Intriguing Properties of Vision Transformers

Abstract

An important question is how such flexibility (in attending image-wide context conditioned on a given patch) can facilitate handling nuisances in natural images e.g., severe occlusions, domain shifts, spatial permutat

We show and analyze the following intriguing properties of ViT:
(a) Transformers are highly robust to severe occlusions, perturbations, and domain shifts, e.g., retain as high as 60% top-1 accuracy on ImageNet even after randomly occluding 80% of the image content.
(b) The robustness towards occlusions is not due to texture bias, instead, we show that ViTs are significantly less biased towards local textures, compared to CNNs.
When properly trained to encode shape-based features, ViTs demonstrate shape recognition capability comparable to that of the human visual system, previously unmatched in the literature.
(c) Using ViTs to encode shape representation leads to an interesting consequence of accurate semantic segmentation without pixel-level supervision.
(d) Off-the-shelf features from a single ViT model can be combined to create a feature ensemble, leading to high accuracy rates across a range of classification datasets in both traditional and few-shot learning para
We show effective features of ViTs are due to flexible and dynamic receptive fields possible via self-attention mechanisms.

Introduction

In this paper, we compare the performance of transformers with convolutional neural networks (CNNs) for handling nuisances (e.g., occlusions, distributional shifts, adversarial and natural perturbations) and general
We are intrigued by the fundamental differences in the operation of convolution and self-attention, that have not been extensively explored in the context of robustness and generalization.

Given a query embedding, self-attention finds its interactions with the other embeddings in the sequence, thereby conditioning on the local content while modeling global relationships [7].
In contrast, convolutions are content-independent as the same filter weights are applied to all inputs regardless of their distinct nature.
Given the content-dependent long-range interaction modeling capabilities, our analysis shows that ViTs can flexibly adjust their receptive field to cope with nuisances in data and enhance expressivity of the represen

Our systematic experiments and novel design choices lead to the following interesting findings:
1 ViTs demonstrate strong robustness against severe occlusions for foreground objects, non-salient background regions and random patch locations, when compared with state-of-the-art CNNs.

2 When presented with texture and shape of the same object, CNN models often make decisions based on texture [9].
In contrast, ViTs perform better than CNNs and comparable to humans on shape recognition.
This highlights robustness of ViTs to deal with significant distribution shifts e.g., recognizing object shapes in less textured data such as paintings.

3 Compared to CNNs, ViTs show better robustness against other nuisance factors such as spatial patch-level permutations, adversarial perturbations and common natural corruptions (e.g., noise, blur, contrast and p
However, similar to CNNs [10], a shape-focused training process renders them vulnerable against adversarial attacks and common corruptions.

4 Apart from their promising robustness properties, off-the-shelf ViT features from ImageNet pretrained models generalize exceptionally well to new domains

To this end, we propose an architectural modification to DeIt to encode shape-information via a dedicated token that demonstrates how seemingly contradictory cues can be modeled with distinct tokens within the sa

Related Work

CNNs have shown state-of-the-art performance in independent and identically distributed (i.i.d) settings but remain highly sensitive to distributional shifts; adversarial noise [11, 12], common image corruptions [13], a

Our analysis indicates that large ViT models have less texture bias and give relatively higher emphasis to shape information.

Our findings are consistent with a concurrent recent work that demonstrates the importance of this trend on human behavioural understanding and bridging the gap between human and machine vision [22].

In comparison, we show how shape-focused learning can impart similar capability in the image-level supervised ViT models, without any pixel-level supervision.

In a similar spirit, we study the generalization of off-the-shelf features of ViT in comparison to CNN.

The receptive field of Transformer based models covers the entire input space, a property that resembles handcrafted features [25], but ViTs have higher representative capacity.
This allows ViT to model global context and preserve the structural information compared to CNN [26].
This work is an effort to demonstrate the effectiveness of flexible receptive field and content-based context modeling in ViTs towards robustness and generalization of the learned features.

