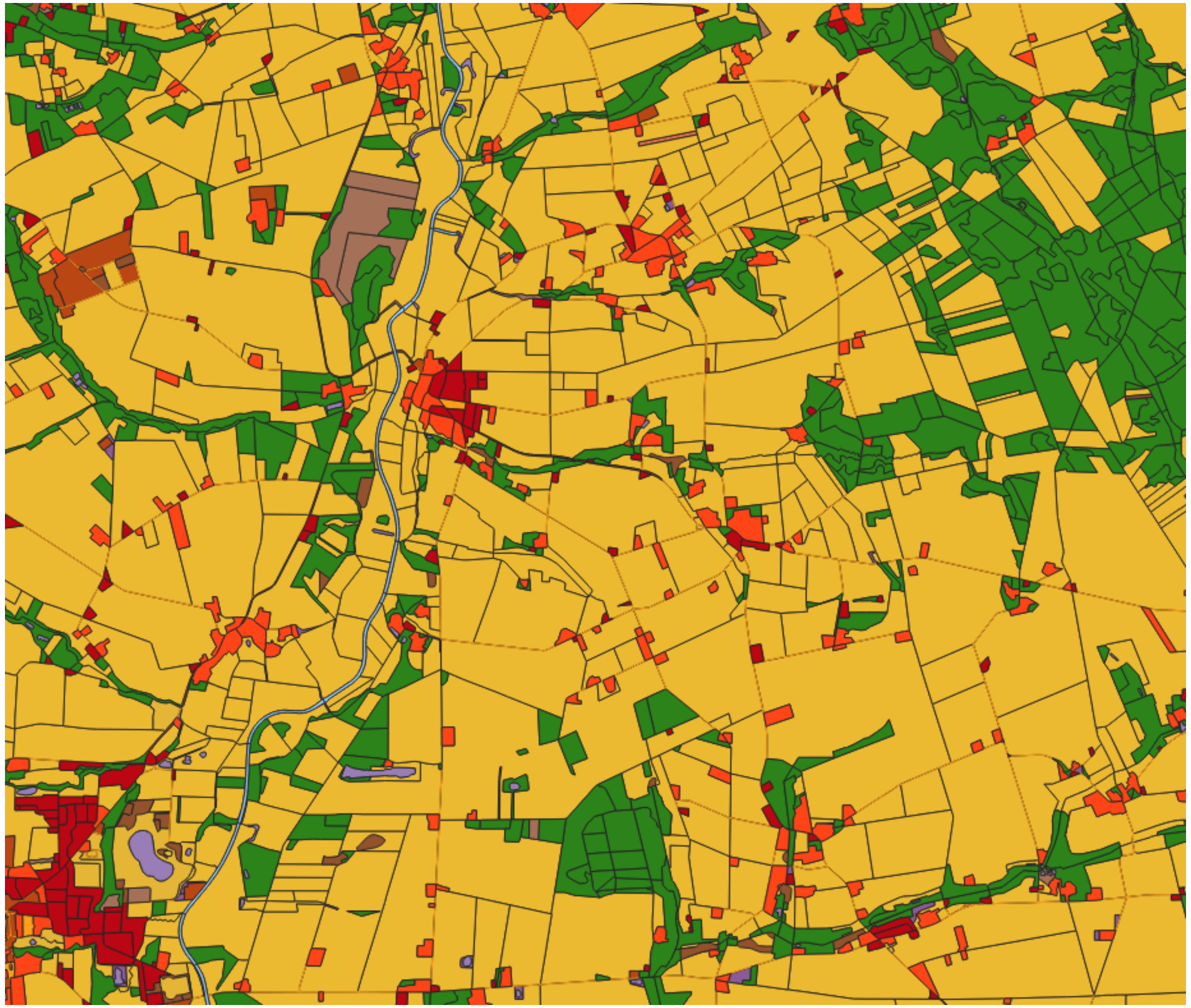
- Hierarchical Vision Transformer
- Automatic mask generator (point grid)



Topographic database - ATKIS

- Authorative Topographic-Cartographic Information System
- Contains polygons and lines with attributes

➢ Here: derive prompts and classes



Related Work

- Other works predominantly
  - Use SAM as auxiliary for other networks [Mei2024]
  - Do fine-tuning of parts of the network [Luo2024, Liu2025]
  - Insert adapters for their need [Chen2023, Xie2024, Zhang2025]
  - Work on automatic prompting and prompt refinement [Luo2024, Ren2024, Diab2025]
    - ➢ Pre-connect networks to generate prompts
- Focus on images with smaller ground sampling distance than 10 m

➢ Use prior knowledge directly as prompts and generate complete semantic segmentation without any adaptation or training

Research Question

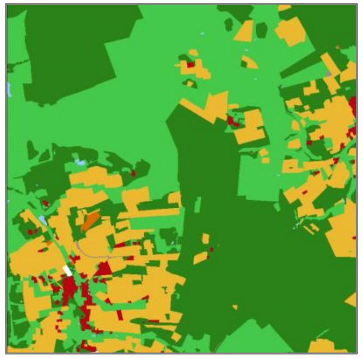
- How can we use (pretrained) Foundation Models for semantic segmentation and/or change detection for topographic database updates?
  - Is this working in a zero-shot / few-shot manner with no or very less training data?
  - How can we process image time series?
  - How can conditions about land cover classes be taken into account?
  - How do we transfer semantic segmentation to polygons?
    - ➢ This presentation: How good can **semantic segmentation** of **Sentinel-2** images be performed using “Segment Anything Model 2” in a **zero-shot** manner together with **ATKIS** information?

Content

- Introduction
- Methodology
- Experiments
- Conclusion
- Future work and research direction

