UNIVERSITY OF COPENHAGEN

novo nordisk foundation | PIONEER CENTRE FOR AI | Danmarks Grundforskningsfond Danish National Research Foundation

Beyond Building Footprints: Probing DINOv3 to Map Roof Material and Geometry

Venkanna Babu Guthula[1,*], Jakob Johannes Ålbæk Kehlet[1], Ankit Kariryaa[1], Nico Lang [1], Stefan Oehmcke[2], Christian Igel[1]

[1]University of Copenhagen, Denmark [2]University of Rostock, Germany

*vegu@di.ku.dk

Motivation

- Meta released several DINOv3 [1] variants trained on web and satellite data
- It remains unclear which model is best suited for a given downstream task
- This work evaluates whether the latest DINOv3 variants can capture fine-grained details of buildings from satellite imagery

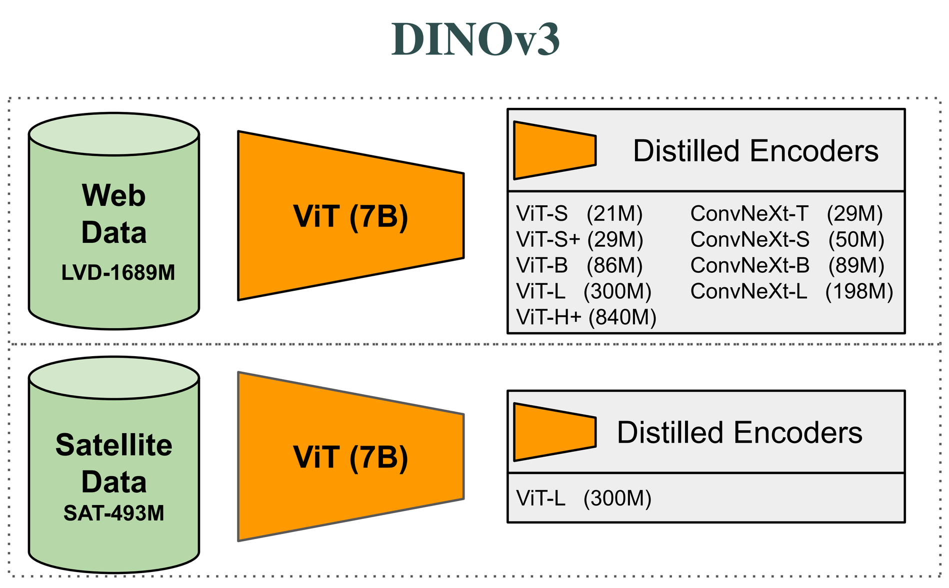

Figure 1: List of datasets (left side) used for training a very large ViT with 7 billion parameters (middle) and list of distilled encoders (right side).

Data

- Two tasks: Classifying roof material and geometry of individual buildings

Table 1: Roof material dataset

Material	Train Count	Test Count
Roof tiles	11,158	1,817
Tar paper	4,784	823
Metal	117	32
Concrete	28	2
Glass	17	6
Gravel	58	10

Table 2: Roof geometry dataset

Material	Train Count	Test Count
Gabled	22,664	3,616
Flat	22,743	3,537
Skillion	919	125
Hipped	747	111
Gambrel	120	40
Half-hipped	238	40
Pyramidal	150	19
Mansard	36	20

Method

- Compute feature maps F_{up} for image I, where F_{up} is the original feature maps F up-sampled to image resolution
- Generate a feature vector by averaging all spatial features within a single building
- Apply nearest-neighbour classification