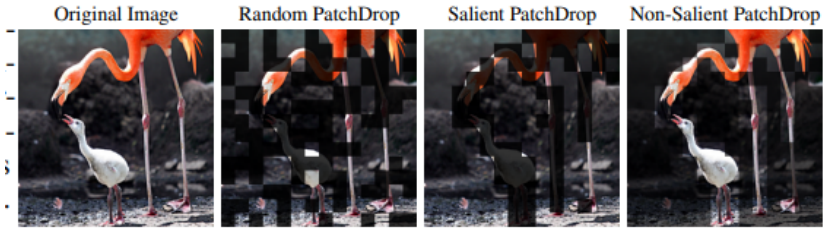
Intriguing Properties of Vision Transformers

Are Vision Transformers Robust to Occlusions?

The receptive field of a ViT spans over the entire image and it models the interaction between the sequence of image patches using self-attention [26, 27].

Occlusion Modeling

While there can be multiple ways to define occlusion, we adopt a simple masking strategy, where we select a subset of the total image patches, M < N, and set pixel values of these patches to zero to create an occlu
We refer to this approach as PatchDrop.
We experiment with three variants of our occlusion approach, (a) Random PatchDrop, (b) Salient (foreground) PatchDrop, and (c) Non-salient (background) PatchDrop.

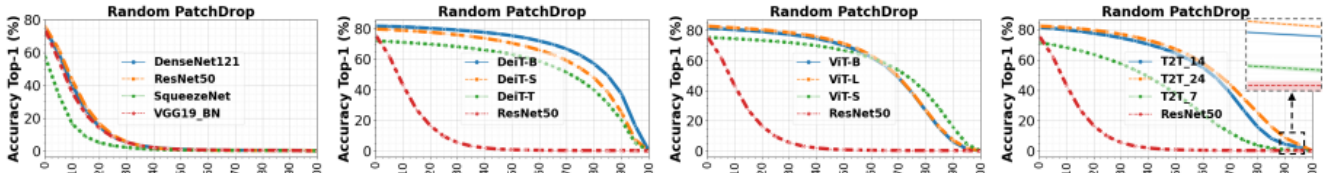


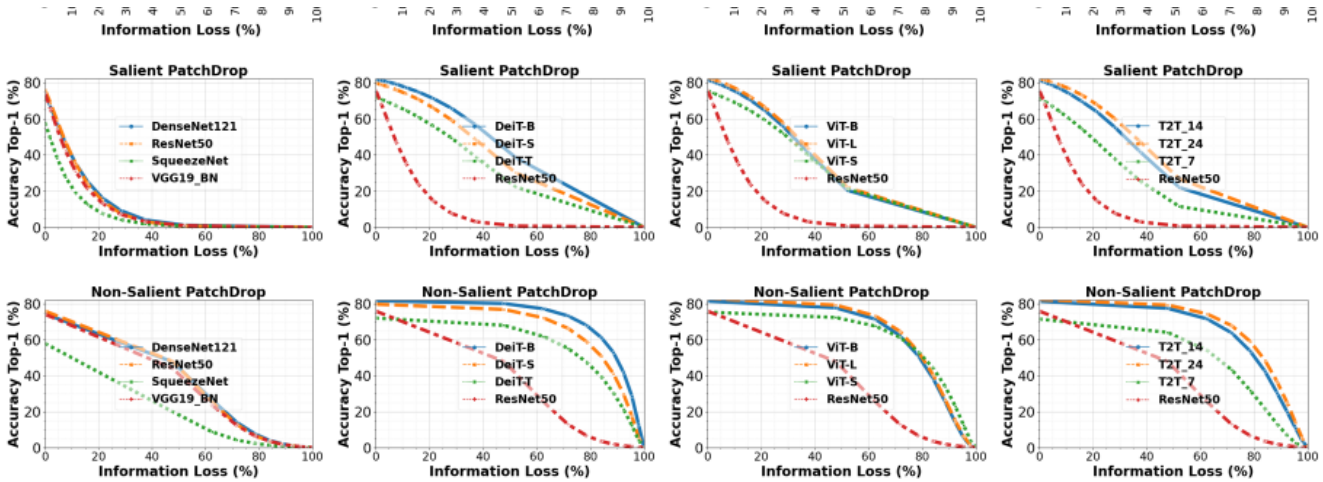
Random PatchDrop
A subset of M patches is randomly selected and dropped (Fig. 2).

