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

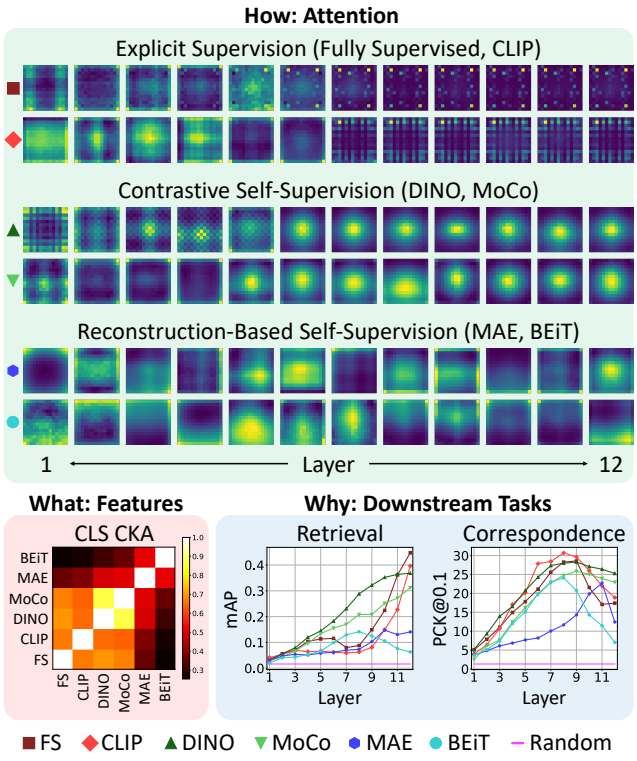

Figure 1. **ViTs exhibit highly varied behaviors depending on their method of training.** In this work, we compare ViTs through three domains of analysis representing the How, What, and Why of ViTs. **How** do ViTs process information through attention? (Top) Attention maps averaged over 5000 images 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.

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

from CNNs. For the *What*, we examine the **Features** of ViTs, as these are typically *what* practitioners take away from them. Finally for the *Why*, we focus on **Downstream Tasks**, which are *why* we care about using ViTs.

Our work unveils that a powerful aspect of the ViT architecture is its local-global dual nature, which plays a role in all three aspects of our analyses. While standard CNNs are restricted to building representations hierarchically from local to global, in a ViT each token can attend to information from any other image region at any time. And unlike popular CNN modifications like Spatial Pyramids [20, 28, 33, 35] and top-down strategies [6, 47, 56], ViTs have the freedom to decide when and where global information should be integrated. In this study, we show that the order and the relative ratio of local and global attention in ViTs varies dramatically based on the method of supervision. We also find clearly different trends in the allocation of attention in the mid-to-late layers of these networks, as highlighted in Figure 2. This local-global dual nature is also embedded into the structure and features of the ViT, which encodes both local spatial tokens and a non-local classifier (CLS) token throughout its entire depth. We analyze the features of ViTs for both the CLS and spatial tokens, and assess how they align with semantics at the image, object, part, and pixel-level. We perform this analysis at every layer of the ViT to show the emergence of different levels of semantic information. Finally, we assess ViTs on a number of local and global downstream tasks.

Overall, our contributions are: **[1]** A detailed comparison of ViTs trained with six different methods, including both fully supervised and self-supervised training. **[2]** A cross-cutting analysis spanning three major domains: Attention, Features, and Downstream Tasks. **[3]** Multiple insights into the inner workings of ViTs to guide future development of ViT variants, training strategies, and applications.

In addition, we summarize some of our key observations about ViT behavior: **[1]** The attention maps of explicitly supervised ViTs devolve into **Sparse Repeating Patterns** in the mid-to-late layers, but the quality of features continues to improve in these layers (Section 4.1). **[2]** All ViTs studied learn to use **Offset Local Attention Heads**, suggesting they are fundamentally necessary in ViTs (Section 4.2). To the best of our knowledge, no prior work has brought attention to this phenomenon. **[3]** ViTs learn to process local and global information in different orders depending on their method of supervision (Section 4.3). **[4]** All ViTs studied differentiate salient foreground objects by the early-to-mid layers (Section 4.4). **[5]** Reconstruction-based self-supervised methods can learn semantically meaningful CLS representations, even when the CLS token is only a placeholder (Section 5.1, 5.2). **[6]** Supervised method's features are the most semantically rich, but contrastive self-supervised methods are comparable or even superior in some cases (Section 5.2, 5.3). **[7]** For localized tasks, the best performance often comes from a mid-to-late layer (Section 6.2). **[8]** There is no single "best" training method or layer for all downstream tasks (Section 6.3).