Method - Overview

ATKIS objects



Sentinel-2 image

SAM 2

per class segmentation

prompt generation

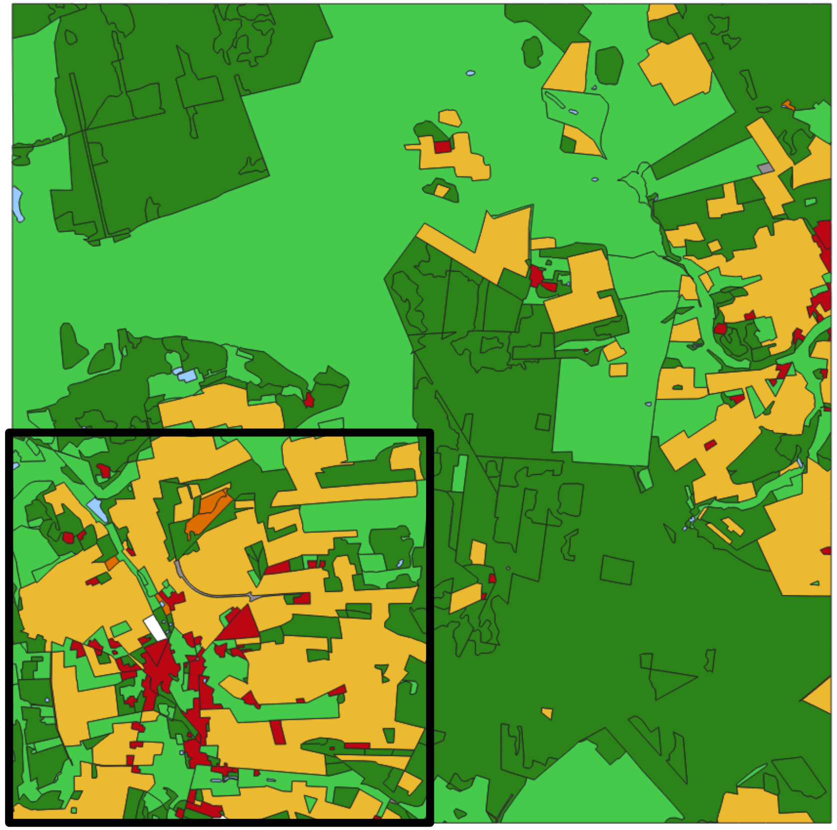
prompts per class

semantic segmentation as a combination of all class-wise segmentations

Comparison between ATKIS and semantic segmentation

Prompt generation

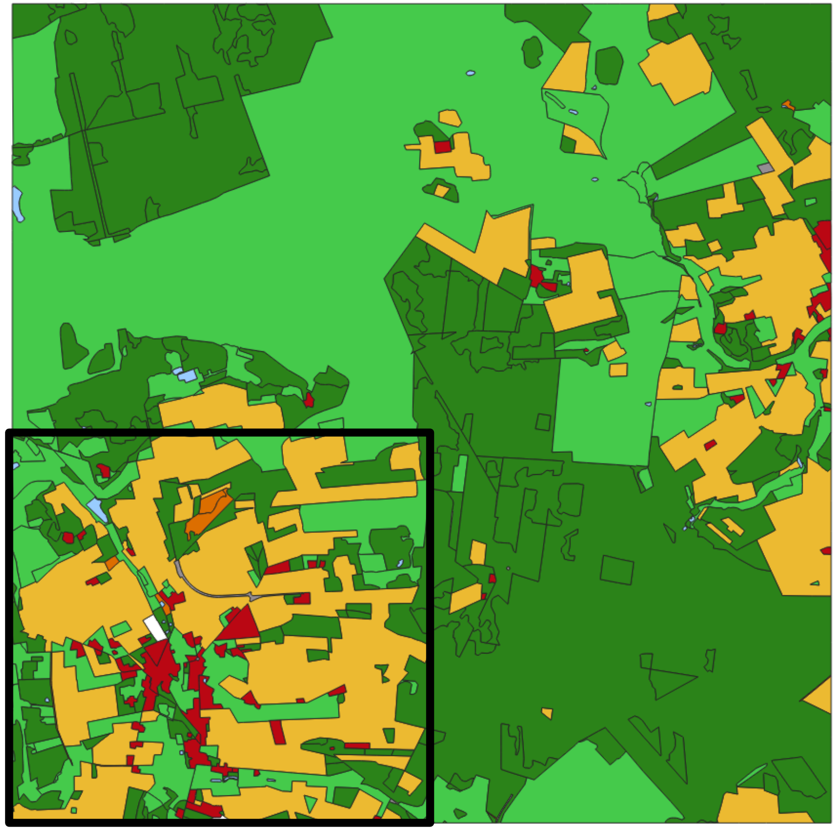
- (Prior) Location information for prompting from database
- Class wise polygon selection

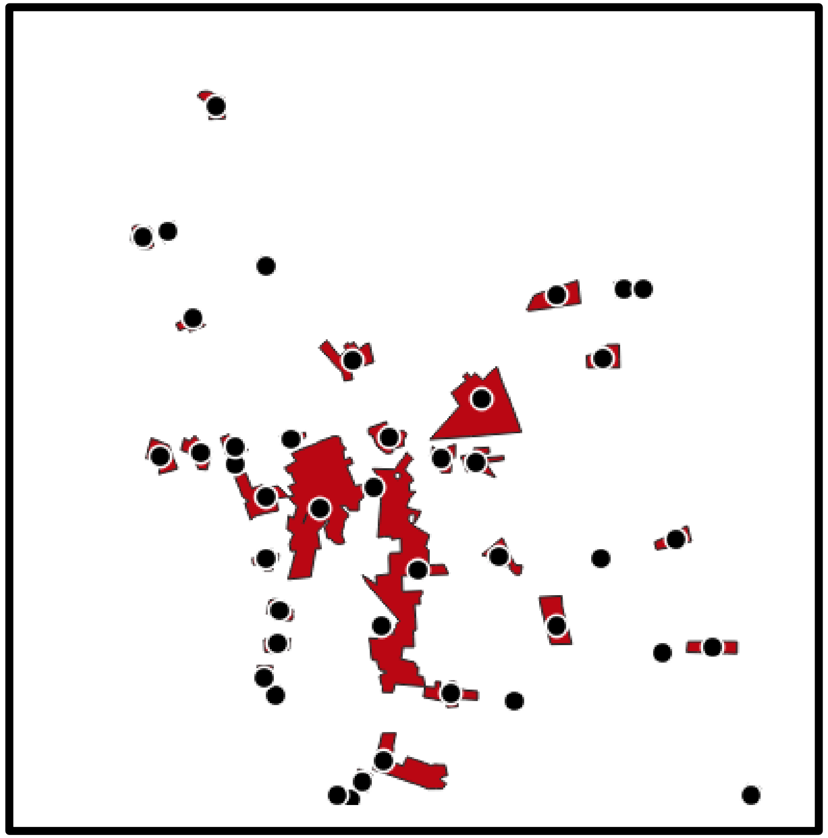


Other
Settlement
Sealed area
Agriculture
Vegetation
Forest
Water – flow
Water-stand
Sea
Barren land

Prompt generation: points

- Representative point of each polygon
  - Guaranteed to be within geometry

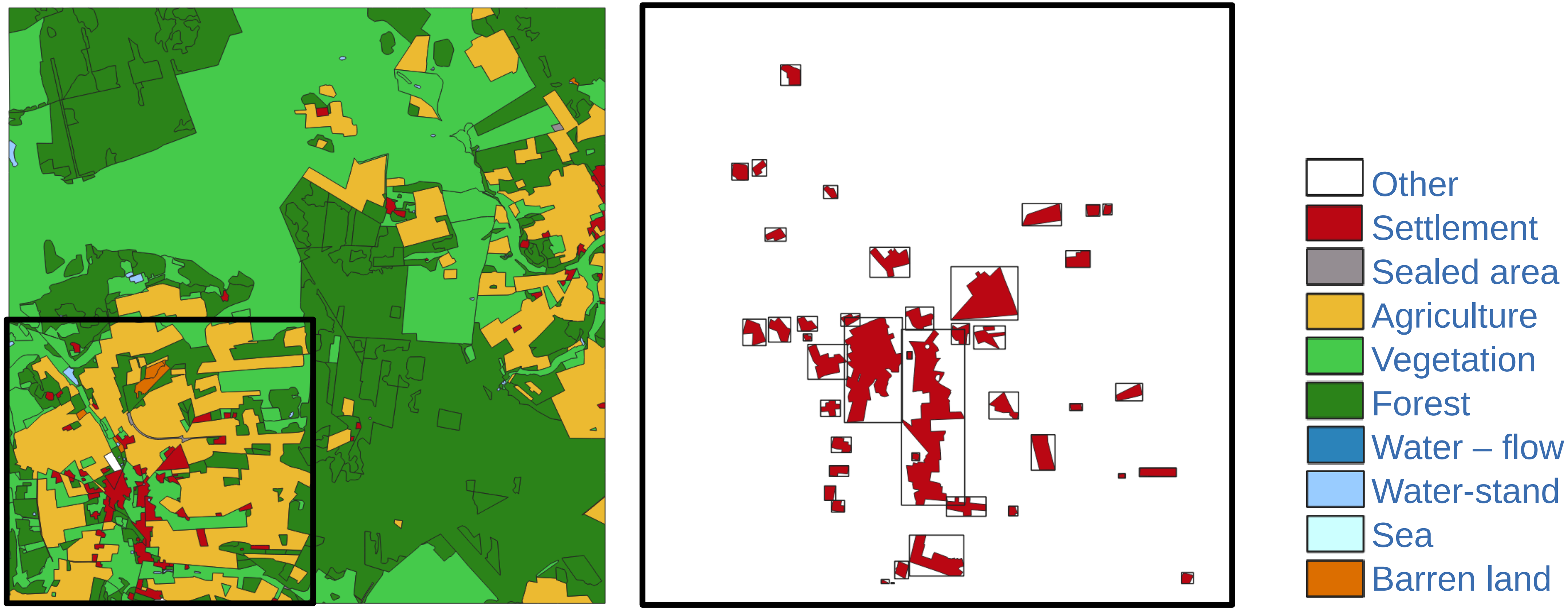




- Other
- Settlement
- Sealed area
- Agriculture
- Vegetation
- Forest
- Water – flow
- Water-stand
- Sea
- Barren land

