Teaching Matters: Investigating the Role of Supervision in Vision Transformers

Matthew Walmer*, Saksham Suri*, Kamal Gupta, Abhinav Shrivastava

*Equal Contributors, Narrators

TUE-PM-321

UNIVERSITY OF MARYLAND 18 56

UMIACS

JUNE 18-22, 2023

CVPR

VANCOUVER, CANADA

ViTs: A New Black Box

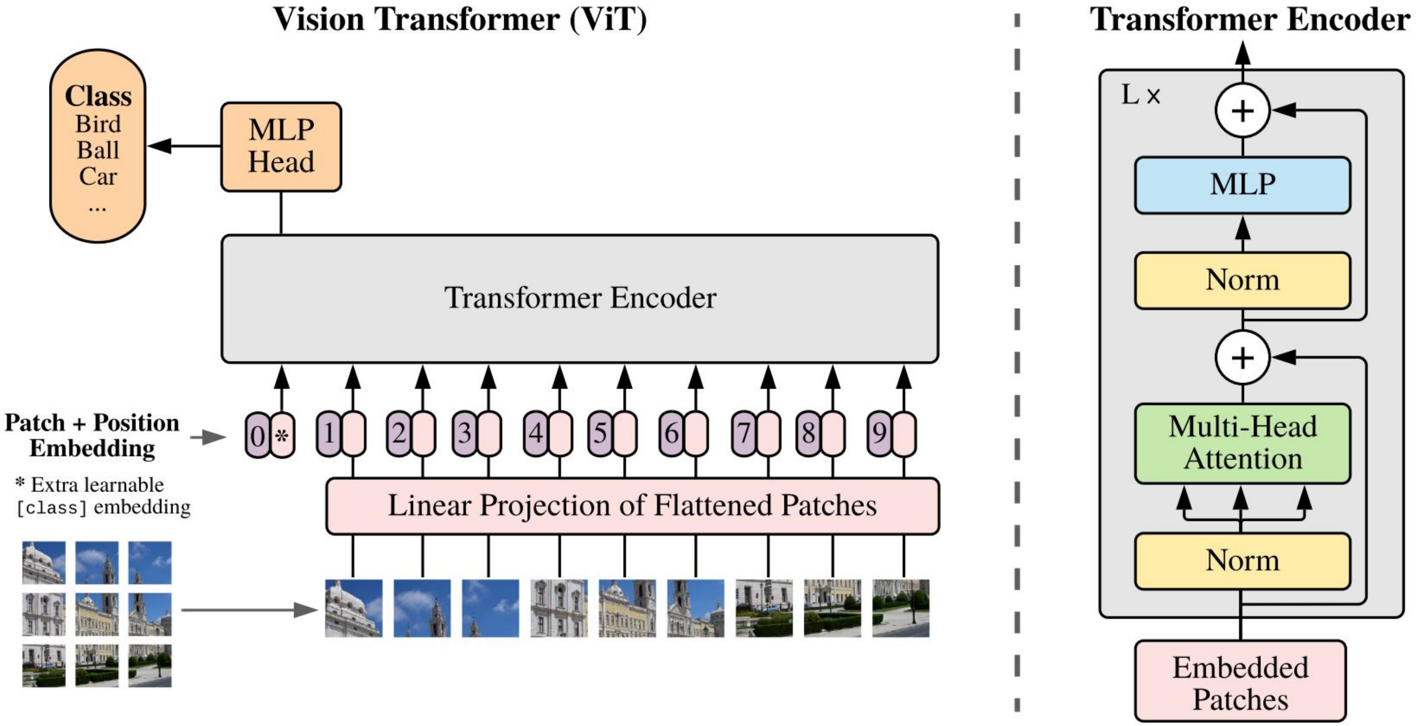
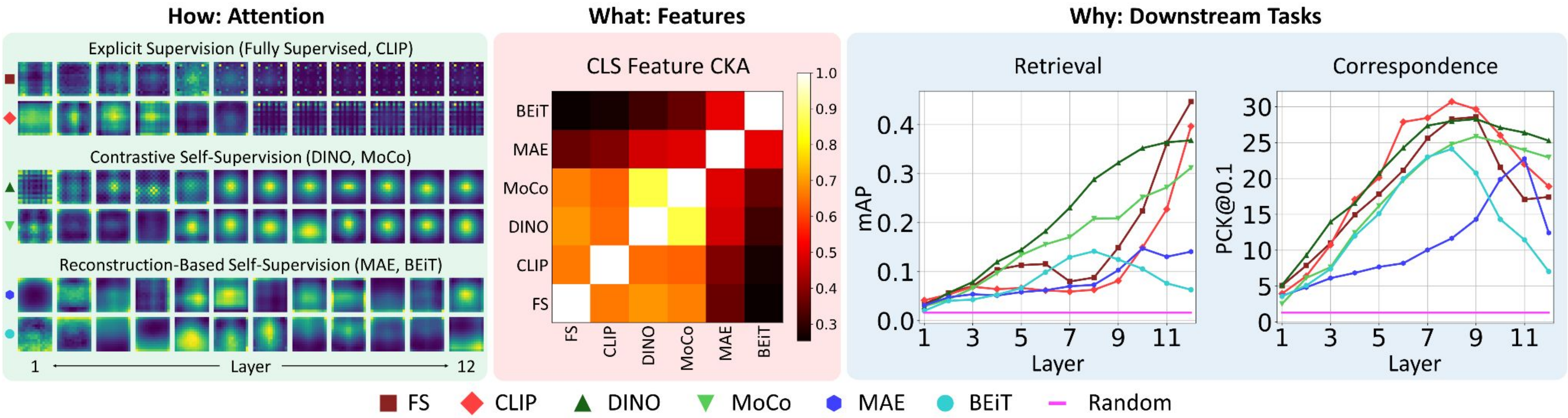


Figure from:
Dosovitskiy et al. (2020)

- ViTs: a new go-to model for vision tasks
- Less structural bias → more flexible learning
- But what are they learning under different supervision?

Teaching Matters



- First in depth comparison of ViTs trained with different supervision
- Identify commonalities and key differences
- Analysis covering Attention, Features, and Downstream Tasks

Additional Information

Poster Session: **TUE-PM-321**

Full Presentation Includes:

- Overview of models
- Summary of experiments
- Key observations

website

code

Teaching Matters: Investigating the Role of Supervision in Vision Transformers

Matthew Walmer* Saksham Suri* Kamal Gupta Abhinav Shrivastava
University of Maryland, College Park

Abstract

Vision Transformers (ViTs) have gained significant popularity in recent years and have proliferated into many applications. However, their behavior under different learning paradigms is not well explored. We compare ViTs trained through different methods of supervision, and show that they learn a diverse range of behaviors in terms of their attention, representations, and downstream performance. We also discover ViT behaviors that are consistent across supervision, including the emergence of Offset Local Attention Heads. These are self-attention heads that attend to a token adjacent to the current token with a fixed directional offset, a phenomenon that to the best of our knowledge has not been highlighted in any prior work. Our analysis shows that ViTs are highly flexible and learn to process local and global information in different orders depending on their training method. We find that contrastive self-supervised methods learn features that are competitive with explicitly supervised features, and they can even be superior for part-level tasks. We also find that the representations of reconstruction-based models show non-trivial similarity to contrastive self-supervised models.

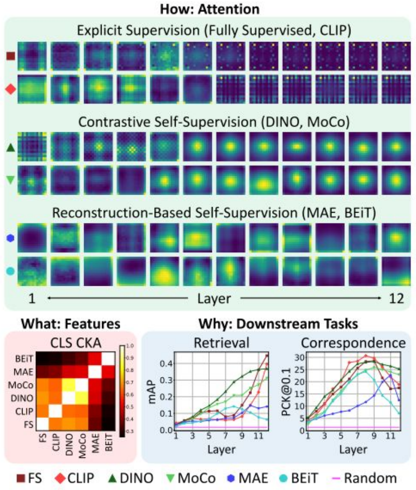


Figure 1. **ViTs exhibit highly varied behaviors depending on their method of training.** In this work, we compare ViTs through three domains of analysis representing the How, What, and Why of ViTs. **How** do ViTs process information through attention? (Top) Attention maps averaged over 5000 images show clear differences in the mid-to-late layers. **What** do ViTs learn to represent? (Left) Contrastive self-supervised ViTs have a greater feature similarity to explicitly supervised ViTs, but also have some similarity with ViTs trained through masked reconstruction. **Why** do we care about using ViTs? (Right) We evaluate ViTs on a variety of global and local tasks and show that the best model and layer vary greatly.

