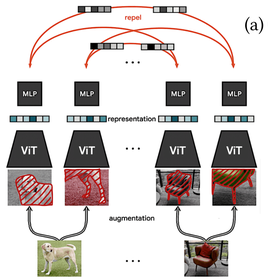
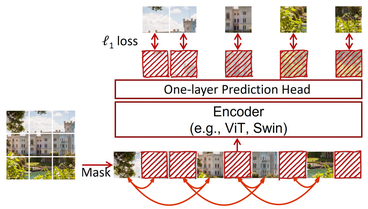
BACKGROUND CL Is Image-Level Approach and MIM is Token-Level Approach



(a) Contrastive Learning (CL)

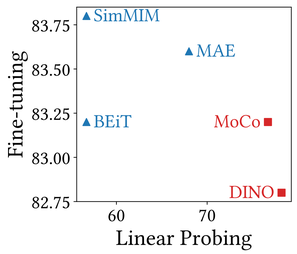


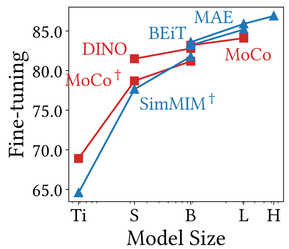
(b) Masked Image Moodeling (MIM)

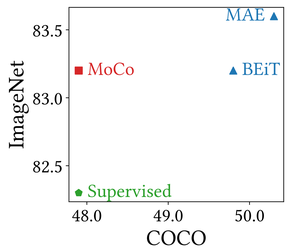
CL aims to learn the **invariant semantics** for two random views by making global projections. CL can be deemed as "*image-level*" self-supervised learning.

MIM trains ViTs by reconstructing the **correct semantics** (similarities) of masked input patches. Since it learns semantics of patch tokens, it can be deemed as "*token-level*" self-supervised learning.

BEHAVIOUR CL and MIM May Not Be a Silver Bullet for All Tasks



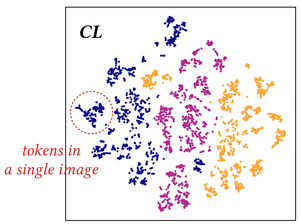


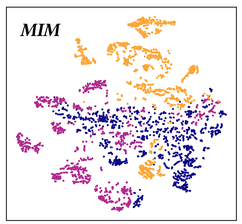


CL ■ works well in linear probing tasks, small model regimes, and classification tasks.

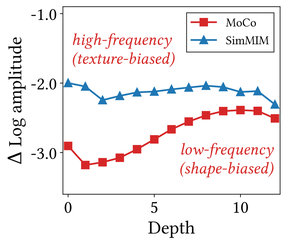
MIM ▲ outperforms CL in fine-tuning tasks, large model regimes, and dense prediction tasks.

REPRESENTATION CL Distinguish Images and MIM Distinguish Tokens





(a) Token-level t-SNE visualization (18 images from 3 classes)

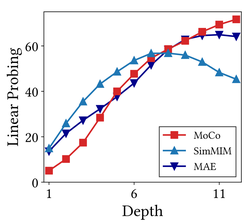


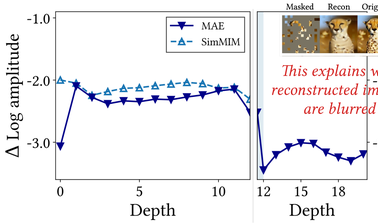
(b) Fourier Analysis

Left: The tokens of CL form a cluster for each image, while those of MIM are intermingled.

Right: CL exploits low-frequency, but MIM exploits high-frequency. However, a few last layers of MIM reduce the high-frequencies even though they capture local patterns, because they behave like decoders.

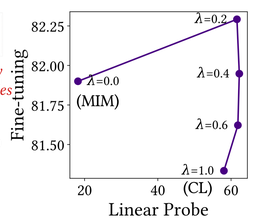
ARCHITECTURE CL Focuses on Later Layers and MIM Focuses on Early Layers





(a) Linear Probing

(b) Fourier Analysis of MAE



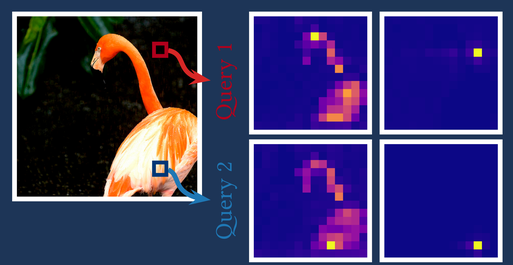
(c) Hybrid model

Left: Later layers of CL and early layers of MIM play a key role.

Middle: MAE helps ViTs fully leverage MIM by decomposing decoders from backbones.

Right: "CL + MIM" outperforms CL and MIM in both linear probing and fine-tuning accuracy.

WHAT DO SELF-SUPERVISED VISION TRANSFORMERS LEARN?



(a) CL (b) MIM

Contrastive Learning	Masked Image Modeling
Linear probing & small model	Fine tuning & large model
Capture globals and shapes	Capture locals and textures
Distinguish images	Distinguish tokens
Focus on later layers	Focus on early layers

NAMUK PARK[1], WONJAE KIM[2], BYEONGHO HEO[2], TAEKYUNG KIM[2], SANGDOO YUN[2]

[1] Prescient Design, Genentech [2] NAVER AI Lab