Salient (foreground) PatchDrop:
Thus, it is important to study the robustness of ViTs against occlusions of highly salient regions.
We leverage a self-supervised ViT model DINO [23] that is shown to effectively segment salient objects.
In particular, the spatial positions of information flowing into the final feature vector (class token) within the last attention block are exploited to locate the salient pixels.
This allows to control the amount of salient information captured within the selected pixels by thresholding the quantity of attention flow.

Non-salient (background) PatchDrop:
The patches containing the lowest Q% of foreground information are selected and dropped here.

Robust Performance of Transformers Against Occlusions



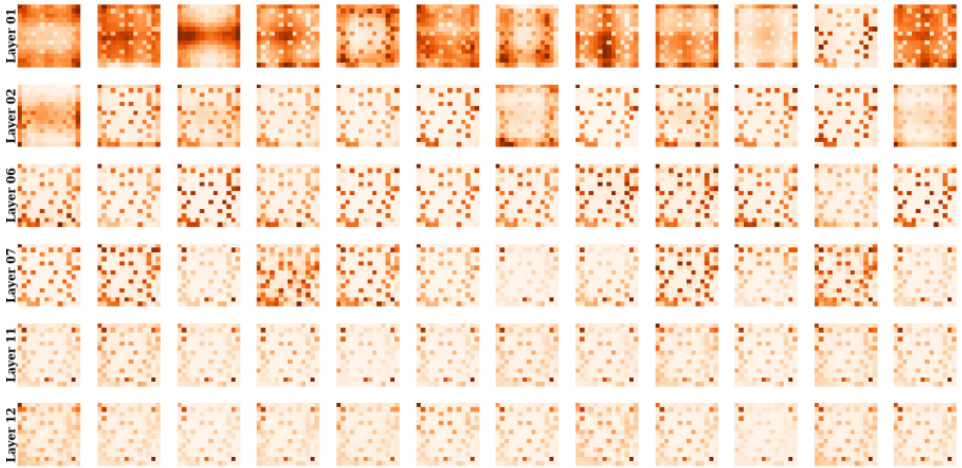


CNNs perform poorly when 50% of image information is randomly dropped.

This finding is consistent among different ViT architectures [2, 3, 4].

Similarly, ViTs show significant robustness to the foreground (salient) and background (non-salient) content removal.

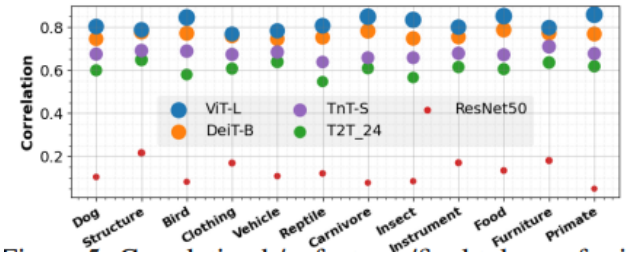
ViT Representations are Robust against Information Loss



While initial layers attend to all areas, deeper layers tend to focus more on the leftover information in non-occluded regions of an image.

We then study if such changes from initial to deeper layers lead to token invariance against occlusion which is important for classification.

Class tokens from transformers are significantly more robust and do not suffer much information loss as compared to ResNet50 features (Table 1).