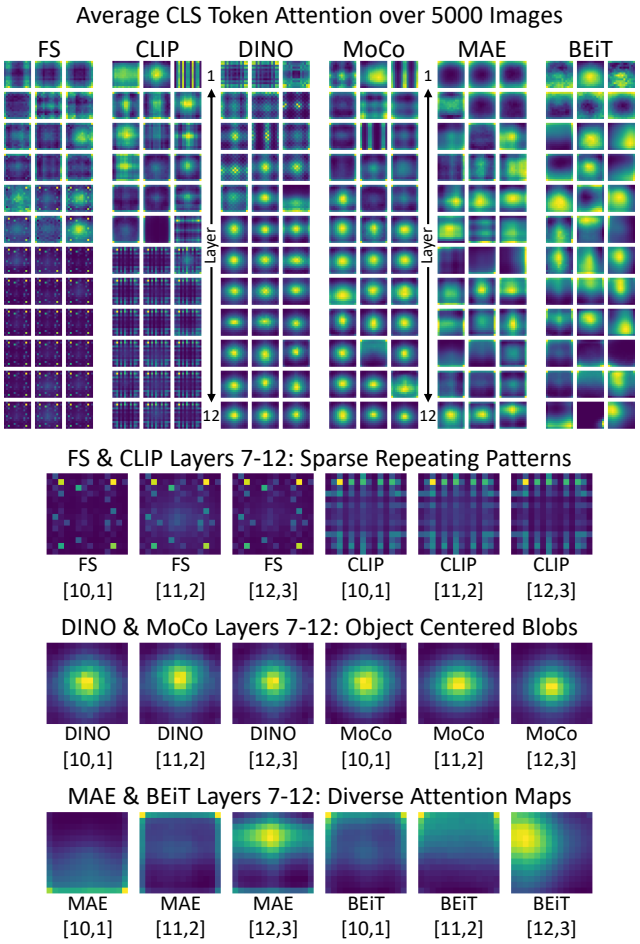

Figure 2. **Clear differences in attention emerge in the mid-to-late layers under different supervision methods**. These plots show the attention maps of CLS tokens averaged over 5000 ImageNet images. Rows indicate layers and columns indicate heads. For brevity, we show only three heads per layer. The bracketed numbers in the lower half denote the layer and head.

2. Related Work

Previous works have attempted to understand the representation quality for both supervised and self-supervised training for Convolutional Neural Networks (CNNs). [5] focuses on understanding the concepts learned by individual neurons while [41] looks at explaining their compositionality in the case of supervised networks. Simultaneously, due to the popularity of self-supervised learning methods [3, 4, 9–11, 26, 27, 29, 40, 43, 60, 73] multiple works have analyzed these representations learned from no labels. Under this umbrella, [13, 62] tried to understand the effect of training data in terms of both the number and type

of samples. Some works [66, 67] analyze the alignment, separability, and uniformity of features while [49] looks at invariance to augmentations like occlusion, illumination, and viewpoint change in the learned representation. [64] looks at the downstream performance of self-supervised networks on fine-grained tasks. Finally, [21, 23, 32] analyze multiple self-supervised methods and compare their performance based on representation similarity and downstream task performance over multiple datasets along with comparisons to supervised methods.

Since the proliferation of ViTs, a number of works have tried to understand and explore the different properties of the representations learned by these networks. A few works [45, 55, 75] have analyzed the robustness of ViT features against corruptions, perturbations, distribution shifts, and adversarial examples while also analyzing the role of self-attention for robustness. [34] benchmarks different pretrained ViTs as backbones for object detection. [7] provides a theoretical understanding of how MAEs work while [53] analyzes attention using convex duality. [61] gives insights to train and use ViTs more efficiently. [44] gives a deeper understanding of how Multi-Headed Attention layers work while comparing and contrasting to how convolution layers behave in terms of the loss landscapes and low-pass/high-pass filtering. [52] compares fully supervised ViTs and ResNets in terms of the local and global information encoded at different depths, the role of skip connections, and the uniformity of representations.

All these prior works either examine the impact of supervision on CNNs or compare CNNs and ViTs trained with full supervision. Some recent and concurrent works have compared the properties of differently supervised ViTs, though typically focused on a particular task and only two methods of supervision at a time. [1] compares the properties of fully supervised and DINO ViT features in the context of dense feature descriptors, and [2] further compares these two across several semantic correspondence tasks. [19] compares fully supervised and CLIP ViTs through feature visualizations. To the best of our knowledge, we present what is to date the broadest and the most in-depth comparison of ViTs with varying supervision, including six different methods covering three supervision subcategories. Additionally, we propose new attention-based analysis methods along with evaluations on multiple downstream tasks focused on both local and global information.

