Learnable Earth Parser: Discovering 3D Prototypes in Aerial Scans

Romain Loiseau[1, 2] Elliot Vincent[1, 3] Mathieu Aubry[1] Loic Landrieu[1, 2]

romain.loiseau@enpc.fr elliot.vincent@enpc.fr mathieu.aubry@enpc.fr loic.landrieu@enpc.fr

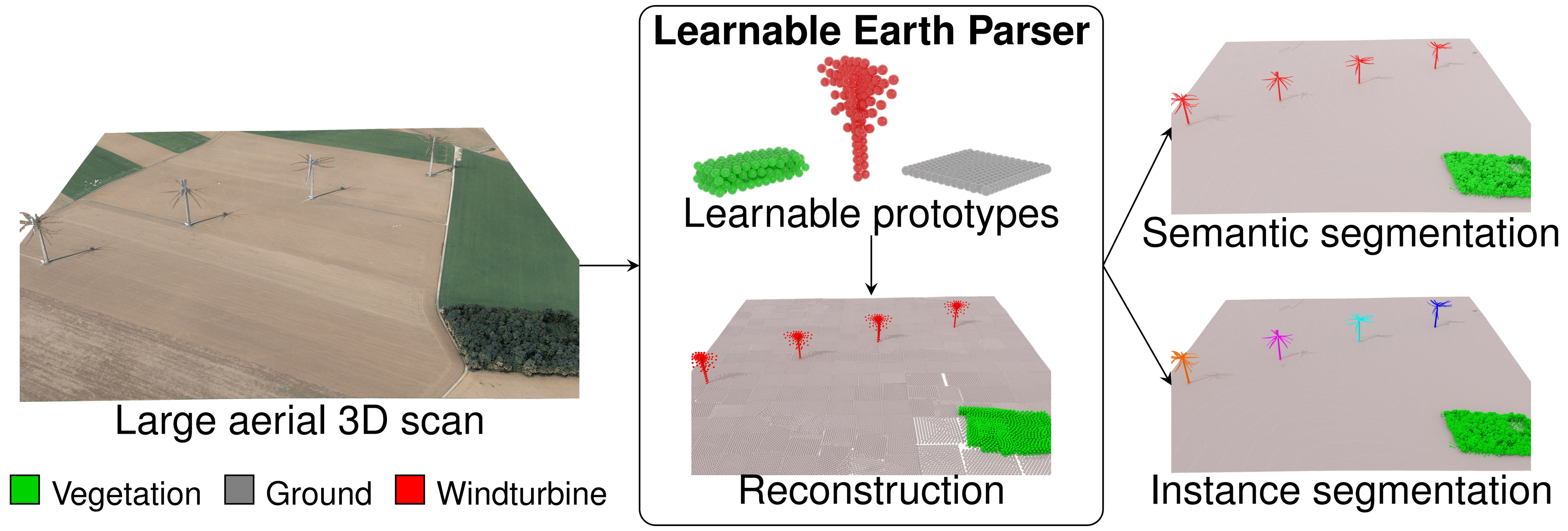
1 École des Ponts ParisTech 2 IGN Institut National de l'Information Géographique et Forestière 3 Inria

