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

**Venkanna Babu Guthula*** University of Copenhagen
**Jakob Johannes Ålbæk Kehlet** University of Copenhagen
**Ankit Kariryaa** University of Copenhagen
**Nico Lang** University of Copenhagen
**Stefan Oehmcke** University of Rostock
**Christian Igel*** University of Copenhagen

## Abstract

Spatially dense image features are needed for most remote sensing tasks. The current generation of computer vision foundation models has improved the extraction of pixel-level features. In particular, the training of DINOv3 introduced several strategies to generate more robust dense features, which can enable the extraction of smaller object-level details. This study evaluates whether the latest DINOv3 variants, trained on both web and satellite imagery, are capable of capturing fine-grained details of smaller objects from remote sensing imagery. We focus on two downstream applications: roof material and roof geometry classification from satellite images. In this context, we analyze the effect of architectural differences (CNNs vs. ViTs) and the influence of pretraining data (web vs. satellite images) on performance. Although the feature maps from these models are spatially coarse, the models can extract information to accurately classify roof characteristics. Simple bilinear upsampling of the input image leads to consistent improvements for the ConvNeXt models, but not for ViTs. Finally, ConvNeXt-Tiny with upsampling input images resulted in good performance on our tasks, matching ViT-Large trained on satellite images.

## 1 Introduction

Recent foundation models, such as DINO [2], DINOv2 [12], and DINOv3 [15], have been making a great impact in different applications by using them as a basic building block, often replacing traditional backbone networks such as trainable CNN encoders. thanks to all the research efforts on self-supervised learning (SSL) and almost unlimited training data. While DINO and DINOv2 trained only Vision Transformers (ViTs) architectures [5], DINOv3 brings back convolutional neural networks (CNNs) to the foundation models paradigm by training the ConvNeXt [10] models along with the ViTs. DINOv3 weights consist of one huge ViT with 7B parameters and several distilled models of ConvNeXts and ViTs. While most of the released weights trained on web images collected from public posts on Instagram, a couple of weights released were trained on satellite images with the resolution of 0.6 meters. The public posts data (web dataset) LVD-1689M consist of 1689M images and and the satellite data SAT-493M consists of 493M images.

DINOv3 introduced several post-hoc strategies for extracting dense features and improve the model's flexibility w.r.t. model size and image resolution. The strategies include Gram anchoring, Multi-Crop startergy [1] and high-resolution adoption step [17]. All DINOv3 ViTs are trained with the patch size of $16 \times 16$ pixels, so all ViTs return a single feature vector for this specific size of patch. All ConvNeXts of DINOv3 generate the feature vector for every $32 \times 32$ pixels on the input image. The

*Corresponding authors: vegu@di.ku.dk, igel@di.ku.dk

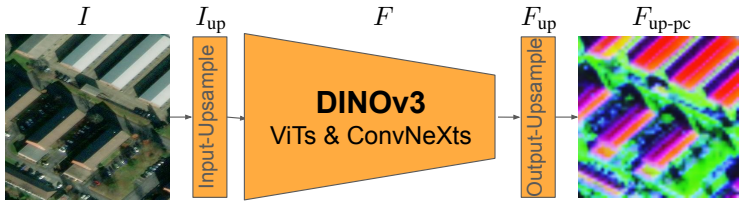


Figure 1: Methodology for generating DINOv3 feature maps. The right-side image was generated with ViT-Large trained on satellite images with SF of *Input-Upsampling* is 2. Where $I = \mathbb{R}^{384\times384\times3}$ is input image, $I_{\text{up}} = \mathbb{R}^{768\times768\times3}$ is image after *Input-Upsampling* with SF=2, $F = \mathbb{R}^{48\times48\times1024}$ is feature maps from ViT-L, $F_{\text{up}} = \mathbb{R}^{348\times348\times1024}$ is feature maps after *Output-Upsampling* to image resolution, and $F_{\text{up-pc}} = \mathbb{R}^{348\times348\times3}$ is projected first 3 principal components. The spatial dimensions and depths of $F$ differs based on the model.

main advantage of using CNNs is that we can extract feature maps at different resolutions, while the ViTs' embeddings have the same spatial dimensions after every transformer layer.

