DINOv2: Learning Robust Visual Features without Supervision

M. Oquab, T. Darcet, T. Moutakanni et al. — Meta AI
April 2023

John Meir Moussaffi
Department of Mathematics
Bar-Ilan University

Introduction to SSL and ViT

Self-supervised learning (SSL) aims to produce transferable features without labels, for the use of other tasks that may require labels.
Meta AI created an SSL "recipe" that improved existing methods, and released the result as **DINOv2**.
The main innovation comes from augmentation in training, which makes pretraining 2 times faster and require 3 times less memory than earlier discriminative SSL methods.
Dataset used was a "small loan" of a 142m curated images. DINO's scalability aims to approach the magnitude of data used in LLMs.
I'll be focusing on the student-teacher networks used in DINOv2.

Self Distillation

X_0 is the given data, and y_0 are the ground truth labels.

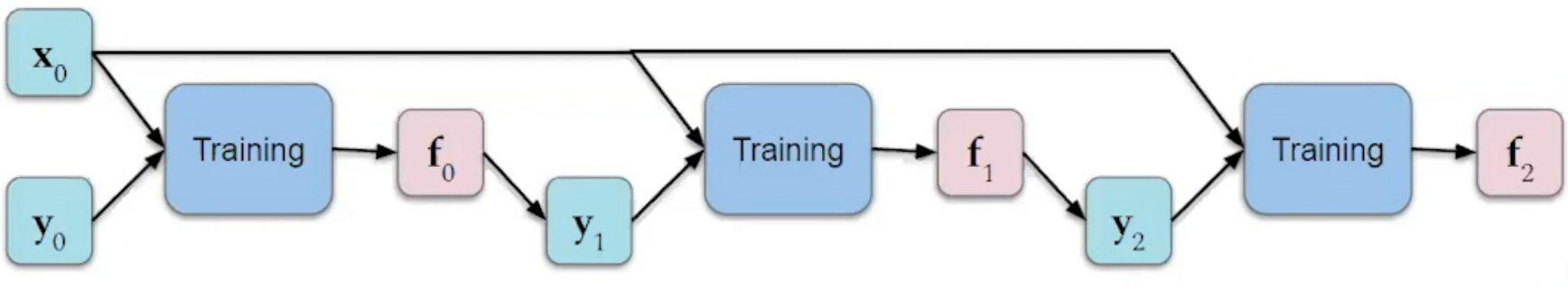
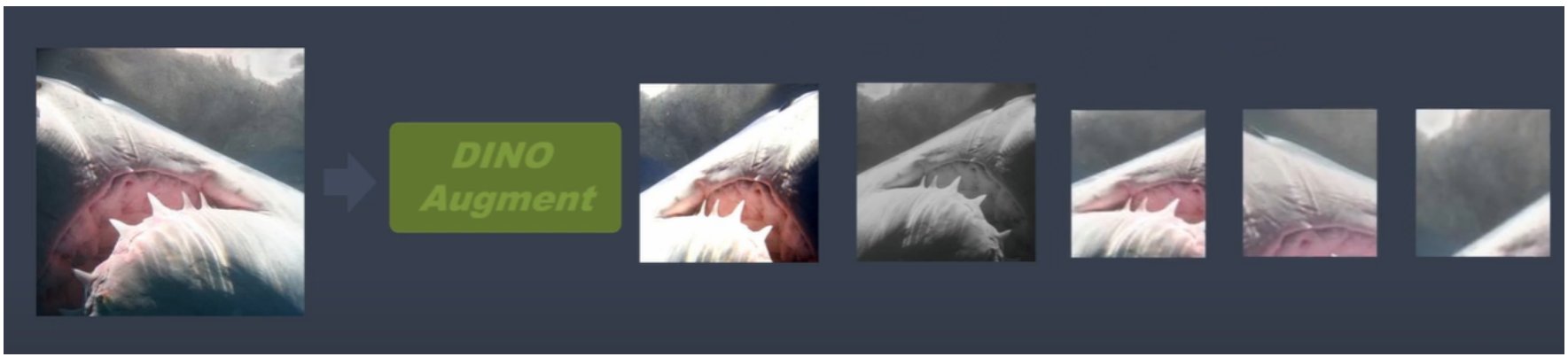


Figure: Simplified Self Distillation Diagram

Batching Process

- A *batch* contains crops from many images processed in one forward/backward step.
- For each image:
  - **2 global crops** (high-resolution views).
  - **8-10 local crops** (smaller random windows).
- Multi-crop strategy enforces scale robustness while keeping memory low.