Prompt generation: boxes

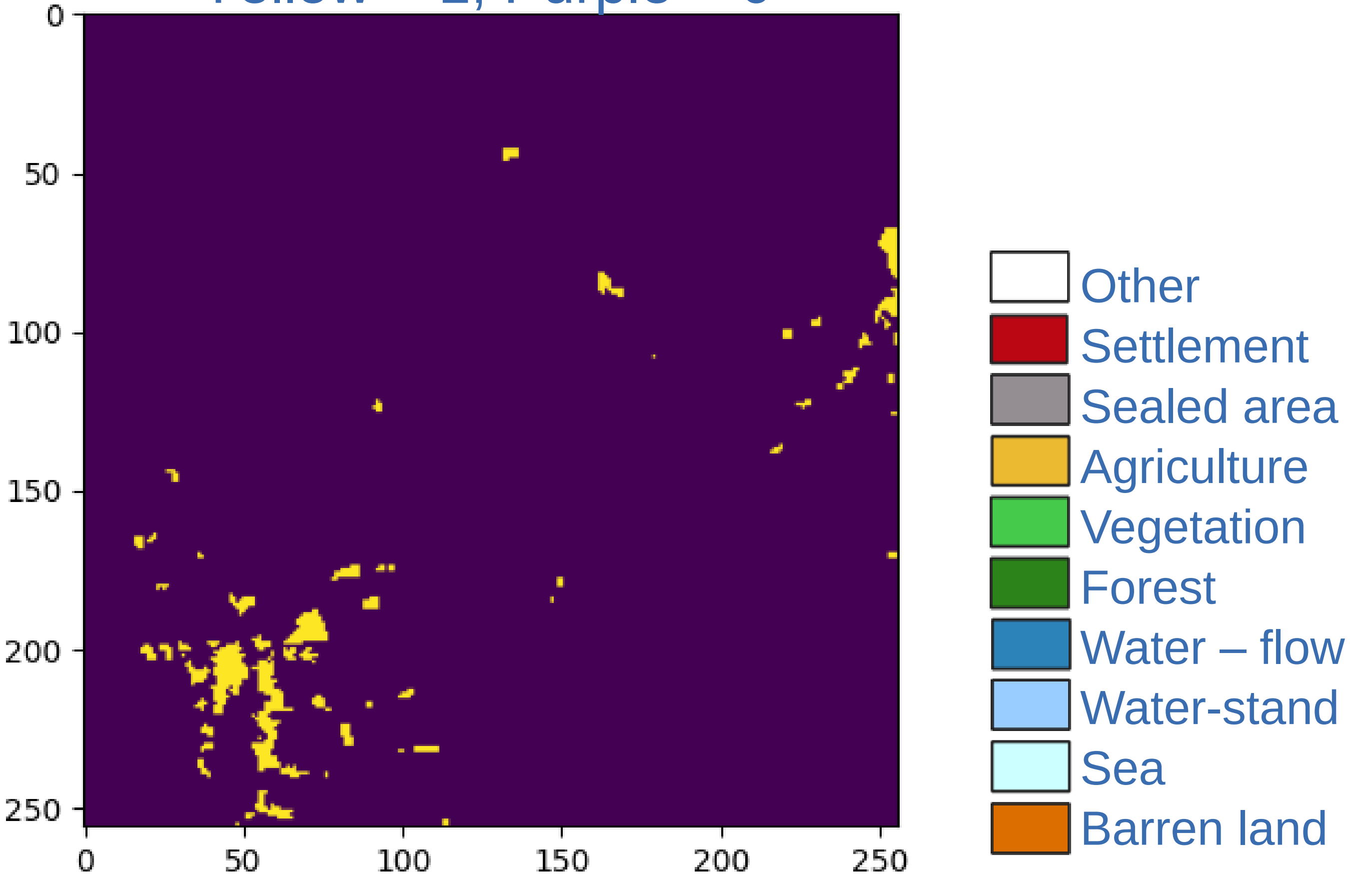
- Boundaries of geometry
  - Minimum and maximum values for x and y



Prompt generation: mask

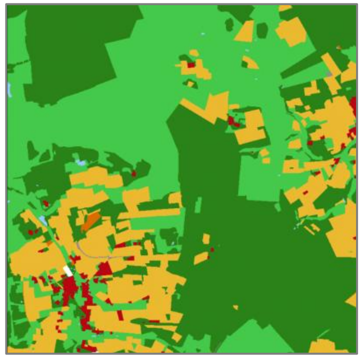
- Binary mask with 4 times lower resolution than input image