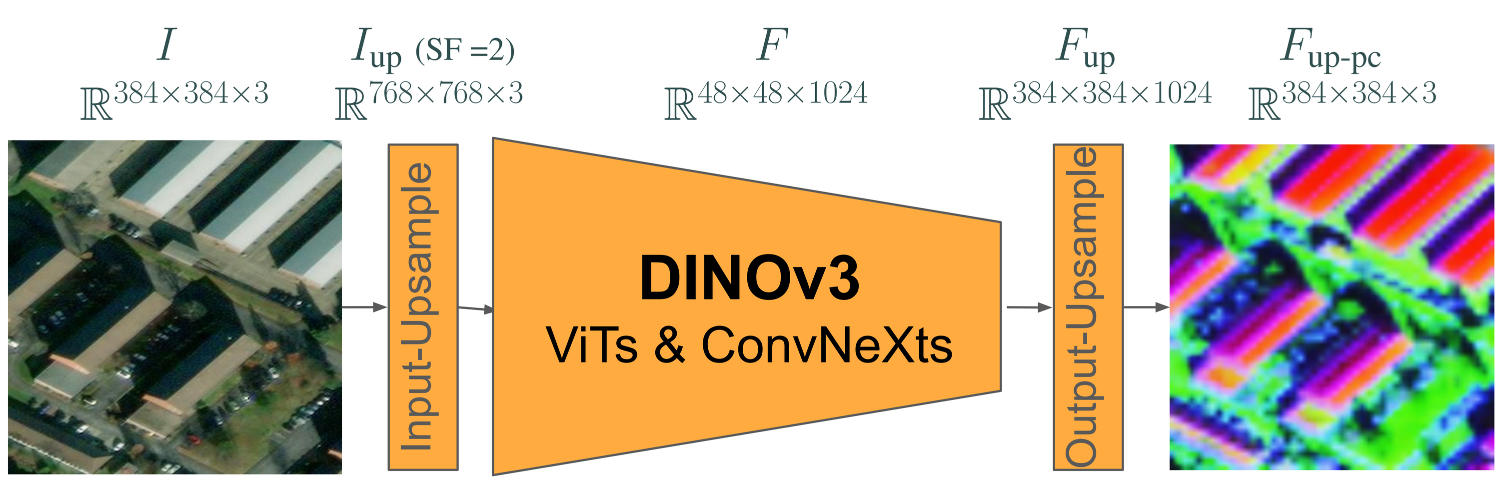

Figure 2: This figure shows the methodology for generating DINOv3 feature maps with exemplary sizes of images and feature maps.

Results

Table 3: Roof material (top) and geometry (bottom) classification results using features from DINOv3 variants. For each input upsampling factor SF, we reported the F1-Score of Micro (Mi) and Macro (Ma) averages. Macro average results are reported to observe average performance when giving equal importance to each class. SF=1 means no upsampling performed. Only the top row shows results when representations are obtained from the model (ViT-L) trained in satellite imagery.

Roof material classification

Model	SF = 1		SF = 2		SF = 4		SF = 8		SF = 16		SF = 32	
	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma
ViT-L	0.89	0.67	0.89	0.71	0.89	0.72	0.89	0.73	0.89	0.73	-	-
ViT-S	0.84	0.52	0.82	0.41	0.79	0.31	0.80	0.42	0.80	0.44	-	-
ViT-S+	0.86	0.53	0.83	0.44	0.81	0.34	0.80	0.34	0.81	0.36	-	-
ViT-B	0.85	0.49	0.82	0.37	0.80	0.31	0.79.	0.32	0.81	0.41	-	-
ViT-L	0.87	0.59	0.84	0.53	0.84	0.46	0.83	0.32	0.81	0.42	-	-
ConvNeXt-T	0.82	0.34	0.86	0.52	0.88	0.69	0.89	0.65	0.90	0.73	0.90	0.64
ConvNeXt-S	0.82	0.32	0.84	0.50	0.88	0.62	0.88	0.64	0.89	0.72	0.90	0.72
ConvNeXt-B	0.83	0.36	0.85	0.45	0.87	0.69	0.88	0.72	0.89	0.71	0.89	0.73
ConvNeXt-L	0.83	0.36	0.84	0.57	0.88	0.66	0.89	0.70	0.90	0.74	0.90	0.69

Roof geometry classification

Model	SF = 1		SF = 2		SF = 4		SF = 8		SF = 16		SF = 32	
	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma
ViT-L	0.80	0.43	0.80	0.42	0.80	0.42	0.79	0.43	0.79	0.43	-	-
ViT-S	0.73	0.31	0.70	0.23	0.67	0.19	0.67	0.22	0.68	0.24	-	-
ViT-S+	0.74	0.35	0.72	0.29	0.69	0.21	0.67	0.21	0.68	0.24	-	-
ViT-B	0.74	0.28	0.71	0.22	0.67	0.19	0.67	0.18	0.68	0.22	-	-
ViT-L	0.80	0.36	0.77	0.29	0.75	0.27	0.67	0.19	0.69	0.23	-	-
ConvNeXt-T	0.72	0.27	0.75	0.33	0.80	0.44	0.81	0.46	0.82	0.46	0.80	0.43
ConvNeXt-S	0.70	0.24	0.73	0.28	0.79	0.41	0.80	0.43	0.81	0.45	0.80	0.44
ConvNeXt-B	0.72	0.23	0.73	0.29	0.78	0.40	0.80	0.46	0.81	0.46	0.80	0.43
ConvNeXt-L	0.74	0.26	0.75	0.28	0.78	0.40	0.80	0.45	0.81	0.44	0.80	0.44