3. Experimental Design

3.1. A Primer on Vision Transformers

Vision Transformers (ViTs) [18] are adapted from Transformers [65] for the Natural Language Processing domain. A ViT consists of an array of tokens, each representing an image patch. In addition, most ViTs include an extra "classifier" or "CLS" token, which is connected to the task-specific output layers during training. ViTs use Multi-Headed Attention (MHA) layers [65], which use a Query-Key-Value system that allows each token to attend to all other tokens with a variable intensity attention map. This is in stark contrast to the limited receptive fields of convolutions. These layers are "multi-headed" because they repeat this process multiple times in parallel, allowing tokens to apply multiple attention strategies concurrently. A ViT architecture includes multiple blocks, each with one MHA layer followed by a position-wise fully connected layer. Unlike CNNs, which usually get narrower in deeper layers, ViTs maintain the same "width" (number of tokens) throughout. There are some transformer variants, like SWiN transformers [37], that introduce a narrowing width, but for our analyses, we focus on only traditional ViTs. Specifically, our primary analysis focuses on ViT-Base models with patch size 16×16 (ViT-B/16) and input size 224×224, which results in a 14×14 spatial token array. ViT-Base has 12 blocks and 12 attention heads per MHA layer. In the Appendix, we provide additional results on a wider range of ViTs, including variations in architecture size and patch size.

3.2. Methods of Supervision

Although a large number of ViT training methods have been proposed in a short span of three years, many of the most popular methods can be loosely categorized into the following three groups. From each group, we select two representative models for in-depth analysis. We further discuss these models' details in Appendix B.2.

Explicit Supervision. These models are trained with an explicit objective that is defined either by human annotations or by labels derived from another source, like paired image captions. For this category, we use a Fully Supervised (FS) ViT pretrained on ImageNet21k and fine-tuned on ImageNet1k [58, 69], as well as a CLIP ViT [51].

Self-Supervision (Contrastive). Self-supervised learning methods broadly attempt to train a model through a pretext task that can be directly derived from the input data. Among the more popular pretext tasks are contrastive learning methods [27, 70] which generally present a model with multiple augmented views of the same image alongside distractor views of other images. The model must learn to identify which of the views came from the same image. For this category, we select DINO [9] and MoCo-v3 [12] which we denote simply as MoCo for the rest of this paper.

Self-Supervision (Reconstruction). Another popular category of self-supervision is reconstruction methods, which train models to predict the missing content from masked or otherwise corrupted images. We select MAE [26] and BEiT [4] for this category. Note that MAE has a separate decoder which is discarded after pretraining,

while BEiT's decoder is learned in the same ViT. This has a strong impact on the behavior of the later layers of BEiT.

3.3. Datasets

We study the ViTs on multiple datasets and downstream tasks. Unless otherwise specified, we use ImageNet-50 [63], a subset of ImageNet [16] which narrows the dataset down to 50 representative categories. We sample 100 images per class to create a diverse collection of 5000 images. We additionally use PartImageNet [25] to measure Attention Saliency and part-level feature purity, as well as COCO [36] to measure object-level feature purity. We use revisited [50] Oxford [46] (ROxford5k) for evaluating image retrieval, DAVIS [48] for video segmentation, and SPair-71k [39] for keypoint correspondence.

3.4. Proposed Analyses

Our analysis is broadly divided into three domains covering the *How*, *What*, and *Why* of ViTs:

***How* ViTs process local/global information (Attention).** Do self-attention heads learn to operate in different ways depending on their method of training? Are there distinctive modes of attention behavior? How does supervision impact the processing order of local and global information?

***What* we take away from ViTs (Features).** How do the final and intermediate representations of a ViT change depending on the method of supervision? Are these trends similar or different for CLS *vs*. spatial tokens?

***Why* we use ViTs (Downstream Tasks).** Which forms of supervision are best suited for different downstream tasks? Which layers of a ViT produce features that are best for different local and global tasks?

4. Attention Analysis

Multi-Headed Attention layers are one of the defining components of the Transformer architecture, and the attention maps they generate can give key insights into what is similar or different about ViTs trained through different methods. We perform an in-depth examination of the self-attention maps of ViT-B/16 models at every layer. Through this study, we uncover a diverse range of attention head behavioral modalities. Additional visualizations are provided in Appendix C for a wide range of ViT variants.

4.1. Attention Visualizations

We start by examining the attention maps of the CLS tokens of each head and layer. To gain a comprehensive understanding of each head's behavior, we compute the average attention maps over 5000 ImageNet images, as shown in Figure 2. For brevity, we display only three heads per layer, but complete plots can be found in Appendix C.1, along with additional visualizations for spatial token attention and individual input images.