Given the intriguing robustness of transformer models due to dynamic receptive fields and discriminability preserving behaviour of the learned tokens, an ensuing question is whether the learned representations in ViTs are biased towards texture or not.

One can expect a biased model focusing only on texture to still perform well when the spatial structure for an object is partially lost.

Shape vs. Texture: Can Transformer Model Both Characteristics?

We first carry out similar analysis and show that ViT models preform with a shape-bias much stronger than that of a CNN, and comparably to the ability of human visual system in recognizing shapes.

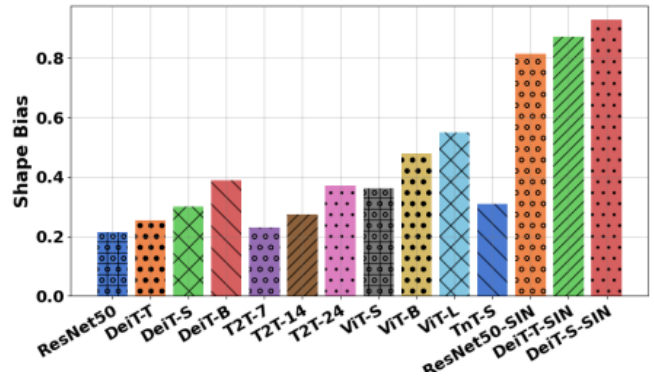
However, this approach results in a significant drop in accuracy on the natural images.

To address this issue, we introduce a shape token into the transformer architecture that learns to focus on shapes, thereby modeling both shape and texture related features within the same architecture using a disti

As such, we distill the shape information from a pretrained CNN model with high shape-bias [9].

Training without Local Texture

In this approach, we first remove local texture cues from the training data by creating a stylized version of ImageNet [9] named SIN.



Thus, we train models on SIN without applying any augmentation, label smoothing or mixup.

Shape Distillation

We introduce a new shape token and adapt attentive distillation [3] to distill shape knowledge from a CNN trained on the SIN dataset (ResNet50-SIN [9]).

We observe that ViT features are dynamic in nature and can be controlled by auxiliary tokens to focus on the desired characteristics.

Model	Distilled	Token Type	ImageNet top-1 (%)	Shape Bias
DeiT-T-SIN	✗	cls	40.5	0.87
DeiT-T-SIN	✓	cls	71.8	0.35
DeiT-T-SIN	✓	shape	63.4	0.44
DeiT-S-SIN	✗	cls	52.5	0.93
DeiT-S-SIN	✓	cls	75.3	0.39
DeiT-S-SIN	✓	shape	67.7	0.47

a single ViT model can exhibit both high shape and texture bias at the same time with separate tokens (Table 3).

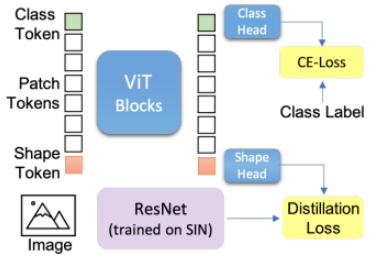


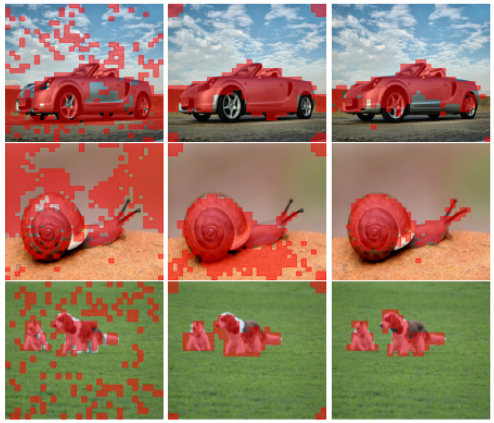
Figure 7: Shape Distillation.

