Meta

DINO(self-DIstillation with NO labels) v3

(2025)

Paper Review

2025.09.17

Changseon Yu

CAU 중앙대학교

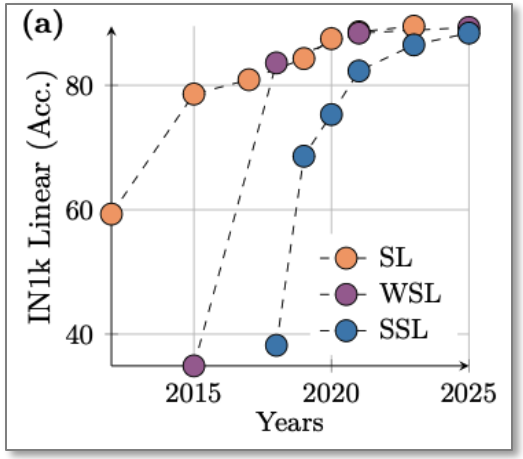
Contents

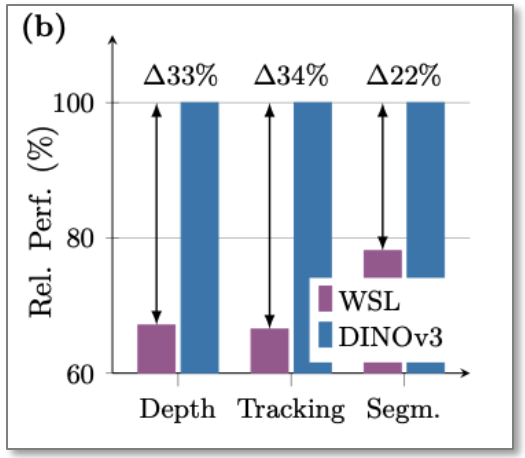
Introduction

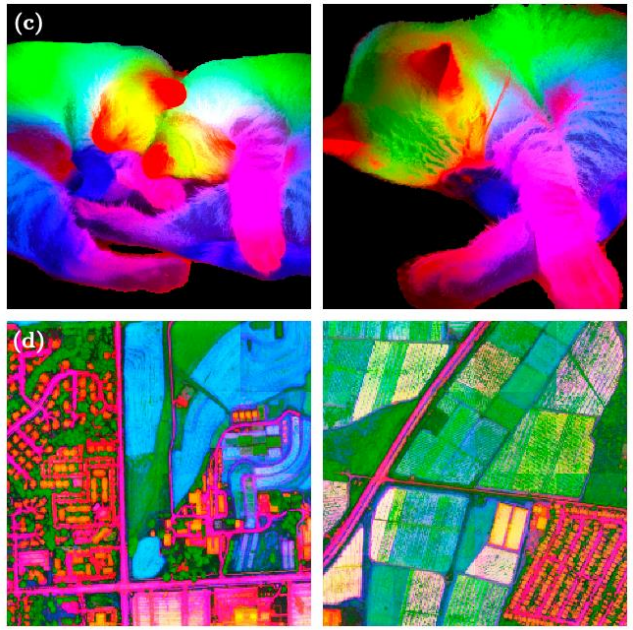
Background(1)

❖ Vision Foundation Model (VFM)

- Foundation models have become a central building block in modern computer vision
- Self-supervised learning (SSL) is a powerful approach (by learning directly from raw pixel data)
- SSL unlocks training on massive, raw image collections (effective for training large-scale visual encoders)







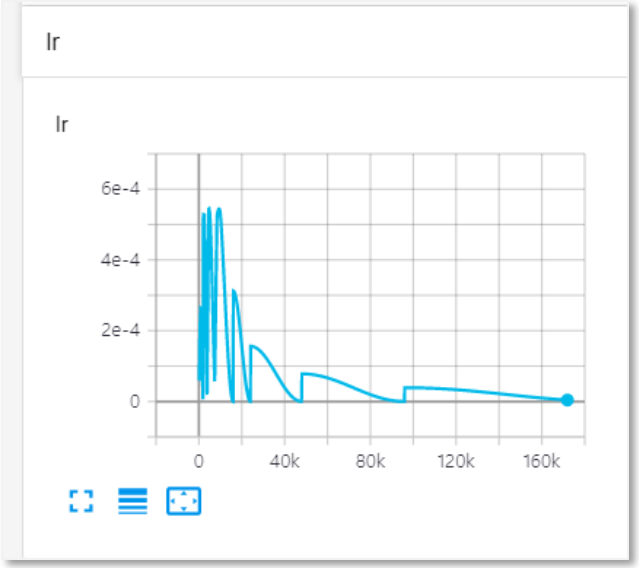
PCA map

Introduction

Background(2)

❖ Challenges of VFM

- Unclear how to collect useful data from unlabeled collections
- Employing cosine schedules implies knowing the optimization horizon a priori, which is difficult when training on large image corpora
- Performance of the features gradually decreases after early training (ViT-Large size (300M parameters), reducing the usefulness of scaling DINOv2)



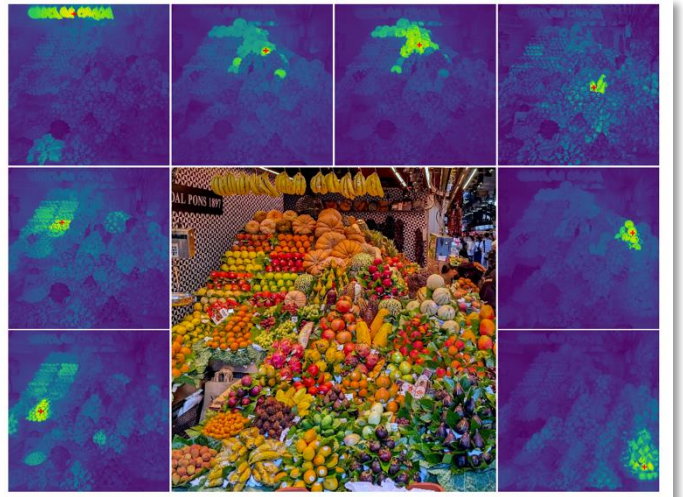


Figure 3: High-resolution dense features. We visualize the cosine similarity maps obtained with DINOv3 output features between the patches marked with a red cross and all other patches. Input image at 4096×4096. *Please zoom in*, do you agree with DINOv3?

