NLS FINNISH GEOSPATIAL RESEARCH INSTITUTE FGI

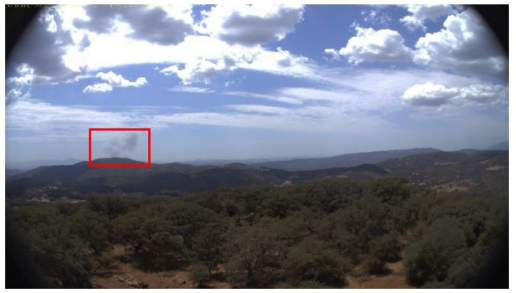
ONERA
THE FRENCH AEROSPACE LAB

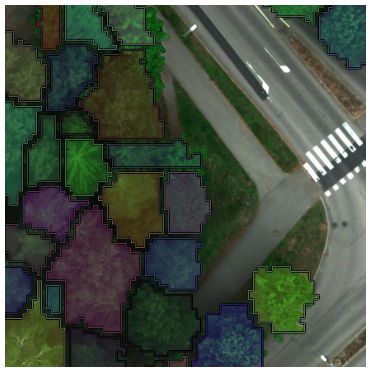
Aalto University

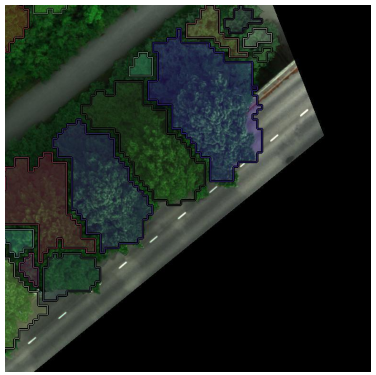
Seeing beyond the labels: Weakly supervised deep learning in remote sensing

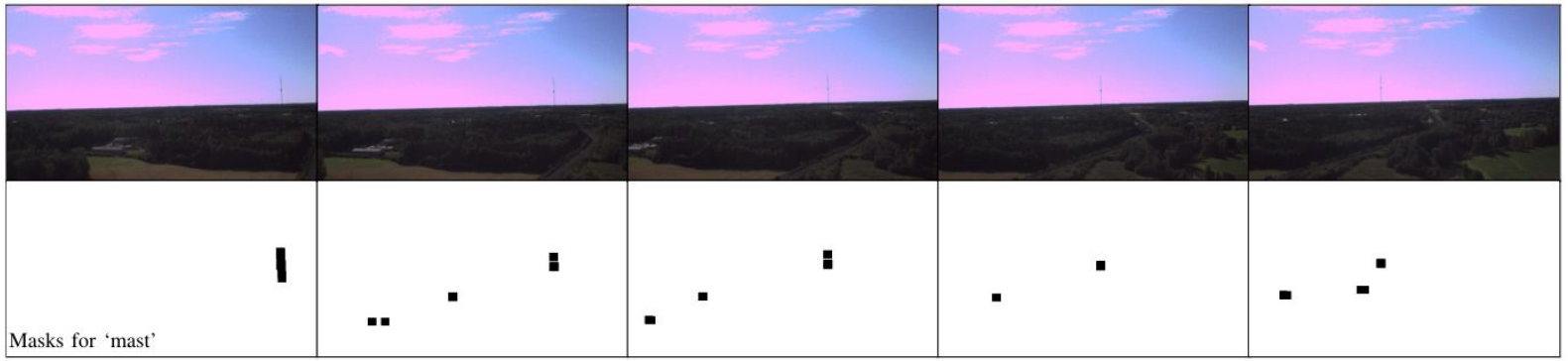
Doctoral research

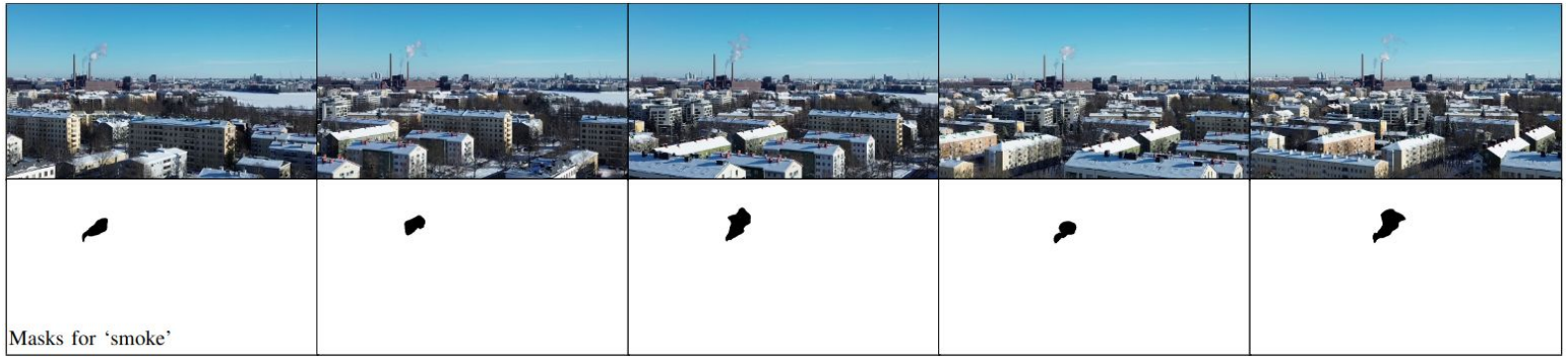
Julius Pesonen
Research Scientist, FGI
Doctoral Candidate, Aalto University
Visiting Researcher, ONERA

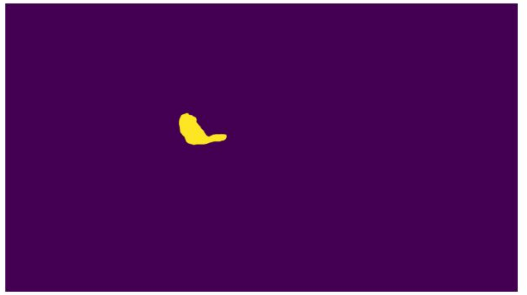


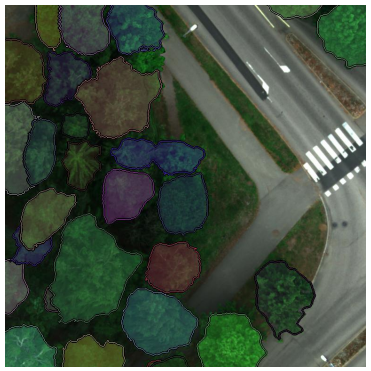


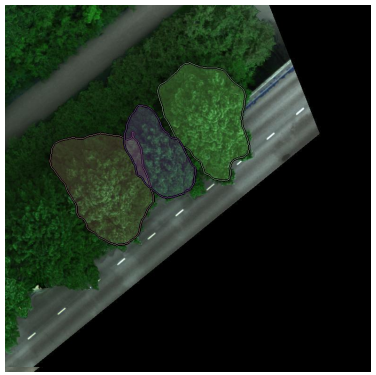


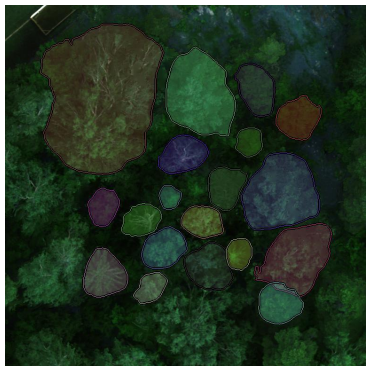

Masks for ‘mast’