We achieve more balanced performance for classification as well as shape-bias measure when the shape token is introduced (Fig. 7).
This confirms our hypothesis on modeling distinct features with separate tokens within ViTs, a unique capability that cannot be straightforwardly achieved with CNNs.

Shape-biased ViT Offers Automated Object Segmentation:

Model	Distilled	Token Type	Jaccard Index
DeiT-T-Random	✗	cls	19.6
DeiT-T	✗	cls	32.2
DeiT-T-SIN	✗	cls	29.4
DeiT-T-SIN	✓	cls	40.0
DeiT-T-SIN	✓	shape	42.2
DeiT-S-Random	✗	cls	22.0
DeiT-S	✗	cls	29.2
DeiT-S-SIN	✗	cls	37.5
DeiT-S-SIN	✓	cls	42.0
DeiT-S-SIN	✓	shape	42.4

Table 4: We compute the Jaccard similarity between ground truth and masks generated from the attention maps of ViT models (similar to [23] with threshold 0.9) over the PASCAL-VOC12 validation set. Only class level ImageNet labels are used for training these models. Our results indicate that supervised ViTs can be used for automated segmentation and perform closer



DeiT-S DeiT-S-SIN DeiT-S-SIN (Distilled)

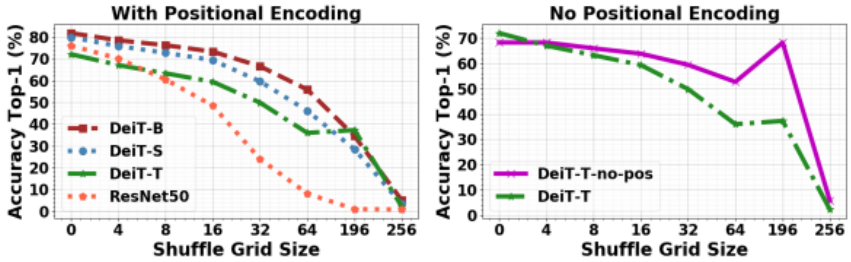
Figure 8: Segmentation maps from ViTs. Shape distil-

be used for automated segmentation and perform closer to the self-supervised method DINO [23].

Figure 8. Segmentation maps from ViTs. Shape distillation performs better than standard supervised models.

Interestingly, training without local texture or with shape distillation allows a ViT to concentrate on foreground objects in the scene and ignore the background (Table 4, Fig. 8).
This offers an automated semantic segmentation for an image although the model is never shown pixel-wise object labels.
That is, shape-bias can be used as self-supervision signals for the ViT model to learn distinct shape-related features that help localize the right foreground object.
The above results show that properly trained ViT models offer shape-bias nearly as high as the human's ability to recognize shapes.

Does Positional Encoding Preserve the Global Image Context?

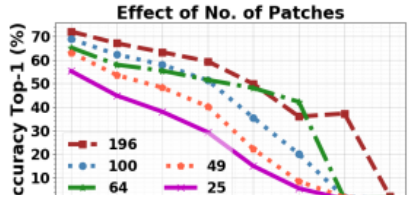


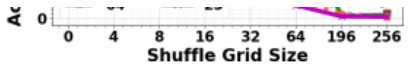
Transformers' ability to process long-range sequences in parallel using self-attention [27] (instead of a sequential design in RNN [31]) is invariant to sequence ordering.
Current ViTs [2, 3, 4, 26] use positional encoding to preserve this context.

Here, we analyze if the sequence order modeled by positional encoding allows ViT to excel under occlusion handling.
Our analysis suggests that transformers show high permutation invariance to the patch positions, and the effect of positional encoding towards injecting structural information of images to ViT models is limited (Fig. 1
This observation is consistent with the findings in the language domain [32] as described below.

Sensitivity to Spatial Structure

Without encoding, the ViT performs reasonably well and achieves better permutation invariance than a ViT using position encoding (Fig. 10).