Purpose

❖ **SSL backbone can serve as a universal visual encoder**

- Training a foundational model versatile across tasks and domains
- Improving the shortcomings of existing SSL models on dense features
- Disseminating a family of models that can be used off-the-shelf

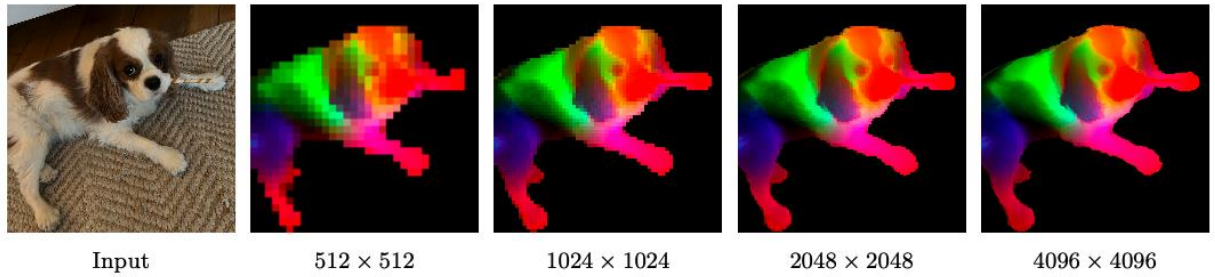


Figure 4: DINOv3 at very high resolution. We visualize dense features of DINOv3 by mapping the first three components of a PCA computed over the feature space to RGB. To focus the PCA on the subject, we mask the feature maps via background subtraction. With increasing resolution, DINOv3 produces crisp features that stay semantically meaningful. We visualize more PCAs in Sec. 6.1.1.

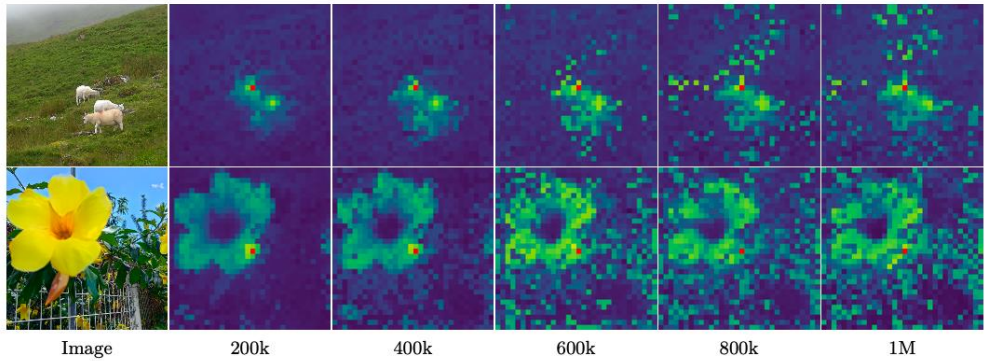
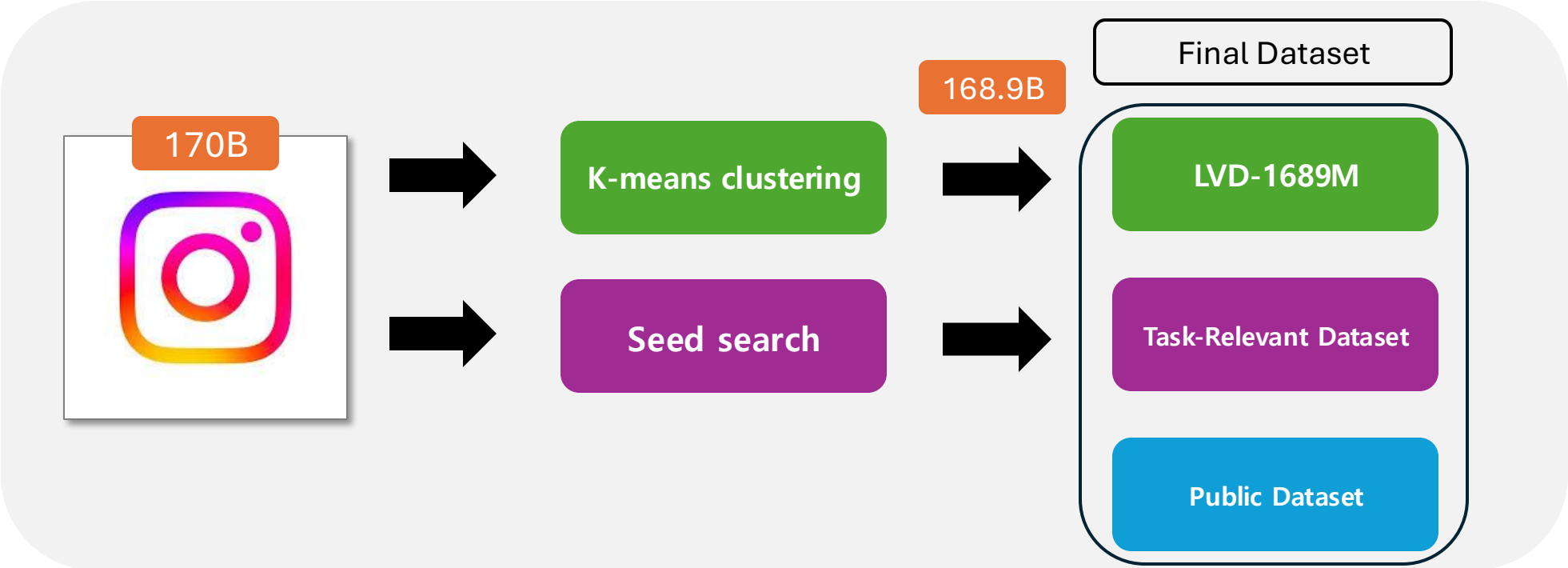


Figure 6: Evolution of the cosine similarity between the patch noted in red and all other patches. As training progresses, the features produced by the model become less localized and the similarity maps become noisier.