Yellow = 1, Purple = 0



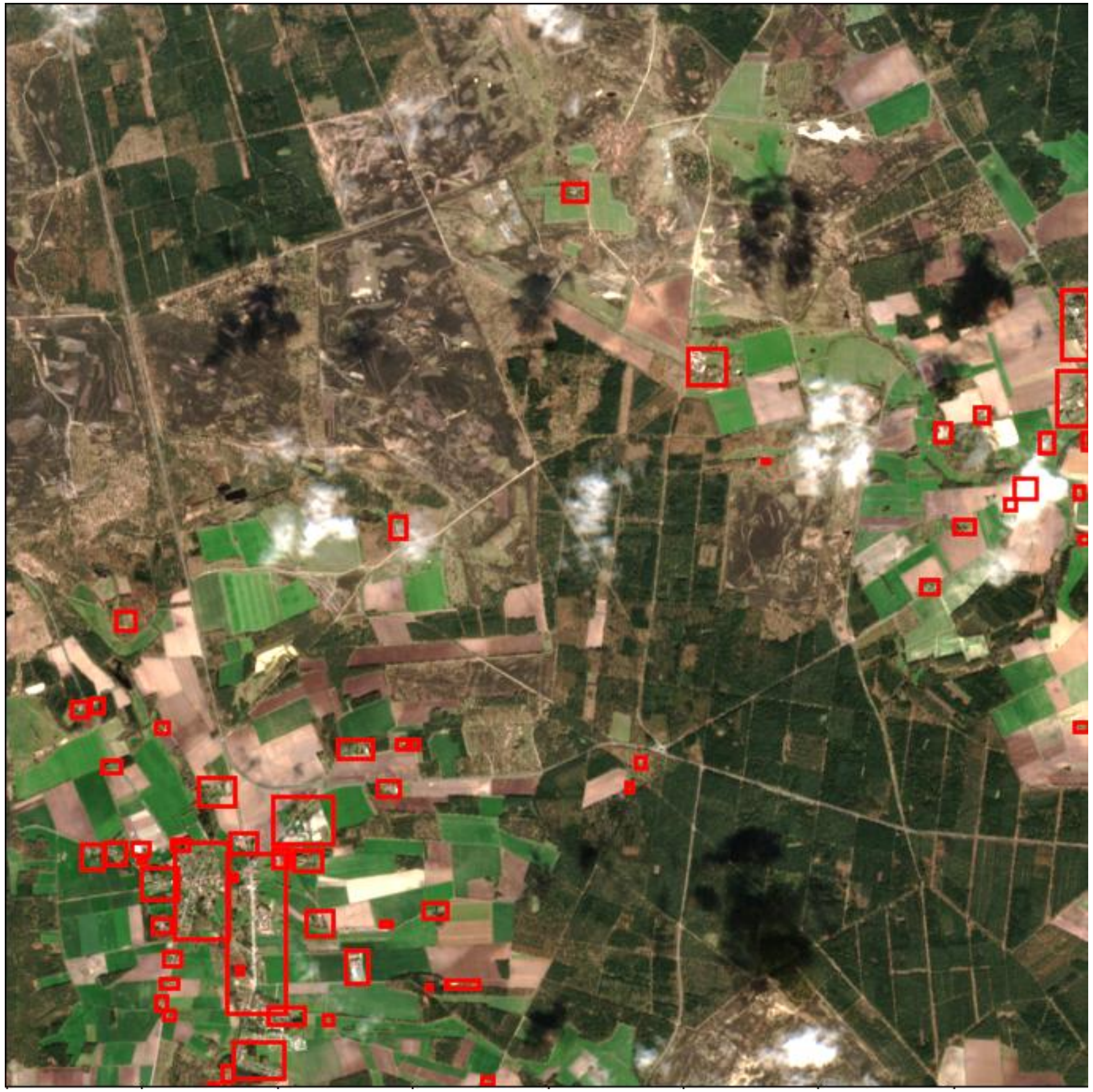
Example: box prompt

ATKIS objects



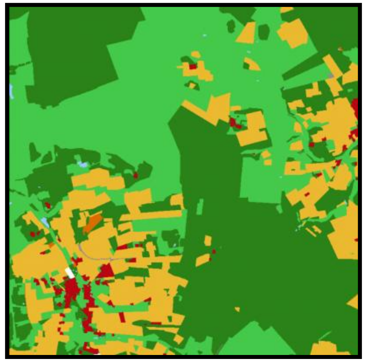
Sentinel-2 image

prompt generation → box prompt per object of a class



Segmentation masks for box prompt

ATKIS objects



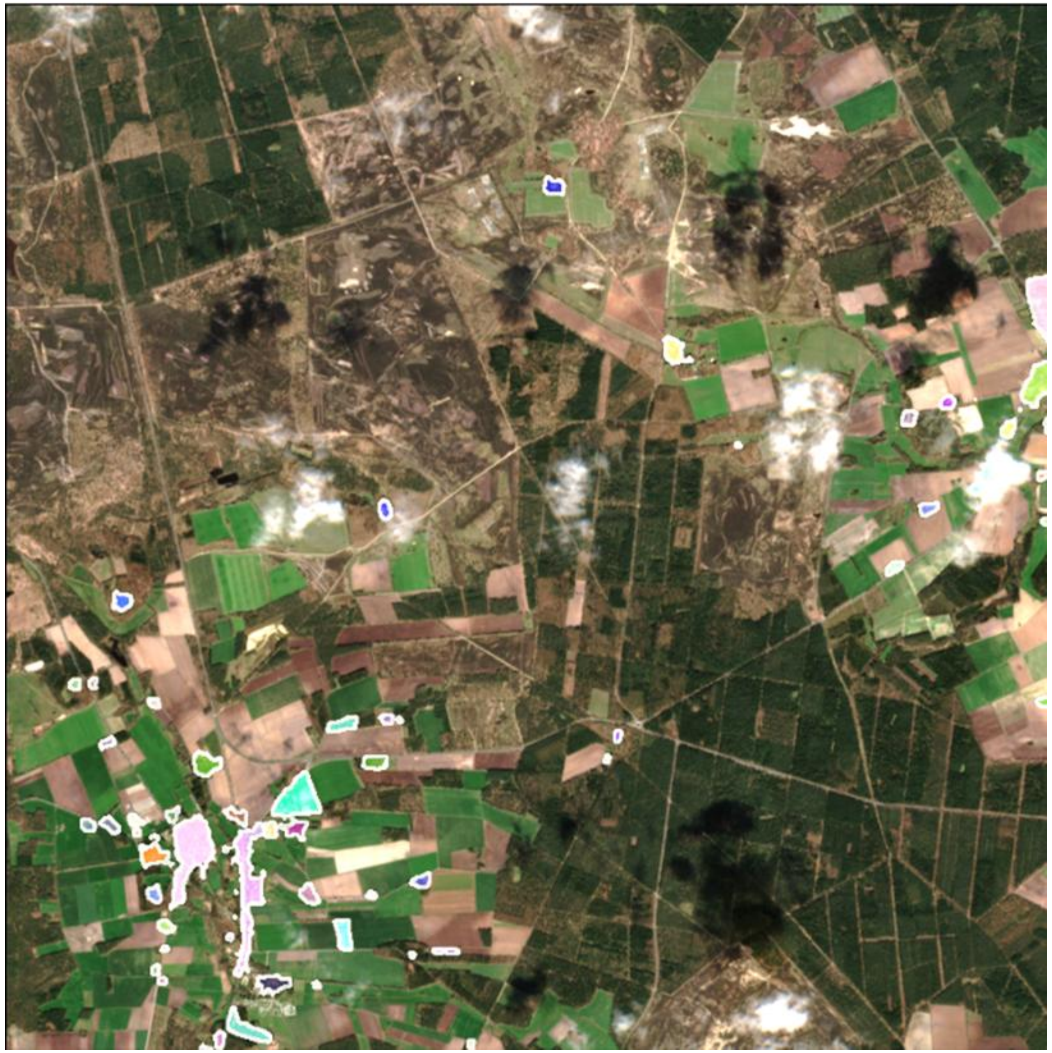
Sentinel-2 image

prompt generation

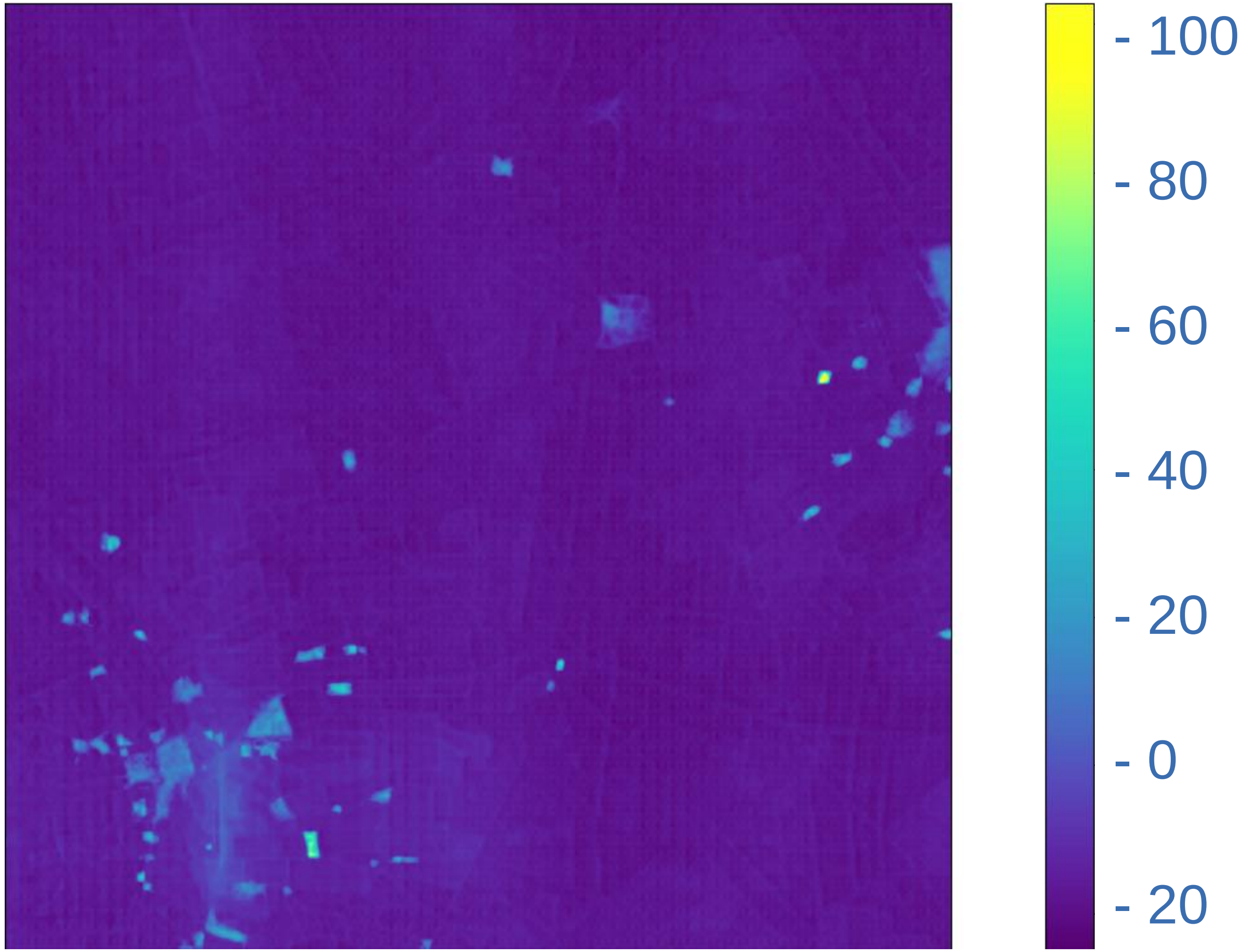
box prompt per object of a class

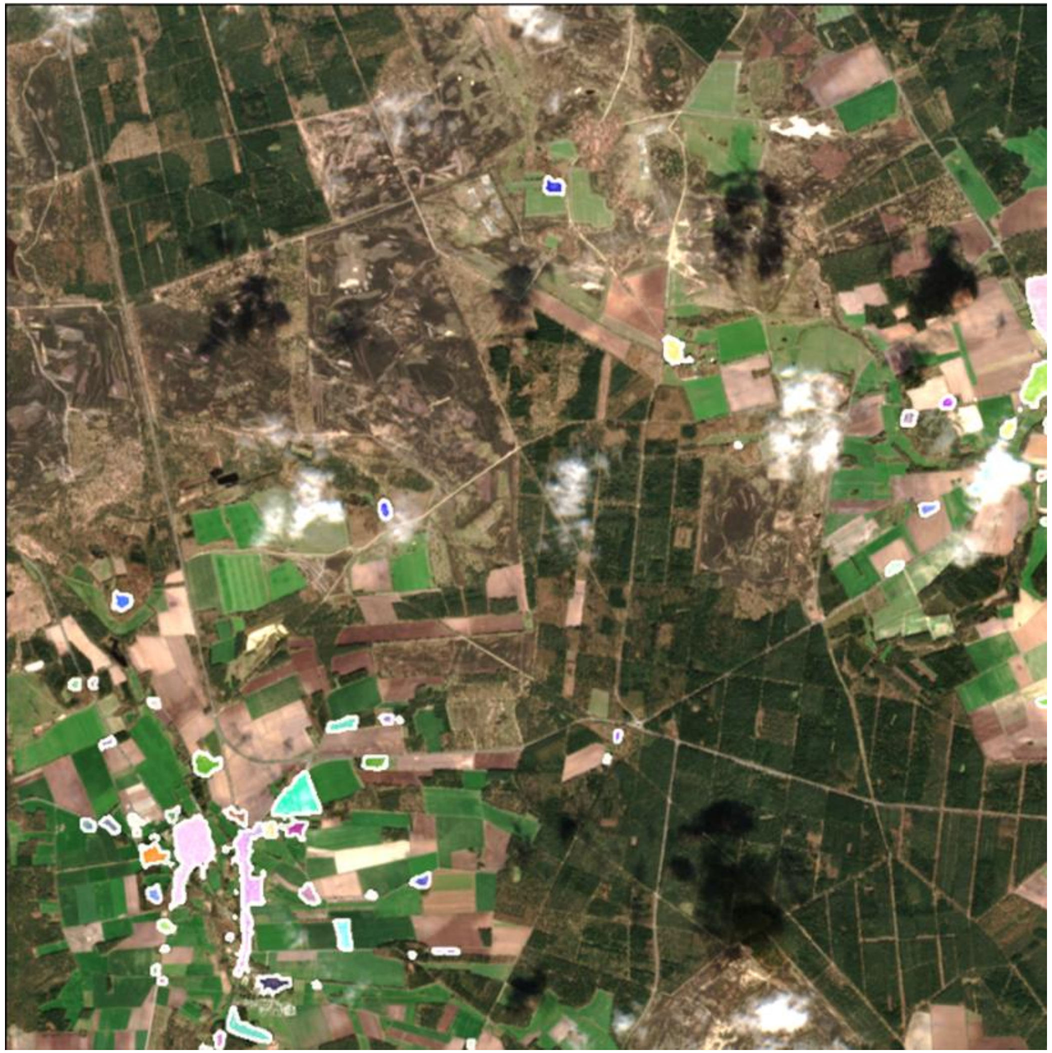
SAM 2

segmentation of this class



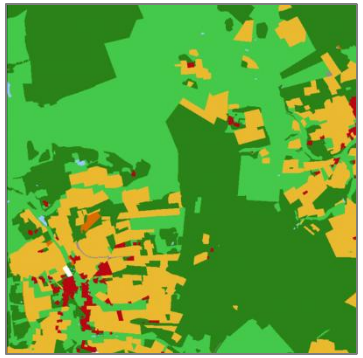
Logits for box prompt





Combination of logits

ATKIS objects



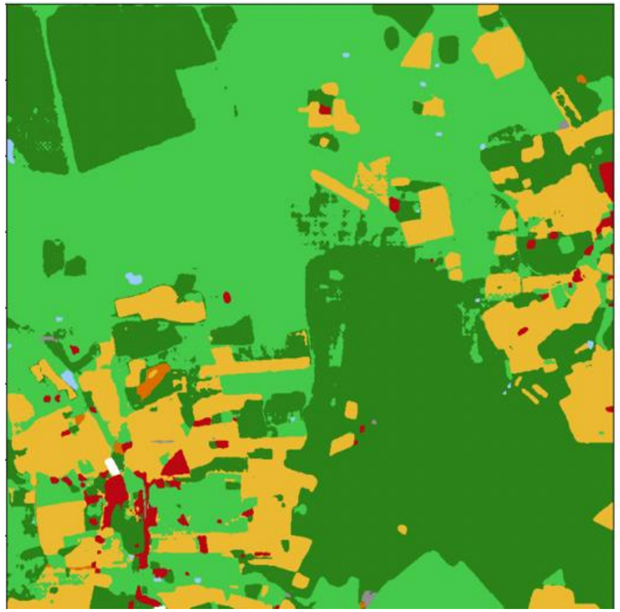
Sentinel-2 image

SAM 2

per class segmentation /logits

per class