1. Introduction

The field of Computer Vision has advanced massively in the past decade, largely built on the backbone of Convolutional Neural Networks (CNNs). More recently, Vision Transformers (ViTs) [18] have shown the potential to overtake CNNs as the go-to visual processing model. Prior works have asked the question *do ViTs see like CNNs do?* [52], but in this work, we ask: *how do ViTs learn under different supervision?* Past examinations of ViTs have largely focused on models trained through full supervision. Instead, we aim to characterize the differences and similarities of ViTs trained through varying training methods, including self-supervised methods. Unlike CNNs, the ViT architecture imposes few structural biases to guide the learning of representations. This gives them the flexibility to learn diverse information processing strategies, and through our analyses, we uncover a wide array of ViT behaviors.

There are countless ways to analyze ViTs, so to guide this analysis we choose three major domains which correspond to the *How*, *What*, and *Why* of ViTs. For the *How*, we focus on *how* ViTs process information through **Attention**. Multi-Headed Attention (MHA) layers are arguably the key element of ViTs, and they most distinguish them

*Equal contributors.
Web: www.cs.umd.edu/~sakshams/vit_analysis
Code: www.github.com/mwalmer-umd/vit_analysis

Experimental Design

Supervision Methods

Three supervision sub-categories:

- Explicit Supervision: Fully Supervised, CLIP
- Contrastive Self-Supervision: DINO, MoCo-v3
- Reconstruction Self-Supervision: MAE, BEiT

Focus on ViT-B/16 models in main work, and more variations in the appendix

Areas of Analysis

How ViTs process information:

→ **Attention Analysis**

What we take away from ViTs:

→ **Feature Analysis**

Why we use ViTs:

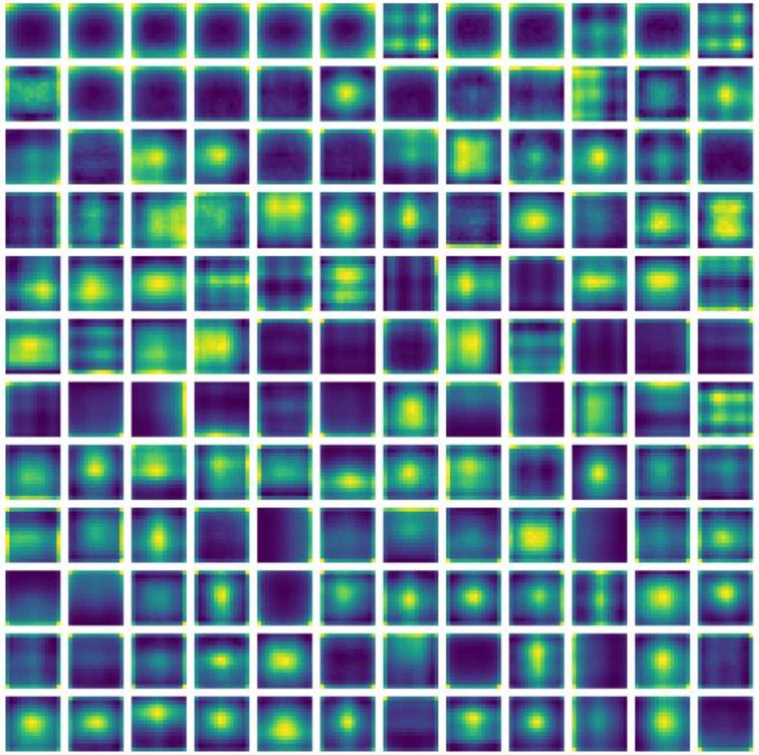
→ **Downstream Task Analysis**

Attention Analysis

The Size of ViT Attention

- Multi-Headed Attention (MHA) layers allow tokens to look anywhere
- 196 spatial tokens and 1 CLS token
- >28,000 attention maps per image

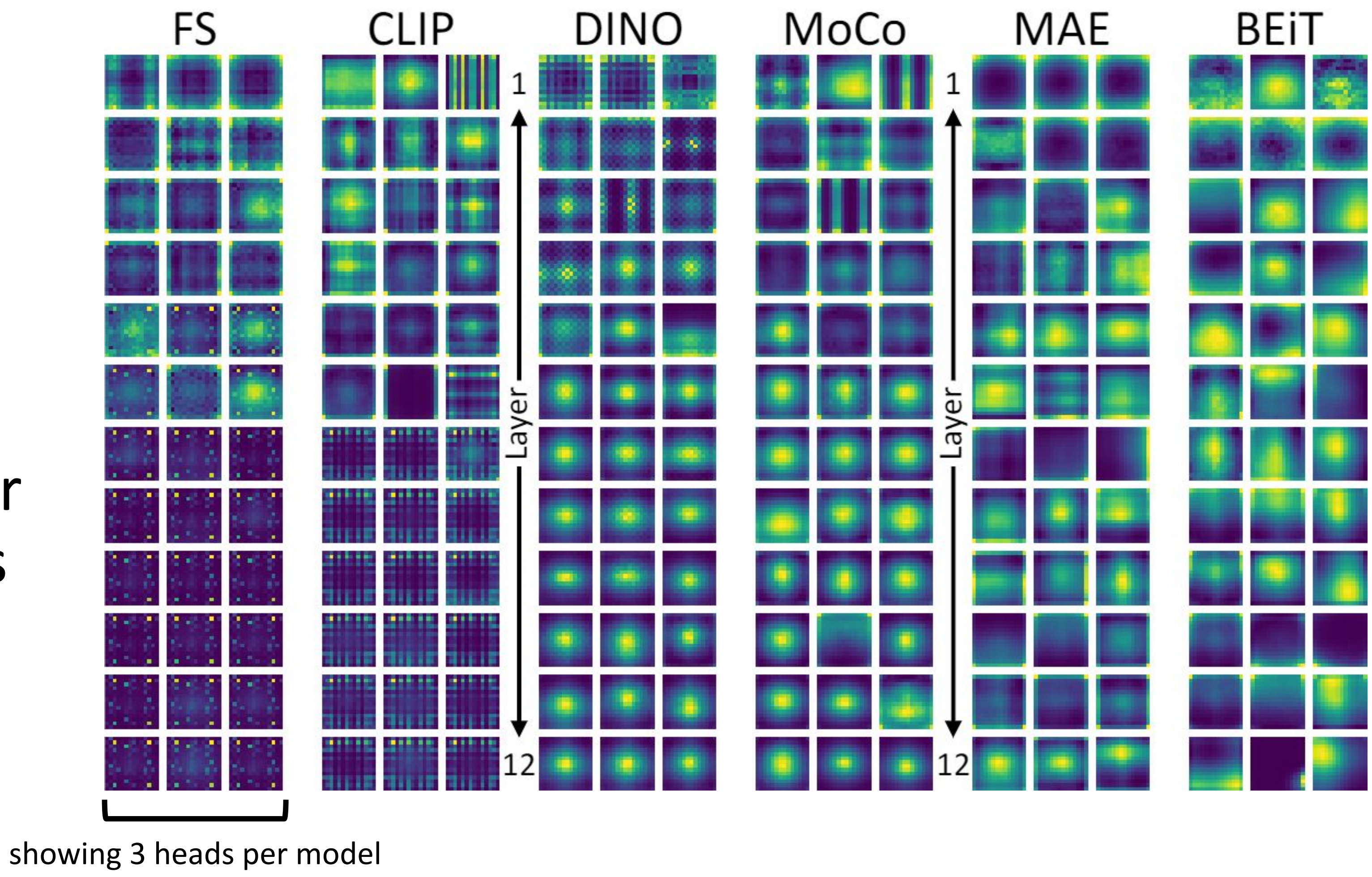
Multiple strategies to summarize ViT attention