Method

❖ Data sampling

1. Raw Data Pool (17B Instagram images) → Automated Curation → Curated Dataset (LVD-1689M)
2. Raw Data Pool → Retrieval-based Curation → Task-Relevant Dataset
3. Public Datasets
4. **Final Training Dataset = Curated Dataset + Task-Relevant Dataset + Publicly Available Datasets**



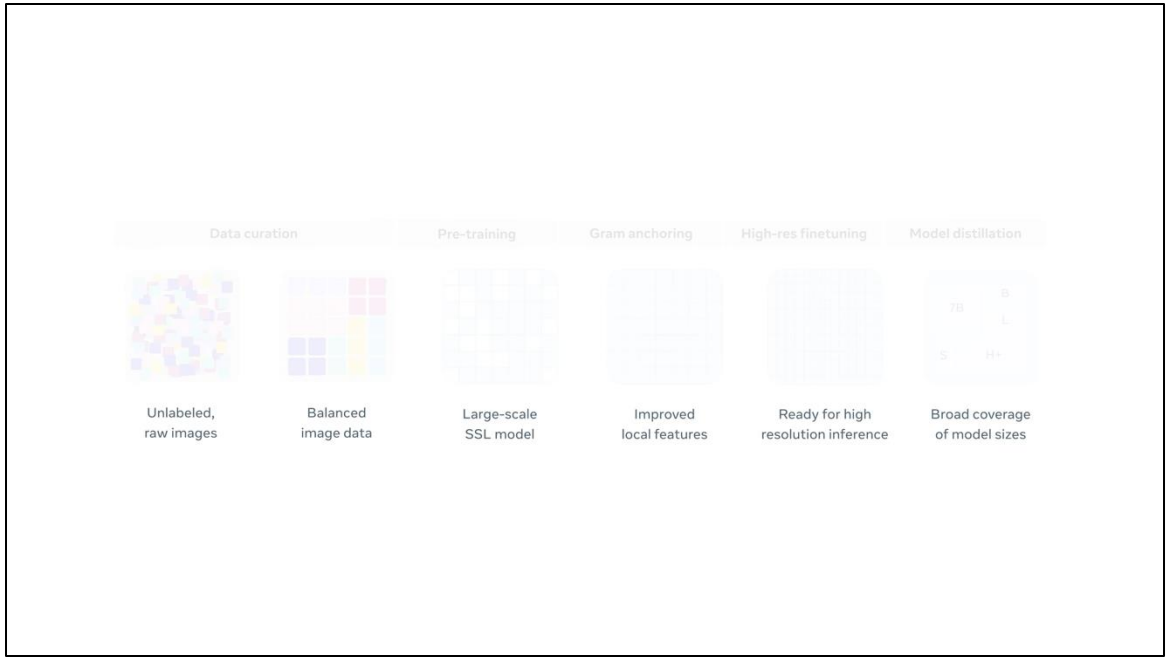
Method

❖ Comparison DINOv2 and DINOv3 models

- Keep the model 40 blocks deep
- Increase the embedding dimension to 4096(Previous : 1536)
- Expand both FFN hidden dimensions, attention heads, and dimensions
- DINOv2 used learnable position embeddings, while DINOv3 adopted RoPE

Table 2: Comparison of the teacher architectures used in DINOv2 and DINOv3 models. We keep the model 40 blocks deep, and increase the embedding dimension to 4096. Importantly, we use a patch size of 16 pixels, changing the effective sequence length for a given resolution.

Teacher model	DINOv2	DINOv3
Backbone	ViT-giant	ViT-7B
#Params	1.1B	6.7B
#Blocks	40	40
Patch Size	14	16
Pos. Embeddings	Learnable	RoPE
Registers	4	4
Embed. Dim.	1536	4096
FFN Type	SwiGLU	SwiGLU
FFN Hidden Dim.	4096	8192
Attn. Heads	24	32
Attn. Heads Dim.	64	128
DINO Head MLP	4096-4096-256	8192-8192-512
DINO Prototypes	128k	256k
iBOT Head MLP	4096-4096-256	8192-8192-384
iBOT Prototypes	128k	96k