There are several ways to extract representations from ViTs, such as using the class embedding that summarizes the entire image or the patch embeddings that capture local information. Representations from ViTs and ConvNeXts can be used for object classification, for example for characterizing buildings from remote sensing imagery, enabling the classification of attributes like roof geometry and roof material. Roof classification is valuable for wide range of applications, including disease and disaster risk assessment, thermal efficiency analysis, roof durability evaluation, heritage conservation, and solar panel installation planning [6, 3, 13, 18, 14, 9]. The classification step typically requires segmented building footprints. These can often be obtained from large-scale building datasets provided by, for instance, Google[16], OpenStreetMap [11], and Microsoft [4].

Here we consider classifying roof material and geometry from satellite imagery. In this setting, we analyze the differences between representations obtained from CNNs and ViTs and the impact of using pretrained models trained on web images versus those trained on satellite imagery. Most importantly, we show how rescaling the input influences the ability to create spatially dense features.

## 2 Data

We collected two recently released datasets from Bonn, Germany: roof material [8] and roof geometry [7]. Both datasets include 50 cm high-resolution satellite imagery from OpenAerialMap. Roof material, roof geometry labels, and building footprints were sourced from OpenStreetMap. Each dataset was originally provided with predefined training, validation, and test splits. The image size is $384 \times 384$ pixels. The **Roof Material** dataset contains six roof material classes and exhibits a strong class imbalance. Because two classes in the test set contain only a single building, the original test set was merged with the training set, and the validation split was used as the new test set. The number of samples in the training and final test sets for each class are: roof tiles (11,158 / 1,817), tar paper (4,784 / 823), metal (117 / 32), concrete (28 / 2), glass (17 / 6), and gravel (58 / 10). In total, the dataset includes 16,162 buildings for training and 2,690 for testing. The **Roof Geometry** includes seven roof geometry classes and also shows class imbalance. Similar to the roof material dataset, the original training and test sets were merged, and the validation set was used as the final test set. The number of samples in the training and final test sets for each class are: 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), and mansard (36 / 20). The total number of buildings with roof geometry annotations is 47,617 for training and 7,508 for testing.

## 3 Methods

We evaluated four ConvNeXts-{Tiny, Small, Base, Large} and four ViTs-{Small, Small+, Base, Large} that were pretrained on web images dataset along with one ViT-Large model pretrained on satellite imagery [15] (the only distilled model trained on satellite data). Larger variants were excluded due to their high computational requirements. To generate feature vector for each building polygon, we first computed the feature map for the entire image. These features were then bilinearly