Université Gustave Eiffel

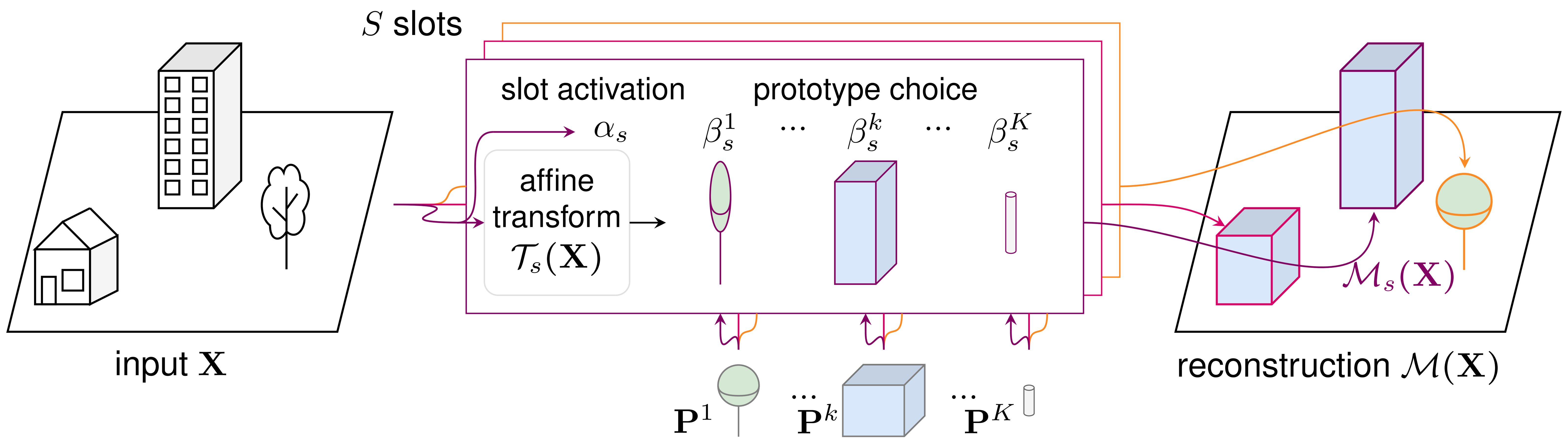
Overview

Goal: provide a **practical and interpretable tool** for **analyzing 3D scenes** for aerial surveying and mapping, without relying on application-specific user annotations.

Approach: a **probabilistic reconstruction model** that decomposes inputs into a set of **learned prototypical 3D shapes**, for **unsupervised instance/semantic segmentation**.



Learnable Earth Parser



Method Overview.

- Our model **approximates an input** point cloud $\mathbf{X}$ with S slot models.
- Each slot maps $\mathbf{X}$ to an **affine 3D deformation** $\mathcal{T}_s(\mathbf{X})$, a **slot activation probability** α_s, and the **joint probabilities** $\beta_s^1, \cdots, \beta_s^K$ of the slot being activated and choosing one of the K prototype point clouds $\mathbf{P}^1, \cdots, \mathbf{P}^K$.
- The output $\mathcal{M}_s(\mathbf{X})$ of an activated slot s is obtained by applying $\mathcal{T}_s(\mathbf{X})$ to its most likely prototype.
- Non-activated slots do not contribute to the output.

Method

Learnable shape prototypes:
Following [4], we define K point clouds $\mathbf{P}^1, \cdots, \mathbf{P}^K$ that we refer to as ***prototypes***. Each prototype is meant to **represent a single instance of a recurring 3D structure** in the considered scene $\mathbf{X}$. The points' coordinates are free parameters of the model.

Scene reconstruction model:

$$\mathcal{M}(\mathbf{X}) = \bigcup_{\substack{s=1\cdots S \\ a_s=1}} \mathcal{M}_s(\mathbf{X}) \text{ , with } \mathcal{M}_s(\mathbf{X}) = \mathbf{Y}_s^k = \mathcal{T}_s(\mathbf{X})[\mathbf{P}^k] \text{ if } b_s = k \text{ .}$$

Probabilistic modeling:
a and b as random variables following (multi)-Bernoulli distributions ;
$p(a_s = 1) = \alpha_s$: probability that the slot s is activated ;
$p(a_s = 1, b_s = k) = \beta_s^k$: probability it is activated and selects prototype k.

Unsupervised training losses:
Slots average of the expected distance between $\mathcal{M}_s(\mathbf{X})$ and $\mathbf{X}$:

$$\mathcal{L}_{\text{acc}}(\mathcal{M}, \mathbf{X}) = \frac{1}{S} \sum_{s=1}^{S} \mathbb{E}_{a_s, b_s} \left[d\left(\mathcal{M}_s(\mathbf{X}), \mathbf{X}\right) \right] \text{ .}$$

Average over all points x of $\mathbf{X}$ of the expected distance between x and its closest point in the reconstruction:

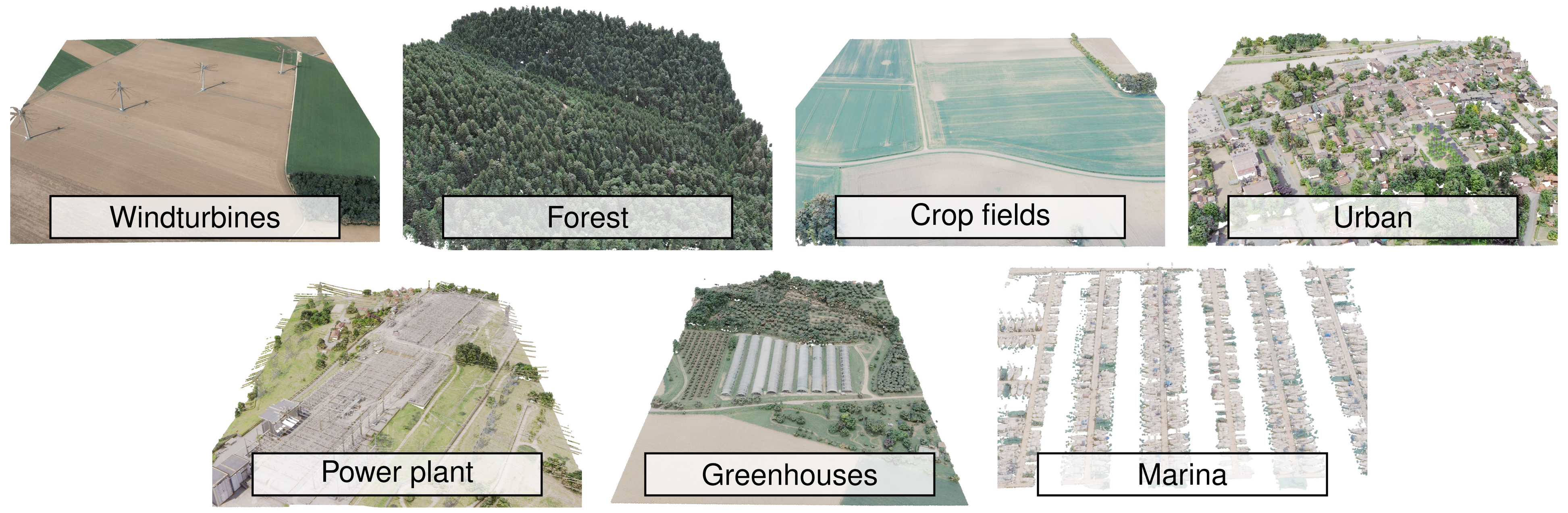
$$\mathcal{L}_{\text{cov}}(\mathcal{M}, \mathbf{X}) = \frac{1}{|\mathbf{X}|} \sum_{x \in \mathbf{X}} \mathbb{E}_{a,b} \left[\min_{s | a_s = 1} d\left(x, \mathcal{M}_s(\mathbf{X})\right) \right] \text{ .}$$

The final unsupervised loss is the sum of reconstruction losses and regularization:

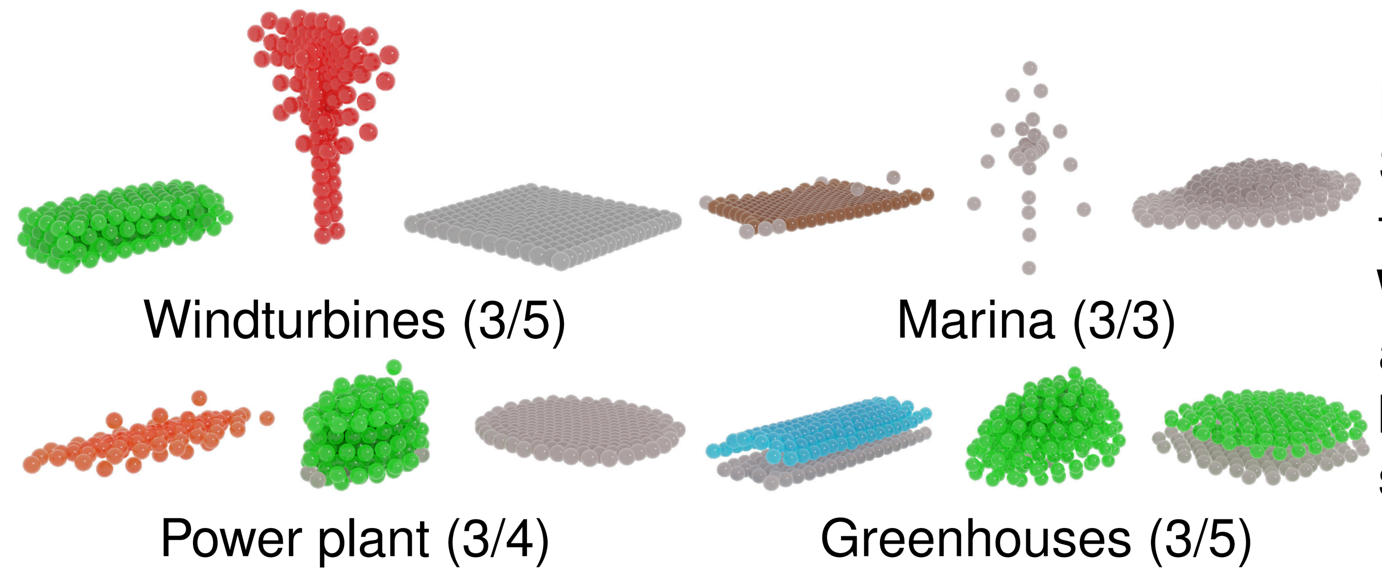
$$\mathbb{E}_{\mathbf{X}} \left[\mathcal{L}_{\text{acc}}\left(\mathcal{M}, \mathbf{X}\right) + \mathcal{L}_{\text{cov}}\left(\mathcal{M}, \mathbf{X}\right) \right] + \lambda_{\text{act}} \mathcal{L}_{\text{act}} + \lambda_{\text{slot}} \mathcal{L}_{\text{slot}} + \lambda_{\text{proto}} \mathcal{L}_{\text{proto}} \text{ .}$$

Earth Parser Dataset

Earth Parser Dataset: Annotated aerial LiDAR scans in diverse real-world environments.