x 197

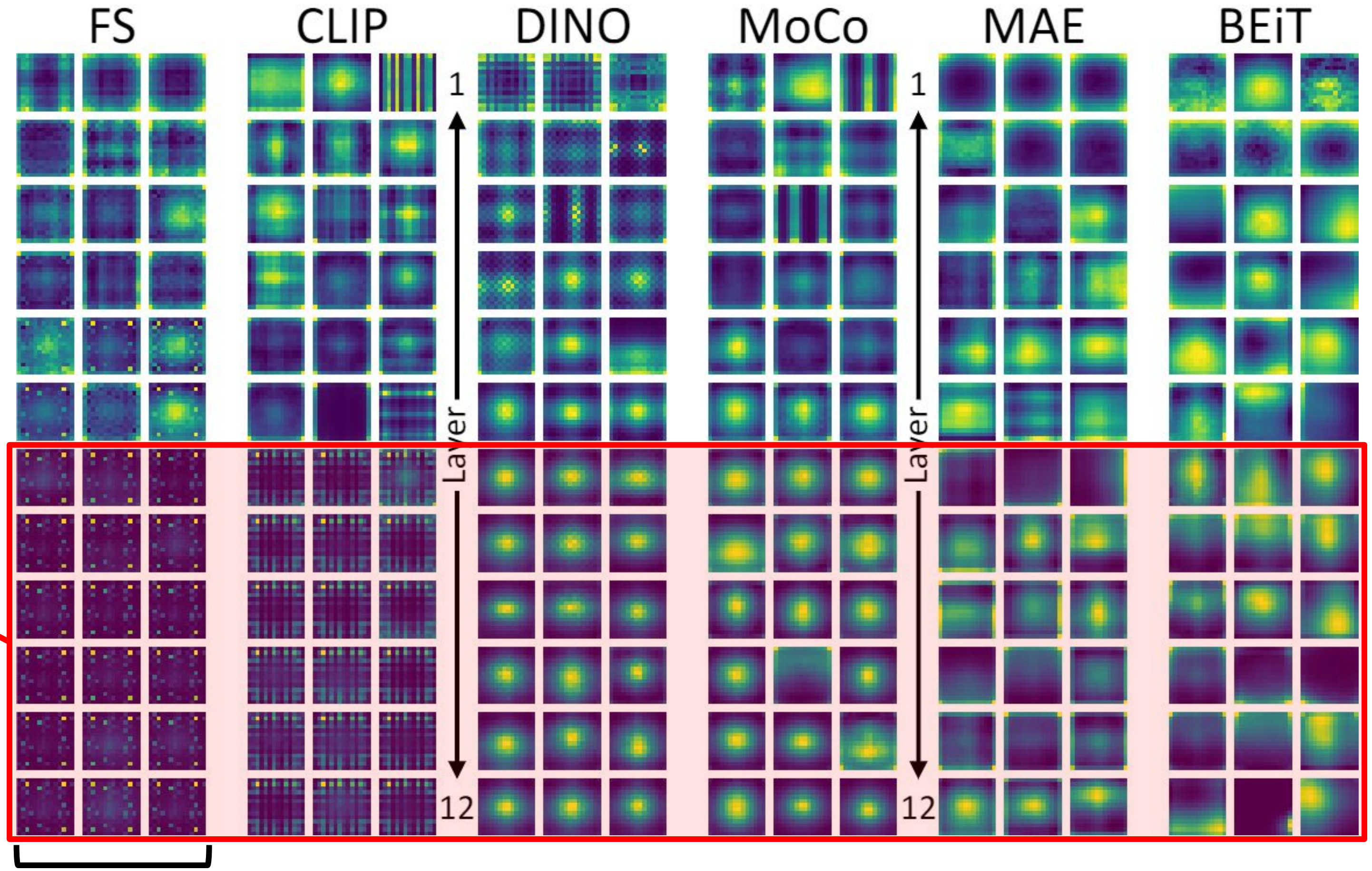
Visualizing CLS Token Attention

- CLS token attention in each layer and head
- Average over 5000 sample images
- Clear differences appear in the mid-to-late layers



Visualizing CLS Token Attention

- CLS token attention in each layer and head
- Average over 5000 sample images
- Clear differences appear in the mid-to-late layers

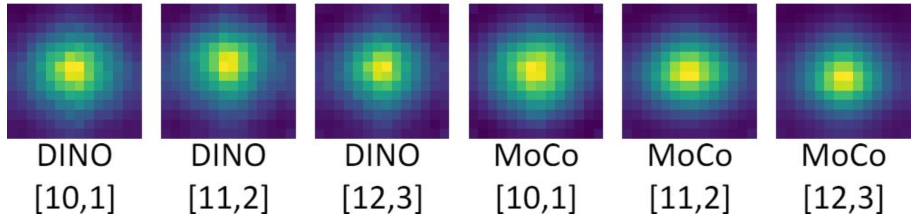


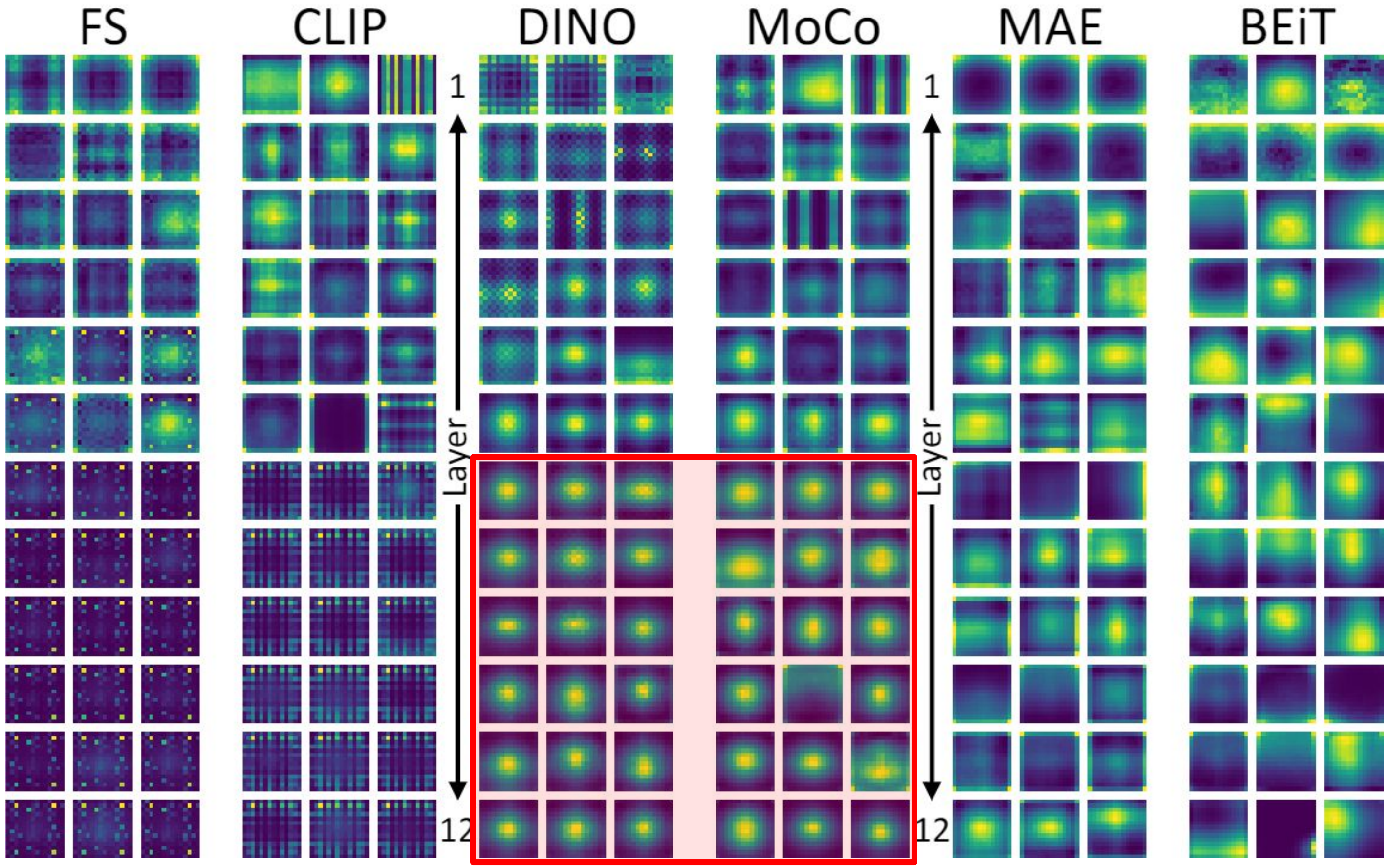
showing 3 heads per model

Visualizing CLS Token Attention

- DINO and MoCo create many centered blobs
- Salient objects are usually centered

