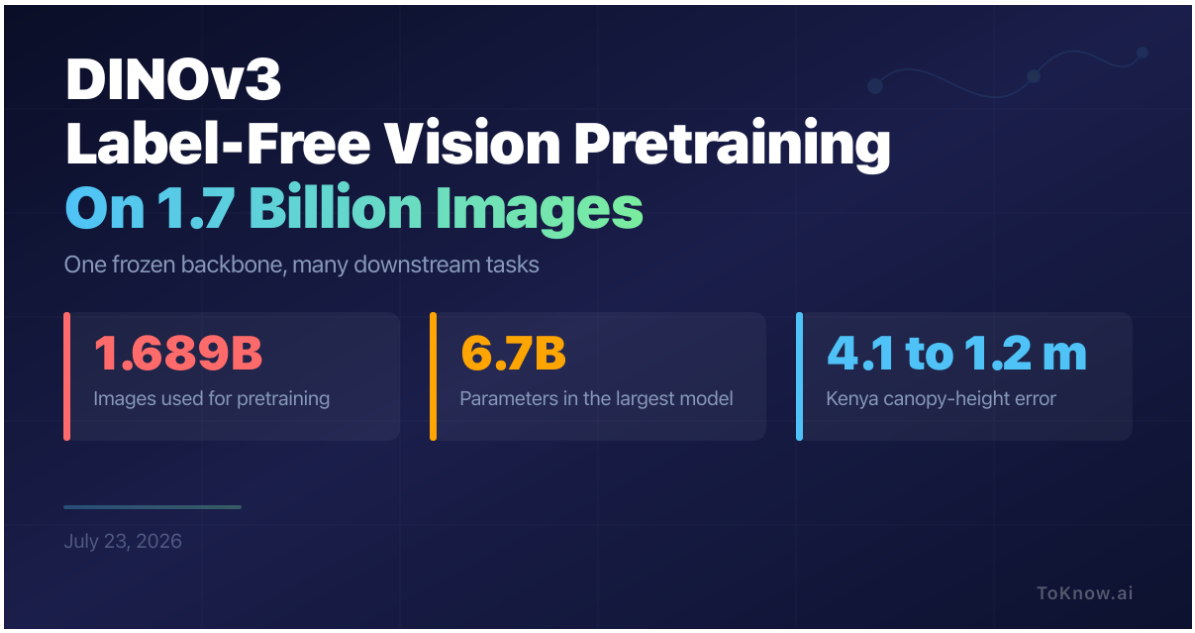
DINOv3: Label-Free Vision Pretraining on 1.7 Billion Images

Kabui, Charles

2026-07-23

Read at ToKnow.ai



Meta released DINOv3 in August 2025. Its largest model is a 6.7-billion-parameter vision transformer pretrained on 1.689 billion curated images selected from about 17 billion public Instagram images. The backbone learned without labels or captions through self-supervised learning, where the training signal comes from the images themselves. A method called Gram anchoring stops the model's fine-grained image features from degrading during long training

runs. With the 7B backbone frozen and task-specific heads trained on top, Meta reports 66.1 mAP on COCO object detection and 63.0 mIoU on ADE20K segmentation. Those scores do not mean downstream labels disappeared. They show that one fixed image encoder can support several labelled tasks without retraining its billions of parameters.

That reuse matters when labels are expensive. In a World Resources Institute project, a satellite-trained DINOv3 model cut average canopy-height error in one Kenyan region from 4.1 metres with DINOv2 to 1.2 metres. The application still trained a task decoder on measured canopy data. Meta released 12 backbones, including distilled models as small as 21 million parameters, but the original 7B training used 61,440 H100 GPU hours. Its 1.689-billion-image corpus is not part of the release, and checkpoint access requires sharing contact details and accepting the custom DINOv3 licence. Most teams can build with the smaller released models, not reproduce the full pretraining run.

Sources:

- DINOv3 paper
- Meta DINOv3 announcement
- DINOv3 model card
- DINOv3 7B checkpoint and access terms
- World Resources Institute application