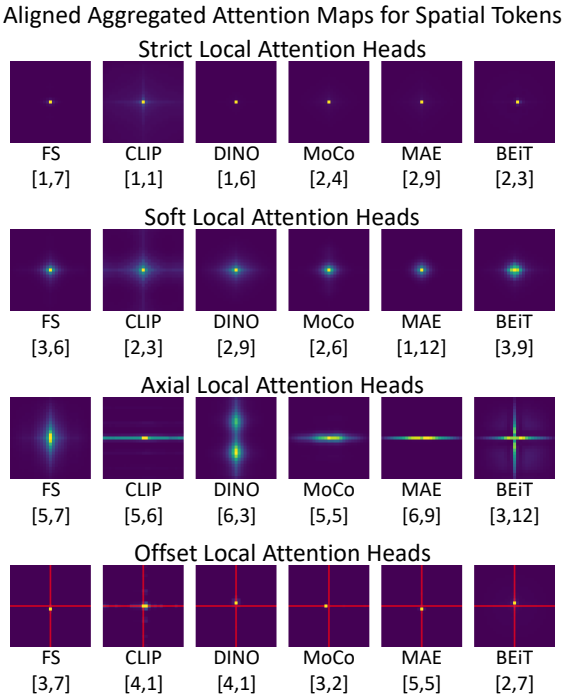

Figure 3. **Multiple distinct forms of local attention exist.** We visualize spatial token attention using Aligned Aggregated Attention Maps, and highlight different types of local attention heads, including Strict, Soft, Axial, and Offset Local Attention Heads. In row 4 we draw the mid-lines in red as a visual aid.

One of the clearest differences can be seen by comparing the mid-to-final layers. For the contrastive self-supervised methods, DINO and MoCo, the attention maps tend to be centered blobs. These heads tend to focus on salient foreground objects, so these blobs simply reflect object-centered photography bias. For the reconstruction-based methods, MAE and BEiT, we see a more diverse group of attention maps. This is likely because these methods must reconstruct all image regions, and thus their attention in the final layers must be more diverse and cover more of the image. Finally, for the explicitly supervised methods, FS and CLIP, the mid-to-final layers do not focus on salient object regions and instead focus on **Sparse Repeating Patterns** with seemingly no spatial meaning. This occurs for both the CLS tokens and spatial tokens, and the patterns are repeated across both heads and layers. We hypothesize that these patterns occur because the mid-to-late layers are no longer focused on parsing the scene structure, and instead are using their processing power to generate their final decisions for their respective tasks. This phenomenon helps to explain why the attention maps of the later layers of fully-supervised ViTs are poorly suited for segmentation tasks, as was observed by [9].

4.2. Emergence of Offset Local Attention Heads

It has been shown that ViTs use a mixture of short-range "local" and long-range "global" attention heads in any given layer [9, 18]. To gain a better understanding of local attention, we propose a visualization strategy of **Aligned Aggregated Attention Maps (AAAMs)**. We extract all spatial token attention maps for 5000 ImageNet images, but before averaging them, we first realigned them so the current spatial token is always in the center of the array. Additional samples of AAAMs are provided in Appendix C.1. Studying these aligned views reveals multiple forms of local attention, shown in Figure 3. We find Strict Local Attention Heads, which attend almost completely to their own position, as well as Soft Local Attention Heads, which attend to a wider neighborhood around them. We also find Axial Local Attention Heads, which are elongated to attend to the local neighborhood along one or both spatial axes.

But perhaps the most noteworthy type of head we observe is the **Offset Local Attention Head**. These are heads that attend locally, but to a point or region offset from the current token in a vertical or horizontal direction. We find instances of Offset Local Attention Heads in all the models examined, suggesting they are fundamentally necessary for ViTs. To the best of our knowledge, ours is the first work to draw attention to this phenomenon. We believe that such heads are absolutely necessary because ViTs, unlike CNNs, do not have an easy built-in way to test if two features occur next to each other *with a particular spatial arrangement*. In a CNN, this type of check is naturally embedded into the convolution operator. But in a ViT, there is no inductive bias to induce such a check. For comparison, Soft Local Attention Heads are able to identify if a certain feature is near another feature, but they cannot identify their specific directional arrangement due to their symmetrical attention pattern. The existence of Offset Local Attention Heads implies one possible path for improvement for the ViT architecture, possibly by adding a self-attention variant that introduces some implicit directional structure.

4.3. Average Attention Distance

We measure the Average Attention Distance [18, 52] of each head to assess if particular heads have a short-range "local" focus or a long-range "global" focus. This metric is computed by measuring the distance from each spatial token to all other tokens and taking a weighted average using the attention map. We normalize the distances so the token grid is embedded on a 1×1 square. [52] observed that for a well-trained fully-supervised ViT, the early layers have a mixture of local and global attention heads, while the later layers have only global attention heads.

Figure 4 (left) shows the Average Attention Distances of all heads organized by layer and model. Like [52], we see that most layers use a mixture of local and global attention heads, however, we also find that the ordering of local and global processing varies greatly with the supervision type. FS, CLIP, DINO, and MoCo all use exclusively global attention heads in the last layers, but the reconstruction-based methods MAE and BEiT use a diverse range of heads in their later layers. Figure 4 (right) compares the combined Average Attention Distances at the per-layer level. In all models, we observe a greater number of global attention heads in the initial layers, followed by decreased distances around layers 3-6. This result is again in contrast to [52]. The behaviors diverge in the mid-to-late layers. For the models trained with explicit or contrastive supervision, the Attention Distance trends upward in the later layers. For the reconstruction-based methods, the Average Attention Distances stay lower. These results show that, unlike CNNs, ViTs can learn a variable local/global processing order depending on the training method used.