Masks for ‘smoke’









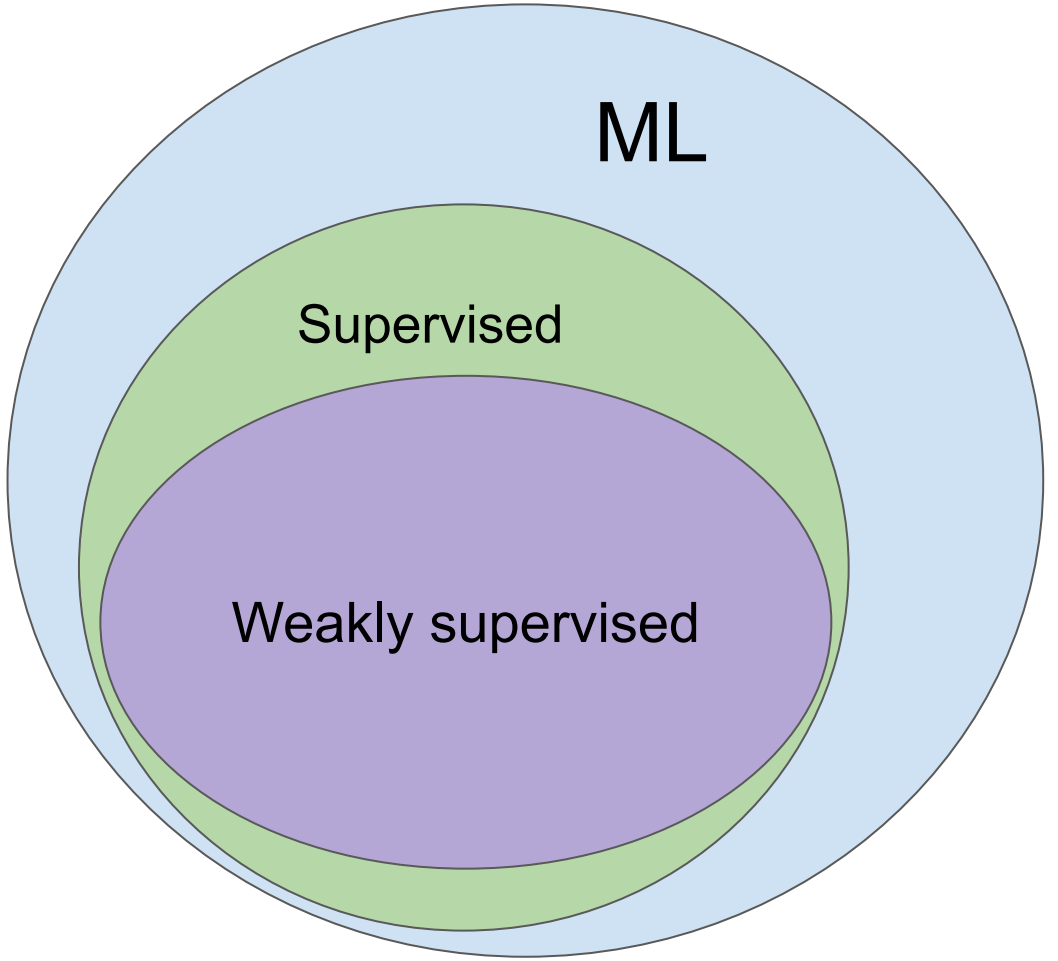
Mast at (200, 50, 400)

Smoke cloud at (500, 150, 1200)

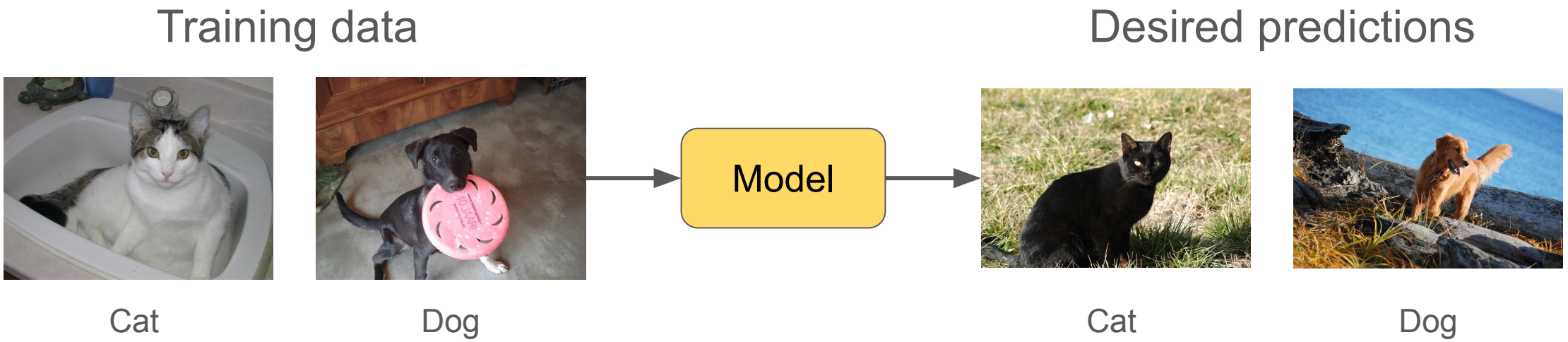
Contents

What is weakly supervised learning

- A (big) subcategory of supervised learning
- Deals with situations where there is discrepancy between training data and desired model behaviour
- Requires extra emphasis on evaluation



Fully supervised learning



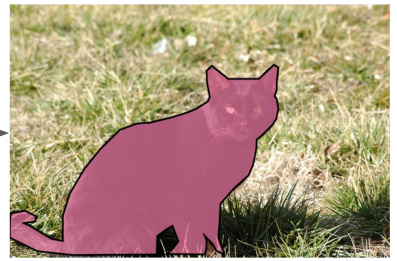
Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

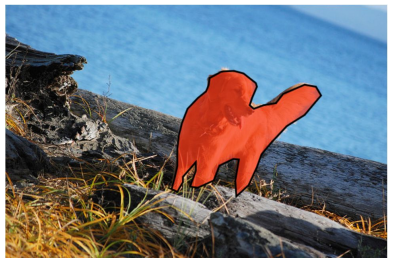
Fully supervised learning

Training data

Model

Desired predictions





Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

Weakly supervised learning examples: Text to segmentation

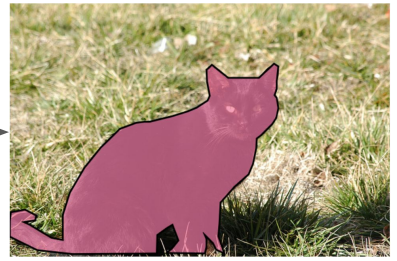
Training data

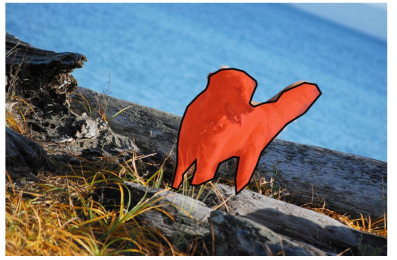
Cat

Dog

Model

Desired predictions





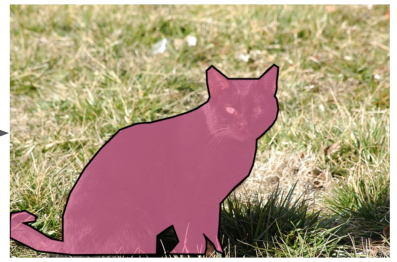
Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