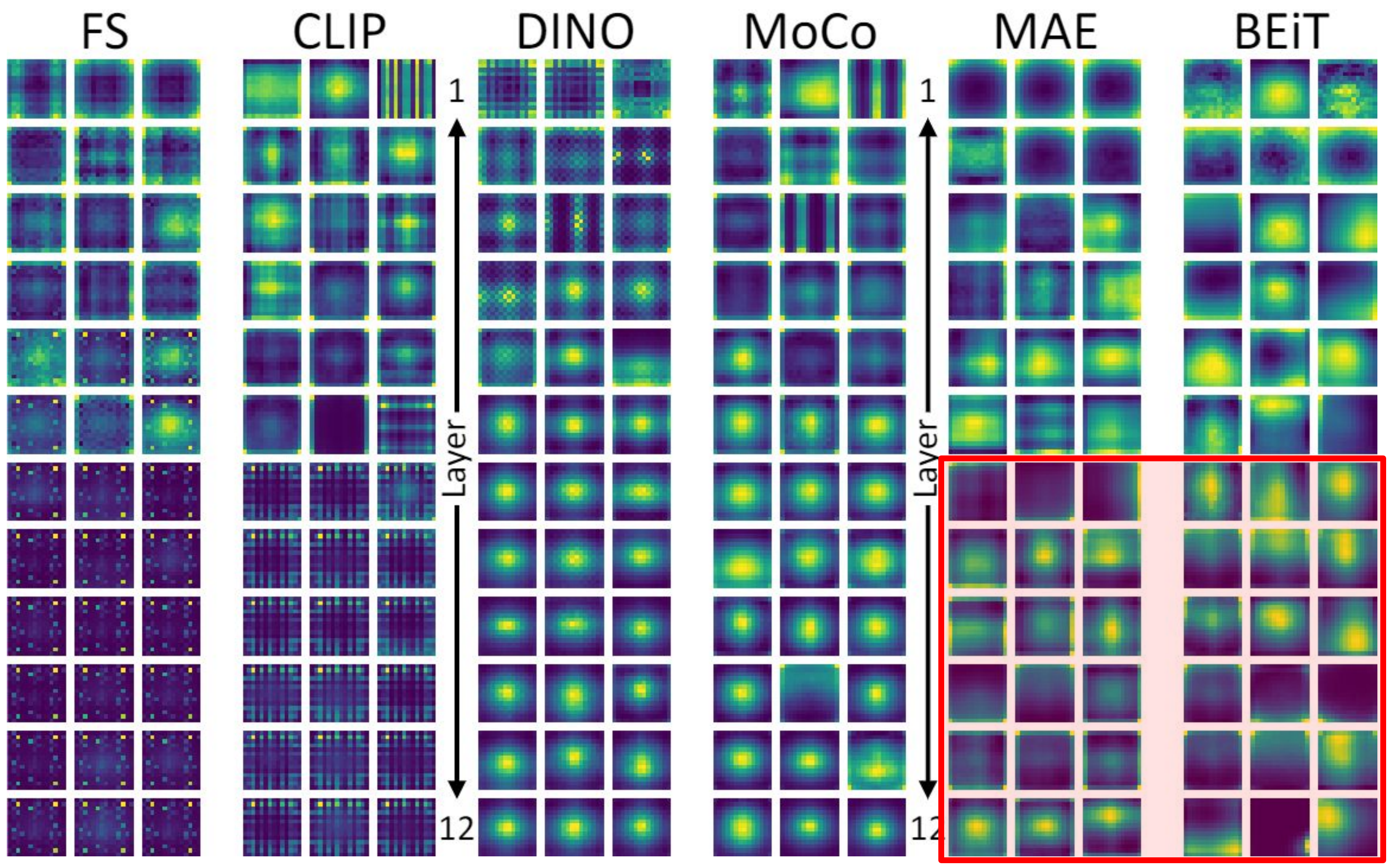
DINO & MoCo Layers 7-12: Object Centered Blobs



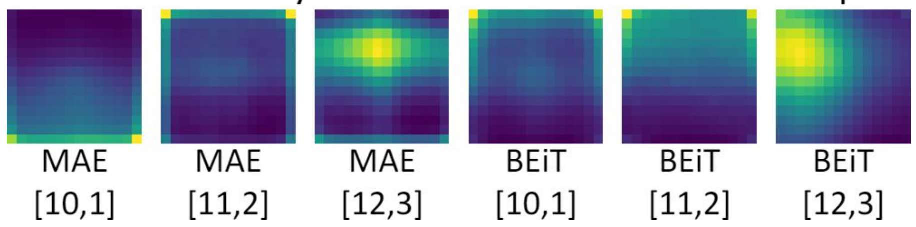


Visualizing CLS Token Attention

- MAE and BEiT have more diverse attention
- They must reconstruct the whole image, so they need wider attend



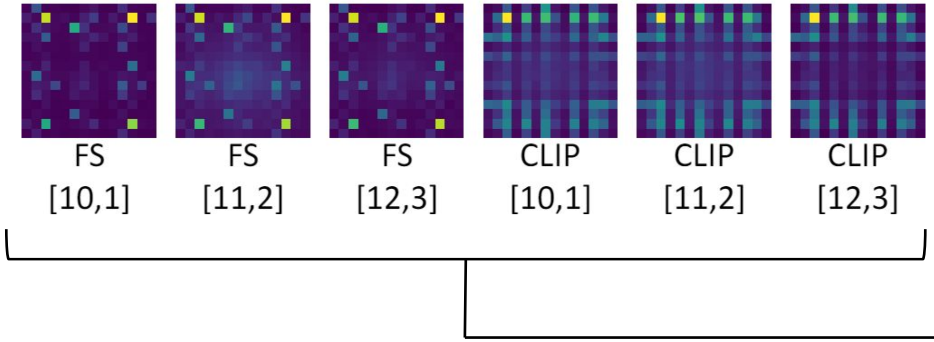
MAE & BEiT Layers 7-12: Diverse Attention Maps