logits



prompt generation

prompts per class

semantic segmentation as a combination of all class-wise segmentations

Comparison between ATKIS and semantic segmentation

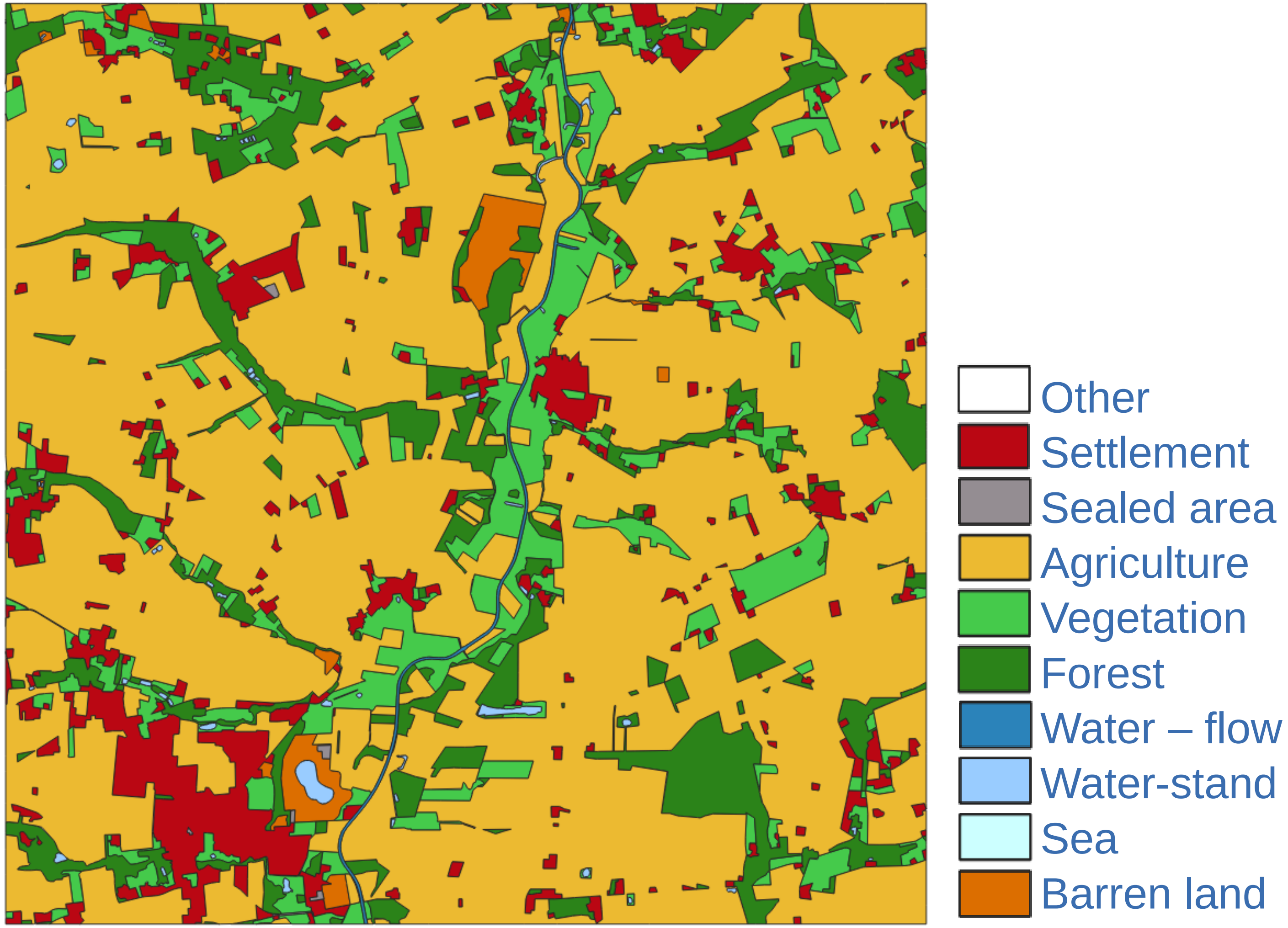
Experiments – data

Sentinel-2 image

- True color images
- 8x8 km tiles
- Cloud free
- Images from the end of March 2019
  - Focus on monotemporal images

Experiments – data

- Corresponding polygon data from ATKIS
- 10 classes
- Data from 31st of March 2019
- Use for
  – Prompt generation
  – Reference for evaluation



ATKIS (reference data)

Experiments – setup

- Different prompt types and combinations
  - Points
  - Boxes
  - Masks
  - Points and boxes
  - Points and masks
  - Boxes and masks