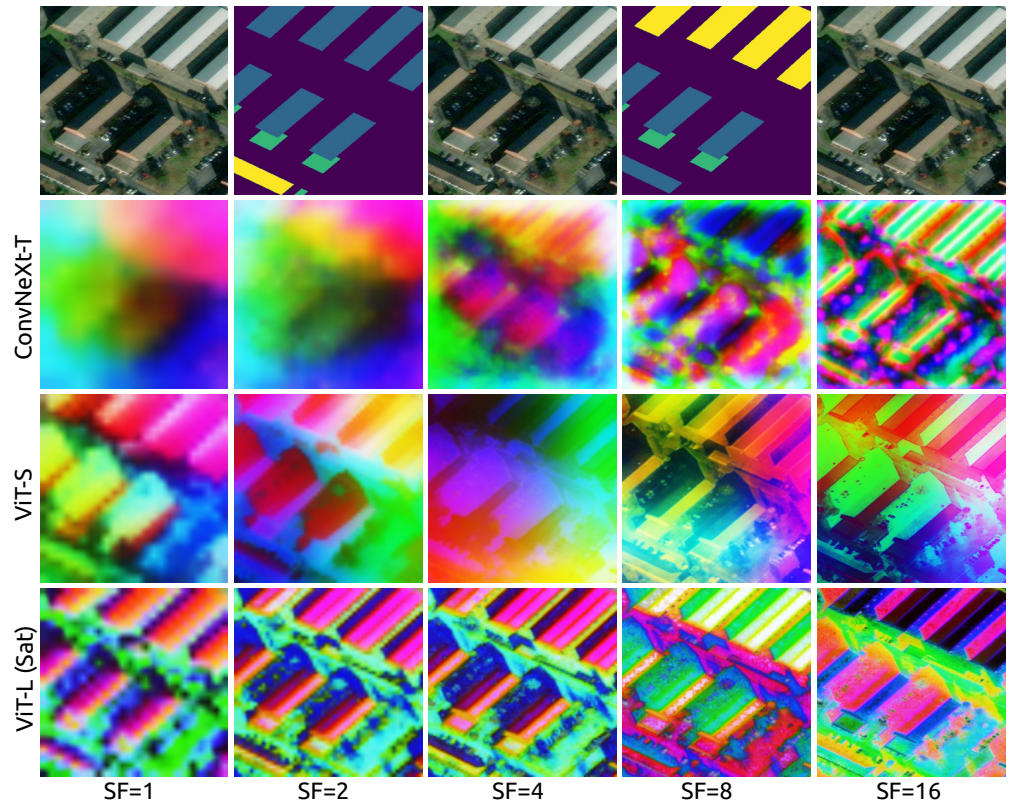


Figure 2: The top row shows the RGB image alongside the corresponding roof geometry (second column) and roof material (fourth column) reference labels. The geometry classes are color-coded, blue: gabled, green: flat, and yellow: skillion. The material classes are blue: roof tiles, green: tar paper, and yellow: metal. The visualization uses the method described in [15]: the first three principal components are scaled by a factor of two and then passed through a sigmoid function.

upsampled to match the original image resolution (see Figure 1), and features outside each building polygon were masked out. For ViTs, the embeddings were reshaped into spatial feature maps prior to upsampling. Finally, the features within each building footprint were averaged over all pixels to obtain a single representative feature vector per building.

During DINOv3 pretraining, input images were upsampled by a scale factor (SF) of 2 to enable flexibility across different image resolutions. To analyze this effect, we experimented with different SFs during feature extraction (see Table 1). We added two steps of input and output upsampling to the feature extraction (see Fig. 1). *Input-Upsampling* with some SF increase the input image resolution, and *Output-Upsampling* increase the resolution of feature maps from DINOv3. The maximum SF for *Input-Upsampling* is chosen to generate feature maps that matches the input image resolution. So, the maximum SFs for ViTs and CNNs are 16 and 32, respectively. This is the reason we only run *Output-Upsampling* (not downsampling) and why there are no results for SF=32 for ViTs in both tables. The final feature maps after upsampling are used to generate a feature vector for each individual building. The F-score is used as an evaluation metric and micro and macro averages of all experiments can be found in Table 1. After extracting the DINOv3 features, we employed a simple $K$-nearest neighbour (KNN) classifier as a baseline for training and validated it on the final test set.

## 4 Results and Discussion

The results from both tasks show that ViTs are performing better when no *Input-Upsampling* performed. As SF in *Input-Upsampling* increases, the performnce of CNNs significantly improved while ViTs performnace is reduced. This can be due to CNNs downscale feature maps iteratively (ConvNeXt with a stride of 4) while ViTs only do it once (patchifying) and fine information can get

Table 1: 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 |

lost easily. Usually in CNNs, we increase the number of filters whenever we downscale the image to reduce information loss. If the downscaling is too strong, we will lose details. With increasing mage size through *Input-Upsampling*, we circumvent that problem. In the end, upscaling does not create more information, it just reduces the negative impact of architectural design choices. The Fig. 2 shows how quality of feature maps consistently increasing with SF in-case of ConvNeXt, while feature maps of ViTs looks similar with increasing SF. One interesting observation is that ViTs are trying to reconstruct the image as the SF increases (see last two columns of ViT-S in Fig. 2).