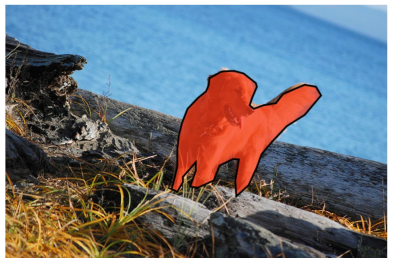
Weakly supervised learning examples: Sparse labels

Training data

Model

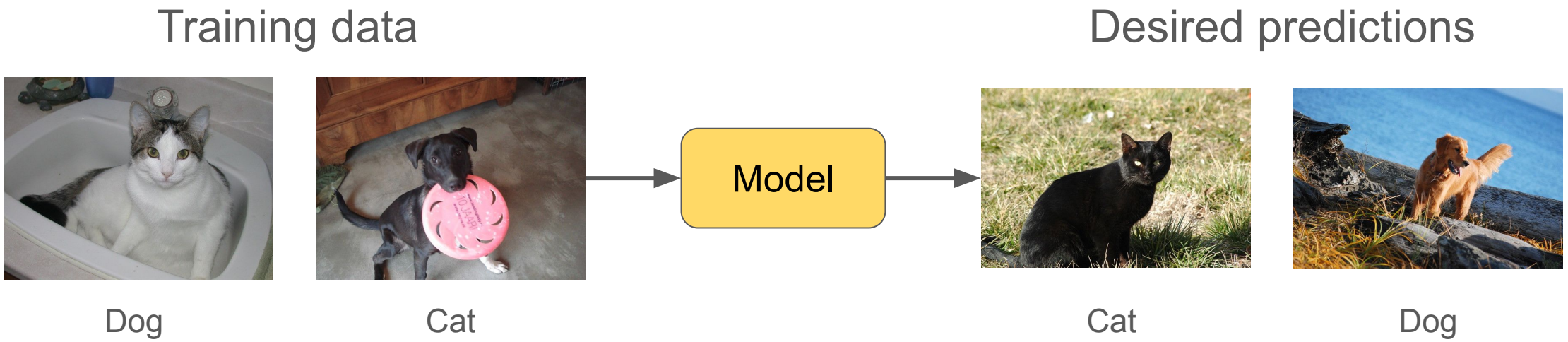
Desired predictions





Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

Weakly supervised learning examples: Noisy labels



Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

Weakly supervised learning examples: Finding 3D positions

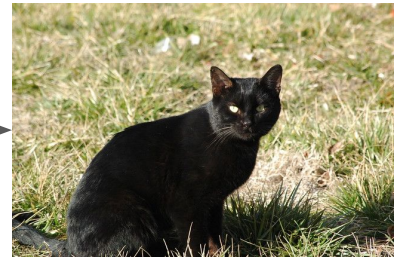
Training data

Cat

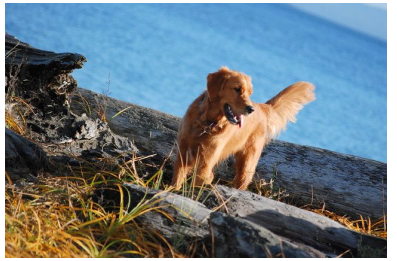
Dog

Model

Desired predictions



Cat at (0, 1, -1)

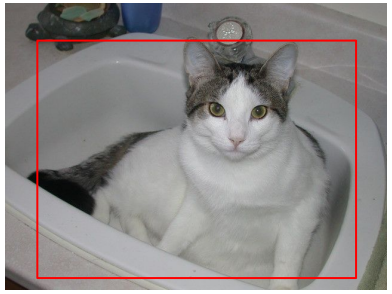


Dog at (0, 2, -2)

Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

How can we actually do it? Explicit example

Training data



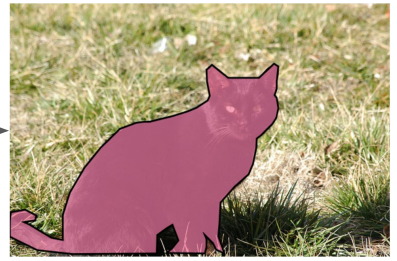
Segment Anything Model

Pseudo-labels

Pseudo-labels

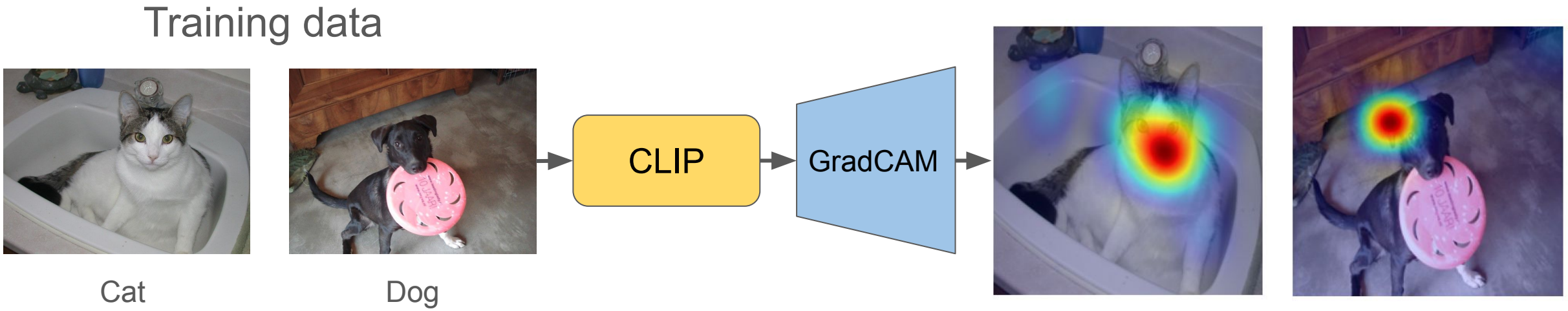
Downstream model

Final model outputs



Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

How can we actually do it? Implicit example



Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

What is weakly supervised learning: conclusion

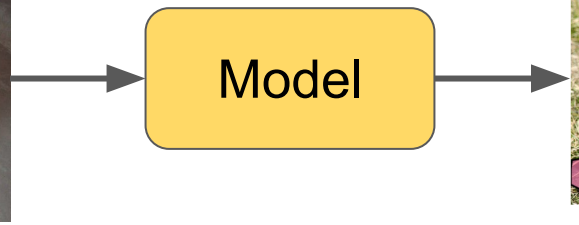
- Training models to predict things which are not directly available in the training labels
  - Text labels -> segmentation
  - Sparse labels -> dense predictions
- Broad term that covers many types of learning methods such as self-supervised and semi-supervised learning
- Can be explicit or implicit
  - Explicit: Specific methods that aim to train segmentation models directly from bounding box labels, e.g. BoxSnake, SAM distillation
  - Implicit: Training a model for one thing and applying it for another, e.g. CLIP for segmentation
- Connections to transfer learning, knowledge distillation, learning from privileged information, etc.

