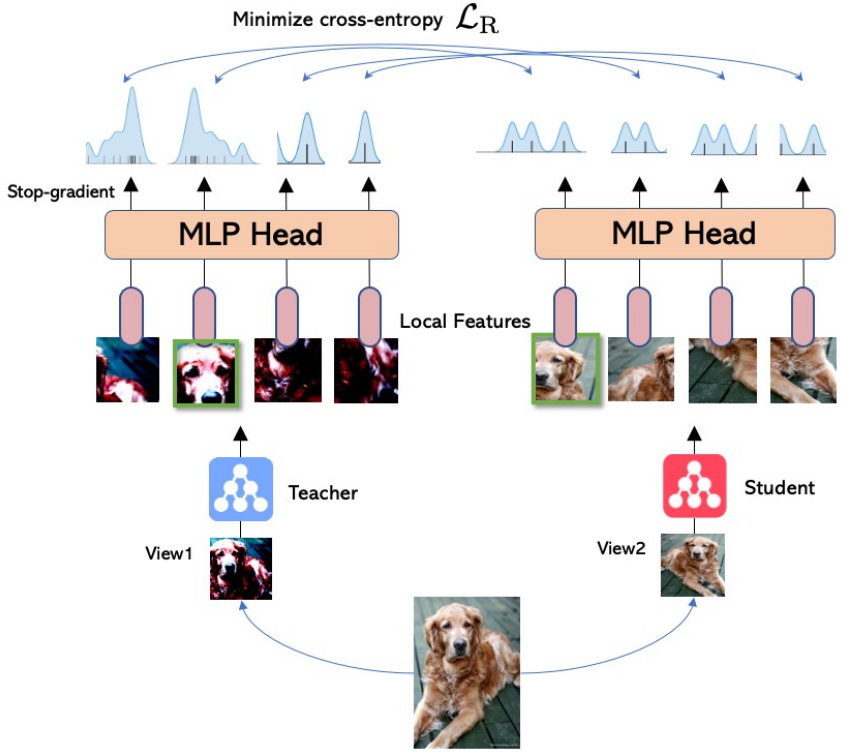
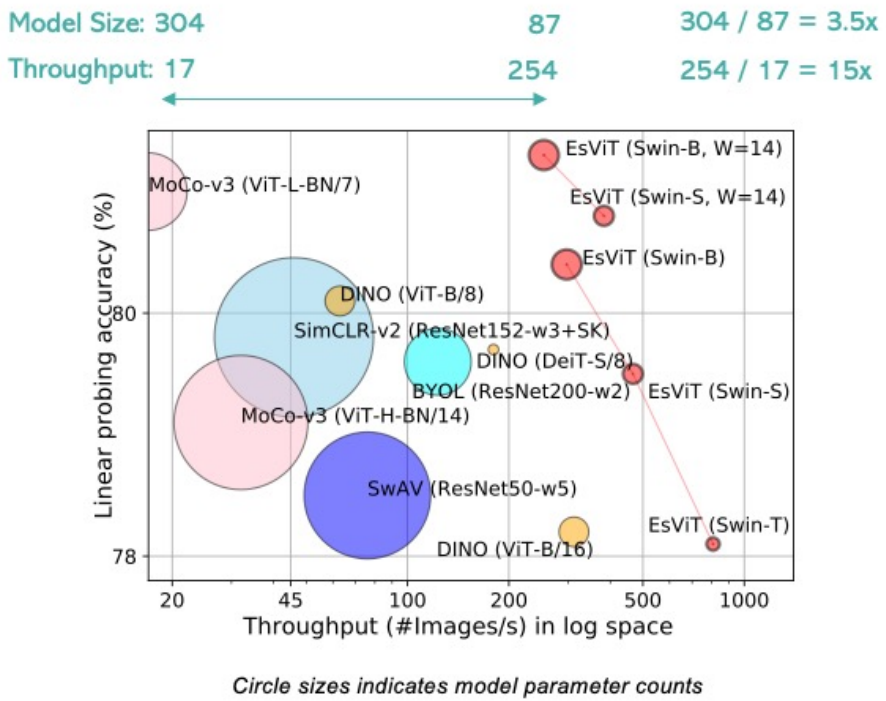
EsViT: Efficient Self-supervised Vision Transformers for Representation Learning

Network architectures: A multi-stage Transformer architecture

Pre-training Objectives: A region-level pre-train task



- SoTA on ImageNet self-supervised performance
- EsViT outperforms the supervised counterpart on 17 out of 18 classification tasks



Circle sizes indicates model parameter counts