  - Points, boxes and masks
- Evaluation metrics
  - F1-score and mean F1-score
  - Overall accuracy

Evaluation

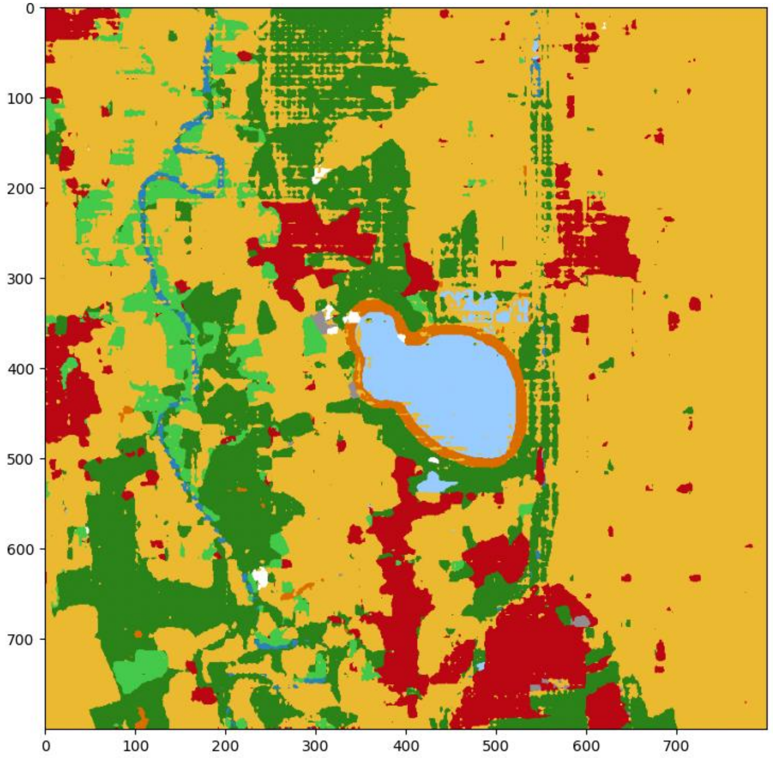
Prompt types	mF1-Score [%]	Overall accuracy [%]
Points	26,4	24,3
Boxes	59,1	70,6
Masks	59,9	75,5
Points and boxes	60,1	71,7
Points and masks	29,3	28,8
Boxes and masks	60,1	72,2
Points, boxes and masks	60,9	73,2

Evaluation: Class specific

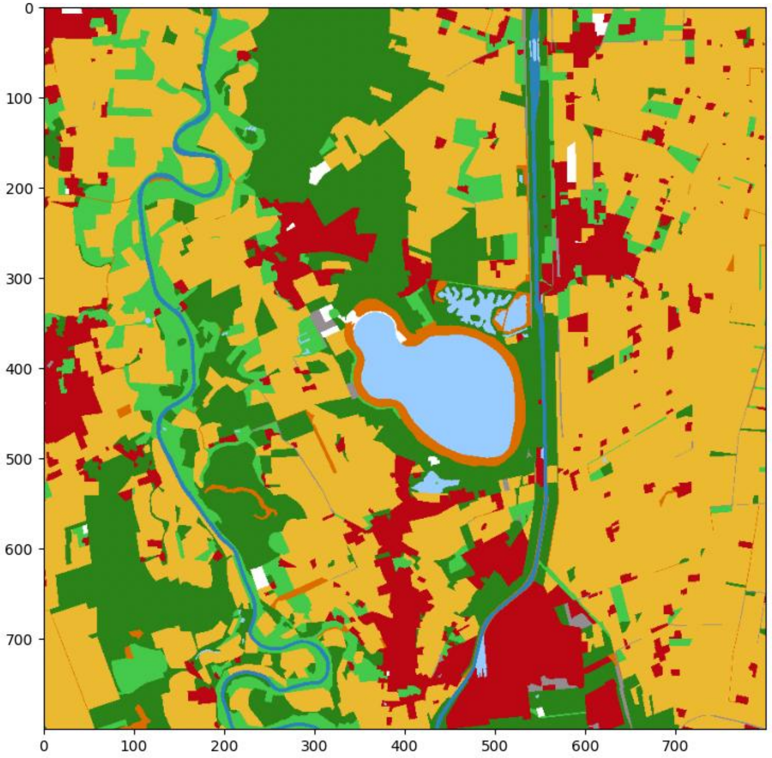
Prompt types	Settlement F1-Score [%]	Forest F1-Score [%]	Water- flow F1-Score [%]
Points	22,5	35,4	40,1
Boxes	78,8	83,5	36,9
Masks	62,8	79,2	50,8
Points and boxes	79,3	84,8	38,5
Points and masks	26,6	38,6	42,1
Boxes and masks	78,6	85,3	40,9
Points, boxes and masks	79,6	85,8	44,8