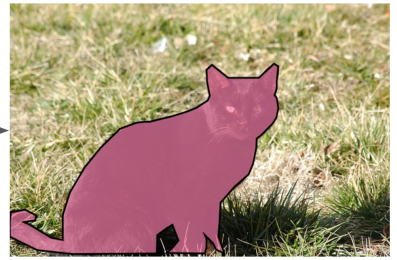
Why do we care about weak supervision?

Doing this:

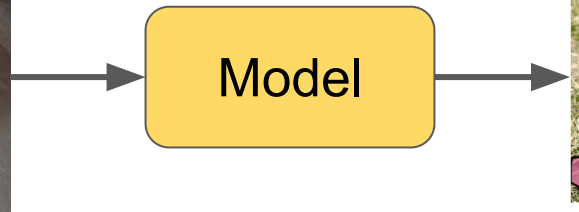
Cat

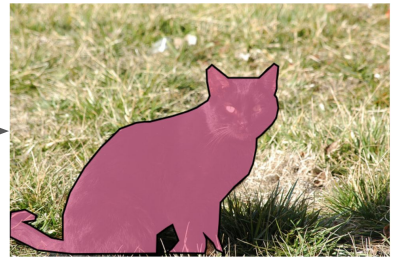
Dog





Is cheaper than doing this:





Images from COCO dataset Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Cham: Springer International Publishing.

Other benefits

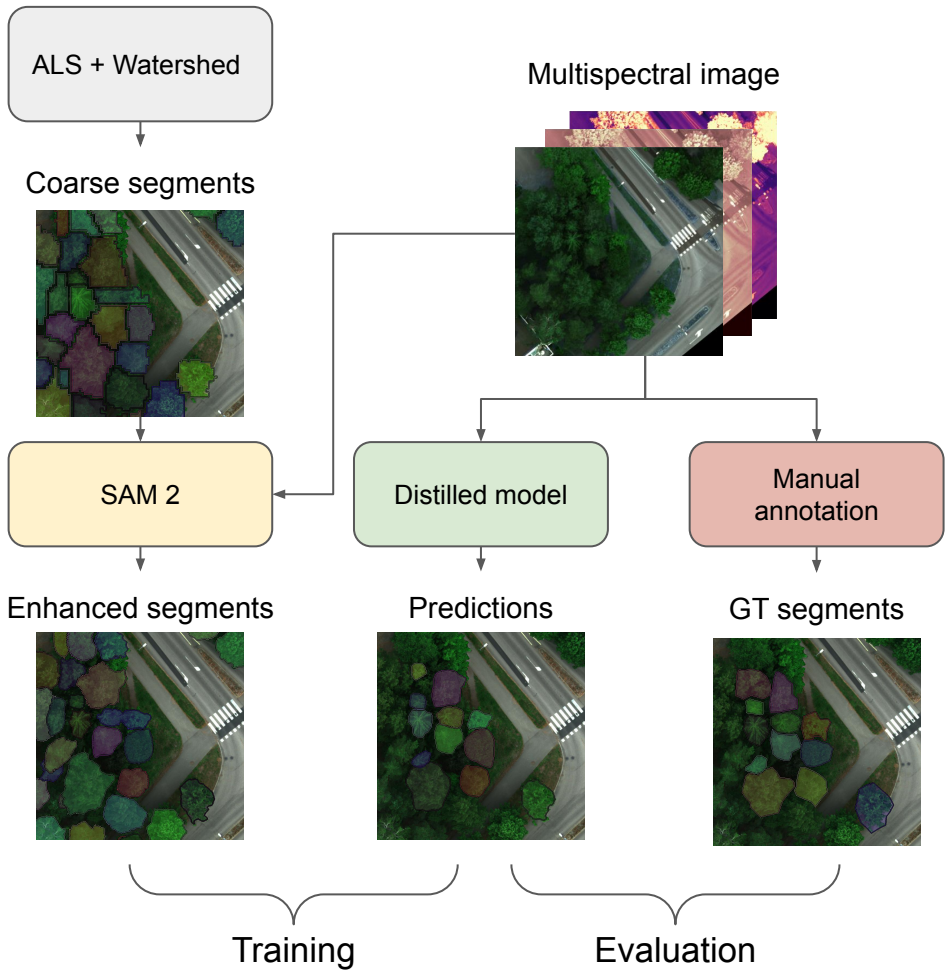
- Alleviating training label noise/enhancing their accuracy
- Distillation: Creating more efficient task-specific models
- Can enable models in new domains or even tasks

	Publication				
Benefit	I	II	III	IV	V
Efficient labelling	✓	✓	✓		✓
Domain extension	✓	✓	✓		✓
Distillation			✓		
Novel task				✓	✓

Weakly supervised learning benefits appearing in the different publications of my thesis.

Domain extensions

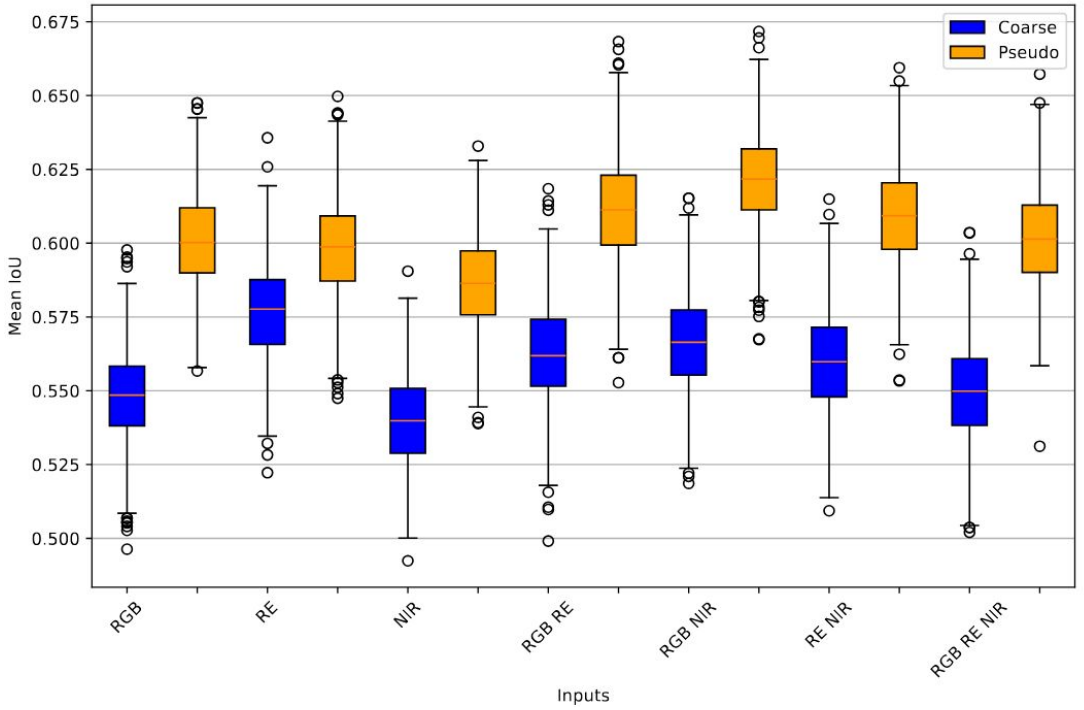
- Efficient labelling -> easy to label data from new domains
- Example: Tree crown instance segmentation
  - Poor domain transferability
  - Required training a new model
  - ALS + aerial imagery provided free model training data
  - Simple enhancement with SAM