Interestingly, the best performing ConvNeXt-Tiny trained on web images matches the performance of the ViT-Large trained on satellite imagery. An interesting observation is that the best performance of ConvNeXt-Tiny trained on web images matches that of ViT-Large trained on satellite imagery. The difference between parameters of these models is huge, ConvNeXt-Tiny and ViT-Large have 29M and 300M parameters, respectively. It would be valuable to further investigate distilled CNNs trained on satellite imagery.

## 5 Conclusion

There is continuous improvement as SF increases when using CNNs. Increasing the resolution by simple upsampling can improve the performance in related downstream tasks. There is not much difference in using different sizes of the distilled models, so using the smallest model is good enough. ConvNeXt-Tiny trained on web images is competative with ViT-Large trained on satellite images, it can be helpful for the remote sensing community if Meta release smaller distilled CNN weights trained on satellite data. Simple upsampling with tiny models can perform better and is more efficient.

## Acknowledgments and Disclosure of Funding

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).

## References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:9912–9924, 2020.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[3] Cecilia N Clark and Fabio Pacifici. A solar panel dataset of very high resolution satellite imagery to support the sustainable development goals. *Scientific Data*, 10(1):636, 2023.

[4] Microsoft Corporation. Microsoft global ML building footprints. `https://planetarycomputer.microsoft.com/dataset/ms-buildings`, 2023. Accessed: 2025-11-15.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[6] Venkanna Babu Guthula, Stefan Oehmcke, Remigio Chilaule, Hui Zhang, Nico Lang, Ankit Kariryaa, Johan Mottelson, and Christian Igel. Drone imagery for roof detection, classification, and segmentation to support mosquito-borne disease risk assessment: The Nacala-roof-material dataset. *Science of Remote Sensing*, 2025.

[7] Julian Huang, Yue Lin, and Alex Nhancololo. Bonn Roof Geometry Dataset. `https://figshare.com/articles/dataset/Bonn_Roof_Geometry_Dataset/28823390`, 2025.

[8] Julian Huang, Yue Lin, and Alex Nhancololo. Bonn Roof Material + Satellite Imagery Dataset. `https://figshare.com/articles/dataset/Bonn_Roof_Material_Aerial_Imagery_Dataset/28713194`, 2025.

[9] Norman Kerle. Disasters: Risk assessment, management, and post-disaster studies using remote sensing. In *Remote Sensing Handbook, Volume VI*, pages 153–198. CRC Press, 2024.

[10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.

[11] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org. `https://www.openstreetmap.org`, 2025.

[12] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

[13] Colin Richardson, Anna T Sullivan, Florence B Berube, Shabnam Jabari, and Travis Moore. Towards urban heat loss modeling using building digital twin. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:35–42, 2025.

[14] Ignacio Rodríguez-Antuñano, Joaquim J Sousa, Matúš Bakoň, Antonio M Ruiz-Armenteros, Joaquín Martínez-Sánchez, and Belen Riveiro. Empowering intermediate cities: cost-effective heritage preservation through satellite remote sensing and deep learning. *International Journal of Remote Sensing*, 45(12):4046–4074, 2024.

[15] 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.

[16] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283*, 2021.

[17] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[18] Zhenyue Xing, Yumeng Ma, Caifeng Wang, Wen Zheng, Baibo Zhang, and Hailiang Wang. Roof damage detection and evaluation using aerial image based on improved DeepLabv3+. *Nondestructive Testing and Evaluation*, pages 1–18, 2025.