The above analysis shows that just like the texture-bias hypothesis does not apply to ViTs, the dependence on positional encodings to perform well under occlusions is also incorrect. This leads us to conclude that ViTs' robustness is due to its flexible and dynamic receptive field (see Fig. 4), which depends on the content of an input image.

Robustness of Vision Transformers to Adversarial and Natural Perturbations

Does higher shape-bias help achieve better robustness?

A ViT with similar parameters as CNN (e.g., DeIt-S) is more robust to image corruptions than ResNet50 trained with augmentations (Augmix [33]).

These findings are consistent with [10], and suggest that augmentations improve robustness against common corruptions.

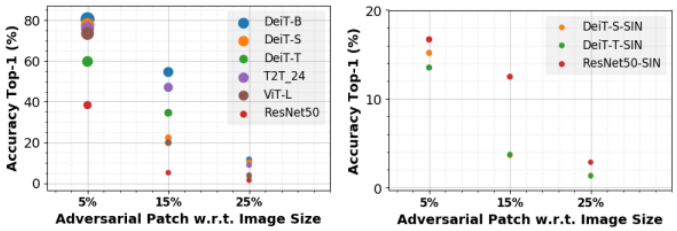
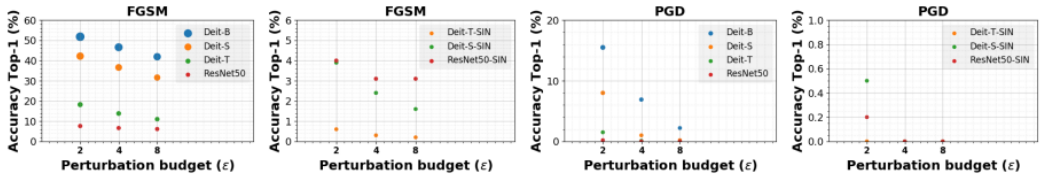
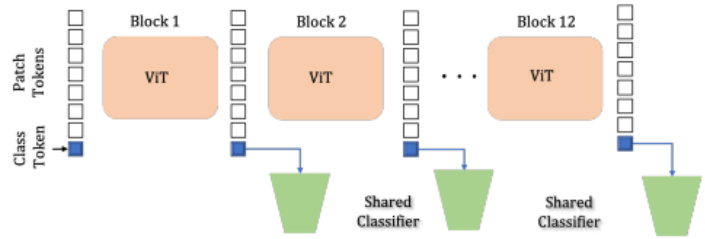
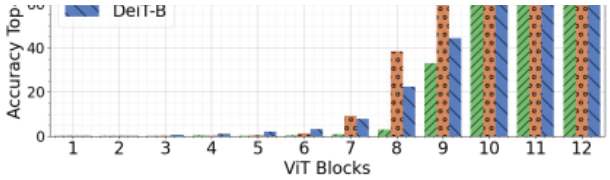


Figure 12: Robustness against adversarial patch attack. ViTs even with less parameters exhibit a higher robustness than CNN. Models trained on ImageNet are more robust than the ones trained on SIN. Results are averaged across five runs of patch attack over ImageNet val. set.



Effective Off-the-shelf Tokens for Vision Transformer





Class tokens generated by the deeper blocks are more discriminative and we use this insight to identify an effective ensemble of blocks whose tokens have the best downstream transferability.

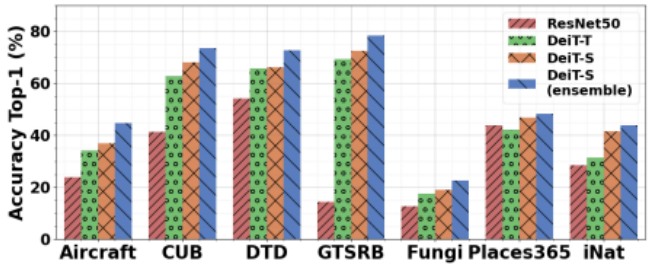
Transfer Methodology:

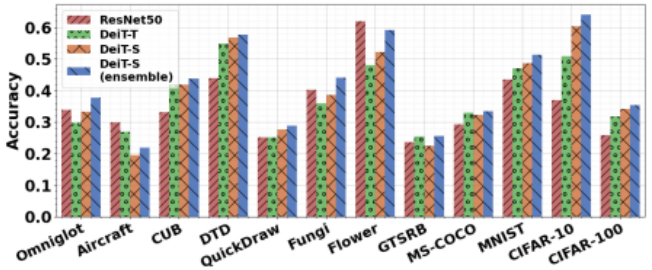
Blocks	Class Tokens	Patch Tokens	CUB [37]	Flowers [38]	iNaturalist [39]
Only 12^{th} (last block)	✓	✗	68.16	82.58	38.28
	✓	✓	70.66	86.58	41.22
From 1^{st} to 12^{th}	✓	✗	72.90	**91.38**	44.03
	✓	✓	73.16	91.27	43.33
From 9^{th} to 12^{th}	✓	✗	**73.58**	90.00	**45.15**
	✓	✓	73.37	90.33	45.12

Here, we concatenate the class tokens (optionally combined with average patch tokens) from different blocks and train a linear classifier to transfer the features to downstream tasks.
The scheme that concatenate class tokens from the last four blocks shows the best transfer learning performance.
Concatenation of both class and averaged patch tokens from all blocks helps achieve similar performance compared to the tokens from the last four blocks but requires significantly large parameters to train.

We find some exception to this on the Flower dataset [38] where using class tokens from all blocks have relatively better improvement (only 1.2%), compared to the class tokens from the last four blocks (Table 5).
However, concatenating tokens from all blocks also increases the number of parameters e.g., transfer to Flowers from all tokens has 3 times more learnable parameters than using only the last four tokens.

Visual Classification
We analyze the transferability of off-the-shelf features across several datasets including Aircraft [40], CUB [37], DTD [41], GTSRB [42], Fungi [43], Places365 [44], and iNaturalist [39].
These datasets are developed for fine-grained recognition, texture classification, traffic sign recognition, species classification, and scene recognition with 100, 200, 47, 43, 1394, 365, and 1010 classes respectively.





The ViT features show clear improvements over the CNN baseline (Fig. 16).
We note that DeiT-T, which requires about 5 times fewer parameters than ResNet50, performs better among all datasets.
Furthermore, the model with the proposed ensemble strategy achieves the best results across all datasets.

Few-Shot Learning:
We consider meta-dataset [45] designed as a large-scale few-shot learning (FSL) benchmark containing a diverse set of datasets from multiple domains.
This includes letters of alphabets, hand-drawn sketches, images of textures, and fine-grained classes making it a challenging dataset involving a domain adaption requirement as well.

On average, the ViT features transfer better across these diverse domains (Fig. 16) in comparison to the CNN baseline.
Furthermore, we note that the transfer performance of ViT is further boosted using the proposed ensemble strategy.

Discussion and Conclusions

In this paper, we analyze intriguing properties of ViTs in terms of robustness and generalizability.
We test with a variety of ViT models on fifteen vision datasets.

We demonstrate favorable merits of ViTs over CNNs for occlusion handling, robustness to distributional shifts and patch permutations, automatic segmentation without pixel supervision, and robustness against adve
Moreover, we demonstrate strong transferability of off-the-shelf ViT features to a number of downstream tasks with the proposed feature ensemble from a single ViT model.

Similarly, we found that ViTs auto-segmentation property stems from their ability to encode shape information.

To highlight a few open research questions:
a) Can self-supervision on stylized ImageNet (SIN) improve segmentation ability of DINO?
b) Can a modified DINO training scheme with texture (IN) based local views and shape (SIN) based global views enhance (and generalize) its auto-segmentation capability?