Domain extension results

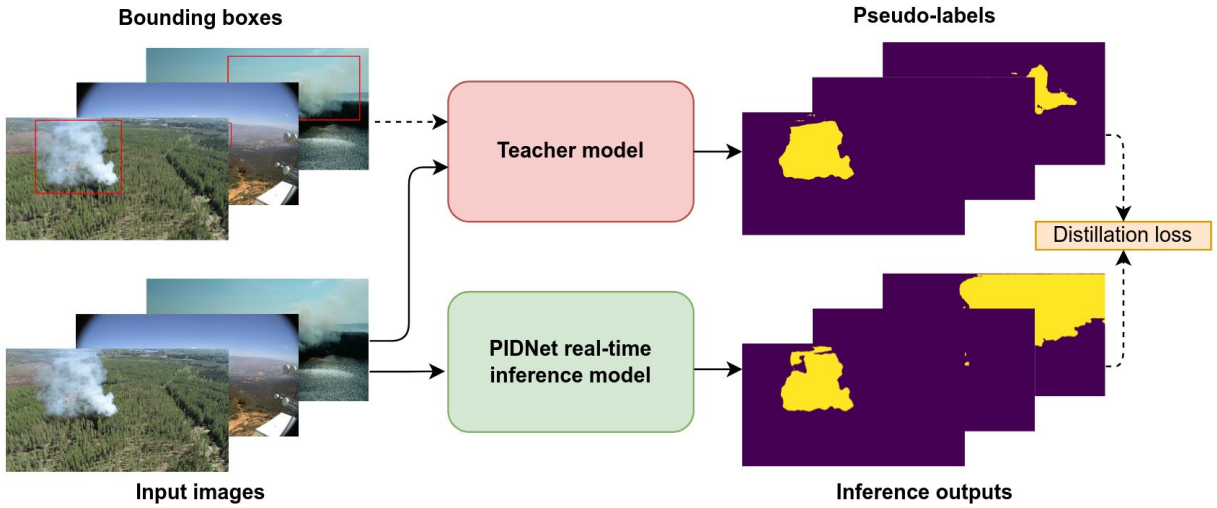
- Automatically generated training data from the correct location is better than similar human annotated data
- Using information from SAM is very useful -> Leverage learning from privileged information whenever it's possible

Method	F1	Precision	Recall	mIoU	mIoU 95% CI
Detectree2*	0.556	**0.911**	0.400	0.324	[0.281, 0.364]
U-net	0.306	0.214	0.536	0.497	[0.472, 0.520]
Grounded SAM	0.166	0.430	0.103	0.094	[0.070, 0.121]
DeepForest	0.501	0.544	0.464	0.396	[0.363, 0.428]
SAM 3	0.558	0.822	0.422	0.352	[0.311, 0.397]
Coarse masks	0.571	0.550	0.594	0.477	[0.447, 0.505]
Pseudo masks	0.597	0.584	0.611	0.513	[0.478, 0.545]
RGB	0.758	0.783	0.733	0.599	[0.566, 0.632]
RGB NIR	**0.778**	0.783	**0.772**	**0.621**	[0.587, 0.652]



Distillation and real-time inference

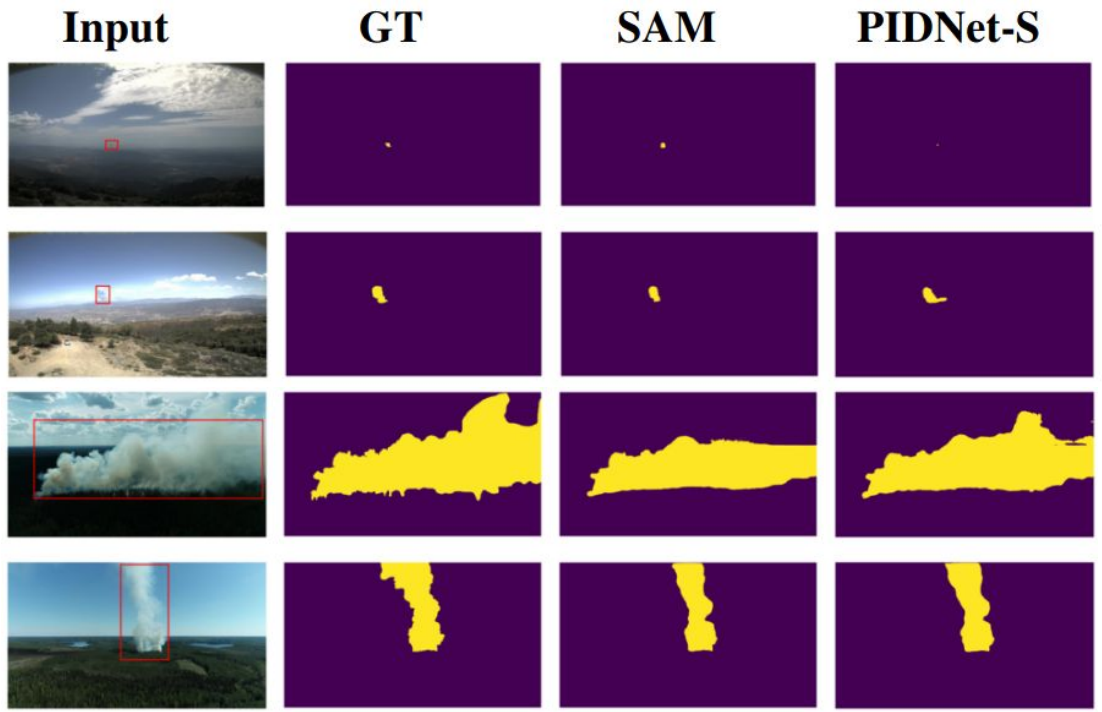
- Pseudo-labels from big models -> training data for smaller models
- Example: On-board wildfire smoke segmentation with drones
  - Segmentation hadn’t been used for the task before
  - Transforming bounding box data to segmentation data (with SAM)