Method

❖ **Con**

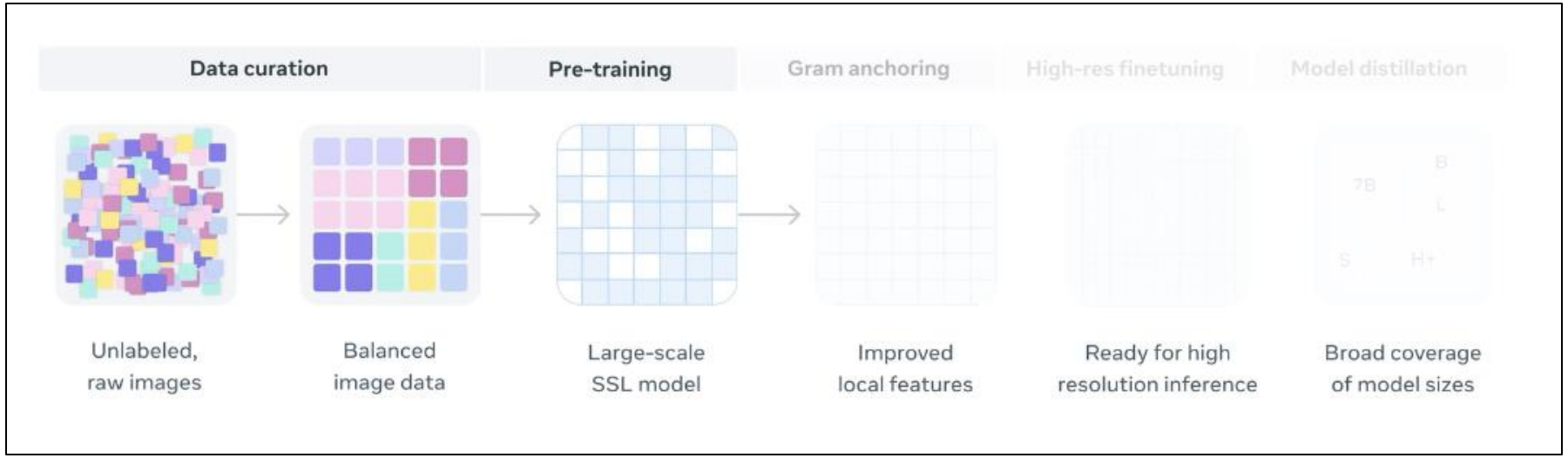
- Keep
- Increa
- Expar
- DINC

Table 2: Cor
40 blocks deep
changing the

Method

❖ Pre-training strategy

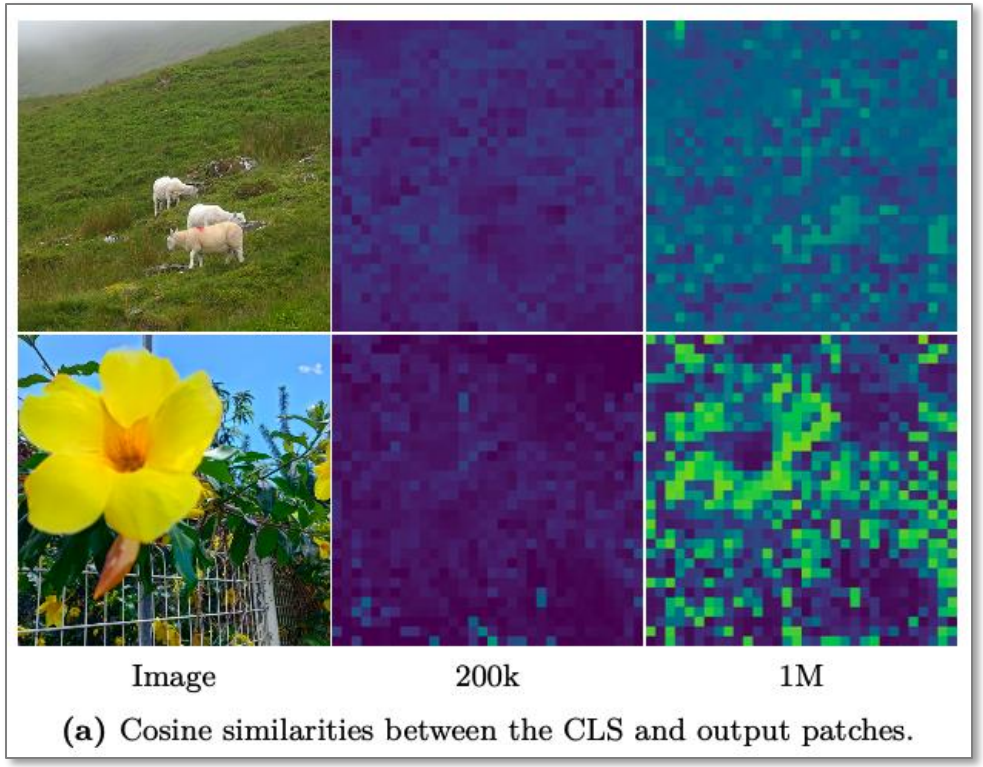
- RoPE-box jittering : improve the robustness of the model to resolutions, scales and aspect ratios
- The coordinate box [−1, 1] is randomly scaled to [−s, s], where s∈[0.5,2]
- Better learn detailed and robust visual features, improving its performance and scalability
- Get rid of all parameter scheduling, and train with constant learning rate, weight decay, and teacher EMA momentum



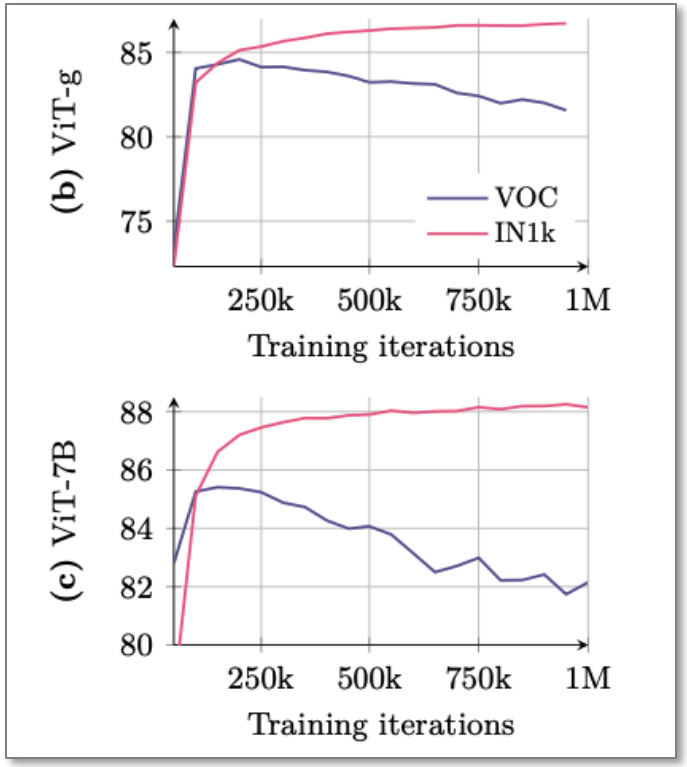
Method

❖ Gram anchoring

- To fully leverage the benefits of large-scale training, aim to train the 7B model for an extended duration
- The performance degrades on dense tasks (Due to the emergence of patch-level inconsistencies in feature representations)



(a) Cosine similarities between the CLS and output patches.



Method

❖ Gram anchoring

- Gram Anchoring is a novel regularization technique for addressing patch-level consistency degradation during training
- Derive the gram matrix of the student model to be similar to the gram matrix of the earlier version of the model with better dense features

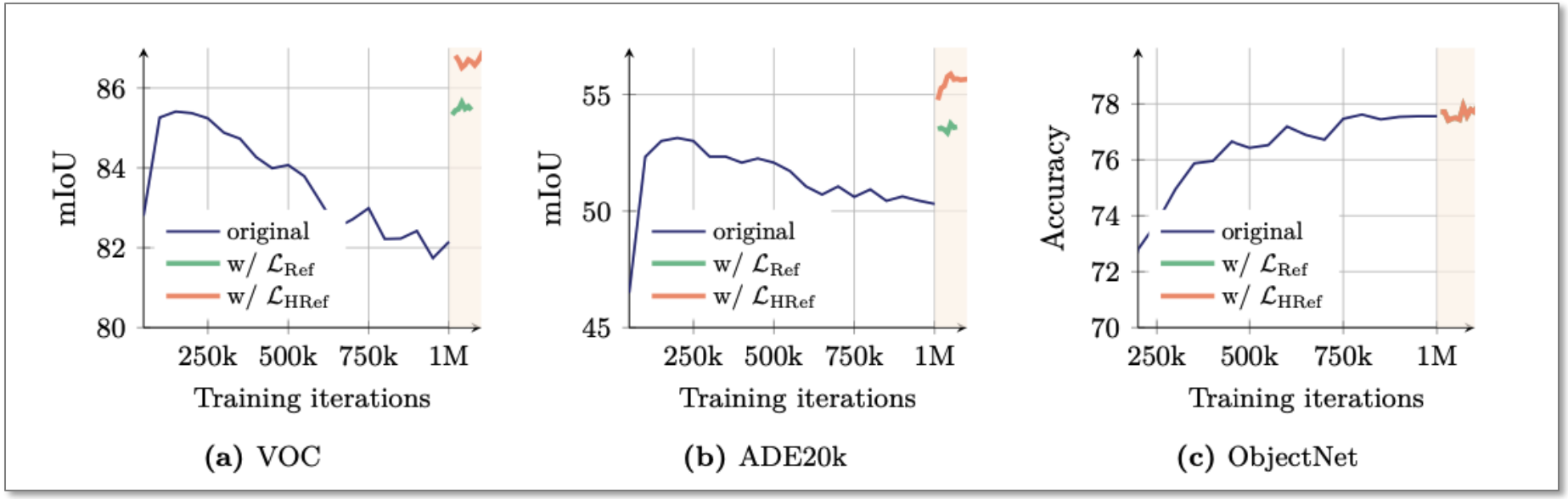
$$\mathcal{L}_{\text{Gram}} = \left\| \mathbf{X}_S \cdot \mathbf{X}_S^\top - \mathbf{X}_G \cdot \mathbf{X}_G^\top \right\|_{\text{F}}^2 .$$