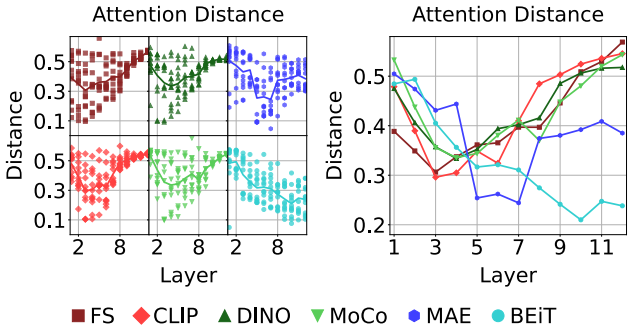

Figure 4. **Different methods of supervision lead to different orderings and ratios of local and global processing.** We show the Average Attention Distance of all ViT attention heads organized by layer (left), and the per-layer averages (right).

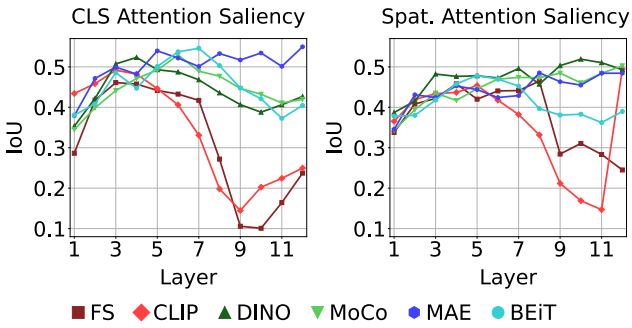

Figure 5. **Attention IoU with salient content plateaus early for all ViTs evaluated.** We calculate the alignment of ground-truth segmentation masks with CLS token attention maps (left) and the average of spatial token attention maps (right).

4.4. Attention Alignment with Salient Content

One of the most desirable (and exploitable) features of DINO is that the CLS token attention maps of the last layer tend to be well-aligned with salient foreground objects [9].

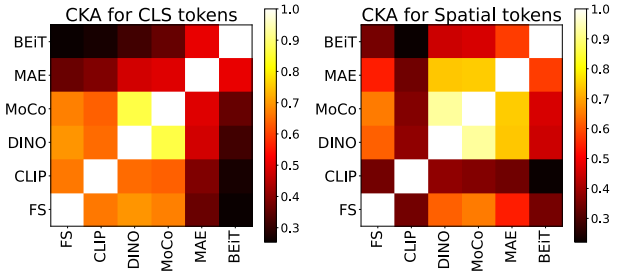

Figure 6. CKA similarity between final layer features of different ViTs for their CLS tokens (left) and spatial tokens (right).

Several methods propose to use DINO attention maps, feature maps, or a combination of the two to generate segmentations in a self-supervised manner [24, 57, 68]. We conduct a quantitative analysis of this property at all layers of the ViTs, both to measure the usefulness of masks and to assess how early the ViTs differentiate salient object regions. Like [9], we threshold the CLS token attention masks keeping 60% of the total attention mass. We then compute the Intersection over Union (IoU) of said masks with ground-truth segmentations from PartImageNet [25]. As an alternative to CLS token attention, we also extract masks using the average of spatial token attention maps. We present results for the single "best" head per layer in Figure 5.

We see a clear drop in FS and CLIP mask IoU around the middle of the network, which directly corresponds to the emergence of the Sparse Repeating Patterns observed in Section 4.1. We also find that the IoUs plateau around layers 3-6 for all networks. This demonstrates that ViT models already have a solid understanding of foreground/background separation by the middle layers. While the later attention maps of FS and CLIP are much worse than their self-supervised counterparts, their early-to-mid layers are more comparable. We find that MoCo, MAE, and BEiT can all produce attention maps with IoUs that are comparable with DINO. In addition, we see that the average of spatial tokens produces maps that are comparable with the CLS token, and for CLIP the IoU increases greatly in the final layer.

5. Feature Analysis

In this section, we directly compare ViT features across models and layers using Centered Kernel Alignment (CKA) [14, 31]. We also study unsupervised clustering performance to compare global and local semantic information in the learned representations. We present addition analysis in Appendix D.4 focused on ViT residual connections.

5.1. Last Block Feature Comparisons

Comparing representations is non-trivial due to varying feature sizes, large feature representations, and lack of alignment between them. To overcome this, we use batched Centered Kernel Alignment (CKA) [14, 31, 42] which can align features and compute a similarity score. We compare the last layer outputs for each model.

Figure 6 (left) shows that the CLS token representations are usually similar for similar supervision strategies (explicit, contrastive, reconstruction). The contrastive methods, MoCo and DINO, show very high similarity to each other, indicating that the CLS token encodes the same type of information for both these methods. There is also an increased level of similarity between the explicitly supervised methods, FS and CLIP, and the contrastive methods. Interestingly, we see that MAE has as high a similarity with DINO and MoCo as it does with BEiT. This result is surprising because MAE's CLS token has no explicit training objective or loss, and the way these approaches are trained is very different. This presents evidence that training autoencoders with a high masking percentage indeed forces the model to learn image-level semantics.