Student-Teacher Networks

- Student and teacher share the *same* ViT architecture.
- Teacher sees *global* crops only; student sees *all* crops.
- Goal: align student probabilities to teacher probabilities for every matching crop.

Exponentially Moving Average (EMA) Teacher

- Temperature-scaled softmax smooths each output:
  $p = \text{softmax}(z/T_s), \;\; q = \text{softmax}((z-c)/T_t)$.
- Teacher weights are an EMA of student weights (also known as Exponensial Smoothing):

$$\bar{\theta}_{t+1} = \tau\,\bar{\theta}_t + (1-\tau)\,\theta_t$$

- No back-propagation through the teacher $\Rightarrow$ stable yet improving target.

We get Self-Supervised Self-Distillation.

DINO Loss + KoLeo Entropy

- Image-level term: $\mathcal{L}_{\text{DINO}} = \mathrm{KL}(q \,\|\, p) = -\sum p_t \log p_s$.
- Patch-level term (iBOT style): $\mathcal{L}_{\text{iBOT}} = -\sum_i q_k \log p_k$.
- KoLeo entropy regulariser:
  $\mathcal{L}_{\text{koleo}} = -\frac{1}{n}\sum_{i=1}^{n} \log d_{n,i}$, where $d_{n,i} = \min_{j \neq i} \|x_i - x_j\|$

Combined objective loss function:

$$\min_{\theta} \mathcal{L} = \mathcal{L}_{\text{DINO}} + \mathcal{L}_{\text{patch}} + \mathcal{L}_{\text{koleo}}$$