Student Teacher

Method

❖ Gram anchoring

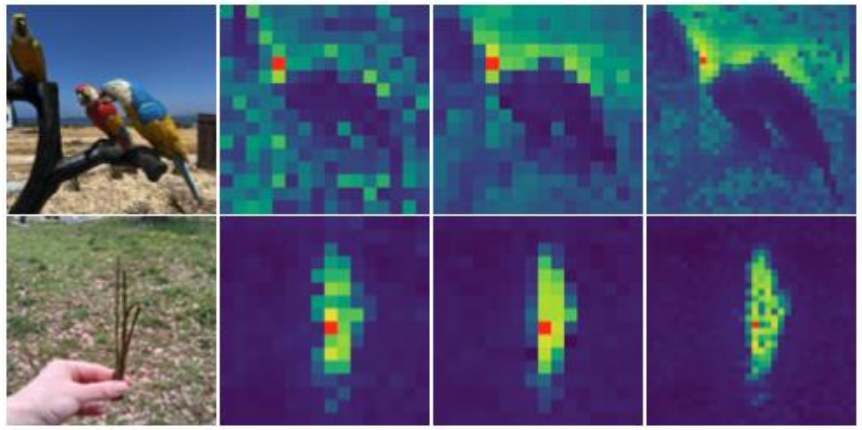
- Gram Anchoring is a novel regularization technique for addressing patch-level consistency degradation during training
- Derive the gram matrix of the student model to be similar to the gram matrix of the earlier version of the model with better dense features



Method

❖ High Resolution Finetuning

- DINOv3 leverages high-resolution features to enhance Gram anchoring
- Enter an image that is twice the normal resolution, and then use bicubic interpolation to reduce the feature map
- Maintain good patch consistency of high resolution features + obtain smooth feature maps to fit the student model's output size



(a) Gram matrices at different input resolutions.

Method	Teacher Iteration	Res.	IN1k Linear	ADE mIoU	NYU RMSE
Baseline	—	—	**88.2**	50.3	0.307
GRAM	200k	×1	88.0	53.6	0.285
GRAM	200k	×2	88.0	**55.7**	**0.281**
GRAM	100k	×2	87.9	**55.7**	0.284
GRAM	1M	×2	88.1	54.9	0.290

(b) Ablation of Gram teachers and resolutions.

Method

❖ High Resolution Finetuning

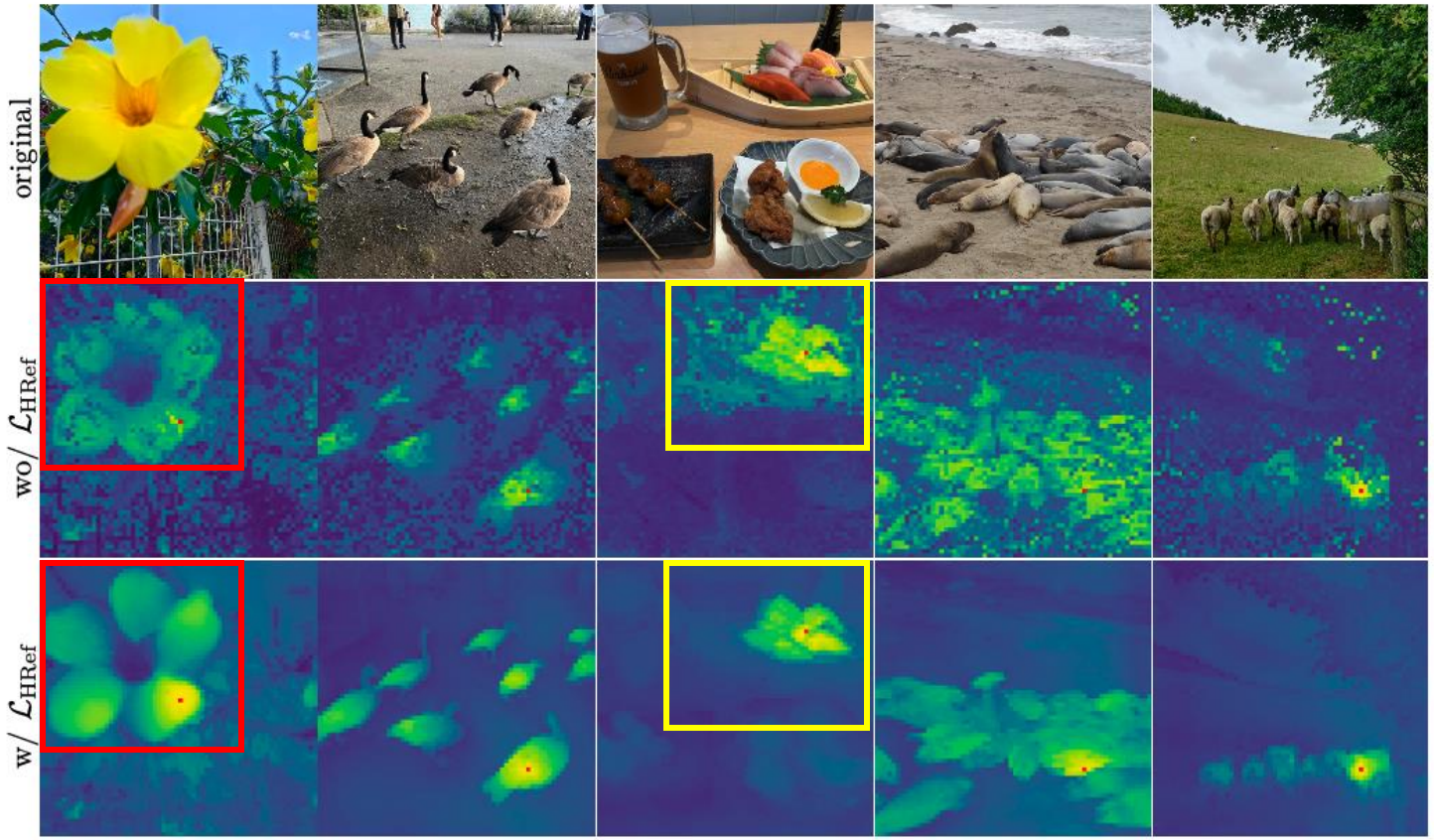
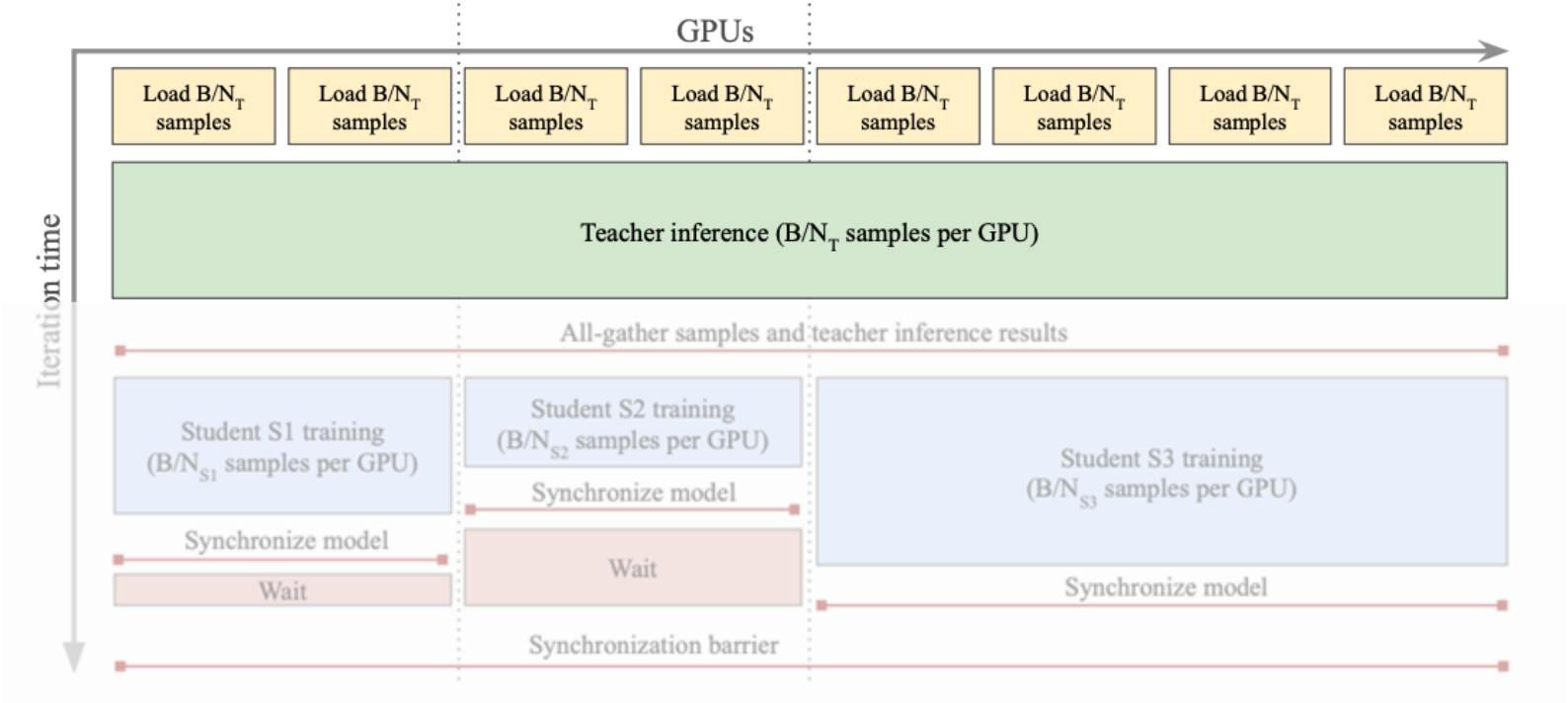