Analysis

- As the scaling factor in *Input-Upsampling* increases, the performance of all ConvNeXts significantly improves compared to ViTs (See Fig. 3)
- The results from ViT-Large (satellite) are consistent when increasing SF
- The best performing ConvNeXt-Tiny (web) matches the performance of the ViT-Large (satellite)

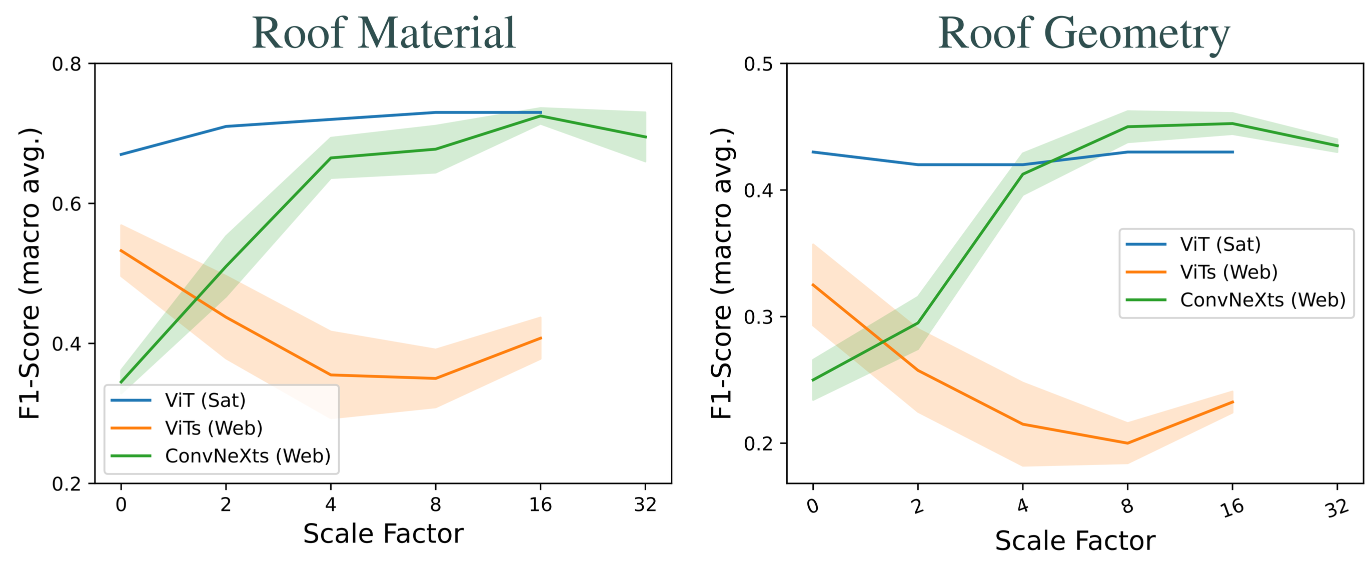

Figure 3: Accuracies of all models at different upsampling factors SF. ViT (sat) is a single model result, while ViTs (web) and ConvNeXts (web) are combined results of all ViTs and ConvNeXts. Combined results presented by mean and standard deviation