In Figure 6 (right), we look at the similarity of the last-layer spatial token representations. Unlike the CLS token representations, CLIP and FS have low similarity in their spatial representations. The self-supervised methods DINO, MAE, and MoCo show a high level of similarity to each other, and a lower level of similarity to BEiT. MoCo and DINO show the highest similarity due to their similar kind of self-supervision. Once again, MAE has a high similarity to MoCo and DINO despite their very different supervision.

5.2. Feature Clustering for Global Semantics

Through this analysis, we aim to test how well the learned CLS and spatial token representations encode global (image-level) semantic information at every layer. We extract the CLS token features from the end of each block for 5000 ImageNet images, and we generate k-Means cluster assignments with $k = 50$. We present results for cluster purity measured with respect to ground truth image labels, but additional clustering metrics are also presented in Appendix D.5. For the spatial tokens, we follow the same process except we average-pool over all positions before clustering. We also compute a random chance score by replacing the ViT features with Gaussian random noise.

For CLS token features, shown in Figure 7 (left), cluster purity improves with depth with the exception of the last layers of 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. Unsurprisingly, FS achieves the best cluster purity, followed by CLIP. The contrastive methods, DINO and MoCo, achieve scores close to the explicitly supervised methods. The reconstruction-based methods, MAE and BEiT, have the lowest cluster purity, but they are still above random chance, which again indicates that they do learn to encode some image-level semantic information in their CLS tokens. Also, we find that semantic information emerges earlier for

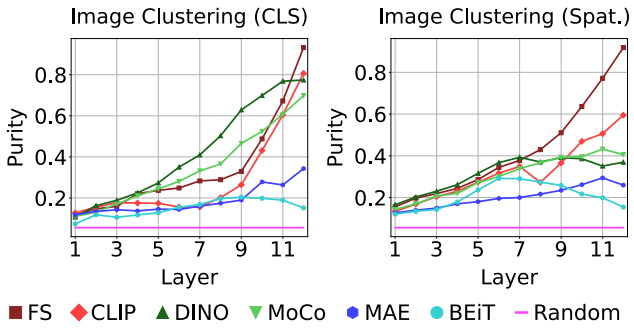

Figure 7. Clustering purity analysis with image-level labels in ImageNet-50 for CLS features (left) and average-pooled spatial token features (right).

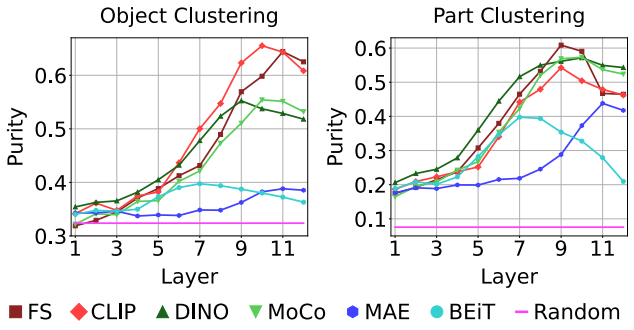

Figure 8. Clustering purity of spatial token features gathered at the object-level in COCO (left) and part-level in PartImageNet (right).

DINO and MoCo. For the spatial token features, shown in Figure 7 (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. For all other ViTs, the spatial feature purity is lower in the final layers, but is comparable in layers 1-7.

5.3. Feature Clustering for Local Semantics

We next measure how well the spatial token features differentiate salient image content at the object or part-level using COCO [36] and PartImageNet [25] respectively. We use a tiling strategy to extract a denser array of features, which we detail in Appendix B.4. Using ground truth segmentation masks, we extract and average the features of the tokens overlapping with the masks. This generates a collection of object-level or part-level features which we cluster just like Section 5.2. The results for object-level features are shown in Figure 8 (left). We see that the supervised methods CLIP and FS have the highest feature purity by far, followed by the contrastive methods DINO and MoCo. The purity is much lower for the reconstruction methods MAE and BEiT. For part-level features, shown in Figure 8 (right), FS achieves the best purity, but the contrastive methods are very competitive in this case, surpassing CLIP completely. In addition, while still being the lowest scoring, MAE and BEiT are much more competitive at the part-level. Like the image-level purity, the object and part-level feature purity tend to improve with depth, but the purity peaks early around layers 9 to 11. The peak for BEiT is even earlier, likely due to its integrated decoder.

6. Downstream Task Analysis

Finally, we analyze the performance of these models on downstream tasks that can be performed directly without any fine-tuning or training. We follow the evaluation protocols of [9, 30] for k-NN classification, image retrieval, and video object segmentation. We also perform keypoint correspondence as a more local-focused task. Again we compute random chance scores by replacing all ViT features with Gaussian noise.