Results

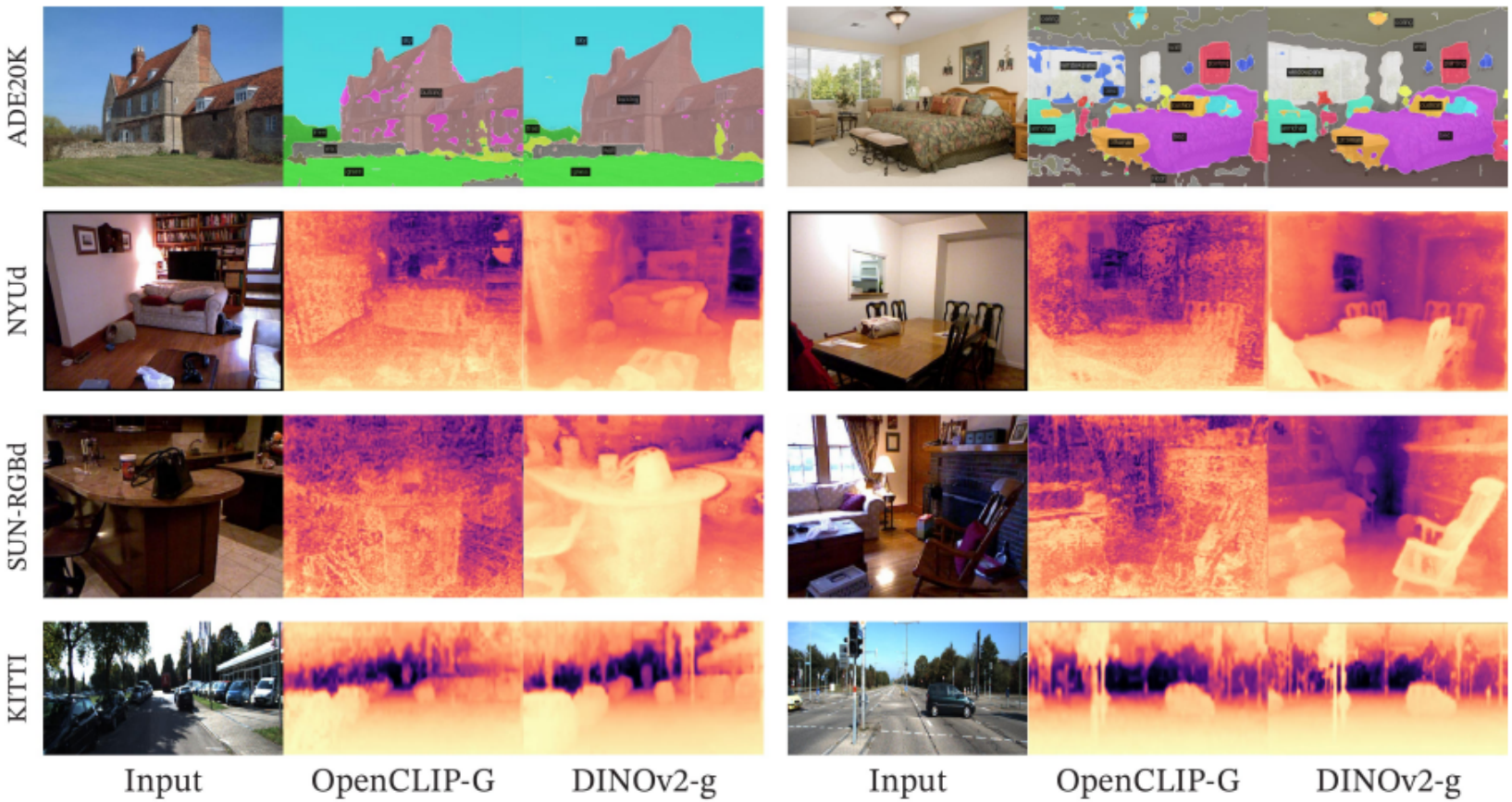


Figure: Segmentation and depth estimation with linear classifiers

Results

Figure: Examples of out-of-distribution examples with frozen DINOv2-g features and a linear probe

Performance

- **ImageNet 1k**: 83-87 % linear-probe top-1. ViT-g achieves 86.7 % *frozen*.
- Same backbones set new SSL records on NYU/KITTI depth and ADE20K segmentation.
- Robustness suites (ImageNet-A/R/Sketch) improve by $>$ 5pp over prior SSL methods.
- All results obtained with *frozen* features + shallow heads; no task-specific fine-tuning.