Qualitative results – mask prompt

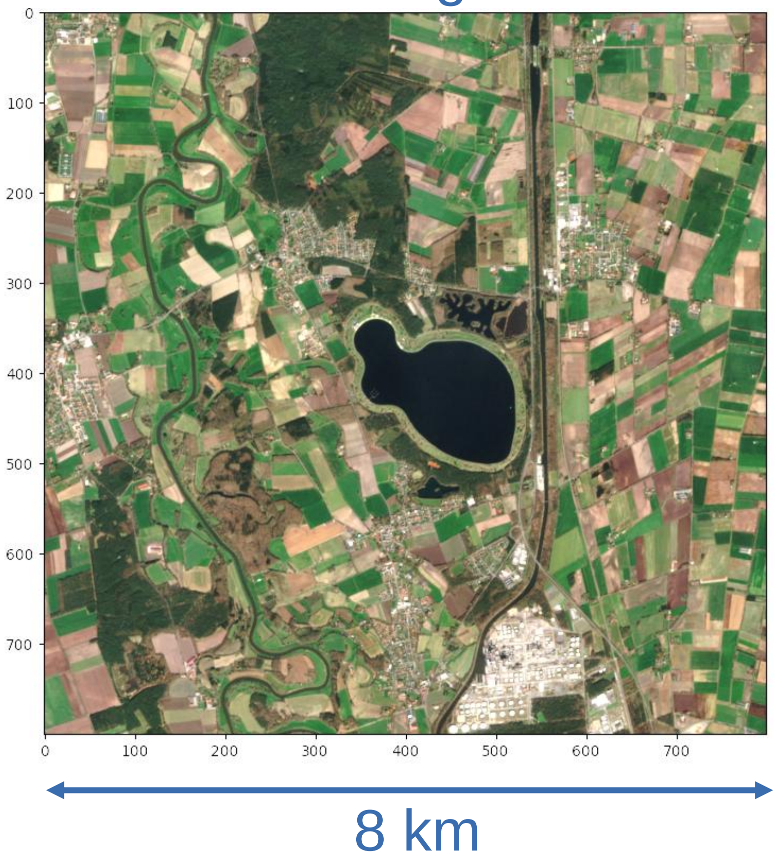
semantic segmentation



ATKIS reference



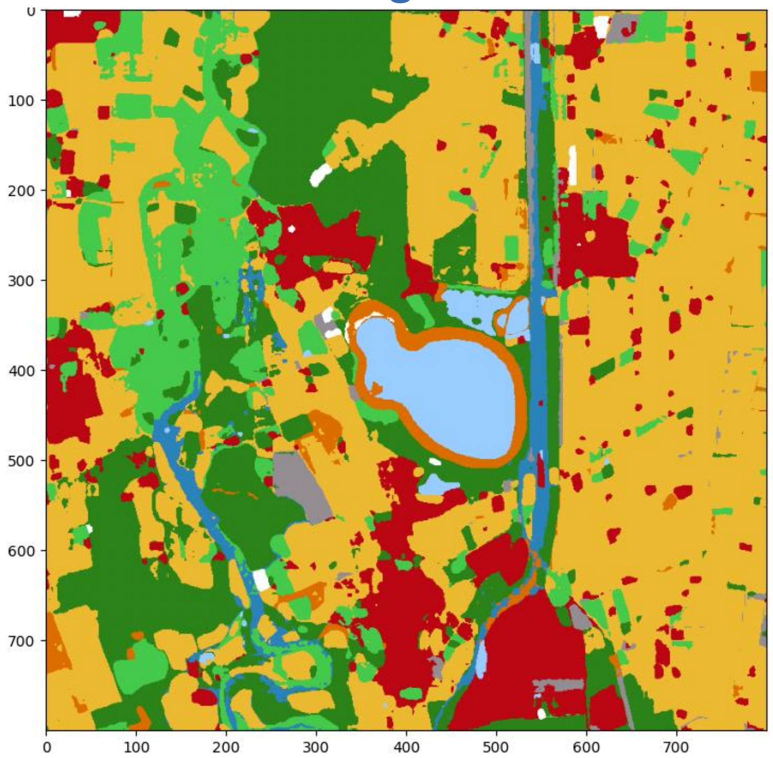
Sentinel-2 image



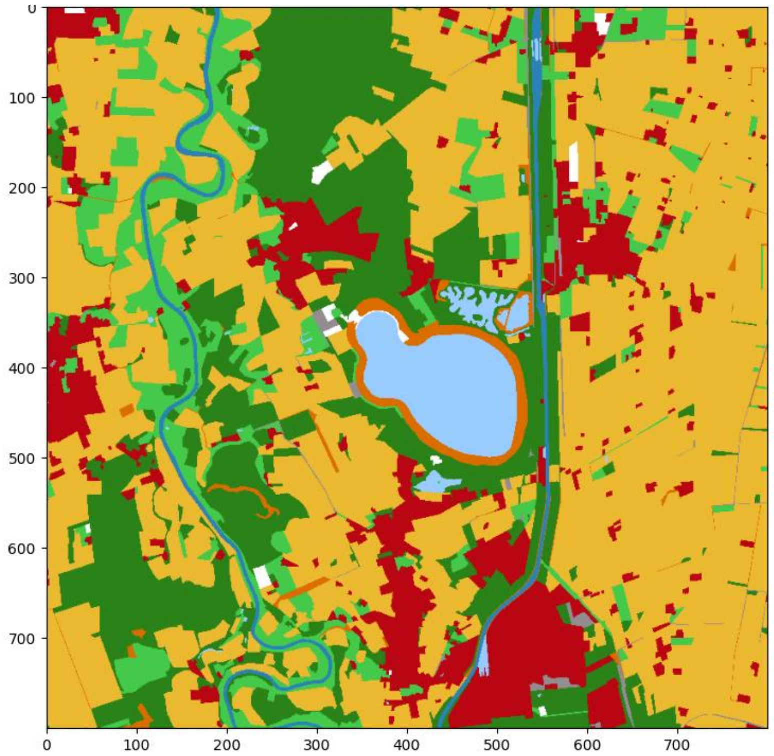
- Other
- Settlement
- Sealed area
- Agriculture
- Vegetation
- Forest
- Water – flow
- Water-stand
- Sea
- Barren land