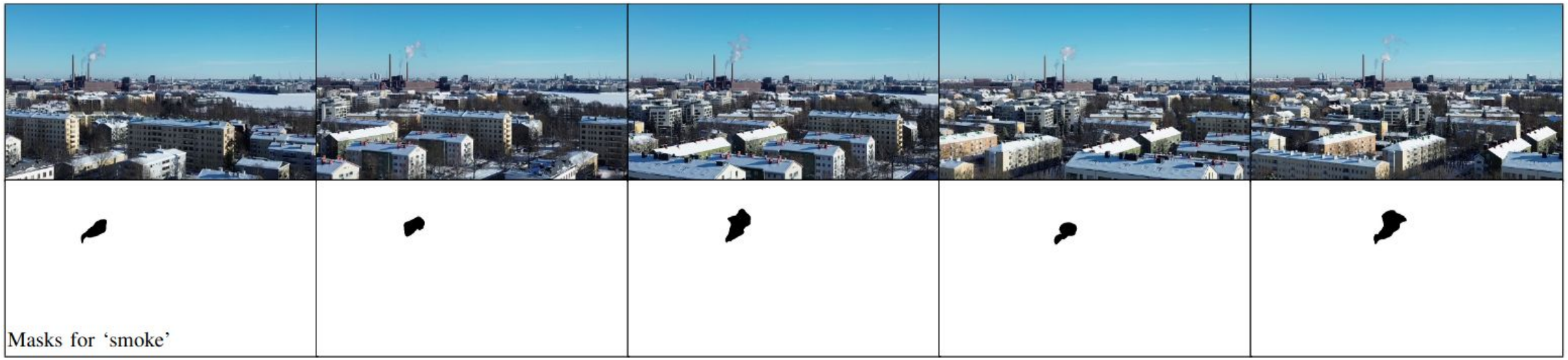
Distillation results

Teacher	Student	mIoU	Accuracy	Precision	Recall	F_1
Swin-L-FPN	-	**0.651**	0.966	0.871	**0.730**	**0.772**
SAM	PIDNet-S	0.594	0.960	**0.891**	0.645	0.707

- Using SAM for training data generation proved useful once again
  - Similar performance to a model trained directly with BoxSnake
- Extremely convenient way to obtain models for real-time segmentation on specific tasks
- Inference video



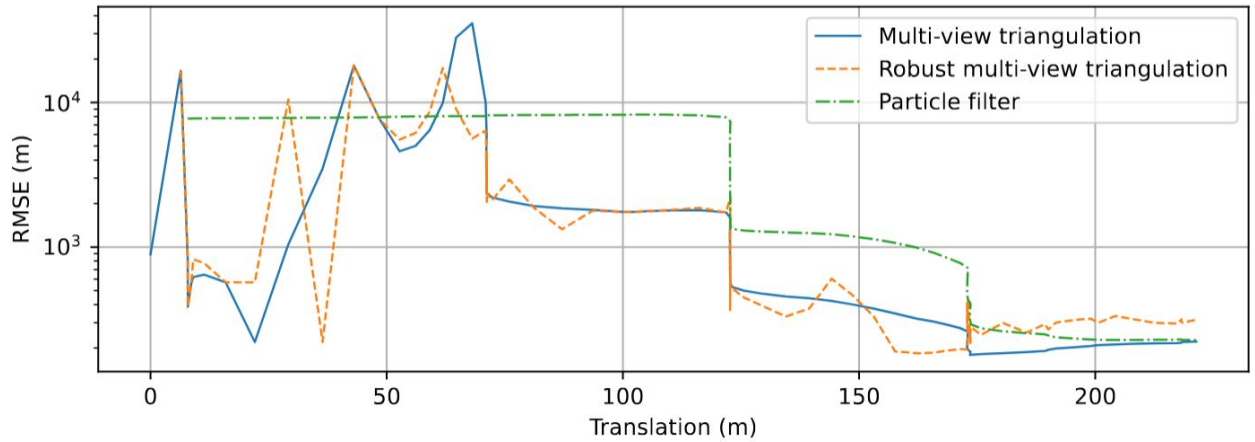
Novel tasks - 3D vision

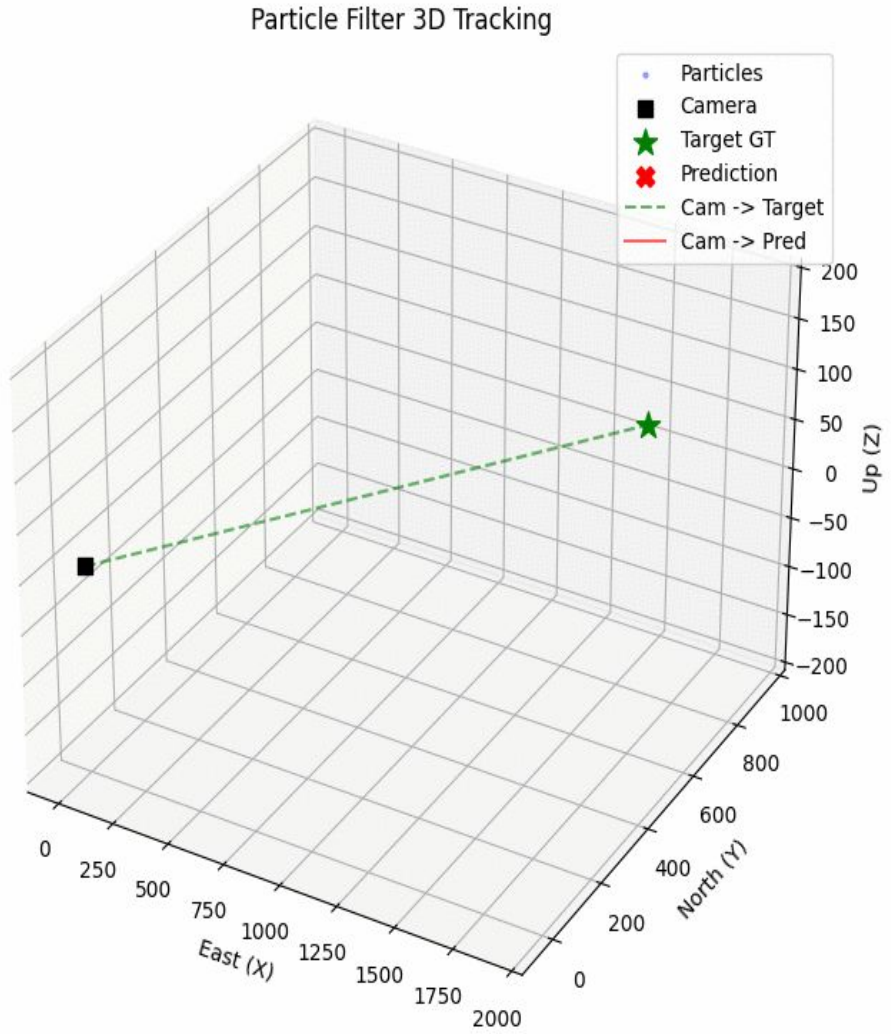


- How do we get from single perspective 2D images to 3D position estimates?
  - By moving the camera
- Sometimes it’s useful to just combine ML with robust non-learning based algorithms