Figure 10: Qualitative effect of Gram anchoring. We visualize cosine maps before and after using the refinement objective $\mathcal{L}_{\text{HRef}}$. The input resolution of the images is 1024×1024 pixels.

Method

❖ Multi student distillation

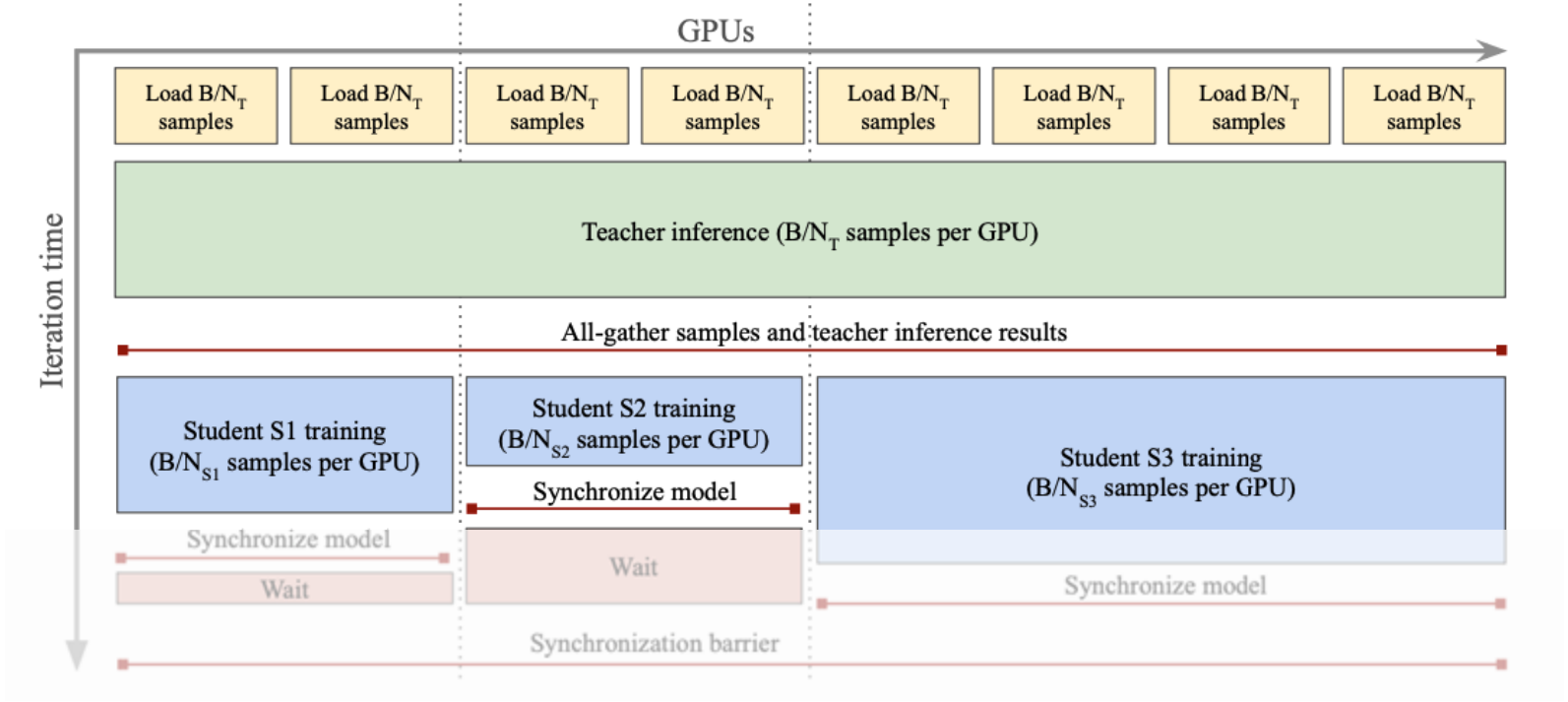
1. Teacher Inference: First, perform teacher inference using the total number of GPUs (NTs)
2. Results share: Teacher's reasoning results are shared with all computing nodes through a collective operation called All-gather
3. parallel student training: Subsequently, each student model (S1, S2, S3) independently trains using the assigned GPU group