6.1. Global Tasks

ImageNet Classification. We perform k-Nearest Neighbor (k-NN) image classification on ImageNet [16] with $k = 20$. We use the CLS token features from each network and assign the label for a test sample based on the training set features and labels. As can be seen from Figure 9 (left), FS performs the best as it has been trained to classify the same dataset. DINO and MoCo follow a similar trend as Section 5.2 and better encode semantic information in the earlier layers. FS and CLIP also follow a similar trend where their performance shoots up in the last few layers. It is also interesting to see how MAE and BEiT, for which the CLS tokens have no explicit objective, do better than chance, although MAE is considerably better than BEiT.

Image Retrieval. Similar to k-NN classification, we utilize the CLS token representation for retrieval. We evaluate on ROxford5k [50] for the Medium (M) split and report the Mean Average Precision (mAP). In Appendix E.2 we also report results for the Hard (H) split and the RParis6k [50] dataset, which follow similar trends. The results, shown in Figure 9 (right), align closely with those for k-NN Classification on ImageNet. FS performs the best followed by CLIP and then DINO and MoCo, and finally by MAE and BEiT with the lowest performance. We hypothesize that the local/global crops used in DINO training help it perform competitively in these global tasks.

6.2. Local Tasks

DAVIS Segmentation Propagation. DAVIS Segmentation Propagation is a dense prediction-based video localized task where frame-by-frame features are used to propagate the first frame segmentation mask to subsequent frames. Like Section 5.3, we use a tiling-based dense feature extraction strategy. Results are shown in Figure 10 (left). The contrastive techniques of DINO and MoCo perform the best while FS and CLIP, which are more image-level approaches, face a drop in performance towards the

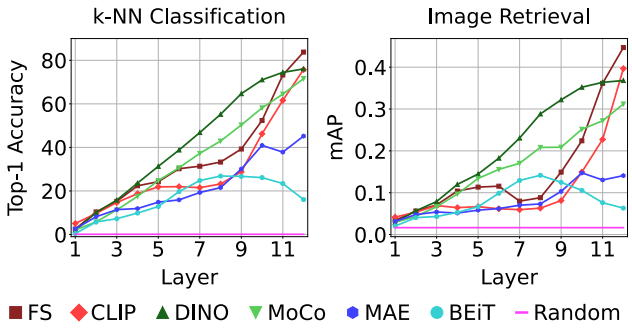

Figure 9. Global (image-level) downstream task analysis using the CLS token. We present k-NN classifier Top-1 Accuracy on ImageNet (left) and image retrieval mAP on ROxford5k (right).

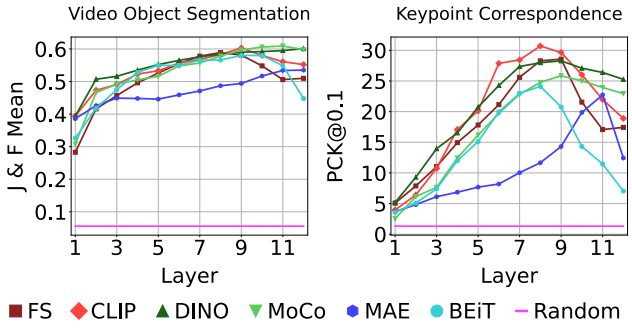

Figure 10. Local (pixel-level) downstream task analysis using the dense spatial token features. We perform DAVIS video segmentation (left) and SPair-71k keypoint correspondence (right).

later layers. The local reconstruction-based methods, MAE and BEiT, are also much more competitive in this task. These results show that for purposes of constructing highly descriptive local features, contrastive methods like DINO and MoCo and reconstructive methods like BEiT can surpass features trained with explicit image-level supervision.

Keypoint Correspondence. We choose keypoint correspondence as an additional local-focused downstream task. Given an image with annotated keypoints, the model must predict the position of corresponding keypoints in a paired image with similar content. We use the SPair-71k [39] dataset and follow the evaluation protocol of [1] and report the Percentage of Correct Keypoints (PCK) [72]. The results are summarized in Figure 10 (right). CLIP excels at this task, closely followed by both FS and DINO. Meanwhile, MoCo, BEiT, and MAE are all very competitive also. The position of the best layer varies significantly, from 8 for CLIP and BEiT to 11 for MAE.

6.3. Summary of Downstream Tasks

We summarize the best results for all downstream tasks in Table 1. We denote the best-performing layers in parenthesis. These results show that ViTs with different supervision methods **[1]** peak at different layers, and **[2]** perform best at different tasks. In image-level tasks like k-NN classification and retrieval, usually, the last layer works the best. For localized tasks like keypoint correspondence and video object segmentation, most models' peak performance happens a few layers before the last one. This shows that always picking the last layer output is not optimal.

Table 1. Best performance for each ViT on each downstream task with the corresponding best layer in parenthesis.