Performance

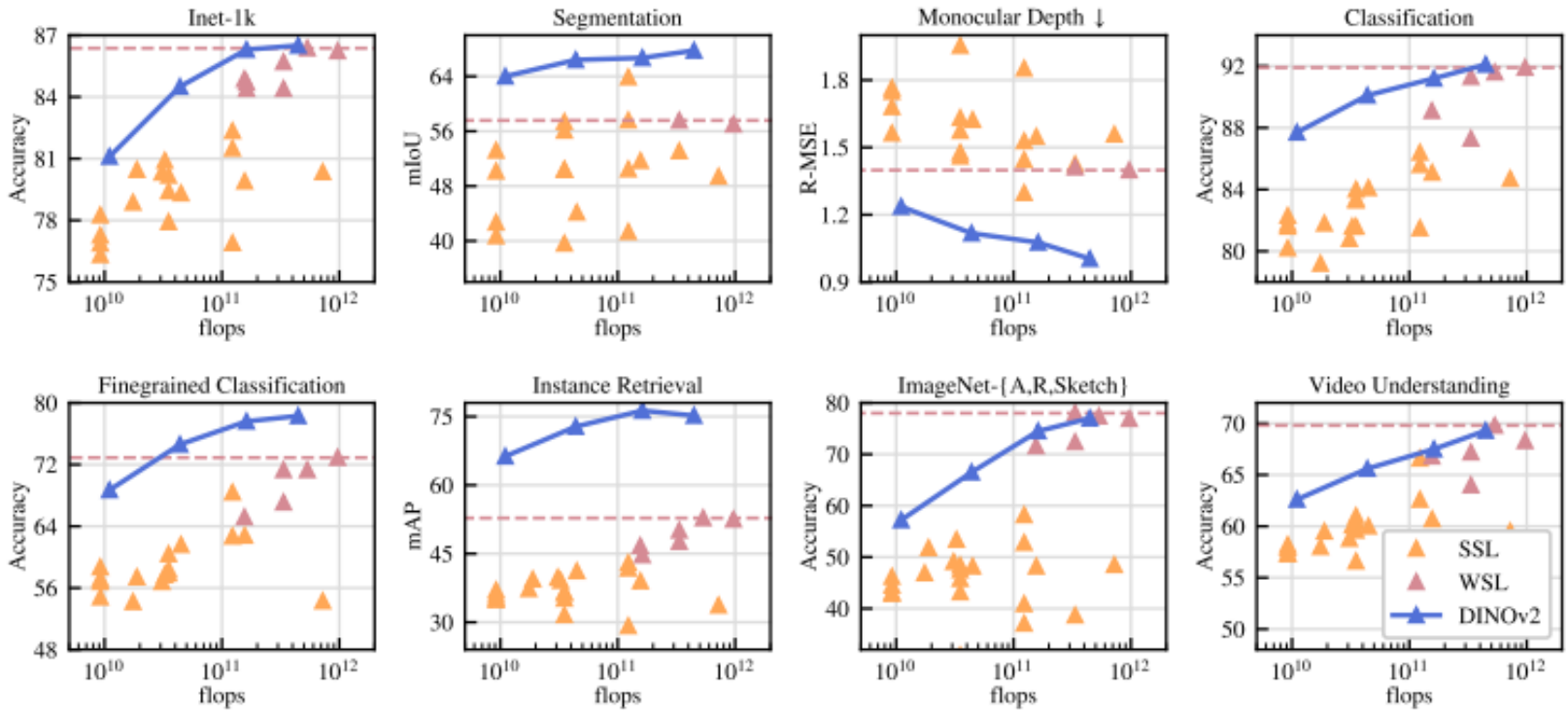


Figure: DINOv2 (dark blue) vs SSL mthods (pale orange) vs WSL (dark pink) on different vision tasks. We can see that this new family of models drastically improves over the previous state of the art in self-supervised learning and reaches performance comparable with weakly supervised features.

Knowledge Distillation

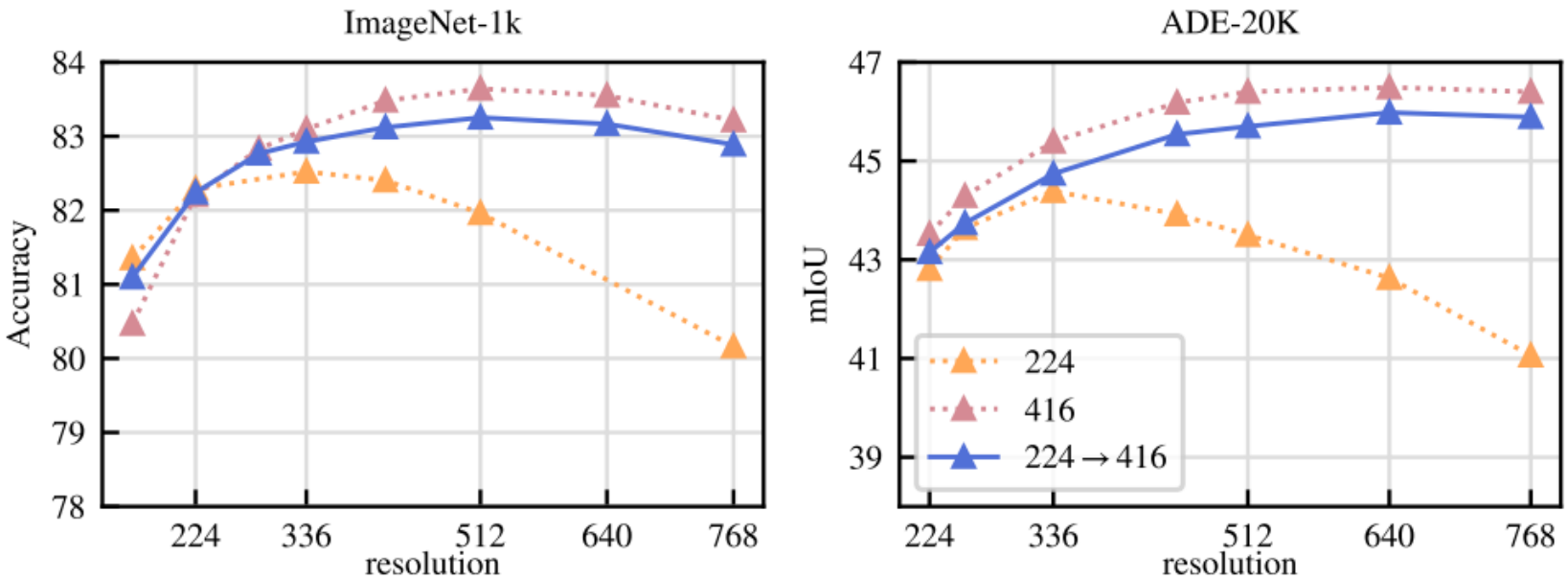


Figure: Effectiveness of knowledge distillation

Comparison between a ViT-L trained from scratch or distilled from DINOv2 using ViT-g/14

Emergent Part Awareness

- PCA of patch embeddings reveals clear separation of semantic parts vs. background.
- Self-attention maps highlight object boundaries without any labels.
- With a tiny segmentation head, performance rivals specialised segmentation networks.
- Provides strong few-shot transfer for downstream dense tasks.

Conclusions

- DINOv2 delivers up to 86.5 % ImageNet accuracy and leads on 10+ downstream tasks without labels or text.
- EMA teacher + multi-crop loss yields scalable, robust ViT backbones.
- Curated data and efficient training make SSL practical at billion-parameter scale.

END