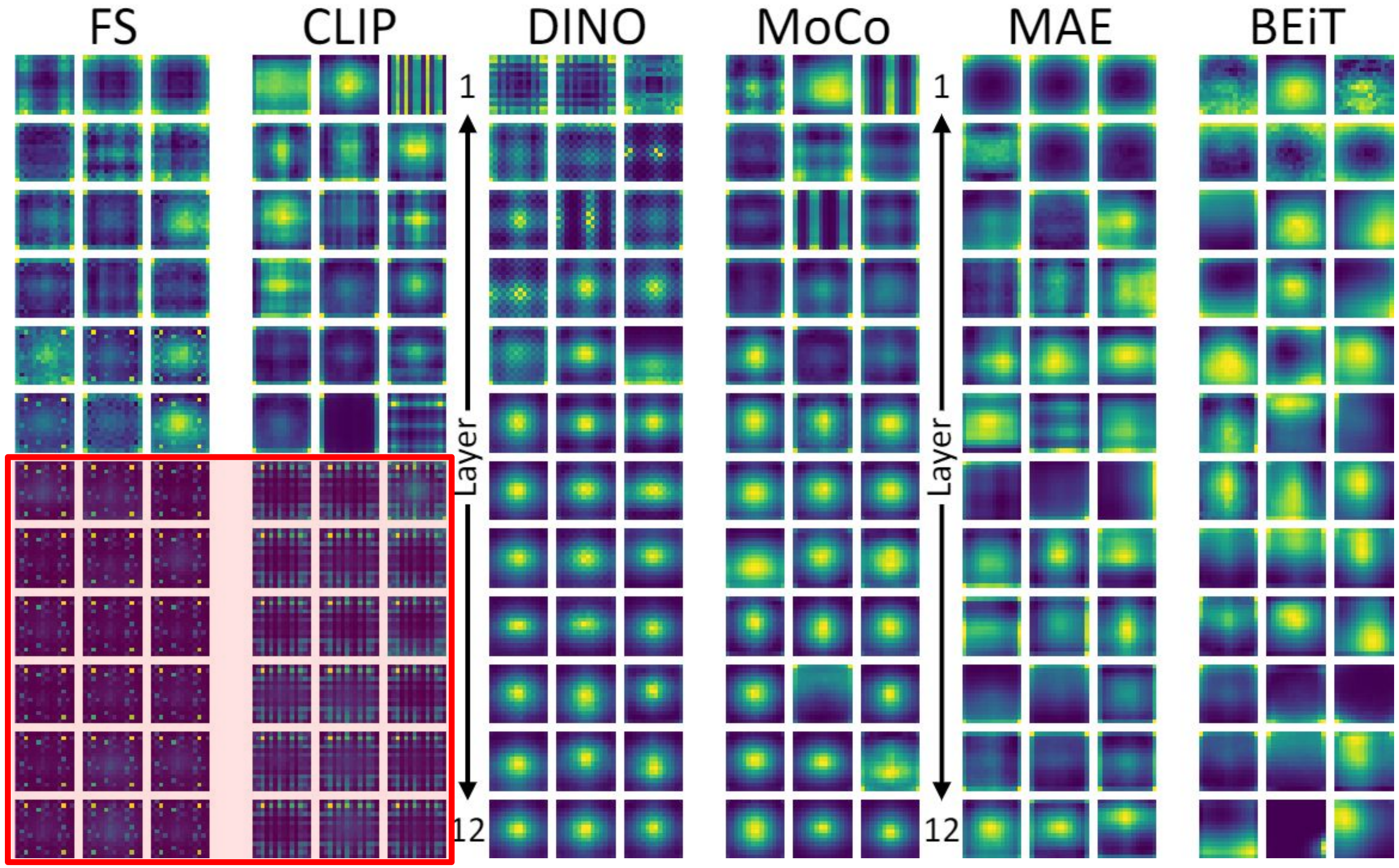


Visualizing CLS Token Attention

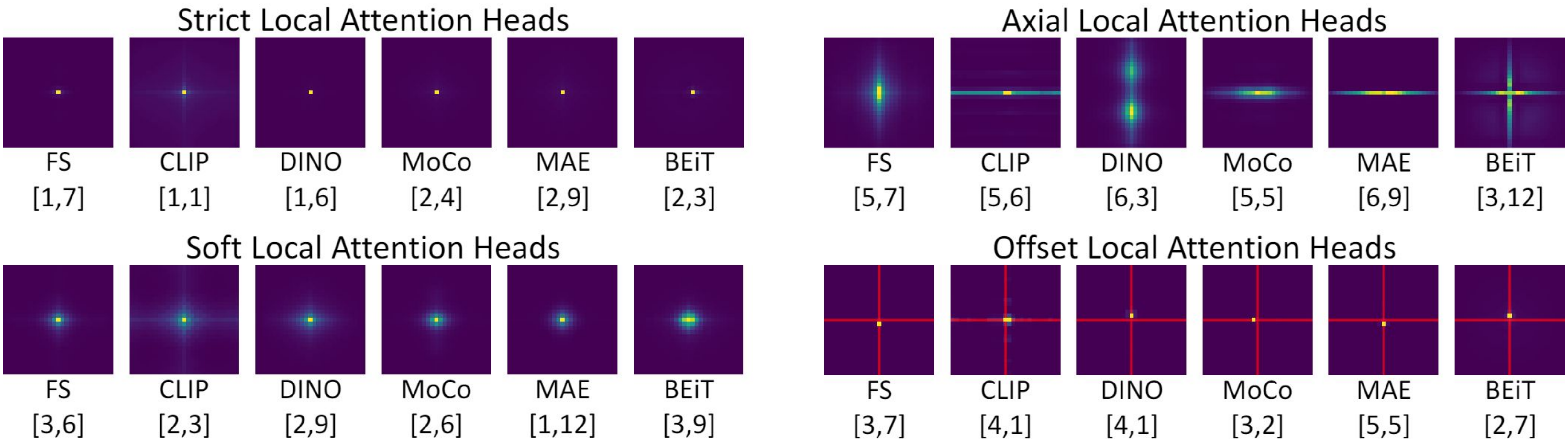
- FS and CLIP ViTs make **Sparse Repeating Patterns**
- Repeated over both layers and heads
- No clear spatial meaning