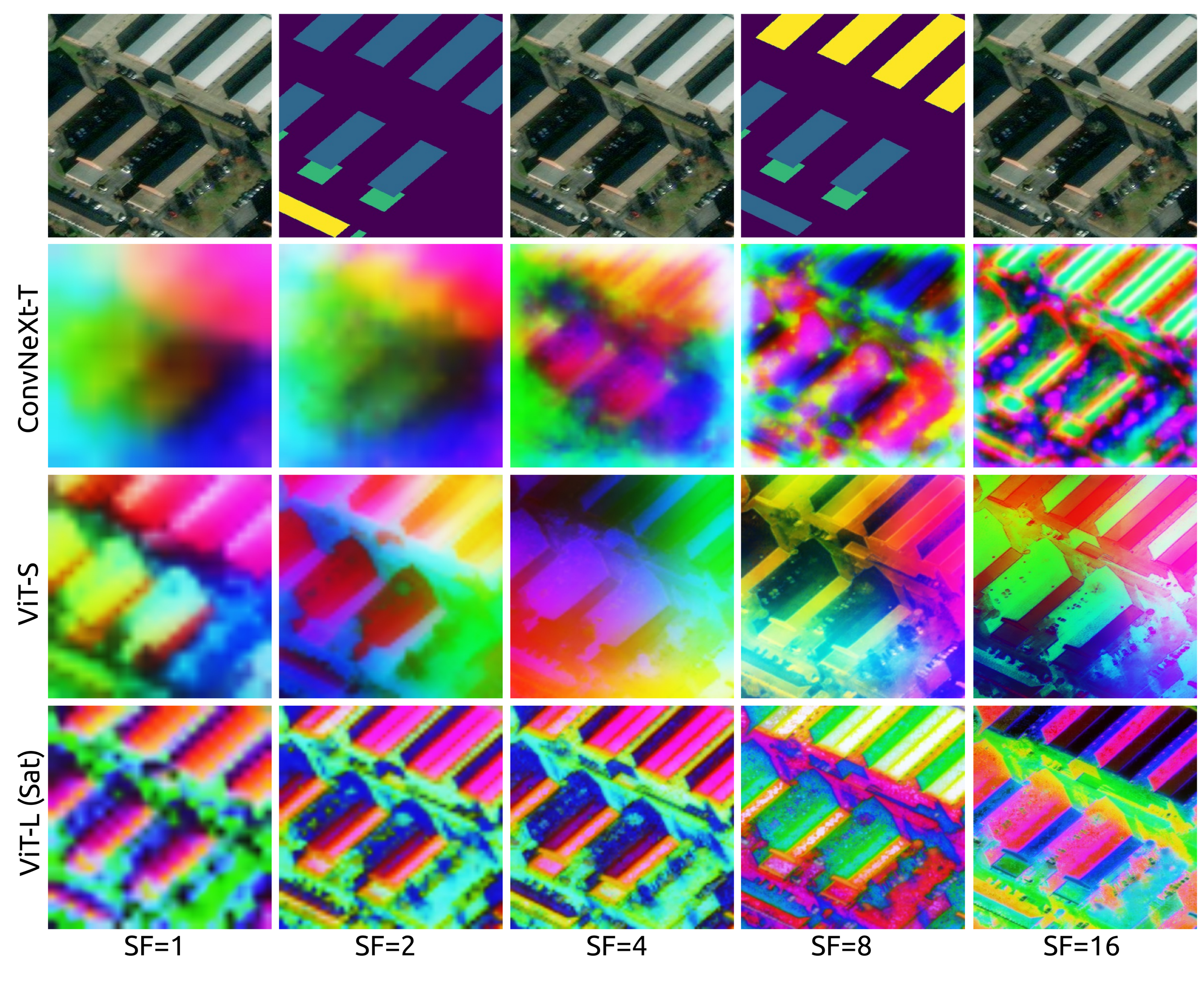

Figure 4: The top row shows the RGB image alongside the corresponding roof geometry (second column) and roof material (fourth column) reference labels. The geometry classes are color-coded, blue: gabled, green: flat, and yellow: skillion. The material classes are blue: roof tiles, green: tar paper, and yellow: metal. For visualization, the first three principal components are scaled by a factor of two and then passed through a sigmoid function.

Conclussion

- Using the smallest distilled model (ConvNeXt-Tiny) is good enough for our tasks
- ConvNeXt-Tiny (web) with increasing SF is competitive with ViT-Large (satellite)
- It would be valuable to further investigate CNNs trained on satellite imagery

Acknowledgements

This work is part of the project Risk-assessment of Vectorborne Diseases in African Cities Based on Deep Learning and Remote Sensing funded by the Novo Nordisk Foundation (grant number NNF21OC0069116). CI and AK acknowledge additional support from the Danish National Research Foundation (DNRF) through TreeSense, the Center for Remote Sensing and Deep Learning of Global Tree Resources (DNRF192). NL and CI acknowledge additional support from the Pioneer Centre for AI (P1) and by the Novo Nordisk Foundation through the Global Wetland Center (grant number NNF23OC0081089).

[1] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Theo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.