Method

❖ Multi student distillation

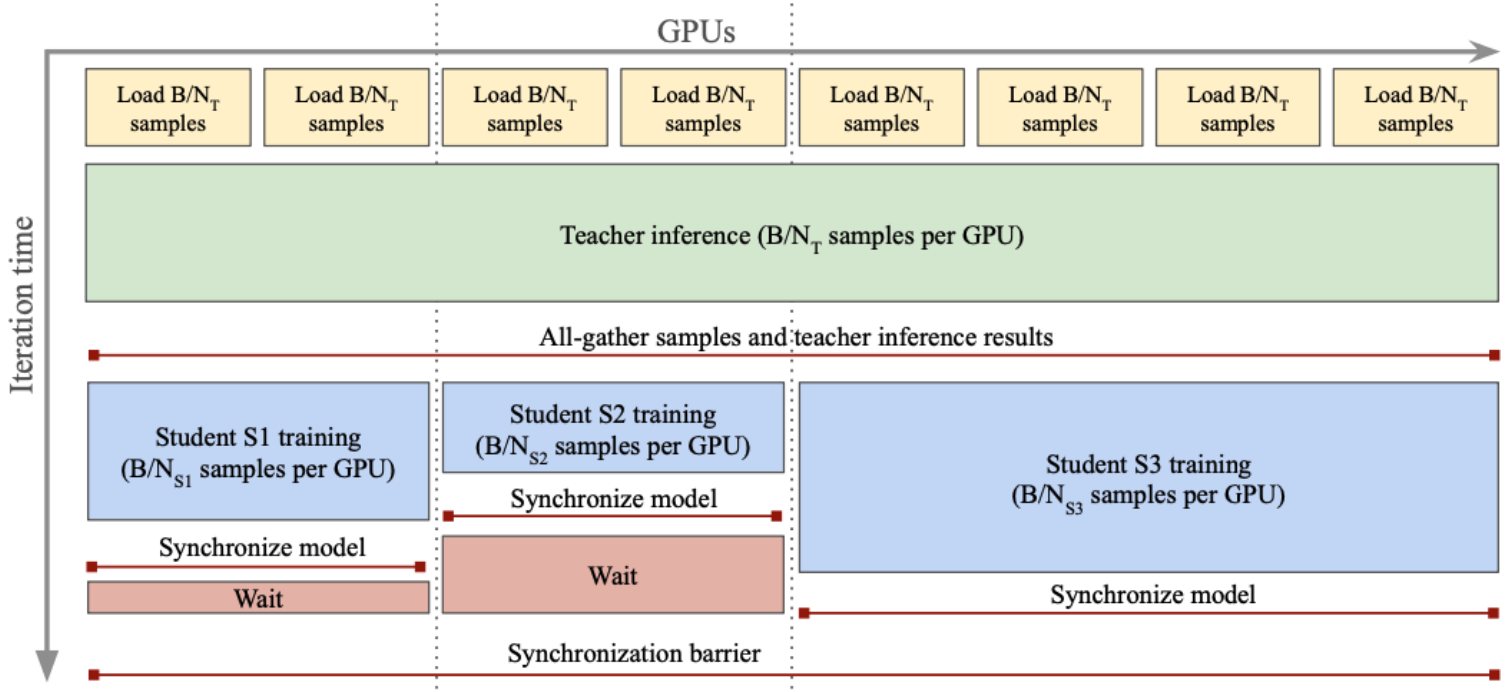
1. Teacher Inference: First, perform teacher inference using the total number of GPUs (NTs)
2. Results share: Teacher's reasoning results are shared with all computing nodes through a collective operation called All-gather
3. parallel student training: Subsequently, each student model (S1, S2, S3) independently trains using the assigned GPU group



Method

❖ Multi student distillation

1. Teacher Inference: First, perform teacher inference using the total number of GPUs (NTs)
2. Results share: Teacher's reasoning results are shared with all computing nodes through a collective operation called All-gather
3. parallel student training: Subsequently, each student model (S1, S2, S3) independently trains using the assigned GPU group