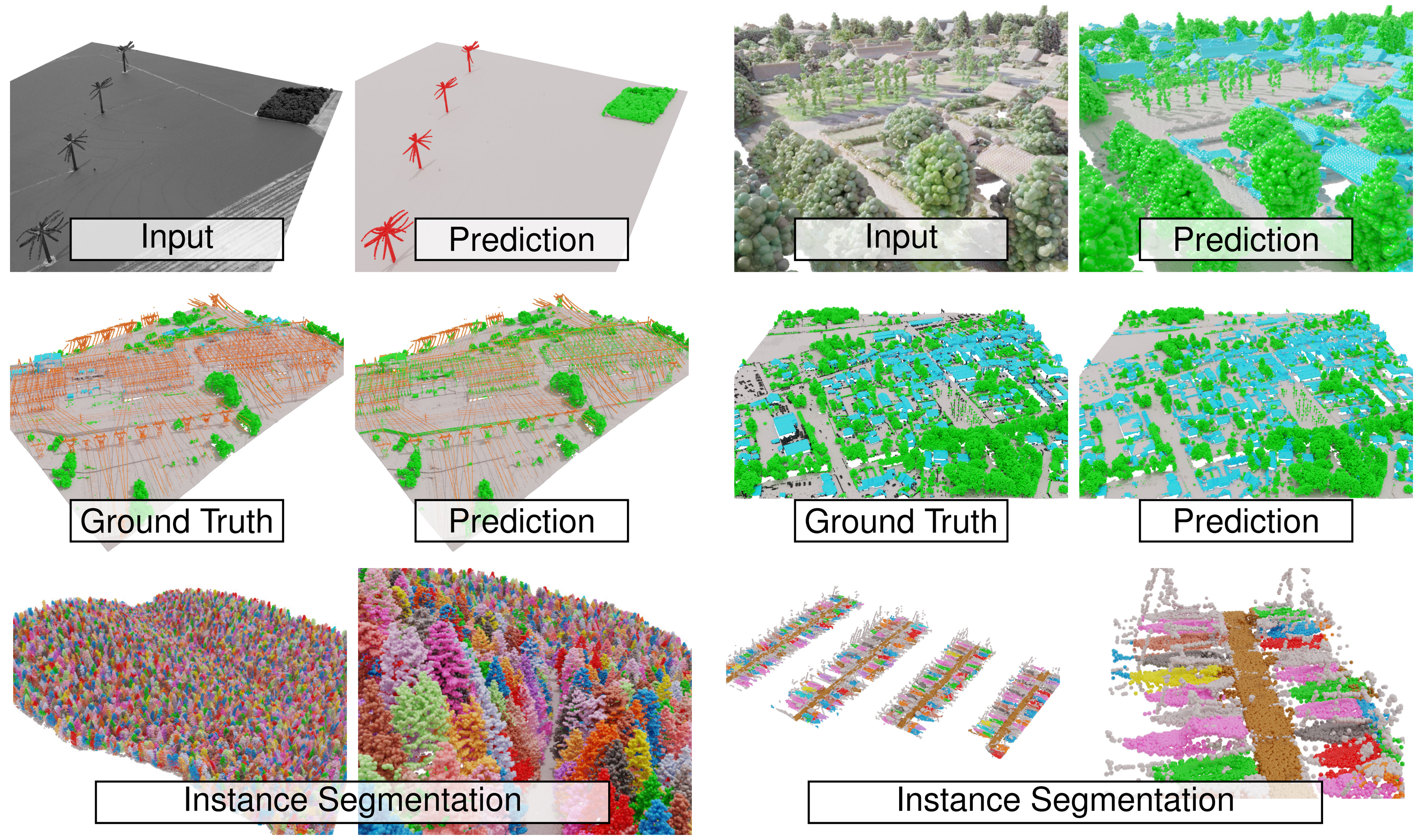


Meaningful and Interpretable Prototypes



Learned Prototypes.
Selected learned prototypes on different scenes. We show three prototypes among those selected by our post-processing selection.

Unsupervised Qualitative and Quantitative Results



	Rec.	Sem.	Crop fields		Forest		Greenhouses		Marina		Power plant		Urban		Windturbines	
			Cham.	mIoU	Cham.	mIoU	Cham.	mIoU	Cham.	mIoU	Cham.	mIoU	Cham.	mIoU	Cham.	mIoU
k-means (i,z)	✗	✓	—	93.8	—	71.5	—	39.3	—	41.4	—	42.8	—	56.5	—	87.6
SuperQuadrics [1]	3D	✗	0.86	—	1.04	—	0.60	—	0.93	—	0.58	—	0.40	—	13.50	—
DTI-Sprites [2]	2.5D+i	✓	6.10	83.2	14.59	40.2	5.36	42.0	6.16	41.4	5.36	29.0	2.99	47.3	36.19	25.9
AtlasNet v2 [3]	3D+i	✓	1.07	43.1	1.58	71.4	0.56	49.1	**0.73**	42.1	0.45	41.6	0.63	48.8	9.47	48.1
Ours	3D+i	✓	**0.72**	**96.9**	**0.88**	**83.7**	**0.40**	**91.3**	0.82	**78.7**	**0.44**	**52.2**	**0.29**	**83.2**	**6.65**	**93.4**

Bibliography: [1] Paschalidou *et al*. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. CVPR19. [2] Monnier *et al*. Unsupervised layered image decomposition into object prototypes. ICCV21. [3] Deprelle *et al*. Learning elementary structures for 3D shape generation and matching. NeurIPS19. [4] Loiseau *et al*. Representing Shape Collections with Alignment-Aware Linear Models. 3DV21. [5] IGN (French mapping agency). LiDAR-HD: A 3D mapping of France's soil and subsoil, 2021.

Acknowledgements: This work was supported by ANR project READY3D ANR-19-CE23-0007, ANR under the France 2030 program (ANR-23-PEIA-0008), and HPC resources of IDRIS made by GENCI. The work of MA was partly supported by the European Research Council (ERC project DISCOVER, number 101076028). The scenes of Earth Parser Dataset were acquired and annotated by the LiDAR-HD project [5]. We thank Z. S. Baltaci, E. Blettery, N. Dufour, A. Guedon, H. Mair Rawsthorne, T. Monnier, D. Robert, M. Petrovich and Y. Siglidis for inspiring discussions and feedback.

CVPR SEATTLE, WA JUNE 17-21, 2024