FS & CLIP Layers 7-12: Sparse Repeating Patterns



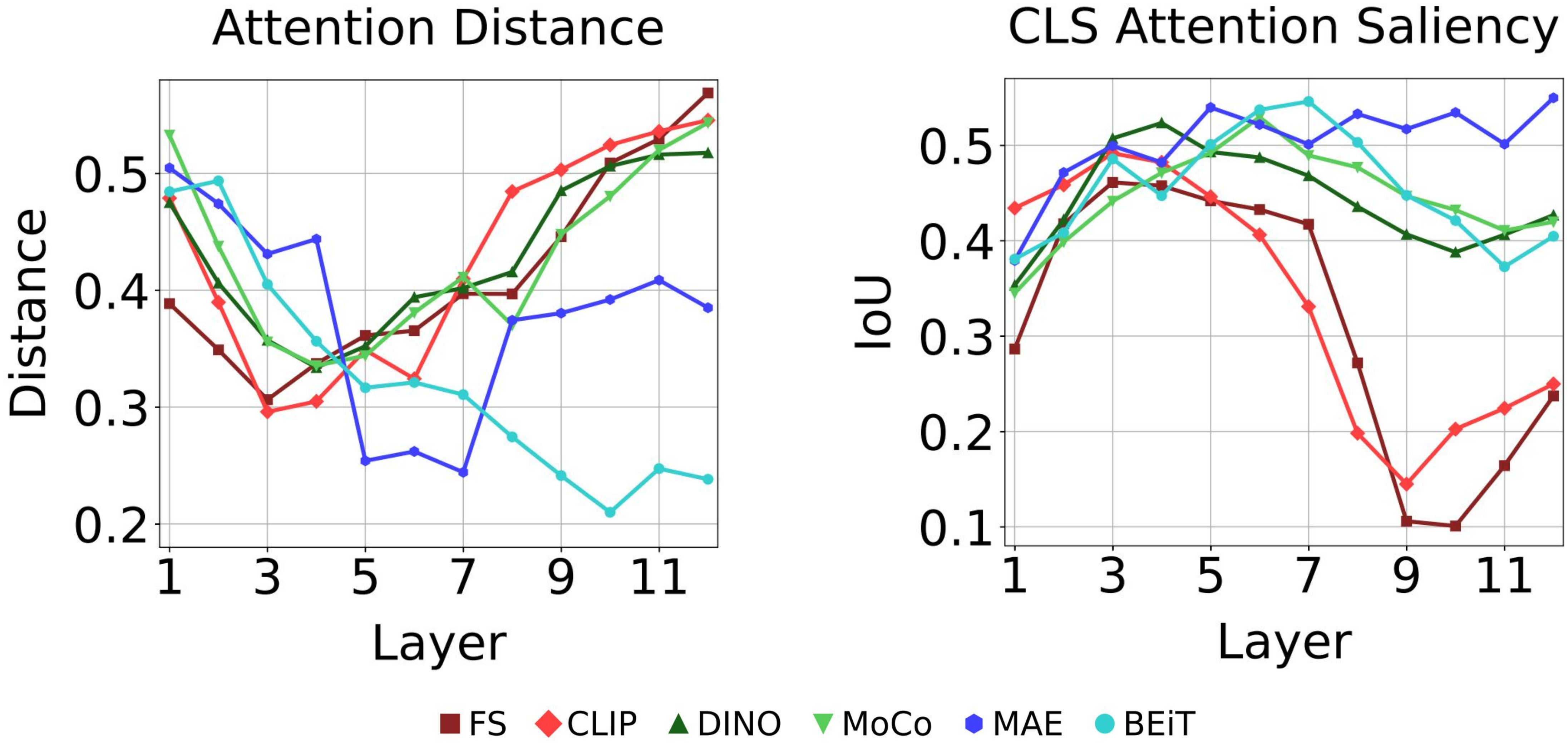


Aligned Aggregated Spatial Token Attention



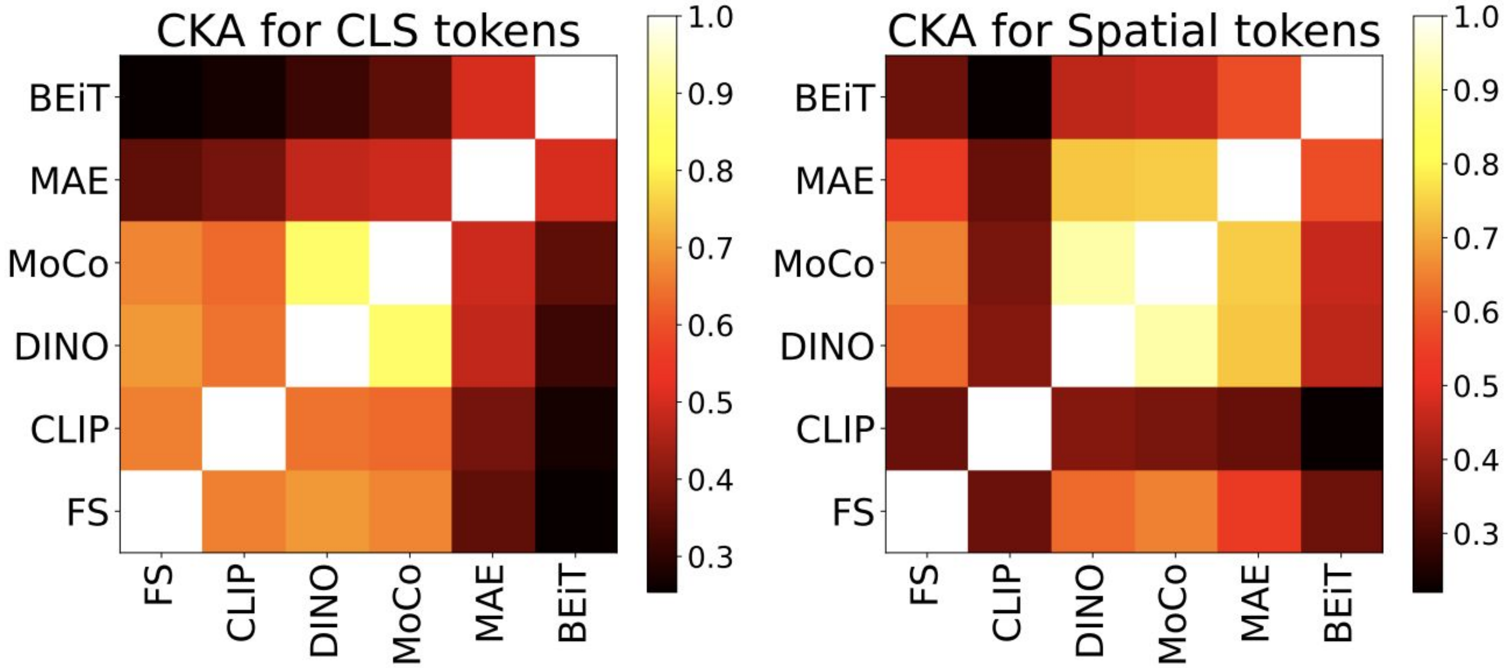
- **Aligned Aggregated Attention Maps** for Spatial Tokens
- We find different forms of local attention
- **Offset Local Attention Heads** with a fixed directional offset

Attention Distance and Saliency



Feature Analysis

Analyzing Last Layer Representations

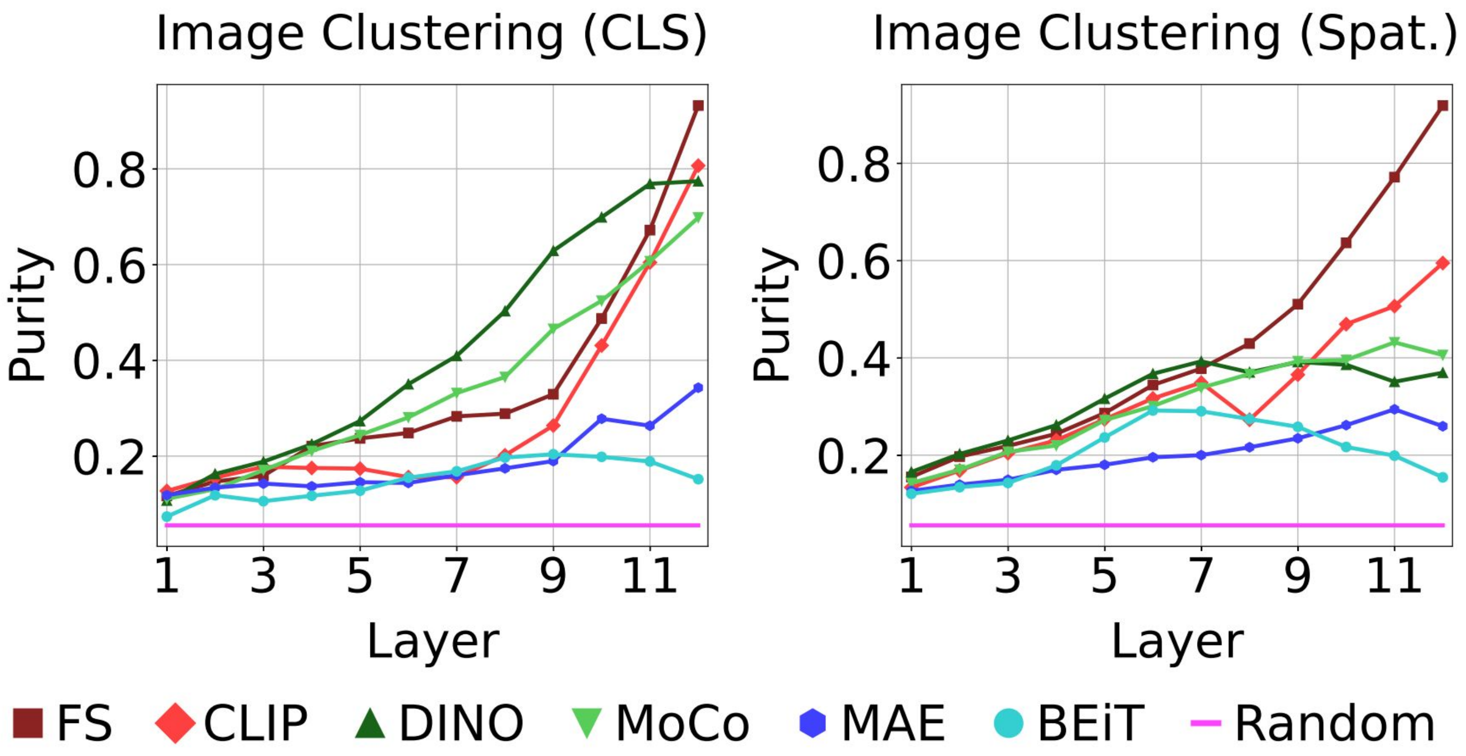


CLS token representations are usually similar for similar supervision strategies (explicit, contrastive, reconstruction).