3D vision results

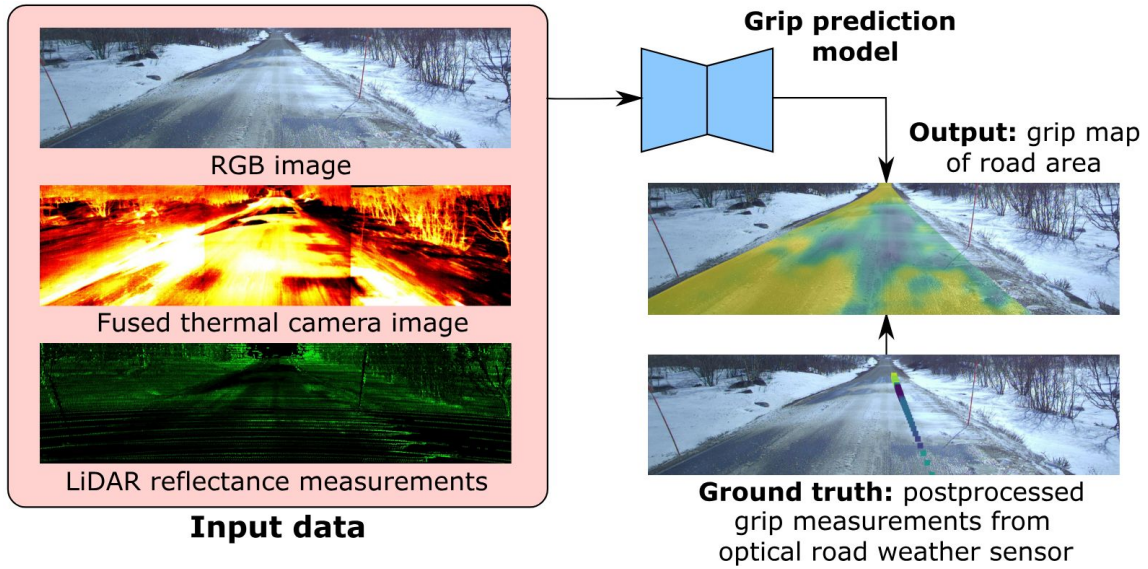
- Weak supervision can also provide minor improvements even when it’s not the main focus
- Using weakly supervised segmentation instead of bounding box prediction in this case provides more reliable estimates of the smoke cloud’s 3D size and shape





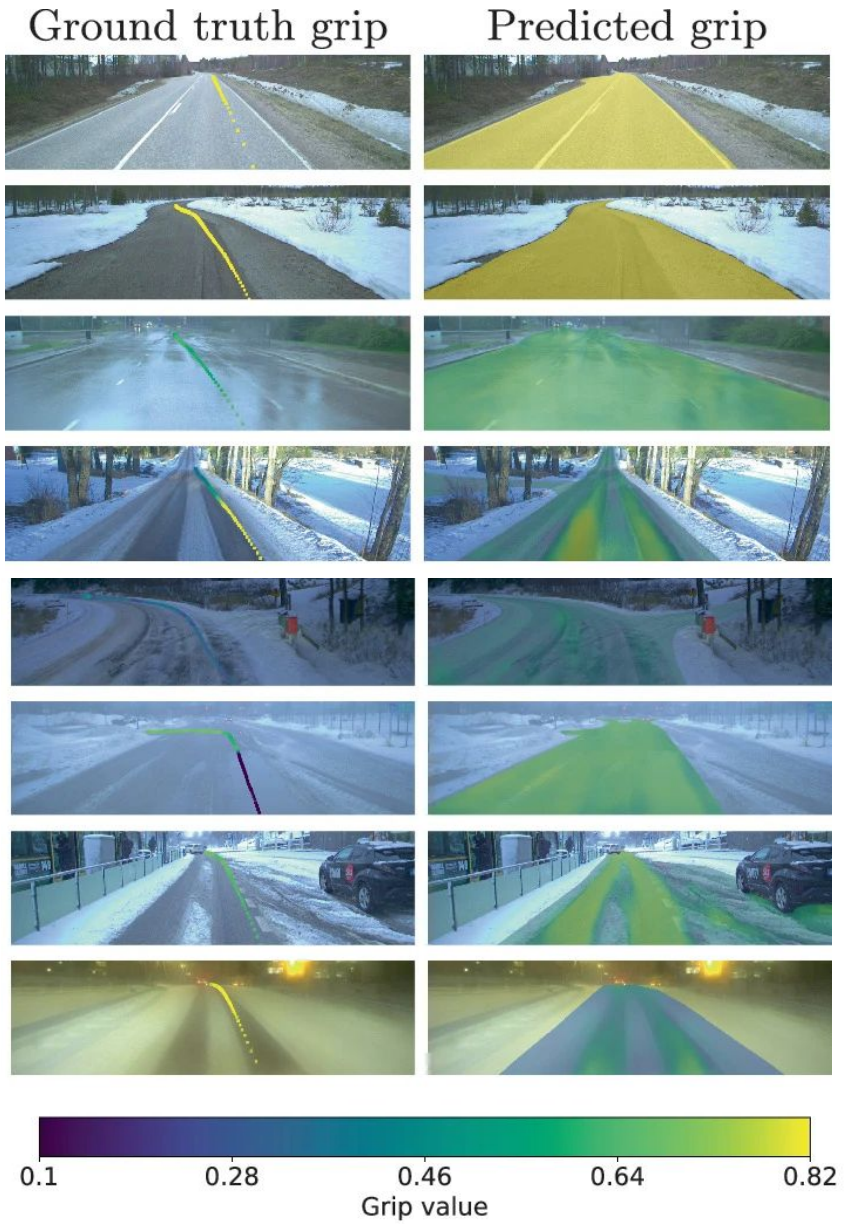
Novel tasks - Road grip estimation

- Prior to our work no data existed to do pixel-wise grip estimation from car mounted long range sensors
- Collected a custom sparsely annotated dataset
- Automatable data annotation



Road grip estimation results

- Convolutional image-to-image models learn from sparse labels with sufficient augmentations
- Evaluation remains challenging
- Weak supervision enables models in practical novel use cases
- Demo video

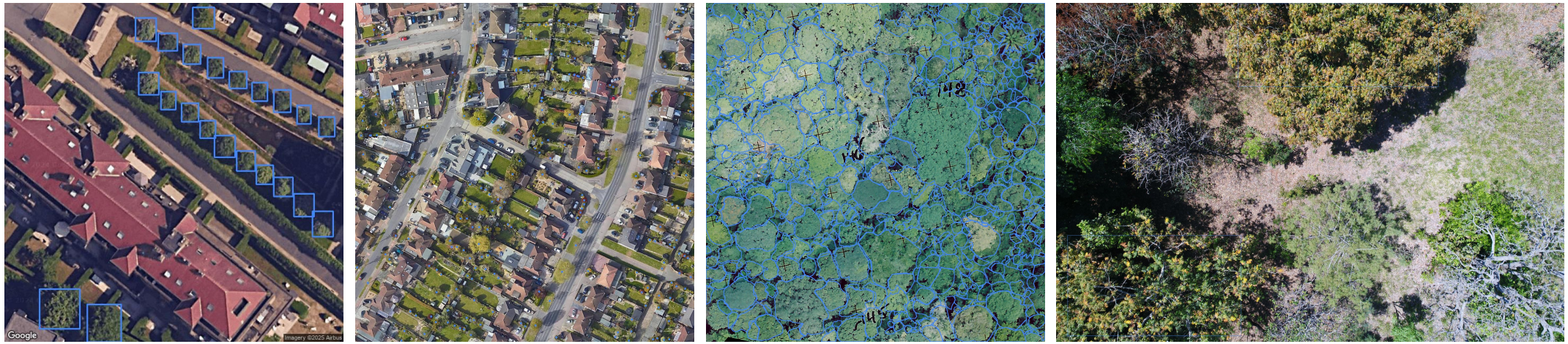


Future prospects

- More analysis on what kind of features are learned implicitly for more efficient use of existing resources
  - Examples: DINO and CLIP models
- Practice: More (complex) applications
  - Examples: Learning-based 3D vision, Hyperspectral sensing for vegetation trait estimation, …
  - Limited by evaluation methods -> building new benchmarks