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

7. Conclusion

In this work, we have performed an in-depth comparison of ViTs trained through different methodologies by examining their attention patterns, learned representations, and downstream task performance. We review some of the key findings of our analyses. First, different methods of supervision lead to ViTs that process local and global information in different orders. All ViTs have heads that align well with salient image content, but for the explicitly supervised models, the late-layer attention maps change into Sparse Repeating Patterns. In addition, all ViTs examined have learned to use Offset Local Attention Heads in multiple layers. While explicitly supervised ViTs have the most semantically rich representations at the image level, contrastive methods are competitive, and reconstruction-based methods can also learn meaningful CLS token representations even though said token is a placeholder and has no explicit supervision. Finally, there is no single best model for all the downstream tasks, and the best layer to extract representations from also varies greatly by task and model, so one should not simply take the last layer representation. ViTs have shown a great deal of potential, and we expect they will become more widely used in the coming years. We hope these insights can help with the future development of losses and architectures for Vision Transformers.

Acknowledgements. This project was partially funded by DARPA SAIL-ON (W911NF2020009), DARPA SemaFor (HR001119S0085), and DARPA GARD (HR00112020007) programs. We would like to thank our collegues Vatsal Agarwal, Matthew Gwilliam, Alex Hanson, Pulkit Kumar, Saketh Rambhatla, and Gaurav Shrivastava for their feedback on this work.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022. 3, 8, 24

[2] Mehmet Aygün and Oisin Mac Aodha. Demystifying unsupervised semantic correspondence estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 125–142. Springer, 2022. 3

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 2

[4] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2, 3, 12, 14

[5] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 2

[6] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2956–2964, 2015. 2

[7] Shuhao Cao, Peng Xu, and David A Clifton. How to understand masked autoencoders. *arXiv preprint arXiv:2202.03670*, 2022. 3

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 14

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 3, 4, 5, 6, 7, 12

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 3, 12

[13] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14755–14764, 2022. 2

[14] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13:795–828, 2012. 6, 18

[15] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 12

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 7

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 14

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 5

[19] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022. 3

[20] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1458–1465. IEEE, 2005. 2

[21] Tom George Grigg, Dan Busbridge, Jason Ramapuram, and Russ Webb. Do self-supervised and supervised methods learn similar visual representations? *arXiv preprint arXiv:2110.00528*, 2021. 3

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 14

[23] Matthew Gwilliam and Abhinav Shrivastava. Beyond supervised vs. unsupervised: Representative benchmarking and analysis of image representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9642–9652, 2022. 3

[24] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022. 6

[25] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, and Alan Yuille. Partimagenet: A large, high-quality dataset of parts. *arXiv preprint arXiv:2112.00933*, 2021. 4, 6, 7, 18

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3, 12, 14

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross

Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 3, 12

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 2

[29] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[30] Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020. 7

[31] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 6, 18

[32] Klemen Kotar, Gabriel Ilharco, Ludwig Schmidt, Kiana Ehsani, and Roozbeh Mottaghi. Contrasting contrastive self-supervised representation learning pipelines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9949–9959, 2021. 3

[33] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006. 2

[34] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 3

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4, 7, 18

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[39] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019. 4, 8, 24

[40] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[41] Jesse Mu and Jacob Andreas. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163, 2020. 2

[42] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020. 6, 18

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[44] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 3

[45] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022. 3

[46] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 4

[47] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European conference on computer vision*, pages 75–91. Springer, 2016. 2

[48] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 4

[49] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *Advances in Neural Information Processing Systems*, 33:3407–3418, 2020. 3

[50] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 4, 7

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 12

[52] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 1, 3, 5, 23

[53] Arda Sahiner, Tolga Ergen, Batu Ozturkler, John Pauly, Morteza Mardani, and Mert Pilanci. Unraveling attention via convex duality: Analysis and interpretations of vision transformers. *arXiv preprint arXiv:2205.08078*, 2022. 3

[54] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 24

[55] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers. *arXiv preprint arXiv:2103.15670*, 2021. 3

[56] Abhinav Shrivastava, Rahul Sukthankar, Jitendra Malik, and Abhinav Gupta. Beyond skip connections: Top-down modulation for object detection. *arXiv preprint arXiv:1612.06851*, 2016. 2

[57] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021. 6

[58] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 3, 12

[59] Anand Subramanian. Torch cka. https://github.com/AntixK/PyTorch-Model-Compare, 2021. 18

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 2

[61] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Jakob Verbeek, and Hervé Jégou. Three things everyone should know about vision transformers. *arXiv preprint arXiv:2203.09795*, 2022. 3

[62] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc V Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *Advances in Neural Information Processing Systems*, 34:16238–16250, 2021. 2

[63] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020. 4

[64] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 3

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[66] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021. 3

[67] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 3

[68] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 6

[69] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019. 3, 12

[70] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[71] Kun Xu, Yong Li, Tao Ju, Shi-Min Hu, and Tian-Qiang Liu. Efficient affinity-based edit propagation using kd tree. *ACM Transactions on Graphics (TOG)*, 28(5):1–6, 2009. 24

[72] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 8, 24

[73] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 2

[74] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 12

[75] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022. 3