Unlike the CLS token representations, CLIP and FS have low similarity in their spatial representations.

There is a surprisingly elevated similarity in CLS representations between MAE and the contrastive models, DINO and MoCo

Clustering on ImageNet50

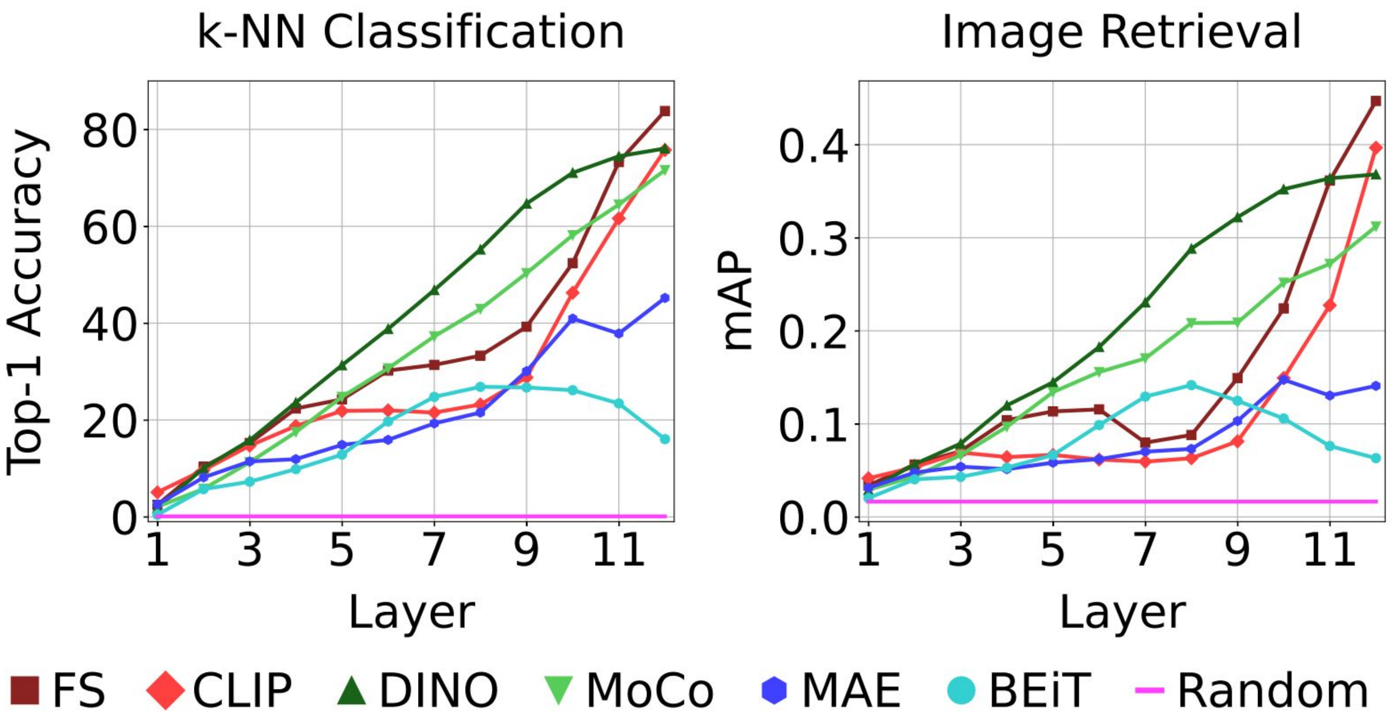


For CLS token features (left), cluster purity improves with depth except for BEiT. This is likely because the last layers of BEiT serve as a task-specific decoder, unlike MAE, where the decoder is separate and discarded after pretraining.

For the spatial token features (right), the cluster purity of FS rises earlier compared with the FS CLS token. This suggests that the FS spatial tokens do more work gathering semantic information in the early layers.

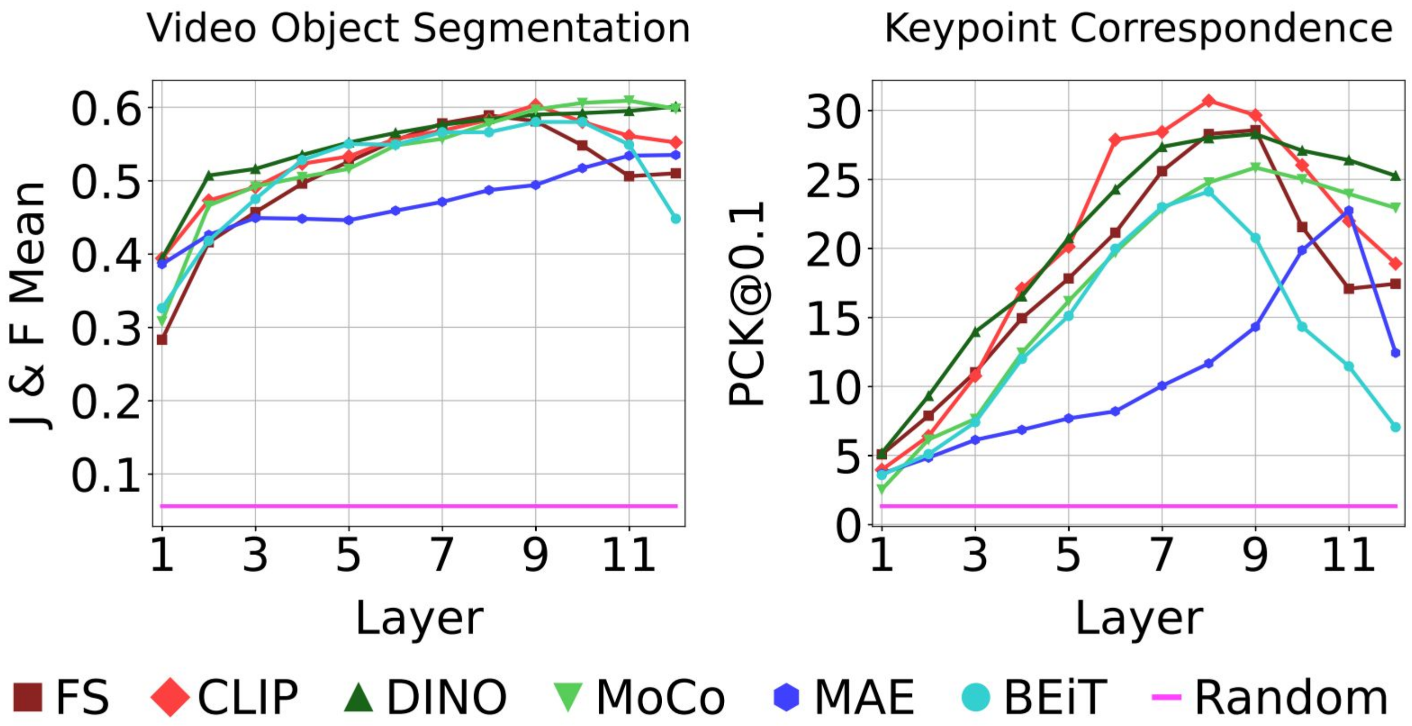
Downstream Tasks

Global Tasks: Classification and Retrieval



For global, image-level tasks like k-NN and image retrieval, methods which have explicit supervision on the CLS token perform better than others. The presence of label/text supervision helps achieve the good performance for FS and CLIP.

Local Tasks: Segmentation and Keypoints



For localized tasks like Video Object Segmentation and Keypoint Correspondence, the best performance occurs in the mid-to-late layers.

Localized supervision methods like MAE and BEiT become much more competitive on these tasks.