Result

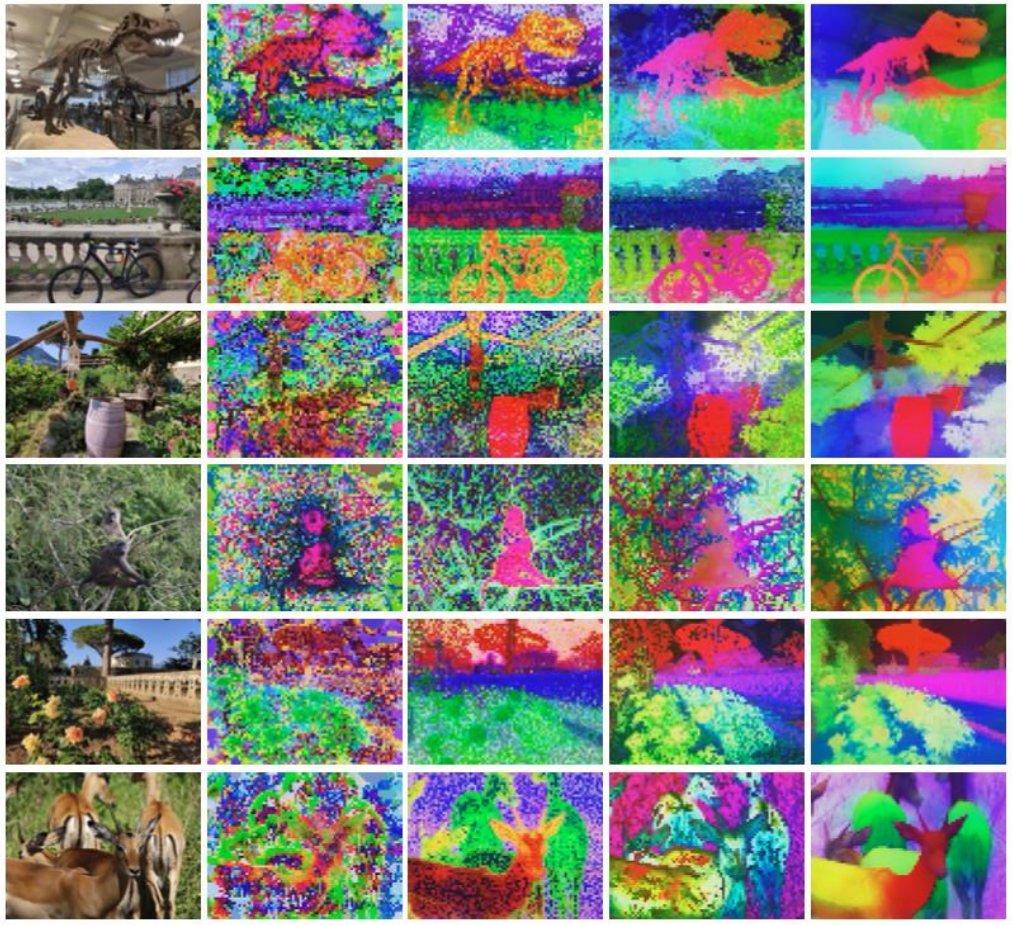
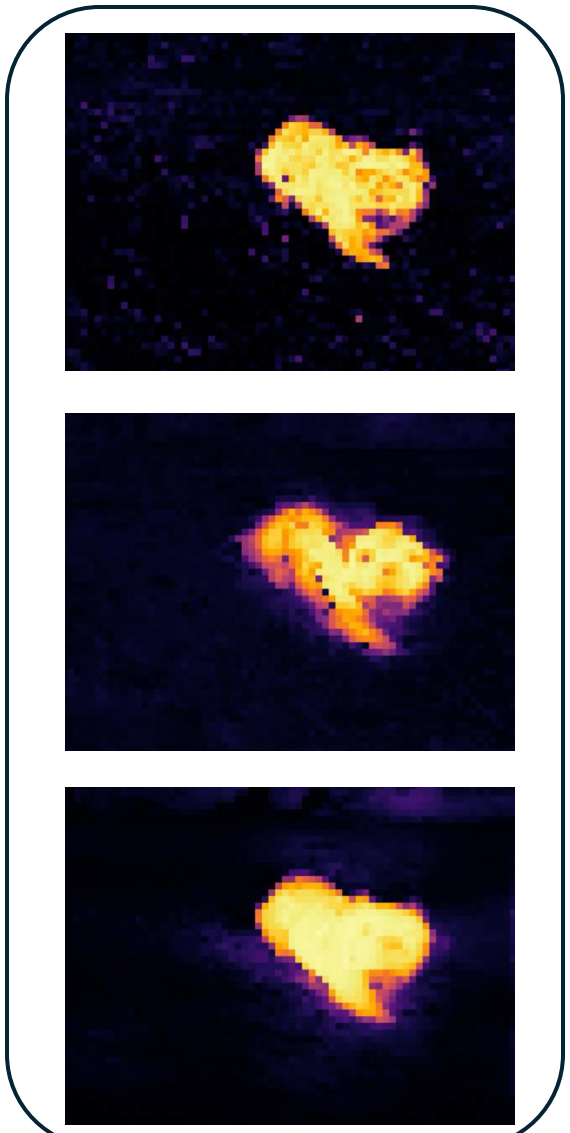


Figure 13: Comparison of dense features. We compare several vision backbones by projecting their dense outputs using PCA and mapping them to RGB. From left to right: SigLIP 2 ViT-g/16, PEspatial ViT-G/14, DINOv2 ViT-g/14 with registers, DINOv3 ViT-7B/16. Images are forwarded at resolution 1280×960 for models using patch 16 and 1120×840 for patch 14, *i.e.* all feature maps have size 80×60.



DINO
80M-parameter models trained on 1M images.

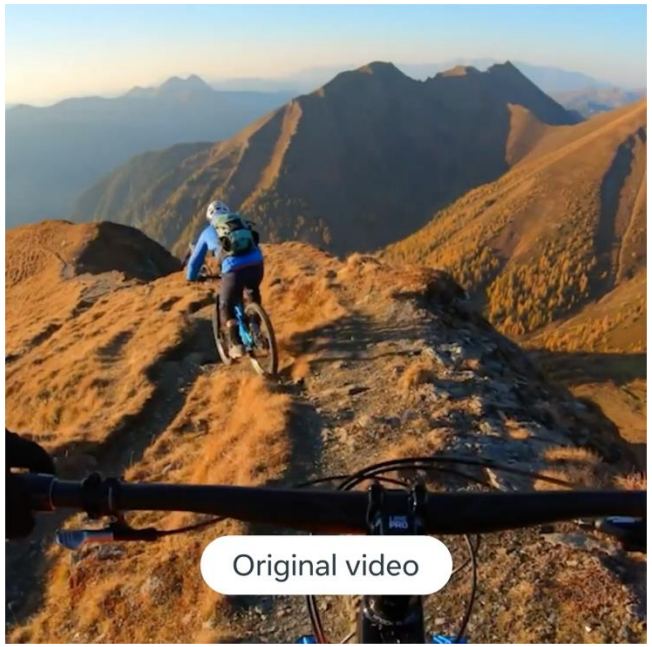
DINOv2
First successful scaling of a SSL algorithm. 1B-parameter models trained on 142M images.

DINOv3
An order of magnitude larger training compared to v2

Result

❖ Improved limitations in SSL

- DINOv3 is an important development in the field of self-supervised learning
- Demonstrate the potential to revolutionize the way visual representations are learned across various domains
- The introduction of the Gram anchoring method effectively mitigates the degradation of dense feature maps
- Leveraging high-resolution post-training and distillation, DINOv3 achieves state-of-the-art performance across a wide range of visual tasks with no fine-tuning of the image encoder.



감사합니다.

End of Document

CHUNG-ANG UNIVERSITY