Qualitative results – box prompt

semantic segmentation



ATKIS reference



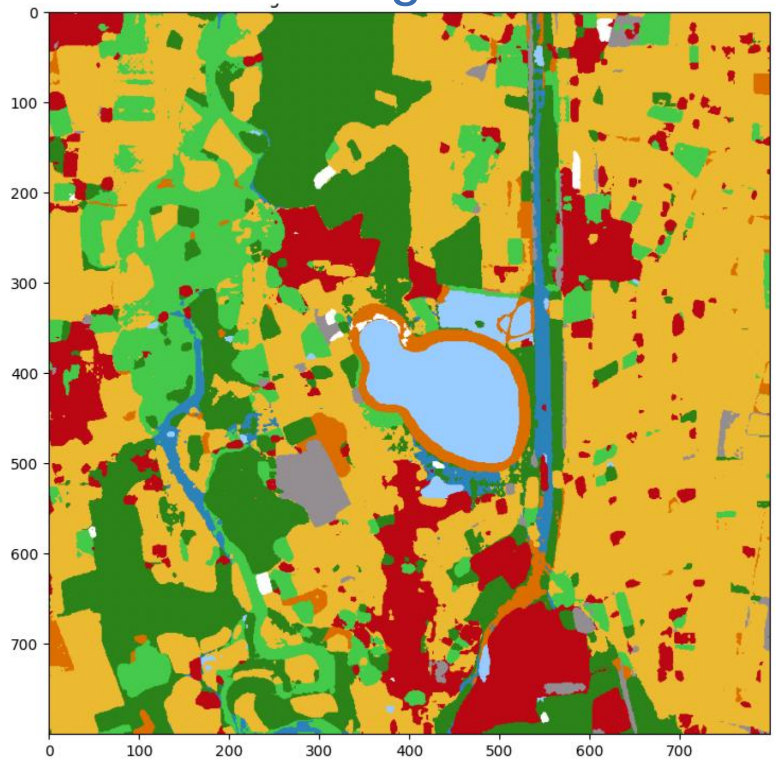
Sentinel-2 image

8 km

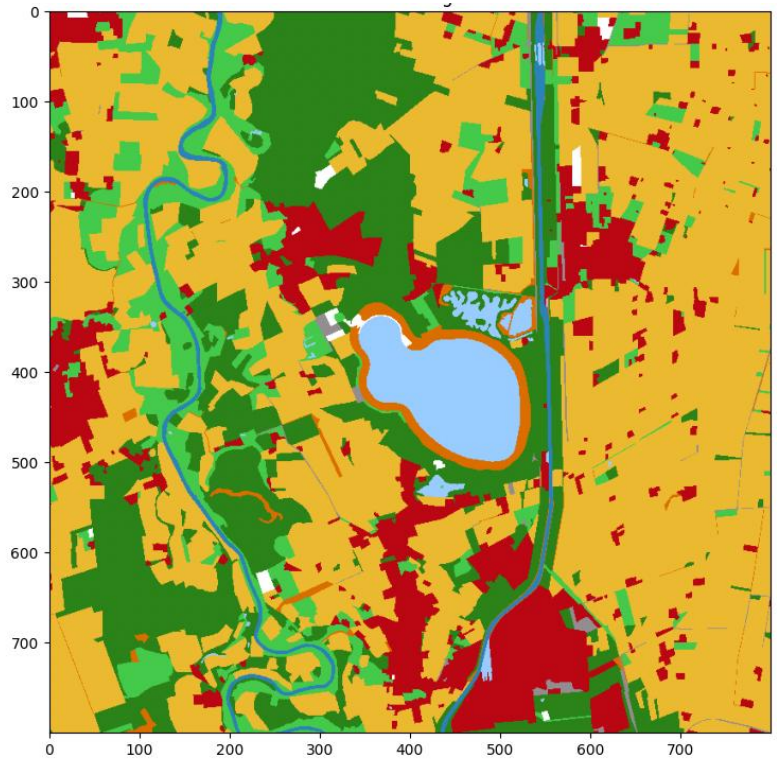
- Other
- Settlement
- Sealed area
- Agriculture
- Vegetation
- Forest
- Water – flow
- Water-stand
- Sea
- Barren land

Qualitative results – points, boxes and masks prompts

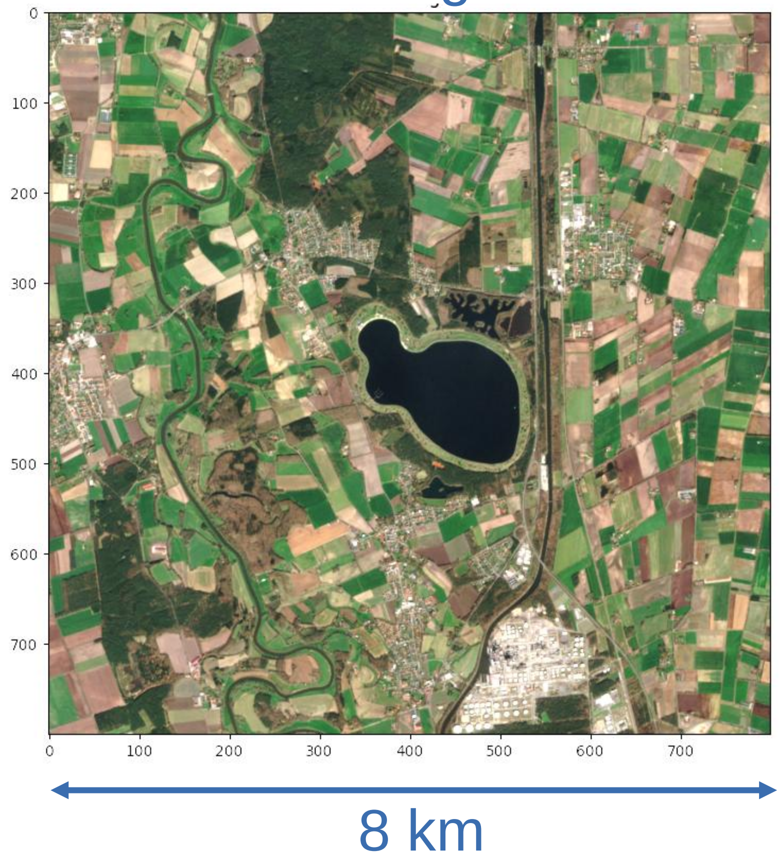
semantic segmentation



ATKIS reference



Sentinel-2 image



- Other
- Settlement
- Sealed area
- Agriculture
- Vegetation
- Forest
- Water – flow
- Water-stand
- Sea
- Barren land

Conclusion and further steps with SAM 2

- Conclusion:
  - Zero-shot segmentation with SAM 2 can reach around 60 % mean F1-score
  - Point prompts perform worst
  - Bounding boxes enlarge “search” area
- Further steps with SAM 2:
  - Extend the current method to “video” segmentation – satellite image time series
  - Optimization of prompts depending on class

Other potential directions of future research

- Change focus from semantic segmentation to change detection
  - Comparison of feature vectors (of objects?)
  - From zero- to few-shot learning?
- Which other Vision Foundation models or Foundation models for remote sensing might be interesting?
  - Use the potential of more channels of remote sensing images
- How can I handle incorrect ATKIS data when used as prior knowledge or training data? How does this affect the segmentation result?
- How do I work with few class changes in time series data?

Literature

- 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., Dollar, P., Feichtenhofer, C., 2024. **SAM 2: Segment Anything in Images and Videos.**
- Luo, M., Zhang, T., Wei, S., Ki, S., 2024. **SAM-RSIS: Progressively Adapting SAM With Box Prompting to Remote Sensing Image Instance Segmentation**. IEEE Transactions on Geoscience and Remote Sensing (TGRS), Vol. 62, p. 1-14.
- 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), p. 3359-3367.
- Mei, L., Ye, Z., Xu, C., Wang, H., Wang, Y., Lei, C., Yang, W., Li, Y., 2024. **SCD-SAM: Adapting Segment Anything Model for Semantic Change Detection in Remote Sensing Imagery**. IEEE Transactions on Geoscience and Remote Sensing (TGRS), Vol. 62, p. 1-13.
- Ren, S., Luzi, F., Lahrichi, S., Kassaw, K., Collins, L. M., Bradbury, K., Malof, J. M., 2024. **Segment anything, from space?** Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), p. 8355-8365.
- Xie, Z., Guan, B., Jiang, W., Yi, M., Ding, Y., Lu, H., Zhang, L., 2024. **PA-SAM: Prompt Adapter SAM for High-Quality Image Segmentation.** IEEE International Conference on Multimedia and Expo (ICME), p. 1-6.
- Diab, M., Kolokoussis, P., Brovelli, M., 2025. **Optimizing zero-shot text-based segmentation of remote sensing imagery using SAM and Grounding DINO.** Artificial Intelligence in Geosciences , Vol. 6, No. 1, p. 100105.
- Liu, N., Xu, X., Su, Y., Zhang, H., Li, H., 2025. **PointSAM: Pointly-Supervised Segment Anything Model for Remote Sensing Images.** IEEE Transactions on Geoscience and Remote Sensing (TGRS) , Vol. 63, p. 1-15.
- 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**. MDPI Remote Sensing , Vol. 17, No. 4.