Large scale tree crown segmentation benchmark

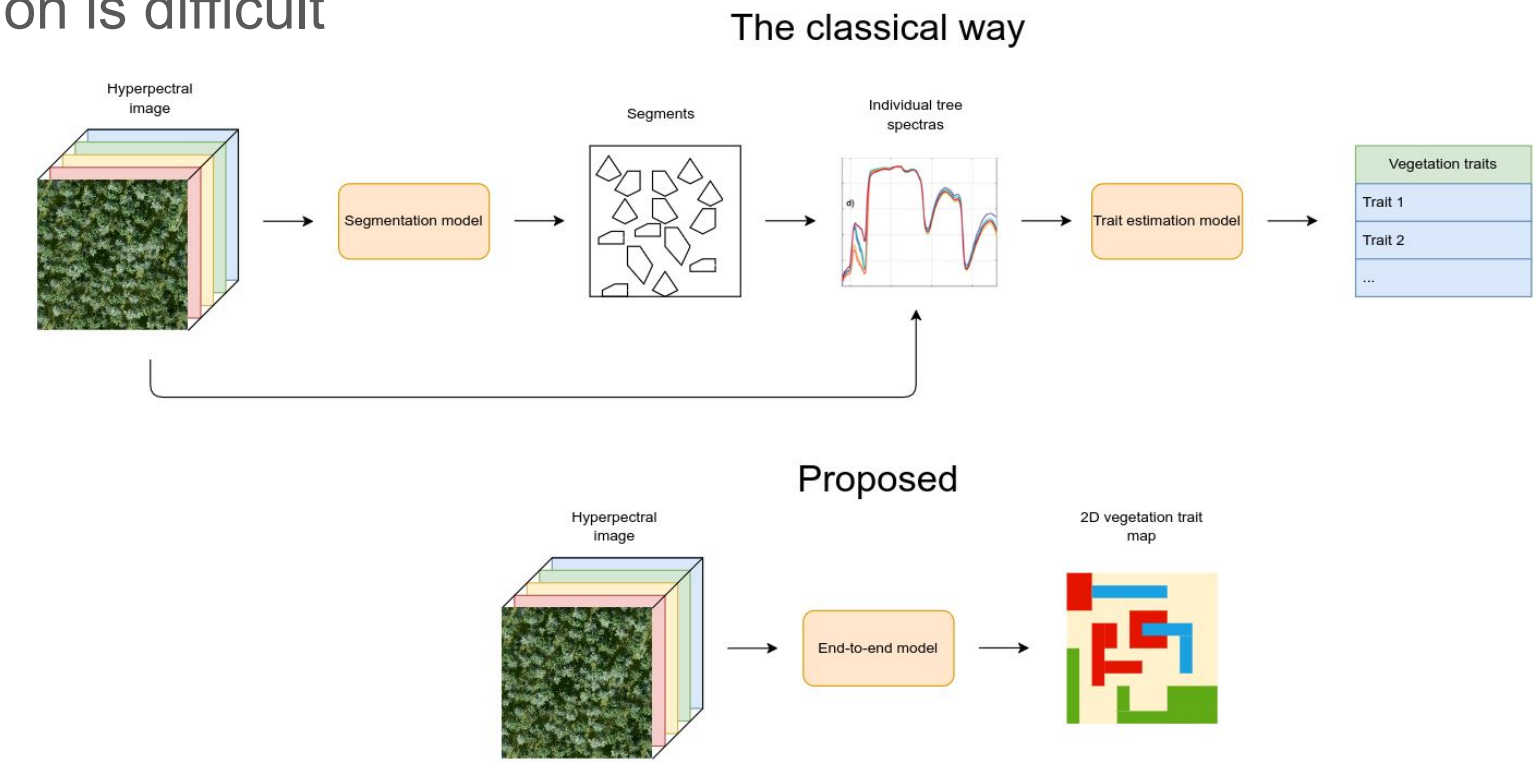
- Improving the tree crown segmentation with more varied data
  - Locations
  - Species
  - Supervision: Hand drawn segments, bounding boxes, tree centre points



Images obtained through the MillionTrees project: https://milliontrees.idtrees.org/en/latest/

Vegetation trait estimation

- Estimating bio-chemical properties of trees directly from hyperspectral imagery
  - Dense to dense prediction
- Evaluation is difficult



Learning-based 3D vision

- What if it would be possible to generate 3D training data for any application?
  - Enables end-to-end learning for 3D tasks
  - Possible data generation methods: Simulation guided diffusion, world models
- Please fund my research

Conclusion

- Very practical data engineering related topic
- Benefits in practically all real world applications of deep learning
- Builds on the creation of meaningful evaluation methods
- Often improved by incorporating information from various sources

Big thanks to:

Eija Honkavaara, Research Professor, FGI
Arno Solin, Professor, Aalto University
Sophie Fabre, Research Director, ONERA
Jyri Maanpää, Research Scientist, FGI
Teemu Hakala, Research Scientist, FGI
Väinö Karjalainen, Research Scientist, FGI
Emma Turkulainen, Research Scientist, FGI
And many others!

Merci!

My homepage with links to publications and other resources →

julppe.github.io

References (unordered)

Our work (links at https://julppe.github.io/)**:**

Pesonen, Julius, et al. "Boreal Forest Fire: UAV-collected wildfire detection and smoke segmentation dataset." Scientific Data 12.1 (2025): 1419.

Pesonen, Julius, et al. "Detecting wildfires on UAVs with real-time segmentation trained by larger teacher models." 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025.

Maanpää, Jyri, et al. "Dense road surface grip map prediction from multimodal image data." International Conference on Pattern Recognition. Cham: Springer Nature Switzerland, 2024.

Pesonen, Julius, Arno Solin, and Eija Honkavaara. "Finding 3D Positions of Distant Objects from Noisy Camera Movement and Semantic Segmentation Sequences." arXiv preprint arXiv:2509.20906 (2025).

Pesonen, Julius, et al. "Learning Image-based Tree Crown Segmentation from Enhanced Lidar-based Pseudo-labels." arXiv preprint arXiv:2602.13022 (2026).

Pesonen, Julius. "Pixelwise road surface slipperiness estimation for autonomous driving with weakly supervised learning." (2023).

Other references:

Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

Ravi, Nikhila, et al. "Sam 2: Segment anything in images and videos." International Conference on Learning Representations. Vol. 2025. 2025.

Carion, Nicolas, et al. "Sam 3: Segment anything with concepts." arXiv preprint arXiv:2511.16719 (2025).

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

Lin, Yuqi, et al. "Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

Xu, Jiacong, Zixiang Xiong, and Shankar P. Bhattacharyya. "PIDNet: A real-time semantic segmentation network inspired by PID controllers." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

Yang, Rui, et al. "Boxsnake: Polygonal instance segmentation with box supervision." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

Caron, Mathilde, et al. "Emerging properties in self-supervised vision transformers." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

We know the Earth
– we secure the future

NLS
FINNISH GEOSPATIAL
RESEARCH INSTITUTE
FGI