No Single “Winner”

Model	**Task Performance (Best Performing Layer)**			
Dataset	ImageNet	ROxford5k (M)	Davis	SPair-71k
Metric	Top-1↑	mAP↑	J and F Mean↑	PCK@0.1↑
FS	**83.79 (12)**	**0.45 (12)**	0.59 (8)	28.56 (9)
CLIP	75.75 (12)	0.40 (12)	0.60 (9)	**30.70 (8)**
DINO	76.06 (12)	0.37 (12)	0.60 (12)	28.28 (9)
MoCo	71.59 (12)	0.31 (12)	**0.61 (11)**	25.85 (9)
MAE	45.19 (12)	0.15 (10)	0.54 (12)	22.74 (11)
BEiT	26.84 (8)	0.14 (8)	0.58 (9)	24.11 (8)
Random	0.10	0.02	0.06	1.32

There is no single “best” model or layer for all downstream tasks.

Key Takeaways

- Sparse Repeating Attention Patterns in late layers of FS and CLIP
- Offset Local Attention Heads in all ViTs studied
- Local and Global information processed in different orders depending on supervision
- ViTs differentiate salient foreground objects by the early-to-mid layers

Key Takeaways

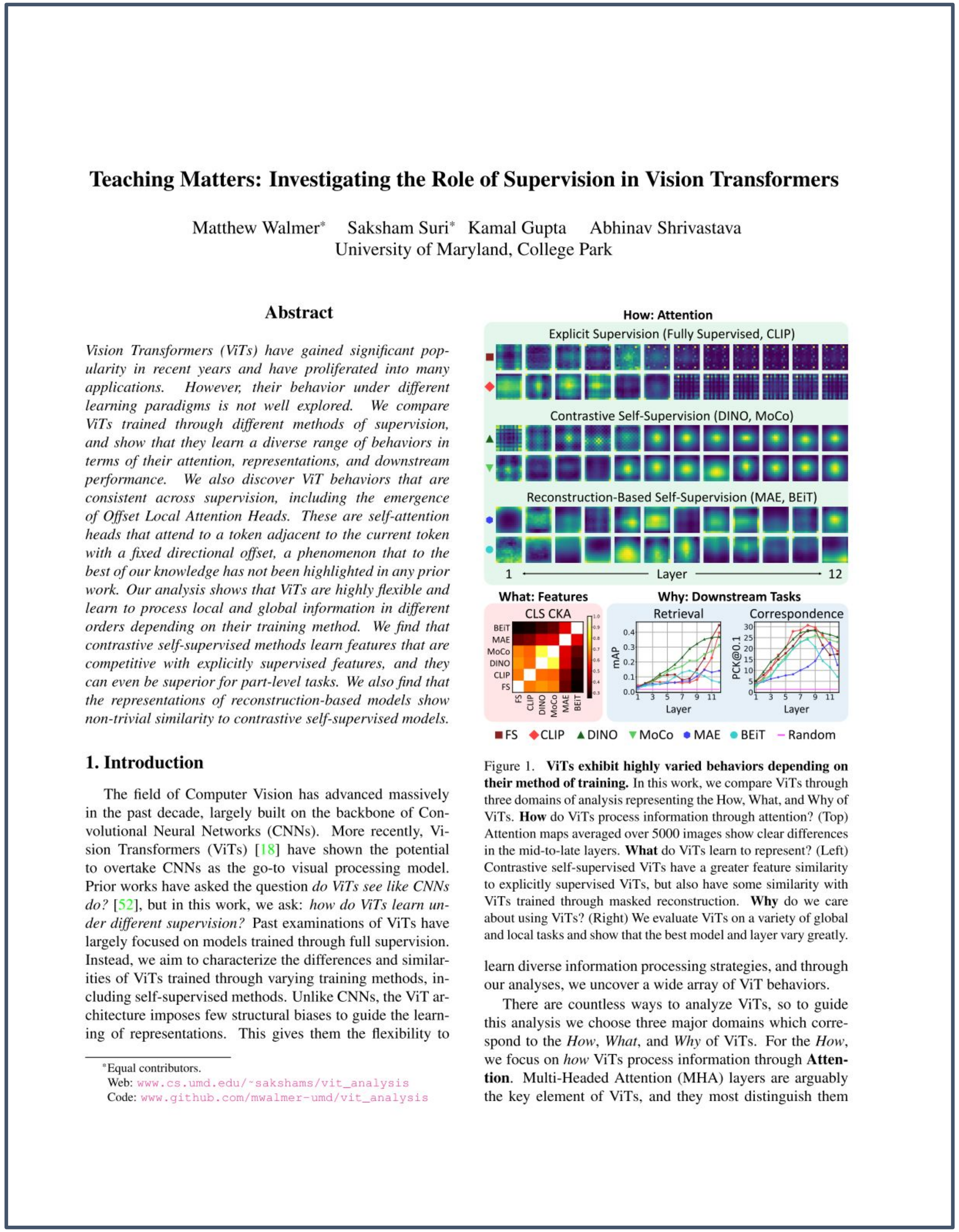
- Surprisingly elevated CLS token feature similarity between DINO and MAE
- Contrastive self-supervised features highly competitive for part-level tasks
- For localized tasks, mid-to-late layer features are better than last layer
- No single “best” training method or layer for all downstream tasks

Thanks for listening!

Poster Session: **TUE-PM-321**

website

code



Teaching Matters: Investigating the Role of Supervision in Vision Transformers

Matthew Walmer* Saksham Suri* Kamal Gupta Abhinav Shrivastava
University of Maryland, College Park

Abstract

Vision Transformers (ViTs) have gained significant popularity in recent years and have proliferated into many applications. However, their behavior under different learning paradigms is not well explored. We compare ViTs trained through different methods of supervision, and show that they learn a diverse range of behaviors in terms of their attention, representations, and downstream performance. We also discover ViT behaviors that are consistent across supervision, including the emergence of Offset Local Attention Heads. These are self-attention heads that attend to a token adjacent to the current token with a fixed directional offset, a phenomenon that to the best of our knowledge has not been highlighted in any prior work. Our analysis shows that ViTs are highly flexible and learn to process local and global information in different orders depending on their training method. We find that contrastive self-supervised methods learn features that are competitive with explicitly supervised features, and they can even be superior for part-level tasks. We also find that the representations of reconstruction-based models show non-trivial similarity to contrastive self-supervised models.

1. Introduction

The field of Computer Vision has advanced massively in the past decade, largely built on the backbone of Convolutional Neural Networks (CNNs). More recently, Vision Transformers (ViTs) [18] have shown the potential to overtake CNNs as the go-to visual processing model. Prior works have asked the question *do ViTs see like CNNs do?* [52], but in this work, we ask: *how do ViTs learn under different supervision?* Past examinations of ViTs have largely focused on models trained through full supervision. Instead, we aim to characterize the differences and similarities of ViTs trained through varying training methods, including self-supervised methods. Unlike CNNs, the ViT architecture imposes few structural biases to guide the learning of representations. This gives them the flexibility to learn diverse information processing strategies, and through our analyses, we uncover a wide array of ViT behaviors.

There are countless ways to analyze ViTs, so to guide this analysis we choose three major domains which correspond to the *How*, *What*, and *Why* of ViTs. For the *How*, we focus on *how* ViTs process information through **Attention**. Multi-Headed Attention (MHA) layers are arguably the key element of ViTs, and they most distinguish them

*Equal contributors.
Web: www.cs.umd.edu/~sakshams/vit_analysis
Code: www.github.com/mwalmer-umd/vit_analysis

Figure 1. **ViTs exhibit highly varied behaviors depending on their method of training.** In this work, we compare ViTs through three domains of analysis representing the How, What, and Why of ViTs. **How** do ViTs process information through attention? (Top) Attention maps averaged over 5000 images show clear differences in the mid-to-late layers. **What** do ViTs learn to represent? (Left) Contrastive self-supervised ViTs have a greater feature similarity to explicitly supervised ViTs, but also have some similarity with ViTs trained through masked reconstruction. **Why** do we care about using ViTs? (Right) We evaluate ViTs on a variety of global and local tasks and show that the best model and layer vary greatly.